Pontifícia Universidade Católica
do Rio de Janeiro

Fernando Alberto Correia dos Santos Junior

Extracting Reliable Information From Large Collections of Legal Decisions

Tese de Doutorado

Thesis presented to the Programa de Pós–graduação em Informática of PUC-Rio in partial fulfillment of the requirements for the degree of Doutor em Ciências – Informática.

Advisor: Prof. Hélio Côrtes Vieira Lopes

Rio de Janeiro
April 2022

Pontifícia Universidade Católica
do Rio de Janeiro

Fernando Alberto Correia dos Santos Junior

Extracting Reliable Information From Large Collections of Legal Decisions

Thesis presented to the Programa de Pós–graduação em Informática of PUC-Rio in partial fulfillment of the requirements for the degree of Doutor em Ciências – Informática. Approved by the Examination Committee.

Prof. Hélio Côrtes Vieira Lopes
Advisor
Departamento de Informática – PUC-Rio

Profª. Simone Diniz Junqueira Barbosa
Departamento de Direito – PUC-Rio

Prof. Marcus Vinicius Soledade Poggi de Aragão
Departamento de Informática – PUC-Rio

Prof. Ivar Alberto Martins Hartmann
Insper Instituto de Ensino e Pesquisa

Prof. Guilherme da Franca Couto Fernandes de Almeida
Yale University

Rio de Janeiro, April 8th, 2022

Fernando Alberto Correia dos Santos Junior

Bachelor's in Computer Engineer (2013) at Universidade Estadual de Feira de Santana (UEFS). Master's in Informatics (2016) at Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio). Started his doctorate at PUC-Rio in 2016, focusing his research on Machine Learning, Natural Language Processing, and Information Extraction applied to the Legal Domain.

Bibliographic data

Correia, Fernando Alberto

Extracting Reliable Information From Large Collections of Legal Decisions / Fernando Alberto Correia dos Santos Junior; advisor: Hélio Côrtes Vieira Lopes. – Rio de Janeiro: PUC-Rio, Departamento de Informática, 2022.

v., 104 f: il. color. ; 30 cm

Tese (doutorado) - Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Informática.

Inclui bibliografia

1. Extração de Informação;. 2. Domínio Jurídico;. 3. Direito;. 4. Reconhecimento de Entidade Nomeada;. 5. Modelagem Dinamica de Tópicos;. 6. Decomposição Tensorial;. I. Lopes, Hélio Côrtes Vieira. II. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Informática. III. Título.

CDD: 004

Acknowledgments

First, I would like to thank my family: my parents, Ivone, Arielson, and Fernando (*in memoriam*), and my sister Ariane. I would never get that far if not for them. Thanks for all the unquestioning support, trust, comprehension, and inspiration that have helped me keep up with my purposes, even in the most troubling times.

I am eternally grateful to my advisor Hélio Lopes for all the support, trust, guidance, and friendship. I learned so much from his lectures, conversations, attitude, and example. I will never forget how you inspired and still inspires me to go further in my career as a researcher.

I am also grateful to all the wonderful professors who I was lucky enough to have while on this journey. Especially Professors Simone and Clarisse, for all the great lectures about Human-Computer Interaction, Semiotics, and Ethics. Their lectures helped me better understand my role as an engineer in a plural society.

Also, I would like to thank Professor Arndt, my former advisor in the master's degree. He inspired me to keep in academia and pursue a doctoral degree. I will never forget the excellent lectures and long conversations about science, society, and life.

I would like to thank the FGV Law School for all the support. I am also very thankful for Ivar, who opened this door and introduced me to the legal domain. He inspired me to keep studying law throughout.

I will never be thankful enough to all the friends who helped me finish this journey in so many different ways. I'm especially grateful to Paulo for reviewing much of my writing and for discussions about studies, work, and life. I'm also grateful to my friends José Luis, Guilherme, Alexandre, Felipe, and Kaline. Also, to the friends that have been closer to me since my first days in Rio de Janeiro: Thayne, Natalia, Diego, and Juliano. Finally, I would like to thank the friends who are geographically distant but remain close: Jhielson, Flávia, and André.

Also I thank Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) for the scholarship. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

Abstract

Correia, Fernando Alberto; Lopes, Hélio Côrtes Vieira (Advisor). **Extracting Reliable Information From Large Collections of Legal Decisions**. Rio de Janeiro, 2022. 104p. Tese de doutorado – Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

As a natural consequence of the Brazilian Judicial System's digitization, a large and increasing number of legal documents have become available on the Internet, especially judicial decisions. As an illustration, in 2020, 25 million decisions were produced by the Brazilian Judiciary. Meanwhile, the Brazilian Supreme Court (STF), the highest judicial body in Brazil, alone has produced 99.5 thousand decisions. In line with those numbers, we face a growing demand for studies focused on extracting and exploring the legal knowledge hidden in those large collections of legal documents. However, unlike typical textual content (e.g., book, news, and blog post), the legal text constitutes a particular case of highly conventionalized language. Little attention is paid to information extraction in specialized domains such as legal texts. From a temporal perspective, the Judiciary itself is a constantly evolving institution, which molds itself to cope with the demands of society. Therefore, our goal is to propose a reliable process for legal information extraction from large collections of legal documents, based on the STF scenario and the monocratic decisions published by it between 2000 and 2018. To do so, we intend to explore the combination of different Natural Language Processing (NLP) and Information Extraction (IE) techniques on legal domain. From NLP, we explore automated named entity recognition strategies in the legal domain. From IE, we explore dynamic topic modeling with tensor decomposition as a tool to investigate the legal reasoning changes embedded in those decisions over time through textual evolution and the presence of the legal named entities. For reliability, we explore the interpretability of the methods employed. Also, we add visual resources to facilitate interpretation by a domain specialist. As a final result, we expect to propose a reliable and cost-effective process to support further studies in the legal domain and, also, to propose new strategies for information extraction on a large collection of documents.

Keywords

Information extraction; Legal Domain; Law; Named Entity Recognition; Dynamic Topic Modeling; Tensor Decomposition;

Resumo

Correia, Fernando Alberto; Lopes, Hélio Côrtes Vieira. **Extraindo Informações Confiáveis de Grandes Coleções de Decisões Judiciais**. Rio de Janeiro, 2022. 104p. Tese de Doutorado – Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

Como uma consequência natural da digitalização do sistema judiciário brasileiro, um grande e crescente número de documentos jurídicos tornou-se disponível na internet, especialmente decisões judiciais. Como ilustração, em 2020, o Judiciário brasileiro produziu 25 milhões de decisões. Neste mesmo ano, o Supremo Tribunal Federal (STF), a mais alta corte do judiciário brasileiro, produziu 99.5 mil decisões. Alinhados a esses valores, observamos uma demanda crescente por estudos voltados para a extração e exploração do conhecimento jurídico de grandes acervos de documentos legais. Porém, ao contrário do conteúdo de textos comuns (como por exemplo, livro, notícias e postagem de blog), o texto jurídico constitui um caso particular de uso de uma linguagem altamente convencionalizada. Infelizmente, pouca atenção é dada à extração de informações em domínios especializados, como textos legais. Do ponto de vista temporal, o Judiciário é uma instituição em constante evolução, que se molda para atender às demandas da sociedade. Com isso, o nosso objetivo é propor um processo confiável de extração de informações jurídicas de grandes acervos de documentos jurídicos, tomando como base o STF e as decisões monocráticas publicadas por este tribunal nos anos entre 2000 e 2018. Para tanto, pretendemos explorar a combinação de diferentes técnicas de Processamento de Linguagem Natural (PLN) e Extração de Informação (EI) no contexto jurídico. Da PLN, pretendemos explorar as estratégias automatizadas de reconhecimento de entidades nomeadas no domínio legal. Do ponto da EI, pretendemos explorar a modelagem dinâmica de tópicos utilizando a decomposição tensorial como ferramenta para investigar mudanças no raciocinio juridico presente nas decisões ao lonfo do tempo, a partir da evolução do textos e da presença de entidades nomeadas legais. Para avaliar a confiabilidade, exploramos a interpretabilidade do método empregado, e recursos visuais para facilitar a interpretação por parte de um especialista de domínio. Como resultado final, a proposta de um processo confiável e de baixo custo para subsidiar novos estudos no domínio jurídico e, também, propostas de novas estratégias de extração de informações em grandes acervos de documentos.

Palavras-chave

Extração de Informação; Domínio Jurídico; Direito; Reconhecimento de Entidade Nomeada; Modelagem Dinamica de Tópicos; Decomposição Tensorial;

Table of Contents

1 Introduction 14

1.1 Problem Statement 15

1.2 Research Goal 16

1.3 Contributions 18

1.4 Thesis Outline 18

2 Background 19

2.1 The Brazilian Supreme Court 19

2.1.1 The Massive Production of Decisions 22

2.1.2 Legal Decisions are Public Documents 23

2.2 The Named Entities Recognition in the Legal Domain 24

2.2.1 The STF's Decisions and its Legal Entities 26

2.2.1.1 The Precedent 27

2.2.1.2 The Academic Citation 28

2.2.1.3 The Legislative Reference 29

2.3 Tensor Decomposition and Dynamic Topic Modeling 30

2.3.1 Notations 31

2.3.2 Tensor Decomposition 32

2.3.2.1 NMF for Matrices 32

2.3.2.2 NMF for Tensor 33

2.3.2.3 The CP Decomposition and the NNCPD 34

3 Related work 36

3.1 Legal Named Entity Recognition 36

3.2 Temporal Data Analysis over a Collection of Documents 39

3.2.1 The Dynamic Topic Modeling 40

3.2.2 Approaches Based on Tensor Decomposition 41

4 Legal Named Entity Recognition 44

4.1 A Preliminary Step with a Pattern-Matching Approach 45

4.2 Selection of Decisions 45

4.3 The Annotation Task 47

4.3.1 Training Tasks 49

4.3.2 Inter-annotator Agreement 49

4.4 The Annotated Corpus 50

4.4.1 The Annotations Merge 51

4.4.2 The Corpus 52

4.4.3 Results 55

4.5 Building the Extractor 56

5 Extracting Reliable Information from a Lager Collection of Documents 60

5.1 Decision-Oriented Structure 60

5.1.1 Grouping Monocratic Decisions 61

5.2 Origin-Oriented Structure 63

5.3 The Textual Features Representation 64
5.3.1 Text Processing 65
5.3.2 Textual Feature Selection 65
5.3.3 Tuning Textual Feature Selection with Named Entities 67
5.4 Tensor Decomposition for Latent Topic Identification 68
5.5 Quality Assessment 68
5.6 The Process in Summary 69

6 Experiments and Case Studies **70**
6.1 Exploring Decisions' Content from Different Categories of Legal Case 70
6.1.1 EXP1.1: Comparing the Different Categories 71
6.1.2 EXP1.2: Comparing Decisions from Writs' Legal Cases 74
6.1.3 EXP1.3: Comparing Decisions in different time intervals 77
6.2 The Legal Named Entities Usage over the Time 80
6.2.1 EXP2.1: Legal NE Usage on Different Categories of Legal Cases. 81
6.2.2 EXP2.2: Exploring the precedents citation per Cabinets on Writs cases. 83
6.2.2.1 Scenario I: Using only the legal case citations 84
6.2.2.2 Scenario II: Using only the Reporting Justice citations 86
6.3 The STF's Vocabulary Study Case 88
6.3.1 Exploring the Vocabulary over Time 90
6.4 Concluding Remarks 92

7 Conclusion **94**
7.1 Future Work 96

Bibliography **99**

List of Figures

Figure 2.1 The decisions' production between 2000 and 2021 and proportions by type. The gray area highlights the period not covered by our study. 23
Figure 2.2 Example of nested mentions in legal context. 25
Figure 2.3 Example of complex nested mentions in legal context. 26
Figure 2.4 Example with two precedent long precedents citations. 28
Figure 2.5 Example with one precedent citations in its simplest form. 28
Figure 2.6 Example of academic citations. First scenario, when the reference follows the cited work passage. 29
Figure 2.7 Example of academic citations. Second scenario, when the author's name precedes the cited passage. 29
Figure 2.8 Citations to Legislative References. 30
Figure 2.9 A third-order tensor and its frontal slice (Kolda and Brett, 2009). 31
Figure 2.10 Rank-one third-order tensor, $\mathcal{X} = \mathbf{a} \otimes \mathbf{b} \otimes \mathbf{c}$. 32
Figure 2.11 A visualization of the factor matrices in NMF of $X \approx AB$. The edges of the matrix visualized in blue and red represent the modes of the matrix with dimension n_1 and n_2, respectively (Ahn et al., 2021). 33
Figure 2.12 CP decomposition of a three-way array. Adapted from Kolda and Brett (2009). 34
Figure 2.13 A visualization of the factor matrices in a CP decomposition. The edges of the tensor visualized in blue, red, and green represent the modes of the tensor with dimension n_1, n_2, and n_3, respectively. Adapted from (Ahn et al., 2021). 35

Figure 4.1 Annotation task process. (1) excerpts are randomly distributed among annotators for the coarse-grained annotation; (2) - the coarse annotations are saved into the database and pooled together; (3) - these coarser annotations are randomly distributed among annotators for the fine-grained annotation; and (4) the fine-grained annotations are saved in a specific database. 48
Figure 4.2 Distribution of inter-annotator agreement score for different types of annotated text. 52
Figure 4.3 Decision uniqueness over the time. 57
Figure 4.4 Sentence uniqueness over the time. 58

Figure 5.1 Visualization of the construction of the 4 Newsgroups tensor. Adapted from Ahn et al. (2021). 61
Figure 5.2 Visualization for a possible decision-based structure. 62
Figure 5.3 Production of monocratic decisions per Cabinet. 64
Figure 5.4 Visualization for a possible group-based structure. 64
Figure 5.5 Text processing in summary. 65

Figure 6.1 The tensor structure (frontal slice). 72
Figure 6.2 Comparing decisions from different types of legal case. NNCPD factors A and B. 73
Figure 6.3 The structure of the tensor $\mathcal{W}$ (frontal slice). 75
Figure 6.4 Comparing decisions on writs legal case. 75
Figure 6.5 Decisions ordering per time slice, sorted by legal case classes. 76
Figure 6.6 zooming in over decisions on *Writs* case. 78
Figure 6.7 Frontal slice of the tensor $\mathcal{X}_{Category} \in \mathbb{R}^{c \times b \times t}$. 81
Figure 6.8 Legal named entities usage for different types of legal case. 82
Figure 6.9 The tensor structure (frontal slice). 85
Figure 6.10 Legal named entities usage for writs legal cases. 85
Figure 6.11 Legal named entities usage for writs legal cases. 87
Figure 6.12 The STF's Vocabulary Evolution. 90
Figure 6.13 The STF's Vocabulary Evolution. 91
Figure 6.14 Exploring the vocabulary through latent topics. 91

List of Tables

Table 2.1 Succession line of Justices per cabinet for the period 2000-2021. 21
Table 2.2 The fine-grained entities mapped for precedents. 27
Table 2.3 The fine-grained entities mapped for academic citations. 29
Table 2.4 The fine-grained entities identified for legislative references. 30

Table 4.1 Distribution of decisions per origins. 46
Table 4.2 Selected annotators. 52
Table 4.3 Counting summary of sentences, tokens, coarse and fine-grained annotations. 53
Table 4.4 Coarser-grained entities count. 53
Table 4.5 Fine-grained entities count. 54
Table 4.6 Example of tagged content with a Precedent. 55
Table 4.7 Example of tagged content without fine-grained annotation. 55
Table 4.8 Extraction Summary. 59

Table 5.1 The method in summary. 69

Table 6.1 Distribution of decisions grouped according with the case's type. 71
Table 6.2 Correlation between topics, categories, and terms. 74
Table 6.3 Terms and topics. 76
Table 6.4 Terms and topics for the period 2000-2011. 79
Table 6.5 Terms and topics for the period 2012-2018. 80
Table 6.6 Topics defined by Legal Named Entities. 82
Table 6.7 Distribution of the Top 100 per entity. 83
Table 6.8 Topics represented by legal cases and bind precedents citations. 86
Table 6.9 Topics represented by Reporting Justice citations. 88
Table 6.10 Summary of the NNCPD decomposition. 92

"Science is more than just knowledge of the natural world. It is a view of life, a way of living, a collective aspiration to grow as a species in a world filled with mystery, fear, and wonder. Science is the blanket we pull over our feet at night, the light we turn on in the dark, the beacon reminding us of what we are capable of doing when we work together in pursuit of a common goal."

Marcelo Gleiser, *The Island of Knowledge.*

1
Introduction

The Brazilian Judicial System issues a massive and increasing amount of decisions every year, in the range of dozen of millions. The sum-up of decisions produced by all Courts in 2020 is more than 31 million, 7% more than in 2018[1]. The Brazilian Supreme Court (STF, or Supremo Tribunal Federal), constituted of only eleven Justices, has produced 115 thousand decisions in 2019 (Supremo Tribunal Federal, 2020) – adding up to 1.2 million during the last decade[2]. For comparison purposes, the United States Supreme Court (SCOTUS), constituted by nine Justices, issued 64 decisions (or *"slip" opinions*[3]) in 2019.

Unlike the SCOTUS, even though the STF is also a multi-member court, not all its decisions are a product of a collective decision-making procedure, by which the views of each individual judge will be aggregated into a single collective decision. In fact, more than 80% of STF's decisions are produced by one single Justice. These decisions, dubbed "monocratic" in Brazilian juristic parlance, can be questioned in an internal appeal to one of STF's collegiate bodies, but they are mostly affirmed (Almeida and Hartmann, 2022). That means that, most of the time, each Justice acts as an individual court on its own, a phenomenon that has a crucial impact on the role played by the STF as a supreme and constitutional court (Arguelhes and Ribeiro, 2018).

Over the last few years, monocratic decisions have been the object of much empirical research in the legal community. These studies seek to understand the different effects of this phenomenon in the public debate or the future of the STF itself (Falcao et al., 2017; Arguelhes and Ribeiro, 2018; Hartmann et al., 2018; Hartmann, 2020). Unfortunately, a common burden

[1] According to the National Council of Justice (Conselho Nacional de Justiça, 2021), in 2020, 25 million decisions were produced by the Brazilian Judiciary, 20% less than in 2019. This is mainly due to the COVID-19 pandemic, which reduced the courts' productivity nationwide and the influx of new cases. In 2020 25.8 million cases were filled 15% less than in 2019.

[2] Even in 2020, during the pandemic, the STF has produced 99 thousand decisions, 16% less than in 2019 (Supremo Tribunal Federal, 2021).

[3] A "slip" opinion consists of the majority or principal opinion, any concurring or dissenting opinions written by the Justices, and a prefatory syllabus prepared by the Reporter's Office that summarizes the decision. Extracted from: `www.supremecourt.gov/opinions/slipopinion/19`, last access on 01/18/2022.

among those studies is the cost of extracting relevant information from this ever-increasing volume of decisions. Therefore, Information Systems (IS) and Natural Language Processing (NLP) are vital to the process and digest the massive amount of decisions coming out of Brazilian courts.

Technology applied to the legal domain in many countries has increased due to the online availability of the court's decisions worldwide, not only in Brazil. This, in turn, inspired the development of sophisticated NLP techniques for legal information extraction that are useful for various purposes, including legal search (Dadgosari et al., 2021), legal document summarizing (Merchant and Pande, 2018; Kanapala et al., 2019), quantifying citation relevance (Galgani et al., 2015; Correia et al., 2019), legal contracts inspection and management (Chalkidis et al., 2017), and legal judgment prediction (Chalkidis et al., 2020; Medvedeva et al., 2020). However, very few studies in NLP for legal information extraction focus in the temporal aspects of the text within the legal document.

1.1 Problem Statement

Seen as a primary task for many NLP approaches, Named Entity Recognition (NER) aims to identify mentions to certain elements in free text and classify them into discrete types called Named Entities (NE). Most of the existing studies in NER are based on general texts (*e.g.*, newspaper text, and social media messages). Legal texts (*e.g.*, opinions, decisions, and contracts), however, constitute "*a special case of highly conventionalized language that strives, and often struggles, to find an optimal balance between rigorousness and flexibility, formality, and understandability*" (Höfler and Piotrowski, 2011), and traditional approaches to NER might fail to capture common patterns of writing and the fine-grained semantic information underlying legal texts.

Also, a text written in natural language presents temporal characteristics embedded in itself, and this is a subject that has attracted the attention of researchers, such as Kulkarni et al. (2015); Hamilton et al. (2016a); Li and Tian and Wang (2021), that have focused on capturing the semantics changes of the words throughout time (concept drift). Nevertheless, our work seeks the big picture by trying to capture the *relevant moments* on the history of a collection of legal documents by presenting dynamic topics that emerge, evolve, and fade across time. Moreover, we expanded the topic idea by also relating it to legal named entities (topics defined by a set of terms and legal named entities) and, by doing so, exploring the collection's history correlating the topics with legislative changes (from outside the court) and reasoning changes (within the court).

Our goal is to propose a reliable process for legal information extraction from large collections of legal documents based on monocratic decisions from STF Justices issued between 2000 and 2018. To do so, we explored the combination of different NLP techniques for NER and text transformation with tensor decomposition for information extraction and dynamic topic modeling. To ensure the process's reliability, we focus on methods and strategies capable of providing interpretability to its results. Interpretability that we explored through visualizations in a multidimensional perspective, where each perspective gives a different point of view of the same result.

1.2 Research Goal

Given the massive output of decisions from the Brazilian Judicial Systems, we focused on the highest court. Since the STF is the highest court, its final decision is also the final one that a legal case can get in the Brazilian Judicial System — no other court can question it. Furthermore, the STF decisions effectively influence the lower court's decision-making process and reasoning. For instance, the "*súmula vinculante*" (binding ruling) represents the STF's formalization of how the court itself and any other lower court must decide over a particular issue — the lower courts must be aware of the binding rulings issued by the STF and consider it in their decision-making process.

Given the relevance of the STF's decisions, our main research question is:

> ***MRQ***: *How to extract reliable information from large collections of legal decisions issued by the STF?*

Any valuable information extracted or any conclusion drawn from exploring the STF's collection must present a high reliability level. So, *MRQ* sets boundaries regarding the methods and approaches for the legal information extraction process: they must provide reliable and interpretable results. In Chapter 2 we present a further discussion over the methods employed in the process.

From the NLP perspective, we explored the Named Entity Recognition (NER) task in the legal domain seeking to answer the following research question and sub-questions:

> ***RQ1***: *Which named entities can we find in a legal decision?*
>
> ***RQ1.a***: *How relevant are those entities within the legal domain?*
>
> ***RQ1.b***: *How to extract those entities?*

In order to answer *RQ1* and *RQ1.a*, we took into account findings in studies as Falcão et al. (2019) and Pereira et al. (2020) to map a initial set of nested entities. While most NER works deal only with flat categories of named entities (*e.g.*, date, person, and location), we mapped two levels of nested legal entities: four coarser legal named entities and twenty-four nested ones (fine-grained). Section 2.2 presents a detailed description of these entities. To ensure reliability and capture the nuances of legal reasoning, we conducted an annotation task with 95 law students as annotators, with the collaboration of the *Escola de Direito do Rio de Janeiro da Fundação Getúlio Vargas* (FGV-Direito Rio Law School). As a result, we consolidated the first version of the largest corpus known in Portuguese dedicated to legal named entity recognition (detailed description in Chapter 4).

To answer sub-question *RQ1.b*, Chapter 4 presents the strategy employed for NER based on annotated corpus produced in the previous task, taking into account the text standardization to reduce the processing cost.

From the Information System perspective, we explored the dynamic topic modeling seeking to answer the following research question and sub-questions:

RQ2*: How to explore patterns of events in legal documents from a time perspective?*

RQ2.a*: How reliable are the results of the process?*

RQ2.b*: How coherent are the results of the process?*

RQ2.c*: How reliable is the legal knowledge extracted from the processed collection?*

To answer *RQ2*, we present in Section 3.2 a literature review of strategies for temporal data analysis of a large collection of documents, with a focus on dynamic topic modeling. Based on this review, we present in Chapter 5 our process, which employs dynamic topic modeling as a tool to reveal patterns of events embedded in those decisions over time, both from the textual point of view and from the point of view of the usage of legal named entities.

Regarding sub-questions **RQ2.a** and **RQ2.b**, in Chapter 6, we present a series of experiments in order to verify the process capability to produce reliable and coherent results. We also explore the results of the experiments in an attempt to answer sub-question **RQ2.c**.

1.3
Contributions

Our main contribution is to support further studies in NLP and the Legal area with a focus on the Brazilian Legal System, by:

(i) contributing with the first version of the largest corpus known in Portuguese dedicated to legal named entity recognition;

(ii) introducing a strategy to deploy a NER strategy over a large collection of legal documents;

(iii) presenting a novel process to employ a strategy for dynamic topic modeling over a large collection of legal decisions, based on tensor decomposition; and

(iv) demonstrating the process capability to produce coherent and interpretable results.

1.4
Thesis Outline

The remainder of this thesis is structured as follows. Chapter 2 presents an overview of essential background information and concepts. Chapter 3 presents a series of related works concerning two specific subjects: (i) NER in the legal context and (ii) the temporal data analysis of a large collection of documents. Chapter 4 presents the employed process for a fine-grained annotation task and the NER over the collection of monocratic decisions. Chapter 5 presents the process and strategies for dynamic topic modeling. Chapter 6 presents the sequences of experiments where we applied the proposed process for dynamic topic modeling. Finally, Chapter 7 presents our final considerations.

2
Background

In the following sections of this chapter, we present an overview of essential background information and concepts that we consider relevant for what will be presented in the following chapters. As stated before in Chapter 1, this Thesis bridges two different areas of knowledge, Law and Computer Science, and we hope to present our findings in as clear a way as possible, regardless of the reader's main area of interest.

In the first Section, we present an overview of the STF organization and its decision-making procedures. In Section 2.2, we address NER in the legal domain, with a focus on the Brazilian context. Finally, in Section 2.3 we present some concepts regarding the tensor decomposition and how it can be used for Dynamic Topic Modeling.

2.1
The Brazilian Supreme Court

The STF is constituted of eleven Justices. Each Justice, except for the Chief Justice, sits in one of two different Panels (collegiate bodies), each comprising five justices, and in the Plenary Court, where the Justices decide cases on the full bench led by the Chief Justice[1].

To become an STF's Justice, first, one of the eleven benches must be vacant. Then, the citing President appoints a person who needs to be confirmed by a majority in the Senate. Once approved, the nominated Justice is allowed to remain in the office until the age of 75 when their retirement is mandatory[2]. As an illustration, in 1990, Marco Aurélio was appointed to the STF by then-president Fernando Collor. Months later, Marco Aruélio's appointment was confirmed by the Senate and started as Justice in July 1990. He then sat on the court for 31 years, retiring in July 2021, when he completed 75 years. [3] The current-serving President appointed André Mendonça to fill the vacancy,

[1] Their peers elect the Chief Justice for a term of two years by secret ballot. However, there is a strong tradition of voting for the most senior Justice of the Court that has not yet served as Chief Justice. Also, the Chief Justice does not participate in any of the Panels.

[2] In 2015, a constitutional amendment (PEC 457/2005) increased the age for mandatory retirement from 70 to 75 years old.

[3] Traditionally, the Senate confirms the President's appointment. The only five rejections occurred during Floriano Peixoto's term in 1894.

and, months later, the Senate confirmed the nomination in November 2021. Nevertheless, the court's history also shows cases of early retirements, like that of Justice Joaquim Barbosa, who left the Court 11 years after his nomination at 59. As presented in Chapter 1, our works focus on the collection of decisions produced during the period from 2000 to 2018 and comprises the production of 27 different Justices (the Table 2.1 presents the complete list of Justices).

Every new case filed in the STF is evaluated by the Judiciary Secretary, who may send it to the Chief Justice (those cases related to subjects that are the exclusive competence of the Chief Justice) or randomly assign it to any of the other ten Justices. Once assigned to a case, the Justice might take an individual (monocratic) decision or send it straight to one of the collegiate bodies of the Court (Supremo Tribunal Federal, 2021). Therefore, a decision published by the STF at a given time may have fourteen different origins (eleven Justices, two Panels, and the Plenary) and be classified as a monocratic decision (single Justice opinion) or collegiate decision (produced by one of the three collegiate bodies). It is important to state that, regarding monocratic decisions, we excluded from the experiments opinions written by the Justice when acting as Chief Justice, dues to the scope of competencies associated with the role and the pace of production that is often in disarray with the other Justices (Almeida and Almeida, 2020).

The Justice assigned with a case is called the Reporting Justice and is responsible for reporting the relevant facts of the case and issuing the first vote in cases decided by a Panel or the Plenary. Every other Justice in that collegiate body then gets his/her turn in seniority order and might subscribe to any previous opinions or render their own dissenting or concurring opinion. The Court's final decision (or *per curiam* decision) results from aggregating the opinions submitted individually. The published collegiate decision is posted online as a PDF file created based on the concatenation of the individual opinions and the final decision summary.

Every Justice has his/her own Cabinet with a staff of law clerks who help them in the decision-making process. Very recently, in September 2020, the Court made easily accessible some information regarding the Cabinet staff[4]. In the same period, the number of clerks per Cabinet varied from 12 to 14. An interesting observation regarding the law clerks is related to the Cabinet renovation when a new Justice assumes.

Between September 2020 and July 2021, two Justices retired, Justices Celso de Mello and Marco Aurélio, and two new Justices took these vacancies, Nunes Marques and Andre Mendonça. Comparing the staffs' compositions of

[4] More details in `https://egesp-portal.stf.jus.br/`, last accessed on 01/20/2022.

both Cabinets one month before Celso de Mello retirement (September 2020) and one month after Andre Mendonça has assumed his Cabinet (January 2022), we observed changes in the Cabinet composition. Still, a significant portion of the old members were kept by the new Justice[5]. Even though the monocratic decisions reflect Justice's opinion regarding a case, some influence from the former Justice might still be present due to the presence of part of their clerks. Unfortunately, we were unable to track in detail the Cabinets compositions before September 2020.

Table 2.1: Succession line of Justices per cabinet for the period 2000-2021.

Cabinet	Justice's Name	Tenure		As Chief Justice	
		Start	End	Start	End
1	Nelson Jobim	1997	2006	2004	2006
	Cármen Lúcia	2006	curr.	2016	2018
2	Celso de Mello	1989	2020	1997	1999
	Nunes Marques*	2020	curr.	—	—
3	Marco Aurélio	1990	2021	2001	2003
	André Mendonça*	2021	curr.	—	—
4	Maurício Corrêa	1994	2004	2003	2004
	Eros Grau	2004	2010	—	—
	Luiz Fux	2011	curr.	2020	curr.
5	Néri da Silveira	1981	2002	1989	1991
	Gilmar Mendes	2002	curr.	2008	2010
6	Ilmar Galvão	1991	2003	—	—
	Ayres Britto	2003	2012	2012	2012
	Roberto Barroso	2013	curr.	—	—
7	Calos Velloso	1990	2006	1999	2001
	Ricardo Lewandowski	2006	curr.	2014	2016
8	Sepulvida Pertence	1989	2007	—	—
	Menezes Direito	2007	2009	—	—
	Dias Toffoli	2009	curr.	2018	2020
9	Moreira Alves	1975	2003	—	—
	Joaquim Barbosa	2003	2014	2012	2014
	Edson Fachin	2015	curr.	—	—
10	Sydney Sanches	1984	2003	—	—
	Cezar Peluso	2003	2012	2010	2012
	Teori Zavascki	2003	2012	—	—
	Alexandre de Moraes	2017	curr.	—	—
11	Octavio Gallotti	1984	2000	1993	1995
	Ellen Gracie	2000	2011	2006	2008
	Rosa Webber	2011	curr.	—	—

[5] Justice Andre Mendonça kept five clerks from the previous composition and hired 7, one of them coming from Justice Gilmar Mendes's Cabinet. Meanwhile, Justice Nunes Marques kept only three from the previous compositions, hired three from the Marco Aurélio's Cabinet, and seven new ones.

At the Court's foundation in 1890, 15 cabinets were created (each one identified by a number), but 9 of them were extinct, such as Cabinet 1 in 1969. However, Cabinet 2 still exists and is occupied by Justice Cármen Lúcia at this date. Another 5 were created in 1965, leaving the Court with 11 seats. Nevertheless, the Court uses the creation date for a numbering identification of each seat. To facilitate the reading and future result discussions, we renamed the Cabinet identification with a numbering sequence in ascending order, preserving the original precedence, *e.g.*, the Cabinet 11 in Table 2.1, is, in fact, the Cabinet 20[6]. Table 2.1 presents the succession line of each of the eleven STF Justice Cabinet since 2000, also indicates when the Justice assumed the role of Chief Justice. From that table, only decisions from Justices Nunes Marques and Andre Mendoça (in emphasis with '*') are not present in our study.

2.1.1
The Massive Production of Decisions

As presented by Arguelhes and Ribeiro (2018), considering the STF as a single entity can be misleading, since it is a sum of individuals with their own views on how cases should be decided, whose opinions might not necessarily coincide with those of the court, and who even may issue an individual decision. The decision-making procedure to form collegiate decisions reflects those different views: obtained through persuasion, deliberation, and bargaining, or, when all else fails, by counting votes, which might lead to problems (Almeida and Chrismann, 2019).

In 2020, the STF has produced 99.5 thousand decisions (Supremo Tribunal Federal, 2020) — adding up to 586 thousand during the last five years. For comparison purpose, the SCOTUS has issued 68 "slip" opinions in 2020. This huge difference between these two Supreme Courts is due to their differences in jurisdiction and decision-making process. Both STF and SCOTUS are ordinary appeals courts of last resort that exert concrete constitutional review, but the STF fills two additional roles: it is also a constitutional court that deals with the abstract constitutional review (appeals that seek to reverse decisions from lower courts on constitutional grounds), and also an ordinary trial court in specific issues (*e.g.*, cases regarding federal politicians and Cabinet members(Falcao et al., 2017)[7].

Figure 2.1 presents the distribution of decisions issued by the STF

[6] More about the succession line of the court in `http://www.stf.jus.br/arquivo/cms/sobreStfComposicaoMinistroApresentacao/anexo/linha_sucessoria_quadro_atual_dez_2021.pdf`, last access in 01/15/2022.

[7] For issues decided based on Federal Law, without Constitutional repercussions, the highest court in Brazil is the Superior Court of Justice (STJ).

according to their type, between 2000 and 2021. As shown in Figure 2.1, most of the decisions published by the STF are monocratic decisions. As stated before, the monocratic decision reflects the sole view of one Justice — a decision that can be questioned in an appeal to one of STF's collegiate bodies, but they are most often affirmed. That means that, most of the time, the STF is a monocratic court of appeals where each Justice acts as an individual court on its own, a phenomenon that has a crucial impact on the role played by the STF as a constitutional court (Arguelhes and Ribeiro, 2018).

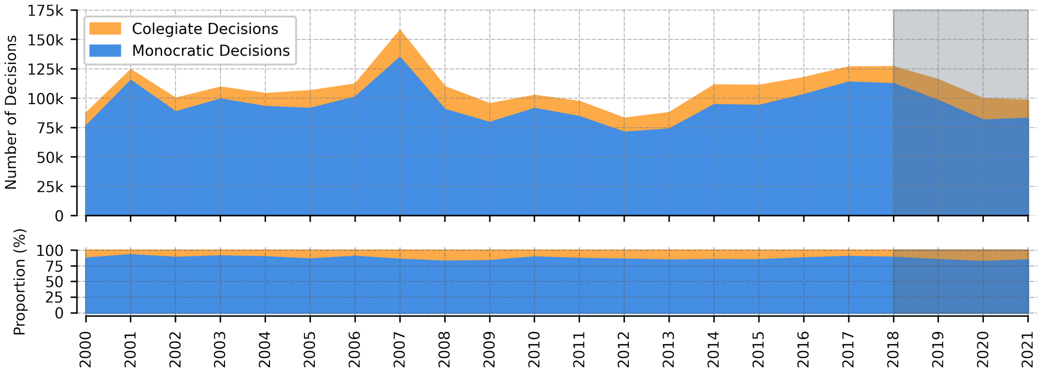


Figure 2.1: The decisions' production between 2000 and 2021 and proportions by type. The gray area highlights the period not covered by our study.

Nevertheless, those numbers also reflect the high demand for the court's attention. In 2020 alone, 75 thousand new cases were filled, 19.4% less than in 2019 — the numbers of new cases have been decreasing since 2017 (when 103 thousand new cases were filled). Moreover, the numbers for 2020 reflect the impact of the COVID-19 pandemic.

One strategy to deal with such demand is to employ the "general repercussion" rule, a filter created by constitutional amendment in order to enable the court to select the concrete constitutional review cases it deems relevant (Mendes, 2017)[8]. In a monocratic decision, a Justice may use the "general repercussion" as a part of their argument denying the case review.

2.1.2
Legal Decisions are Public Documents

Along with the judicial system's digitization starting around 2006 (*Law 11,419*), the decisions published by the STF started to become available on the court's website[9], but only years after (around 2014) all published decisions

[8] The vast majority of the lawsuit filed in the STF are related to concrete constitutional review — appeals that seek to reverse rulings from lower courts on constitutional grounds. In 2020, there were 50 thousand cases related to constitutional review, 66% of the total issued in that year Supremo Tribunal Federal (2021).

[9] `http://portal.stf.jus.br`

became fully accessible on the Internet. Nevertheless, even before Law 11,419, decisions were already public documents available in printed versions of the Judiciary Journals.

According to Art. 93, IX, of the Brazilian Constitution, as a rule, all judicial decisions are public. Furthermore, paragraph XIV, Art. 5, of the Brazilian Constitution guarantees the right to access any of that information.

Nevertheless, our work draws from the results of projects maintained by the FGV-Direito Rio Law School (*Escola de Direito do Rio de Janeiro da Fundação Getúlio Vargas*) focused on legal documents exploration. For instance, FGV developed a project to create and maintain the digital versions of monocratic decisions collection issued by the court since 1989.

Unfortunately, most of the old documents are rudimentary scanned versions of the original printed version. A vast amount of monocratic decisions issued from the later 90s are in RTF (*Rich Text Files*) format, and the most recent ones are PDF files with easily accessible textual content. In order to provide a high-quality standard level, our collection is restricted to the monocratic decisions published from 2000 onward.

For this thesis, we used a collection of 1.130.661 monocratic decisions published between 2000 and 2018. This number already excludes monocratic decisions issued by the Chief Justices.

2.2 The Named Entities Recognition in the Legal Domain

Seen as one of the most important sub-tasks for information extraction, the NER task aims to identify entities mentioned in free text and classify them into types of information elements, called Named Entities (NE). This task became popular in the mid-90s and, since then, several studies have been published, and many annotated corpora have been made available openly on the Internet. However, despite the relevance of NER, this area has received little academic attention for languages not researched worldwide, like Portuguese.

Most of NER's works focus on flat categories of named entities mentions which ignore essential information that can be useful for downstream tasks. We focused on mentions with a nested structure, or nested named entities, to capture the fine-grained detail in legal documents. As benefits of this strategy, Ringland et al. (2019) enlist three phenomena possible to be embedded in this type of annotation:

- **The Entity-entity relationships.** For instance, in Figure 2.2, the location of the "*1^a Turma Recursal Criminal*" from the "*Estado do Rio de Janeiro*".

- **Entity attribute values.** For instance, the title is the embedded ROLE "*Ministro*", which also encodes the employment relation of the person "*Herman Benjamin*" being a justice at the court "*STJ*".
- **Part-whole relationships.** For instance, the panel "*Primeira Turma*" is part of the court STF.

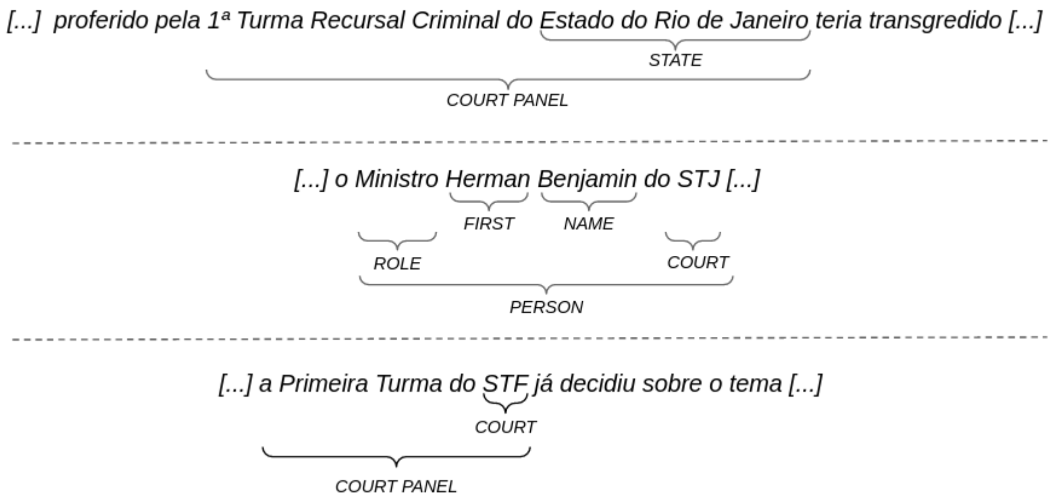

[...] proferido pela 1ª Turma Recursal Criminal do Estado do Rio de Janeiro teria transgredido [...]
STATE
COURT PANEL
[...] o Ministro Herman Benjamin do STJ [...]
FIRST
NAME
ROLE
COURT
PERSON
[...] a Primeira Turma do STF já decidiu sobre o tema [...]
COURT
COURT PANEL

Figure 2.2: Example of nested mentions in legal context.

There are no limits to the number of levels for this kind of annotation. For instance, Figure 2.3 illustrates two other types of multilevel annotations. The first one is an example of expanded annotation of a precedent, identified by one decision and its court of origin. The decision, in turn, is identified by its type and the legal procedure to which it is related. Its class and number usually identify the legal procedure. However, both class and number may vary according to the standard defined by its court of origin[10]. The second citation in Figure 2.3 is related to an academic citation, which a peculiar example. The citation refers to the book entitled "*Mandado de segurança*" (in English, Writ of Mandamus) written by Luiz Fux. The peculiarity is that the author is also an STF's Justice, who has previously been a Justice of the Superior Court of Justice (STJ), and "*Mandado de segurança*" is also a name of a typical legal procedure class.

Most of those legal entities could be extracted with a partner-match approach, if not for the variety of patterns and levels that turns those techniques too error-prone for the automatic identifications of the different phenomenons embedded in the annotation. Like the entity attribute values:

[10] The STF has 72 different classes and two different number formats: "*CNJ ID*" is one of then. The National Council of Justice (CNJ) defines a standard for both information, but the standard wasn't fully internalized by Brazilian Courts as of yet.

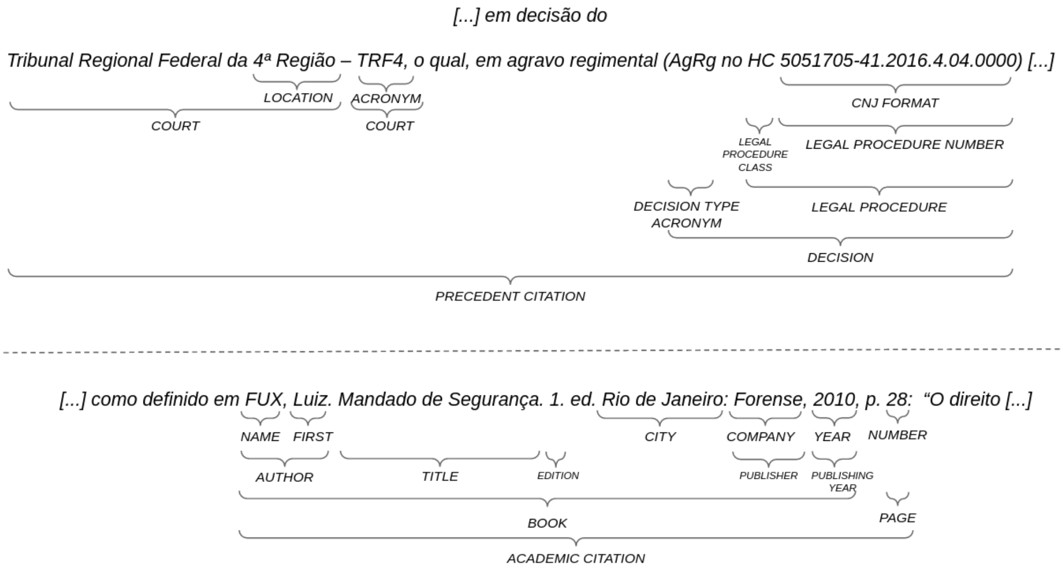


Figure 2.3: Example of complex nested mentions in legal context.

the "*ROLE*" of the person "*Luiz Fux*" may be ambiguous if not explicit in the text.

2.2.1 The STF's Decisions and its Legal Entities

In order to capture the detailed semantic information underlying the legal documents, we conducted a preliminary study to map the entities presented in this work. In that study, we considered findings in previous legal research produced with our team's support, such as Falcão et al. (2019) and Pereira et al. (2020). These entities were chosen in consonance with (i) the legal relevance of each entity in a broad perspective and (ii) the intellectual effort required for the classification. The first factor is related to how often the entity might appear in legal documents, how many values it can assume, and how valuable it is in the context of legal reasoning. The second factor addresses the intellectual effort in the annotation process. To avoid conceptual conflicts between entities, or major disagreements between annotators, we have considered only entities with a meaning that is common sense in the legal domain.

We mapped four coarser legal entities: the **precedent**, the **academic citation**, the **legislative reference**, and the **person**. The entity "*person*" is the only one with no inner elements linked to it. This entity's primary purpose is disambiguation, as both precedent and academic citations have inner entities related to personal identification. The tag "*person*" registers a person's occurrence in the text out of the context of these other coarser entities. In the following subsections, we present each of these entities, along with the inner entities related to them.

2.2.1.1 The Precedent

The precedent is a textual citation of a prior court decision. This undoubtedly offers great value in a common-law based judicial system Galgani et al. (2015); Leibon et al. (2018) — where courts are bound to their previous rulings, such as the United States, Canada, and India. In the U.S., the citation of a precedent in a legal document follows a very standardized format, at least in the Supreme Court. Due to that level of standardization, search databases for legal data, based on precedents, such as Shepard's Citations[11] and Westlaw[12], have existed since the early 90s. Unfortunately, references to precedent in STF decisions do not follow a formal standard.

Since a specific lawsuit in Brazil may have multiple decisions related to it, the reference must inform some further element, temporal or legal, to identify which decision is being referenced. To do so, the citation usually comes with the legal procedure identification and the judgment date or the decision's publication date — or sometimes both. It often contains the rapporteur and the decision type (*e.g.*, merit, injunction, internal appeal).

Table 2.2: The fine-grained entities mapped for precedents.

Fine-grained Entity	Type	Description
Legal Procedure Number	Number	The number that identifies the legal procedure in court.
Legal Procedure Class	Text	Signals the kind of legal procedure. It is often used along the case number to uniquely identify a legal procedure within STF.
Legal Procedure Origin	Text	Indicates from what state (or federation unit) the legal procedure came, usually just the acronym (*e.g.*, RJ stands for Rio de Janeiro).
Decision type	Text	Indicates if the decision referred is related to an internal appeal or motion.
Reporting Justice	Text	Identifies the Justice responsible for the decision (if a monocratic decision, the Justice is also the origin).
Court	Text	The court which rendered the decision.
Judgment date	Date	When the decision was taken.
Publication date	Date	When the decision was introduced to the official record.

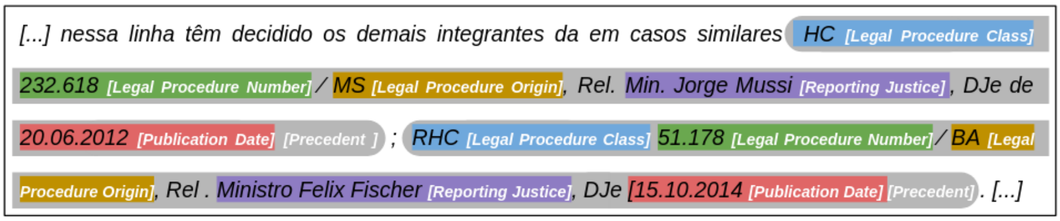
The legal procedure identification is a number with the format regulated by the Conselho Nacional de Justiça (2009). Still, every court in Brazil has its own internal identification system, usually composed of a number and the procedure classification according to classes. The STF alone has 72 different

[11] `https://www.lexisnexis.com/en-us/products/lexis/shepards.page` Accessed on 09/20/2020.

[12] `https://www.westlaw.com` Accessed on 09/20/2020.

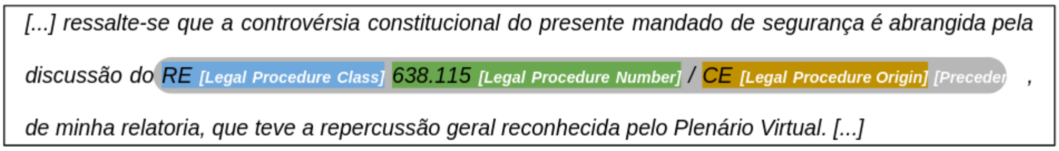
procedural classes. Table 2.2 presents the complete list of fine-grained entities mapped for precedents.

Put together, this information identifies a unique decision, even if the case referred to in the citation has multiple decisions. To illustrate, Figure 2.4 shows tags for two different precedents and their inner elements, enough for their accurate identification.



[...] nessa linha têm decidido os demais integrantes da em casos similares HC [Legal Procedure Class] 232.618 [Legal Procedure Number] / MS [Legal Procedure Origin], Rel. Min. Jorge Mussi [Reporting Justice], DJe de 20.06.2012 [Publication Date] [Precedent] ; RHC [Legal Procedure Class] 51.178 [Legal Procedure Number] / BA [Legal Procedure Origin], Rel . Ministro Felix Fischer [Reporting Justice], DJe [15.10.2014 [Publication Date] [Precedent]. [...]

Figure 2.4: Example with two precedent long precedents citations.



[...] ressalte-se que a controvérsia constitucional do presente mandado de segurança é abrangida pela discussão do RE [Legal Procedure Class] 638.115 [Legal Procedure Number] / CE [Legal Procedure Origin] [Preceder , de minha relatoria, que teve a repercussão geral reconhecida pelo Plenário Virtual. [...]

Figure 2.5: Example with one precedent citations in its simplest form.

However, not all precedent citations have all their inner elements explicitly discriminated. Figure 2.5 presents an example. Even in this incomplete form, the citation represents valuable information that can tell us how relevant a given legal procedure was, is, and will be for the court, as explored in our prior work Correia et al. (2019).

2.2.1.2 The Academic Citation

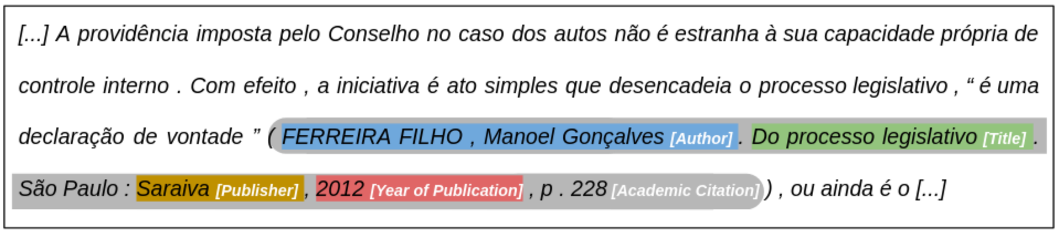
An academic citation is a direct citation to a book, book chapter, or journal article. In the STF rulings, its presence is not as common as the precedents citation or legislative references, but this is not a sign of the lesser importance of scholarly citations in Brazilian law. Instead, it is a result of the types of legal issues brought to the STF, and the court's strategy in dealing with its workload. An academic publication can provide data, concept definitions, or arguments to support a ruling. Recently, it became a study object in the legal community where researchers have explored the academic citations in order to capture the influence of certain authors in the legal debate (Lorenzetto and Kenicke, 2013; Carvalho and Roesler, 2019).

We targeted mentions made within the decision to published academic works, such as studies, scientific articles, and books, as an element of judicial reasoning. Thus, we tried to map all and any case where such a mention occurred. Table 2.3 presents the complete list of fine-grained entities mapped for the academic citation.

Table 2.3: The fine-grained entities mapped for academic citations.

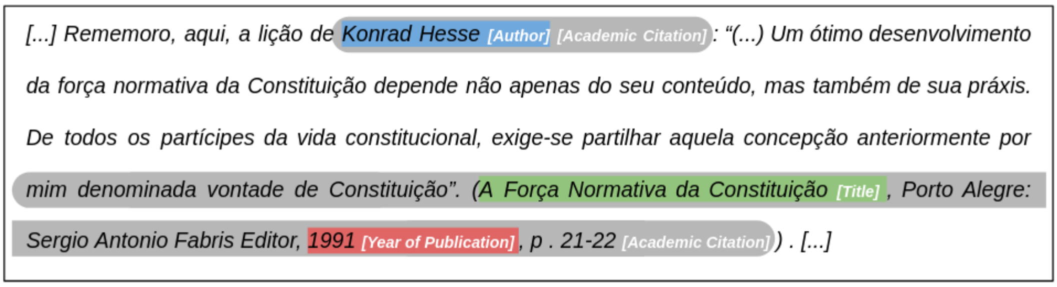
Fine-grained Entity	Type	Description
Title	Text	The work published title.
Collection Title	Text	The collection title, if the publication is part of a collection (e.g., journal title).
Author	Text	The publication first author.
Co-author	Text	The publication co-author(s).
Publisher	Text	The work's publisher.
Year of Publication	Number	Year of publication of the work.

We observed that the citation of academic works mostly follows two styles. Either the reference follows a direct quote, as illustrated in Figure 2.6, or the author's name precedes the quote, and the rest of the reference is provided following it, as shown in Figure 2.7.



[...] A providência imposta pelo Conselho no caso dos autos não é estranha à sua capacidade própria de controle interno . Com efeito , a iniciativa é ato simples que desencadeia o processo legislativo , " é uma declaração de vontade " (FERREIRA FILHO , Manoel Gonçalves [Author] . Do processo legislativo [Title] . São Paulo : Saraiva [Publisher] , 2012 [Year of Publication] , p . 228 [Academic Citation]) , ou ainda é o [...]

Figure 2.6: Example of academic citations. First scenario, when the reference follows the cited work passage.



[...] Rememoro, aqui, a lição de Konrad Hesse [Author] [Academic Citation]: "(...) Um ótimo desenvolvimento da força normativa da Constituição depende não apenas do seu conteúdo, mas também de sua práxis. De todos os partícipes da vida constitucional, exige-se partilhar aquela concepção anteriormente por mim denominada vontade de Constituição". (A Força Normativa da Constituição [Title] , Porto Alegre: Sergio Antonio Fabris Editor, 1991 [Year of Publication], p . 21-22 [Academic Citation]) . [...]

Figure 2.7: Example of academic citations. Second scenario, when the author's name precedes the cited passage.

2.2.1.3
The Legislative Reference

The legislative reference is a fundamental feature of legal reasoning, consisting of the legal dispositions referenced in each decision. This court ruling element is even more relevant in a civil law system, where statutory law

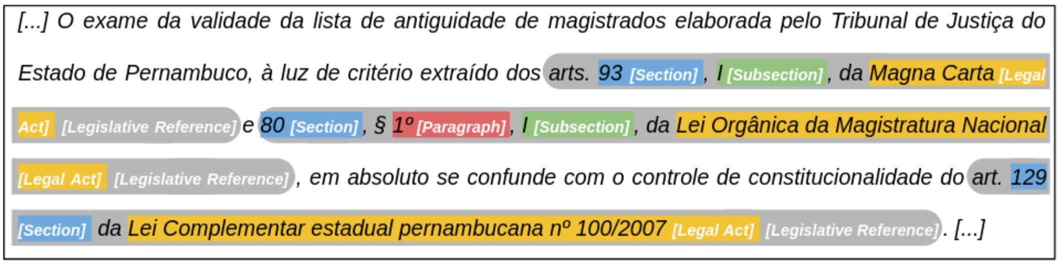
Table 2.4: The fine-grained entities identified for legislative references.

Fine-grained Entity	Type	Description
Legal Act	Text	The legislative act that was cited (*e.g.*, Federal or State Constitution, Legal Statutes).
Institution	Text	When the act is not legislative, such as regulations or internal rules, which institution issued it (*e.g.*, STF internal rules or Federal Reserve resolution).
Origin	Text	The Federation entity that issued regulation, municipality, state, or the Union.
Section	Number	The legal act section.
Paragraph	Number	The legal act paragraph.
Subsection	Letter	The legal act subsection.
Clause	Letter	The legal act clause.

supports most decisions. We expect the STF to cite the constitution heavily since the most significant element of its jurisdiction is constitutional review.

Unlike other coarser entities, we spent efforts defining the fine-grained entities that could be present in a legislative reference. After an experimentation activity, we mapped seven entities that are presented in Table 2.4.

In its shortest form, the citation of a legislative reference may accrue to an indication of a statute or the constitution itself, such as "*in accordance with the Federal Constitution of 1988*" where we consider "*Federal Constitution of 1988*" to be the legal act. Figure 2.8 shows a small excerpt of a decision with three different citation styles citation for three different statutes.

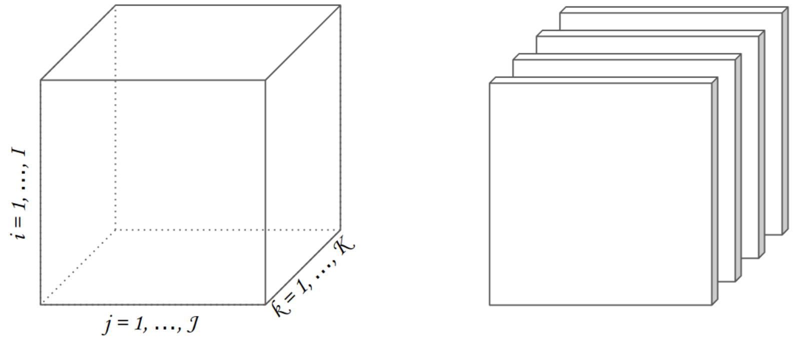


[...] O exame da validade da lista de antiguidade de magistrados elaborada pelo Tribunal de Justiça do Estado de Pernambuco, à luz de critério extraído dos arts. 93 [Section], I [Subsection], da Magna Carta [Legal Act] [Legislative Reference] e 80 [Section], § 1º [Paragraph], I [Subsection], da Lei Orgânica da Magistratura Nacional [Legal Act] [Legislative Reference], em absoluto se confunde com o controle de constitucionalidade do art. 129 [Section] da Lei Complementar estadual pernambucana nº 100/2007 [Legal Act] [Legislative Reference]. [...]

Figure 2.8: Citations to Legislative References.

2.3 Tensor Decomposition and Dynamic Topic Modeling

A tensor is a multidimensional array, and the order of a tensor is the number of dimensions, also known as ways or modes. More formally, Nth-order tensor is an element of the tensor product of N vector spaces, each of which has its own coordinate system. A first-order tensor is a vector, a second-order tensor is a matrix, and tensors of order three or higher are called higher-order tensors Kolda and Brett (2009), as presented in Figure 2.9.a.



(a) A third-order tensor: $\mathcal{X} \in \mathbb{R}^{i \times j \times k}$

(b) Frontal Slice: $\mathcal{X}_{::k}$ or $\mathcal{X}_k$

Figure 2.9: A third-order tensor and its frontal slice (Kolda and Brett, 2009).

Usually, a collection of documents is often represented as third-order tensors in dynamic topic modeling. For instance, given the tensor $\mathcal{X} \in \mathbb{R}^{i \times j \times k}$, the first, second, and third modes represent documents, words, and time, respectively. Since our future discussions and presentation will be regarding a third-order tensor, the concepts presented in this section are limited to that context.

Nevertheless, for those who intend to explore further this topics, most of the concepts presented in this section are explained in detail in Kolda and Brett (2009), and Ahn et al. (2021).

2.3.1 Notations

As mentioned early, the **order of a tensor** is the number of dimensions. **Vectors** (tensors of order one) are denoted by boldface lowercase letters (*e.g.*, **a**). **Matrices** (tensors of order two) are denoted by boldface capital letters (*e.g.* **A**). The **higher-order tensors** (order three or higher) are denoted by boldface Euler script letters (*e.g.*, $\mathcal{X}$). Scalars are denoted by lowercase letters (*e.g.*, a).

Another form to represent a tensor is through **slices** that are two-dimensional sections of a tensor, defined by fixing all but two indices. There are three possible slices for a third-order tensor $\mathcal{X}$, the horizontal, lateral, and frontal, denoted by $\mathcal{X}_i$, $\mathcal{X}_j$, and $\mathcal{X}_k$. Figure 2.9.b shows the frontal slides of a third-order tensor.

Rank-One Tensors is an N-order tensor $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$ is *rank one* if it can be written as the outer product of N vectors, i.e.,

$$x_{i_1 i_2 \cdots i_N} = a^{(1)}_{i_1} a^{(2)}_{i_2} \cdots a^{(N)}_{i_N}, \ for\ all\ 1 \leq i_n \leq I_n. \tag{2-1}$$

Figure 2.10 illustrates $\mathcal{X} = \mathbf{a} \otimes \mathbf{b} \otimes \mathbf{c}$, a third-order rank-one tensor.

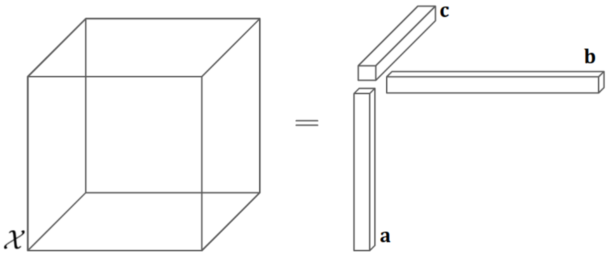


Figure 2.10: Rank-one third-order tensor, $\mathcal{X} = \mathbf{a} \otimes \mathbf{b} \otimes \mathbf{c}$.

2.3.2 Tensor Decomposition

As stated in Ahn et al. (2021), the crucial step of dynamic topic modeling is to decompose high-dimensional tensors into interpretable representations with attention to temporal information. Also, to allow the topic interpretability, decomposition approaches with an additional structure, such as non-negativity — different from traditional approaches like principal component analysis (PCA), where factors often cancel — would be a viable alternative.

Typical methods for such nonnegative tensor decompositions are mainly based on nonnegative matrix factorization (NMF), where a metricized version of the tensor sliced along the temporal dimension is factorized using NMF. There are two basic approaches to NMF-based nonnegative tensor decomposition: Direct NMF and Fixed NMF. To better understand these two approaches, in the following subsection we present NMF for matrices and then their usage for tensor decomposition with Direct NMF and Fixed NMF approach.

In the following subsections, we present the NMF, the tensor decomposition based on NMF approaches, and the nonnegative CANDECOMP / PARAFAC (CP) decomposition (NNCPD) proposed Ahn et al. (2021) for dynamic topic modeling.

2.3.2.1 NMF for Matrices

Nonnegative matrix factorization (NMF) seeks to find an approximate factorization of a nonnegative data matrix $X \in \mathbb{R}_{\geq 0}^{n_1 \times n_2}$ into a nonnegative features of matrix A and a nonnegative coefficients of matrix B, where $r \in \mathbb{N}$ corresponds to the number of latent topics in the data — usually smaller than n_1 and n_2 as presented in Equation 2-2.

$$\mathcal{X} \approx AB, A \in \mathbb{R}_{\geq 0}^{n_1 \times r}, B \in \mathbb{R}_{\geq 0}^{r \times n_2}, \tag{2-2}$$

Moreover, we note that the outer product representation of matrix multiplication lets us rewrite the product AB as presented in Equation 2-3.

$$\mathcal{X} \approx AB = \sum_{l=1}^{r} a_l \otimes b_l, \tag{2-3}$$

In Equation 2-3, $a_l \in \mathbb{R}^{n_1}_{\geq 0}$ is a column of A and $b_l \in \mathbb{R}^{n_2}_{\geq 0}$ is a row of B. See Figure 2.11 for a visualization of NMF as in 2-2. Generally, the factorization is computed by approximately minimizing the reconstruction error. When the minimum of reconstruction error vanishes we say an exact NMF is obtained.

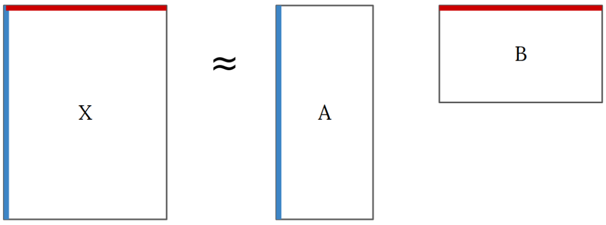


Figure 2.11: A visualization of the factor matrices in NMF of $X \approx AB$. The edges of the matrix visualized in blue and red represent the modes of the matrix with dimension n_1 and n_2, respectively (Ahn et al., 2021).

2.3.2.2 NMF for Tensor

There are two basic approaches to NMF-based nonnegative tensor decomposition: Direct NMF and Fixed NMF. Direct NMF on tensor slices performs NMF independently on each slice of the tensor. In the dynamic topic modeling context, the temporal mode is the third one, so each slice represents the data at a specif moment in time.

Given $\mathcal{X} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$, a frontal slicing gives nonnegative matrices $X_{n_1} \in \mathbb{R}^{n_1 \times n_2}_{\geq 0}$ for $i = 1, 2, ..., n_3$, each of which is factored into nonnegative matrices

$$X_i \approx A_i S_i, \qquad A_i \in \mathbb{R}^{n_1 \times r}_{\geq 0}, S_i \in \mathbb{R}^{r \times n_2}_{\geq 0}, \qquad i = 1, ..., n_3, \tag{2-4}$$

where the A_i's will be referred to as the Direct NMF A factors, and the S_i's the Direct NMF S factors. This form of nonnegative tensor decomposition fails to capture inherent structures along the temporal dimension in the tensor.

As an alternative, the Fixed NMF performs NMF simultaneously on the n_3 slices along mode three (the temporal mode), $X_i, i = 1, ..., n_3$, with the same A. They consider a sequence of nonnegative matrix factorizations (A, S_{n_3}) such that

$$X_i \approx A S_i, \qquad A \in \mathbb{R}^{n_1 \times r}_{\geq 0}, S_i \in \mathbb{R}^{r \times n_2}_{\geq 0}, \qquad i = 1, ..., n_3, \tag{2-5}$$

where A will be referred to as the Fixed NMF common A factor, and the S_i's the Fixed NMF S factors. In other words, Fixed NMF fixes a single dictionary matrix A and searches for the representations S_i for each of the slices X_i. Stacking the products of the Fixed NMF A matrix and S matrices forms an approximation to X, which will be referred to as the Fixed NMF reconstruction.

2.3.2.3 The CP Decomposition and the NNCPD

Unlike the NMF-based nonnegative tensor decompositions, CP decompositions treat the tensor as a whole. The CP decomposition and NNCPD factorize a tensor into a sum of component rank-one tensors without slicing it along the temporal mode.

The CP decomposition factorizes a tensor into a sum of component rank-one tensors. For example, given a third-order tensor $\mathcal{X} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$, we wish to write it as:

$$\mathcal{X} \approx \sum_{r=1}^{R} a_r \otimes b_r \otimes c_r, \tag{2-6}$$

where R is a positive integer and $a_r \in \mathbb{R}^{n_1}$, $b_r \in \mathbb{R}^{n_2}$, and $a_k \in \mathbb{R}^{n_3}$ for $r = 1, ..., R$. Element wise, Equation 2-6 is written as:

$$x_{ijk} \approx \sum_{r=1}^{R} a_{ir} b_{jr} c_{kr}, \; for \; i = 1, ..., n_1, j = 1, ..., n_2, \; k = 1, ..., n_3. \tag{2-7}$$

This is illustrated in Figure 2.12. The factor matrices of CP decomposition refer to the combination of the vectors from the rank-one components, i.e., $A = [a_1, a_2, \cdots, a_r]$, and likewise for B and C. The Figure 2.13 another illustration of the CP decomposition focus in the factor matrices.

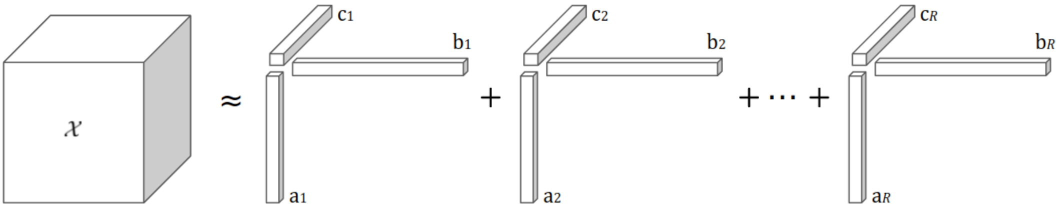


Figure 2.12: CP decomposition of a three-way array. Adapted from Kolda and Brett (2009).

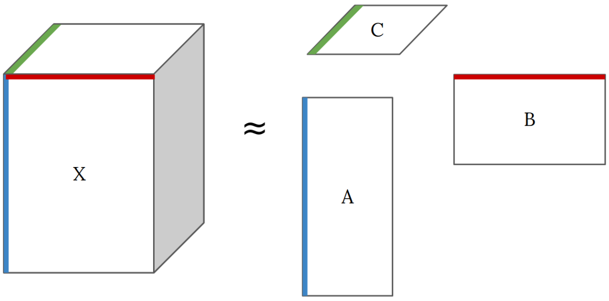


Figure 2.13: A visualization of the factor matrices in a CP decomposition. The edges of the tensor visualized in blue, red, and green represent the modes of the tensor with dimension n_1, n_2, and n_3, respectively. Adapted from (Ahn et al., 2021).

The rank of the tensor $\mathcal{X}$, denoted $rank(\mathcal{X})$, is the smallest integer r so that $\mathcal{X}$ may be expressed as the sum of exactly r rank-one tensors. In other words, this is the smallest number of components in an exact CP decomposition, where "exact" means that there is equality in Equation 2-6. Unfortunately, as presented by Kolda and Brett (2009) in their survey, this is an NP-hard problem, and there is no straightforward algorithm to determine the rank of a specific given tensor. They argue that, in practice, the rank of a tensor is determined numerically by fitting various rank-R CP models.

An approximate CP decomposition may be computed by fixing an r that minimizes the reconstruction error

$$E(\mathcal{X} : A, B, C) = \left\| \mathcal{X} - \hat{\mathcal{X}} \right\|_F , \tag{2-8}$$

where $\|.\|_F$ denotes the Frobenius norm. The solution $\hat{\mathcal{X}} = \sum_{r=1}^{R} a_r \otimes b_r \otimes c_r$ will be referred to as a *rank-r* CP reconstruction of $\mathcal{X}$.

The NMF specializes matrix factorization to factorizing a nonnegative data matrix into the product of two (lower-dimensional) nonnegative factor matrices. In the same way, NNCPD specializes the CP decomposition to decomposing a nonnegative data tensor into the sum of rank-one tensors (Ahn et al., 2020). A nonnegative approximation with fixed r is obtained by approximately minimizing the reconstruction error between $\mathcal{X}$ and the NNCPD reconstruction $\hat{\mathcal{X}} = \sum_{r=1}^{R} a_r \otimes b_r \otimes c_r$ among all the nonnegative vectors.

As mentioned earlier, specifying the rank of a tensor is a problem hard to solve, but, as pointed out by Ahn et al. (2021), the choice of a value to the rank will affect the quality and interpretability of the output topics. They argue that this is a common challenge in determining a choice of rank that both allows for adequate topic representation while not being so large as to start fitting, *e.g.*, noise. We will present further details on this topic in the following chapters.

3
Related work

As previously mentioned, the purpose of this chapter is to present an overview of the existing work concerning two specific subjects: (i) NER in the legal context and (ii) the temporal data analysis of a large collection of documents. Regarding NER, we focus on works directly related to the exploration of NER activities in the legal domain, mainly focused on constructing annotated corpus including, but not limited to, the Brazilian judicial context. Since we also concerned ourselves with the quality of the annotation task, we also present related works which focus on annotation reliability.

3.1
Legal Named Entity Recognition

One of the first works to explore the nuances of the legal entities, Chalkidis et al. (2017) presents a dataset with approximately 3,500 English contracts manually annotated with a flat ontology comprised of 11 categories of contractual elements. Unlike generic NER systems, they proposed specific entities that take such contextual information into account: “start date”, “terminative date”, and “effective date”. Also, they added legal NEs like “jurisdiction” (which specifies the courts responsible for resolving disputes), and “legislation references” (for legal dispositions that the contract relies on). Ten law students performed the annotation. To ensure quality, they conducted preliminary training rounds before the final annotation task. To evaluate the quality of annotations between rounds, they used $\frac{|AB|}{max(|A|,|B|)}$ as an inter-annotator agreement measure, where A and B are sets of contract elements marked by the two annotators, respectively. For the contract entity extraction, they tried different methods combining manually written rules and linear classifiers (Logistic Regression (LR) and Linear Support Vector Machines (SVM)) with hand-crafted features, word embedding, and part-of-speech (POS) tag embeddings. Our work is similar to Chalkidis et al. (2017) in the sense of seeking to give more context to entities typically treated in a generic way. Still, we opted for a nested named entity representation to better capture the contextual information. Besides, we also turn to law students as domain experts for the annotation task.

With regards to legal decisions, Leitner et al. (2019) explored NER in Germany. Due to the complete absence of an annotated dataset, they created and made available their own annotated dataset with 750 decisions manually annotated by one Computational Linguistics student with fine-grained detailed legal entities. The source texts were manually annotated with 19 fine-grained categories of NE — an approach similar to the one presented by Chalkidis et al. (2017) —, which then was generalized into seven more coarse-grained entities. As an illustration: an example of a coarse-grained entity is “person”, which can be one of three different fine-grained entities: “Person”, “Judge”, or “Lawyer”. They applied models based on conditional random fields (CRF) and on bidirectional long short-term memory (BiLSTM) with CRF (BiLSTM-CRF) for both types of grained NE for the NER task. They got the best performance with a BiLSTM-CRF model achieving a 95.46 F1 score for the fine-grained classes and 95.95 for the coarse-grained ones. Their work inspired us in aspects related to the annotation design. Their annotation procedure differs from ours in terms of strategy and the number of annotators involved. Our main concerns were the annotation quality to build a reliable corpus in Portuguese where annotations could capture the nuances of the Brazilian legal documents.

Focusing on the information extraction task, Ji et al. (2020) also explore the legal domain and propose a model for evidence information extraction task from court record documents (named BLACN). Their work is closely related to NER, but instead of a sequence labeling problem, they model the evidence information extraction as a combined task of intermediate paragraph classification and final sequence labeling. The evidence information extracted from paragraphs is divided into five categories, which consist of “evidence provider”, “evidence name”, “evidence content”, “cross-examination party”, “and cross-examination opinion”. To facilitate the evidence information extraction, the paragraphs are classified into two possible categories: “evidence production” and “evidence cross-examination”. The presented model adopts a shared encoder followed by separated encoders for paragraph classification and evidence extraction. To validate their proposal, they used a manually annotated dataset from Chinese courts with 1,128 documents. Documents were collected from courts from several provinces and cities in China from 2013 to 2019 and annotated by experts in the legal field. When compared with other approaches, the model outperforms by a large margin, with 72.36 F1. The *BLACN*, presented by Ji et al. (2020), illustrates the exploration of legal information extraction over a large collection of legal documents. Our focus is not on the extraction task but on the annotation task. Even so, Ji et al. (2020) illustrate the value of a relevant manually annotated dataset for the development of an effective method for legal information extraction.

For any NLP task, languages not widely researched worldwide, like Portuguese, suffer from the lack of annotated corpora. The main barrier is the cost of building large and reliable datasets. Also, the annotation quality might be a limiting factor. According to Galgani et al. (2015), "*the persistent difficulty with building more sophisticated NLP systems is the extraordinary variety in which certain content may be presented in natural language*". Popular works like Luz et al. (2018), in Portuguese, have relied on the collaboration of a few annotators with specialized knowledge in linguistics, but in order to annotate Brazil's court rulings, legal knowledge and expertise are a must. As academic researchers, we decided to stay in the academic environment and invite students with specialized knowledge in law. To make it possible, we developed an approach for fine-grained annotation tasks conducted in a law school.

Another option would be hand-coded techniques like those presented by Dias et al. (2020). They present a study focused on NER for sensitive data discovery in Portuguese (from Portugal). In a study closely related to the legal domain, the objective was to create a system that allows organizations to identify citations to sensitive content, *e.g.*, personal data, in their collection of textual documents. It could enable organizations to have confidence in their data security and comply with protocols and regulations imposed, such as the General Data Protection Regulation (GDPR) (EU, 2018). Their solution relies on hand-coded techniques for named entity identification and classification. Their approach combines several methods such as rule-based models, lexical-based models, and machine learning algorithms for the NER task. The rule-based model is a pattern-matching solution for classifying standardized data, such as postal codes, email addresses, and date formats. The lexicon-based model combines the results of morphological analysis, a set of lexicons (every entity has its own different lexicons), and techniques of stemming and lemmatization for the recognition of entities, such as "person", "role", "medical data", "value" and "time". Since those two modules are hand-coded, most of the effort resides in building the knowledge base, strongly dependent on the patterns found in the Portuguese — written and spoken in Portugal. They also used machine learning algorithms to produce the final result, where they tried algorithms, such as Random Forest, CRF, and BiLSTM-CRF — CRF had the best result with 65.50 F1. However, as will be presented in Section 4.1, a pattern-matching solution fails to capture the legal NE mapped in our work. Thus, we decided to address it in future work as part of a strategy to reduce the annotation effort by suggesting possible annotation to the annotator, *i.e.*, who could validate or not the suggestion.

In regard to the Brazilian legal domain, there are very few studies in this area; we found one dataset made available by Luz et al. (2018) only. This corpus is based on 70 legal documents, where 66 of those are decisions published by at least 10 different Brazilian Courts and four legislation documents. With flat categories of named entities, the dataset is composed of six legal entities: “person”, “organization”, “time”, “location”, “legal case”, and “legislation”. Only the last two categories are closely related to the legal domain, as “legal case” refers to the citation of precedent and “legislation” to statute dispositions. Due to the entities’ flat nature, the corpus tends to represent those categories in a very simplified way that glosses over important fine-grained details, a shortcoming also present in Chalkidis et al. (2017), and Leitner et al. (2019). Even so, their corpus is frequently used in studies related to NER task in Portuguese, such as in Wang et al. (2020), which presents an NER approach based on a pre-training model to improve the task results. Furthermore, the corpus proposed in Luz et al. (2018) has a large breadth, covering a large amount of Courts with few documents in each. On the other hand, our work targeted a more representative dataset in order to provide enough data to develop information extraction.

3.2 Temporal Data Analysis over a Collection of Documents

In NLP, the concept drift or diachronic semantic shift studies how the meaning of a word changes over time (Li and Tian and Wang, 2021). As stated in Hamilton et al. (2016a) understanding how words change their meanings over time is key to models of language and cultural evolution. As illustrated in Hamilton et al. (2016a), the very meaning of a word changes over time. For instance, in the early 20th century, the word *gay* referred to “*cheerful*”; in the 50s shifted to referring to “*frolicsome*”; and in the 90s, the meaning shifted to referring to homosexuality. Nevertheless, in the legal context, the concept drift of a word, beyond the cultural/social change, might also reflect a change in the decision-making process or the decision reasoning.

The rising of the word embedding techniques, such as the *word2vec* Mikolov et al. (2013), inspired many studies focused on concept drift detection in large and historical corpora Kulkarni et al. (2015); Hamilton et al. (2016a,b); Li and Tian and Wang (2021). The basic idea is to construct word embeddings for separate periods (time-windows) and then to compare a word context in those different embeddings representations. As an illustration, if the corpus has documents published between 1900 and 2000, and if the time-windows size is in decades, different embeddings would be trained exclusively with documents published within the same decade, resulting in ten different embeddings. The concept drift is measured based on the word position variance throughout the embedding representations.

However, those approaches based on word embedding representation have some issues that are difficult to overcome. First, the instability of the embedding training process: embedding algorithms are sensitive to factors such as the presence of specific documents, the size of the documents, the size of the corpus, and even seeds for random number generators (Antoniak, M and Mimno). A legal document (as presented in Sections 4.5 and 6.3) has significant variances in terms of the document size (in number of words) and corpus size for each time-window (considering periods of six months). Second, in order to compare word vectors from different time-windows, the process must ensure that the vectors are aligned to the same coordinate axes. However, the stochastic nature of the embedding model training implies that models trained on the very same corpus might produce vector representations where words have the same nearest neighbors but different coordinates.

In an attempt to overcome the instability issue, Antoniak (M and Mimno) suggests testing the statistical confidence of similarities based on word embeddings by training on multiple bootstrap samples. Regarding the difference of the coordinates, Kulkarni et al. (2015); Hamilton et al. (2016a); Li and Tian and Wang (2021) proposes an approach for the embedding coordinates alignment. Even overcoming those issues, those approaches must be very useful for an investigation word-by-word but lacks interpretability for its results due to their nature.

3.2.1 The Dynamic Topic Modeling

Another NLP field that addresses the temporal property within corpora is dynamic topic modeling (Greene and Cross, 2017; Ambrosino et al., 2018; Haddock et al., 2020; Ahn et al., 2021). While topic modeling is a text-mining technique aiming to discover the hidden (latent) thematic structure (topic) in a collection of documents, ignoring the temporal aspects of the corpus, the dynamic topic modeling goes further. It investigates how latent topics emerge, evolve, and fade over time in a collection of historical documents — or any collection of documents that can be ordered by a temporal feature (*e.g.*, creation or published date).

Most studies on topic modeling focus on two methods: the Latent Dirichlet Allocation (LDA) and Nonnegative Matrix Factorization (NMF). These methods ignore the temporal property of the data. However, O'Callaghan et al. (2015), Ambrosino et al. (2018), and Greene and Cross (2017) present approaches that employ these methods for dynamic topic modeling by splitting the corpus in different time-windows — Ambrosino et al. (2018) broke a collection of publications, published between 1800-2016, into time-windows of 10 years —, and them performed LDA or NMF in every window. In Greene and Cross (2017), the authors demonstrate that NMF-based approaches are effective in identifying niche topics with more specific vocabularies. Even so, the result is usually a sequence of tables presenting the different topics generated over each time-window and the set of words related to each topic. It lacks interpretability, turning the evolution tracking of a given topic an additional burden. Our proposal includes a graphic representation of textual evolution such as to aid with interpretability.

3.2.2 Approaches Based on Tensor Decomposition

Kolda and Brett (2009) published a survey providing a long and valuable overview of higher-order tensor decomposition and its applications. The authors highlighted the growing interest in this subject in 2009. They observed that, since the first publication of an application of tensor decomposition for data analysis in the 1970s, the interest had grown substantially, especially by the end of the 1990s. In the late 1970s, the applications were limited to two fields, psychometrics and chemometrics. In the later 1990s, the applications have expanded to many other areas like signal processing, numerical linear algebra, computer vision, numerical analysis, data mining, graph analysis, neuroscience, and more. Recently, Haddock et al. (2020) proposed a new method for dynamic topic modeling based on higher-order tensor decomposition. A method that can overcome the issues observed using NMA and LDA for that type of application, and much of our work is based on the process and results presented in Haddock et al. (2020) and Ahn et al. (2021).

Still, according to Ahn et al. (2021), the crucial step of dynamic topic modeling is to decompose high-dimensional tensors into interpretable representations with attention to temporal information. However, they argue that traditional methods based on NMF fail to capture the data's temporal properties, which usually are represented by the third mode of the tensor. The main reason is that those approaches perform the NMF on tensor slices independently, neglecting the temporal model. They propose the nonnegative CANDECOMP/PARAFAC (CP) decomposition (NNCPD) to overcome this issue. Unlike traditional approaches, the NNCPD treats the tensor as a whole because the method specializes in the CP decomposition to decomposing a nonnegative data tensor into the sum of rank-one tensors (as presented in Section 2.3). The experiments demonstrate the method's capability to present how latent topics emerge, evolve, and fade over time. An important experiment performed by them explores the coherence of the produced results for dynamic topic modeling in a corpus with a collection of approximately 20,000 text documents containing the text of messages from 20 different internet discussion groups, classified into six super-groups: (i) computers, (ii) for sale, (iii) sports/recreation, (iv) politics, (v) science, and (vi) religion. The NNCPD was able to infer topics coherent with these super-groups. Our work used the same strategy to explore a large collection of legal decisions, but we took some steps forward. We also use the method to explore (i) the evolution of the court's vocabulary over time, (ii) the role of different legal named entities over time, and (iii) the differences over collections of decisions written by different Justices.

Haddock et al. (2020) confirms the results observed in Ahn et al. (2021)[1], regarding the NNCPD robustness, by presenting a comparison between approaches for dynamic topic modeling based on NMF and NNCPD. The authors show experimental evidence that with the NMF the noise can have devastating effects on the learned latent topics and obscure the true topics in the data. Meanwhile, the NNCPD is robust to noise in data even when the number of latent topics is overestimated — what is particularly important when the number of topics of the tensor data is unknown. Also, the method's robustness to noise is a welcome feature to overcome possible noise generated by the text extraction process from the file. Despite the legal documents presenting a lower probability of noise generated by mistyping, still, there is a noise generated by the text extraction process (legal documents are usually published as PDF files). Our experiments also demonstrate the method's robustness by comparing results with different scenarios: from least noise-prone to the most noise-prone.

Kassab et al. (2021) present a study over the application of the NNCPD for Short-lasting Topics detection. The paper is still an *e-print* version and, to the best of our knowledge, not peer-reviewed as of yet. Nevertheless, the results present the potential of the NNCPD application for tracking time evolution throughout latent topics and detecting the short-lasting topics. They have performed experiments using two datasets: news headlines and tweets related to the COVID-19 pandemic. With these datasets, they structured

[1] The *e-print* of the paper Ahn et al. (2021) was first released at arXiv on 01/02/2020. The study presented in Haddock et al. (2020) was based on that *e-print* version, still available at `https://arxiv.org/abs/2001.00631`.

them as tensors and applied different strategies for tensor decomposition and dynamic topic modeling. By comparing the results over each dataset, they could demonstrate that NCPD is a powerful dynamic topic modeling technique capable of detecting short-lasting and periodic topics along with long-lasting topics in a dynamic text dataset. The strategy employed for the tensor structure of this experiment is a result of the improvement of a strategy presented by Ahn et al. (2021) in an experiment for dynamic topic modeling in a collection of texts related to four different subjects. As in Kassab et al. (2021), the strategy presented in Ahn et al. (2021) was our starting point to define our strategies for the tensor structuring, where we took into account details regarding monocratic decisions to propose coherent strategies for tensor structuring.

4
Legal Named Entity Recognition

In this chapter, we present the process employed for a fine-grained annotation task of court rulings by law students, in which two levels of nested legal entities were annotated. To ensure reliability and capture the nuances of legal reasoning, the annotation task had the collaboration of 95 law students from the *Escola de Direito do Rio de Janeiro da Fundação Getúlio Vargas* (FGV-Direito Rio Law School). This collaboration was possible thanks to the alignment between teaching and research: the task was performed as part of the coursework of two disciplines focused on law and technology (further details in Section 4.3). The annotation process was composed of two shorter training efforts and one longer final task. To measure the quality of annotations and adjust the guidelines and instructions when needed, we used Cohen's Kappa to measure the inter-annotator agreement throughout the task.

As stated before in Chapter 1, we focus on monocratic decisions for the dynamic topic modeling. However, for the annotation task, the collegiate decisions were included since this task's main goal is to introduce the first version of the largest corpus known in Portuguese dedicated for legal named entity recognition. By doing this, we intend to support further studies in NLP and the legal domain.

In the following Section 4.1 we present a brief review over a preliminary step to evaluate the usage of a simple pattern-matching solution for legal NER. It was also an important prior step for the selection of the decision for the annotation task, which is presented in Section 4.2. Section 4.3 presents the annotation task process employed to build the annotated corpus, presented in 4.4. Finally, Section 4.5 presents how the corpus was used to extract the legal elements from the collection of monocratic decisions.

The discussions and results presented in this chapter are also detailed in our publication, entitled *Fine-grained legal entity annotation: A case study on the Brazilian Supreme Court*, published by the Journal *Information Processing & Management* (Correia et al., 2022). Furthermore, it is important to state that the results presented in this chapter are also a result of a collaborative work of a multidisciplinary team — they are also authors of the above mentioned publication.

4.1 A Preliminary Step with a Pattern-Matching Approach

As a preliminary step, we developed a precedent citation extractor using regular expressions. Unfortunately, there is no formal standard procedure for precedent citation in the Brazilian judiciary. Thus, we tracked mentions of legal procedures on legal decisions by searching for a limited number of combinations of the procedural classes and the number each case receives upon filing. An algorithm limited by common writing patterns used by the Justices of the STF: the legal procedure class (72 possible classes) followed by a sequential number with less than six digits[1]. This resulted in a noisy extraction. For instance, it does not allow us to identify which opinion was cited as precedent if the cited case contained multiple rulings, *e.g.*, a monocratic decision followed by a collegiate decision on appeal. However, in a prior work Correia et al. (2019) we used this method to conduct an exploratory investigation over the precedent network extracted from 1,152,963 decisions published by the STF between 2008 and 2018. Our analysis revealed interesting temporal trends in precedent use. As a result, even though we got a noisy extraction, it was good enough to measure the temporal relevance of the precedents.

We have processed the whole collection of decisions with this algorithm as a preliminary step before the sampling. We considered it essential to ensure that the sample would contain a large number of entities to be identified. That was important not only because we already had direct evidence that one of our coarse-grained entities was present but also because our prior is that there is a correlation between a higher number of precedents citations and the likelihood of the opinion citing academic works. Furthermore, the results of this preliminary step set an important baseline for further results comparison.

An important point to state is that we were unable to recover most of the nested entities mapped for this study with a pattern-matching solution. Furthermore, despite our best efforts, we were unable to create a reliable pattern-matching solution to extract academic works' citations and their inner entities.

4.2 Selection of Decisions

Most of the collegiate decisions published before 2009 were PDF files generated from a rudimentary image scanning process. Unfortunately, this issue was too hard to overcome and assure the quality of the extracted text.

[1] A version of this the algorithm adapted for both courts STF and Superior Tribunal of Justice (STJ) are available in `https://github.com/joseluizn/extrator`

Meanwhile, at that time, the monocratic decisions were published in easy-to-parse text files. However, to keep a balance, we decided to establish 2009 as a starting point for the selection process.

Thus, for the annotation task, we assembled a set of 764 decisions published between 2009 and 2018. They were selected according to the following criteria: (i) At least one precedent was identified in the decision content using the method described earlier; (ii) 50% of them were monocratic decisions, and 50% were collegiate decisions, and (iii) we stratified our sample by its seventeen possible origins: fourteen for monocratic decisions (eleven Justice from Court composition in 2018 plus three other Justices that were in previous composition [2]), and 3 for collegiate decisions (both Panels and the Plenary), see Table 4.1 for the distribution summary.

In addition to those criteria, we were able to ensure a minimum representation of decisions per year of publication. On average, 76 decisions per year, the year 2009 was the least represented with 49 decisions, and 2018 was the most represented with 172 decisions. Aside from these years, the distribution is close to uniform for the remaining years[3].

Table 4.1: Distribution of decisions per origins.

Decision type	Cabinet	Justice/Panel	Qtd.
Monocratic	1	Justice Cármen Lúcia	33
	2	Justice Celso de Mello	40
	3	Justice Marco Aurélio	20
	4	Justice Luiz Fux	40
	5	Justice Gilmar Mendes	29
	6	Justice Ayres Britto	8
	6	Justice Roberto Barroso	27
	7	Justice Ricardo Lewandowski	32
	8	Justice Dias Toffoli	45
	9	Justice Joaquim Barbosa	15
	9	Justice Edson Fachin	19
	10	Justice Teori Zavascki	21
	10	Justice Alexandre de Moraes	14
	11	Justice Rosa Weber	39
Collegiate	-	Second Panel	91
	-	First Panel	138
	-	Plenary	153

The second and third criteria were included to ensure that the corpus was somewhat representative of STF's output. As presented in Table 4.1, the

[2] Justices Ayres Brittos and Joaquim Barbosa who retired in 2012 and 2014, respectively, and Justice Teori Zavascki who died in 2017; Justices Roberto Barroso, Edson Fachin and Alexandre de Moraes were selected to take up those vacancies, respectively.

[3] We opted not to stratify the sample by year since the year has a strong correlation with the reporting Justice, since there were some changes in the composition during the time period (*i.e.*, Justice Alexandre de Moraes joined the Court in 2017) also because it would lead to a large amount of strata which would be hard to balance with the workload we could take for the annotation task.

origin with the fewest decisions was Justice Ayres Britto deciding individually, with only eight decisions, and the Justice with the most extensive collection, was Justice Dias Toffoli, with 45 monocratic decisions. Because the decisions in the scope we were sampling from were usually long, to ensure uniformity in size and workload per annotator, we broke long decisions into smaller text documents[4]. As a result, the initial corpus of 764 decisions became a collection of 1,363 excerpts.

4.3 The Annotation Task

We used two different groups of students for our annotation – one of 54 and the other of 85 students from the third and second years of law school. These students were taking Courses dedicated to providing law students with Python programming knowledge. This initial cohort of 139 annotators enabled us to select only the work of those who performed best, discarding poor quality annotations but still retaining a substantially large dataset. The final annotated corpus used for the tests in this study results from the contribution of only 95 students: 48 from the first group and 47 from the second. They were selected based on the average inter-annotator agreement measured with Cohen's Kappa (further details in section 4.4).

We are committed to the use of open-source annotation tools. After exploring various web-based, open-source annotation solutions, we chose the Doccano project[5]. We considered that Doccano struck the best cost-benefit balance for our goals. Every annotator engaged in the task had their own Doccano account where annotation tasks were automatically assigned. Due to the tool's limitations, we split the annotation task into two specific periods. The first was dedicated to annotating the coarse-grained elements from the ruling excerpts, and the second was oriented toward fine-grained annotation from the previous coarse annotations. Figure 4.1, presents an illustration of the annotation process.

In the first step, all the excerpts were randomly distributed among annotators through Doccano. Also, to ensure a similar workload, each participant received the same amount of excerpts, and every excerpt was delivered to at least two different annotators. We asked the annotators to tag only the four coarse legal entities in the excerpt in the coarse-grained annotation. Once completed, all the coarse annotated entities were saved into the database.

[4] The threshold was 30,000 characters, something close to 12 pages, larger documents were divided in chunks of 30,000 characters

[5] `https://github.com/doccano/`

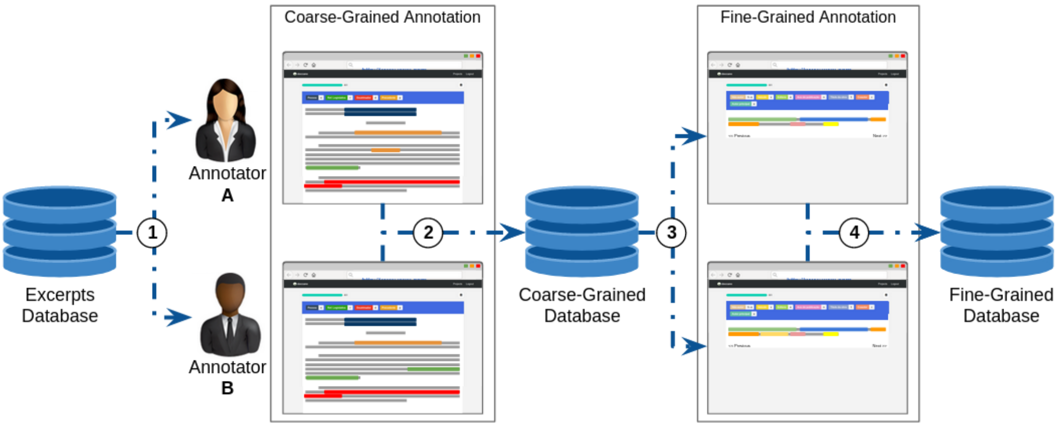


Figure 4.1: Annotation task process. (1) excerpts are randomly distributed among annotators for the coarse-grained annotation; (2) - the coarse annotations are saved into the database and pooled together; (3) - these coarser annotations are randomly distributed among annotators for the fine-grained annotation; and (4) the fine-grained annotations are saved in a specific database.

In the second step, all unique coarse entities were randomly distributed among students for the fine-grained annotation regardless of who performed them. Again, every annotator had the same workload. Also, every coarser annotation was delivered to at least two annotators. In the Doccano interface, these annotations were grouped by types (precedent, academic citation, and legislative reference). We then asked students to perform the fine-grained annotation.

At the end of the second step, all fine-grained annotations were properly saved in different databases. It is important to state that equal coarser annotations taken from different excerpts were treated as a single one in the second step. Two annotations will be considered equal if they have the same sequence of symbols, constituting literal copies. If *Annotator X* tagged the sequence "*§ 4º do art . 103 - B da CF*" as a legislative reference in *excerpt 01*, and *Annotator Y* has also tagged the same sequence as a legislative reference, but in *excerpt 02*, both sequences will be treated as if they were only one regardless of their origins. As a consequence of this strategy, we had a natural reduction in the total amount of coarse annotations to be annotated in the second step.

All annotations carry a reference to their origins, including the identity of the annotator. Hence, by combining saved data from the three databases, as presented in Figure 4.1, it is possible to reconstruct the excerpt with its coarse and fine-grained annotations. Nevertheless, that approach cannot guarantee that all the texts presented in the second step will indeed be annotated. The assigned annotator might refrain from performing a fine-grained annotation for different reasons, *e.g.*, they might have failed to do it within the deadline, ignored that specific annotation altogether, or encountered a portion of text that was incorrectly identified in the previous step, such as when a precedent is tagged as a legislative reference.

4.3.1 Training Tasks

In order to build the corpus, we performed two shorter training annotation tasks, followed by a final and long task. During the whole procedure, follow-up meetings were held to present the results of each phase and to discuss possible difficulties the students might be encountering. Also, the annotator was allowed to make revisions over their annotations until the end of each step. So they could consult our collaborators and fix possible misunderstandings. Such meetings, along with the inter-rater agreement measured between phases, played an essential role in improving the guidelines, student annotation performance, and accuracy, resulting in overall better annotations.

We selected a collection of ten documents for the training phases: four documents used in the first training task and six in the second. Every student performed the training tasks over the same collection of excerpts and was given two weeks to conclude each training task: one week for each annotation step. For these ten documents, experienced researchers produced a golden set from our team, enabling us to track and compare student annotation performance in the training phase.

For the final task, students had four weeks to conclude annotations: two weeks for coarse entities and two weeks for fine-grained entities.

4.3.2 Inter-annotator Agreement

From the very beginning, we closely followed students' progress, proactively identifying problematic issues that surfaced during the training tasks and discussing them with annotators at meetings. To do so, we kept an open communication channel with the annotators throughout, allowing them to share their difficulties and propose guideline improvements. Furthermore, to measure the evolution of annotations' quality of, we have used Cohen's Kappa as an inter-annotator agreement measure.

Cohen's kappa is one of the most common agreement metrics found in the literature, although it is not without its shortcomings (Krippendorff, 2004). Same as in other studies, Chalkidis et al. (2017) use $\frac{|AB|}{max(|A|,|B|)}$ as an inter-annotator agreement measure, where A and B are sets of contract elements marked by the two annotators, respectively. Wyner and Peters and Katz (2013) argue against the use of Cohen's kappa for problems concerning the annotation span and how such a task is difficult to conceptualize as a classification task — instead they measure the precision, recall, and F1 measure between annotators.

The conceptualization as a classification task concerns the idea that not annotating something is different from annotating it as a "negative example" of your standard, which would be needed for the Cohen's Kappa coefficient. To find out how each of these alternatives would affect our final results, we decided to compare all three after the annotation task before building the final corpus presented here. Again, the large number of students participating in the task and the ensuing significant number of total annotations enabled us to discard thousands of annotations and still be left with a substantially sized corpus.

In all cases, we expected that most of each excerpt would not be annotated and reflected this assumption in our calculations. After all, only a fraction of the length of a judicial opinion can be tagged with our entities, regardless of coarse or fine-grained. As expected, 85.21% of the tokens in the corpus were not annotated. Looking for the overlap between two annotations without taking this into account might create distortions in all metrics' result due to highly different marginals and higher agreement than expected. As such, we decided to remove all cases where both annotators left the tokens without any tag. This meant we effectively excluded a large number of tokens from our measurements where both students agreed there was no entity to be marked.

We then calculated the Pearson's correlation between the Cohen's Kappa, the F1 measure, and the percentage of agreement on tokens of all pairs of students at the final task in order to check whether different agreement metrics would render us different results. We found a correlation of over .9 between them, thus showing that the selection among any of these agreement metrics itself was not decisive to our results.

4.4
The Annotated Corpus

As mentioned before, two groups of students over almost a year performed the annotation task under close supervision. At introductory meetings, we explained the activity's academic impact in the legal domain by emphasizing the importance of this annotated corpus to our study and to other studies that will follow it.

Both groups performed annotations on the same collection of 1,363 excerpts. Despite the distribution criteria presented in Section 4.2, there was no guarantee that every excerpt distributed would indeed be annotated since the number of annotations per student varied greatly. Unfortunately, in the second group, there were dropouts at the end of the final task, which reduced the number of effective participants in the fine-grained annotation step to less than half — only 48% of the participants completed the fine-grained annotations.

Not all pairs of students had an excerpt in common, which means that the agreement distribution of a given student with the others did not span everyone else. Each excerpt was annotated on average by 5.4 students.

To build the final corpus presented here, the data annotated by both groups were treated as a single collection. Also, excerpts were filtered based on who made the annotation by selecting the annotators with an average Kappa score above 0.7. We have measured the average inter-agreement score for every annotator considering four different scenarios: the excerpt annotation and the three types of fine-grained annotation. As a result, we could take a closer view of the inter-agreement score and how the annotation quality varies between scenarios. The more frequent the high scores are, the more widespread the understanding of the annotators about the entities.

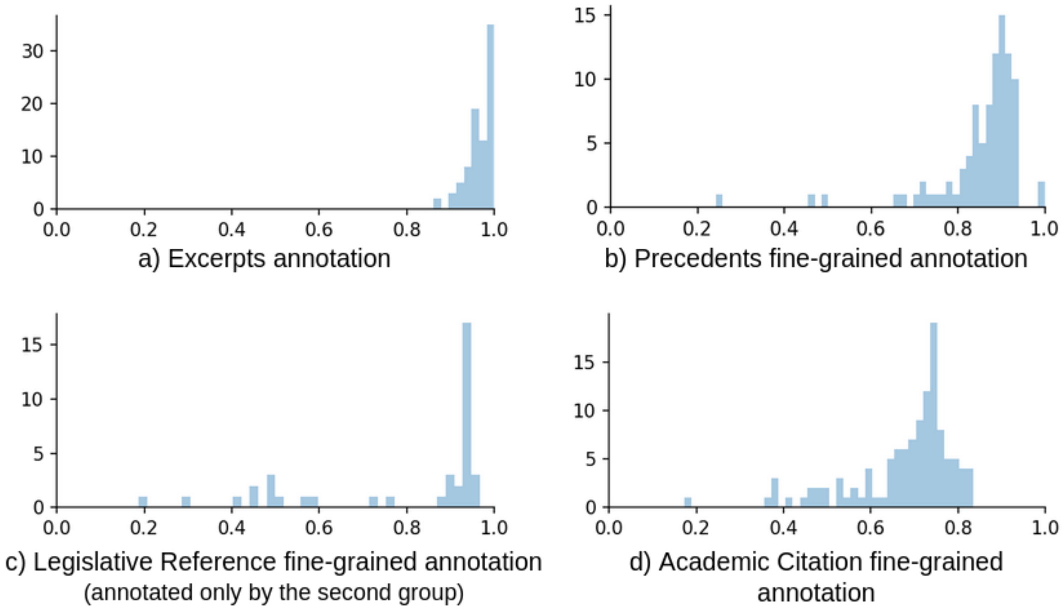
The scenario where we had the lowest agreement was the excerpt annotation for coarser entities, as presented in Figure 4.2.a, something already expected since the coarser annotation demands more effort than the fine-grained annotation performed in the second step. An excerpt has an average of 16,087 tokens and 55.84 coarser annotations. Meanwhile, a precedent (Figure 4.2.b) has an average of 10.28 tokens and 3.69 fine-grained annotations. Even so, the fine-grained annotation in precedents seems more difficult than in legislative references (Figure 4.2.c), which in turn, seems more difficult than in academic citations (Figure 4.2.d).

With this criterion, we selected what we considered to be the best annotators for each scenario. As a result, Table 4.2 presents the number of annotators that passed the criterion. The legislative reference fine-grained annotations were performed only by the second group[6]. In comparison to the other elements, it was the most diverse in structure. Also, as a result, the amount of annotated data dropped after the filtering.

4.4.1 The Annotations Merge

The presented corpus is the result of the combination of all annotations performed by those different selected annotators. Since the annotation task has two different types of texts to be annotated, excerpts, and coarser annotations,

[6] We used the first group annotations (coarser annotations) to understand its format better. We then acted on this knowledge to prepare the fine-grained annotation format used for the second group.



a) Excerpts annotation

b) Precedents fine-grained annotation

c) Legislative Reference fine-grained annotation (annotated only by the second group)

d) Academic Citation fine-grained annotation

Figure 4.2: Distribution of inter-annotator agreement score for different types of annotated text.

Table 4.2: Selected annotators.

	Annotators with Kappa score > .7
Excerpts	67
Precedents	76
Academic Citations	75
Legislative References	28

the annotation merge process has two moments. First, the annotation merges per text, regardless of the type, and then, rebuilding the link between the coarser and fine-grained annotation into one single, and final, representation per excerpt.

The annotation merging strategy employed was majority voting. The label for a word is picked based on whether most annotators agree with it — for the tiebreaker criterion, we used the vote of the annotator with the highest average inter-annotator agreement score. As stated before, we have selected the best annotators and reduced the amount of annotated data to ensure a higher quality of the merging process.

4.4.2 The Corpus

The presented approach's final result is a corpus with 595 annotated excerpts, built from 532 decisions — the initial set contained 1,363 excerpts from 764 decisions. Table 4.3 presents a summary of the count of sentences, tokens, coarse and fine-grained annotations in the corpus. The variation in size of the excerpt is due to the decision-splitter strategy employed after the decision selection (see Section 4.2), but also a consequence of the high number of short decisions in the initial set of 750 decisions – 50% of those were monocratic decisions, which are often very short rulings.

Table 4.3: Counting summary of sentences, tokens, coarse and fine-grained annotations.

	Total	Per excerpt				
		min	max	average	std	median
Sentences	62,933	3	551	105.97	81.33	93.0
Tokens	1,782,395	121	16,087	3,000.66	2,501.79	2,692.5
Coarser-grained	33,055	1	267	55.65	42.44	47.0
Fine-grained	57,573	0	507	96.92	69.21	81.0

As presented in Table 4.4, every excerpt has at least one coarser annotation tagged. The most popular coarser element is the person, followed by the legislative references and the precedent. The rarest is the academic citation, which is also the element with less fine-grained annotations: only 54.37% of the occurrences were annotated. Despite the higher inter-annotator agreement (presented in Figure 4.2), a considerable amount of those elements were not annotated by the participants.

Table 4.4: Coarser-grained entities count.

Legal Elements	Per excerpt	Total	With fine-grained	Avg. tokens
Precedents	15.33	9,108	8,486 (93.17%)	10.27
Acad. Citations	2.99	1,775	965 (54.37%)	24.60
Leg. References	17.22	10,229	8,439 (82.50%)	8.41
Person	20.11	11,943	N/A	3.38

The shorter the text, the greater the chances of it being fully annotated. The legislative reference element illustrates this: it is the smallest text entity, with an average size of 8.42 tokens. As mentioned before, this element's fine-grained annotation task was introduced only to the second group, and 28 annotators have contributed to the final corpus. Even so, the legislative reference had 94.05% of the occurrences in the corpus annotated with the fine-grained annotations.

The table 4.5 shows how diverse is the distribution of fine-grained entities in the corpus. The Reporting Justice (from precedents) and Legal Act (from Legislative References) are the most popular in the corpus. Meanwhile, entities like Institution and Origin (both inner elements of Legislative References) are rare in the corpus – only 79 tokens were tagged as Institution in the whole corpus.

Table 4.5: Fine-grained entities count.

Fine-grained entity	Total of tokens	%	Avg. per coarser annotation
Court	3154	0.18%	0.37
Decision_Date	728	0.04%	0.09
Decision_Type	2561	0.14%	0.30
Legal_Procedure_Class	10970	0.62%	1.29
Legal_Procedure_Number	9562	0.54%	1.13
Origin	4279	0.24%	0.50
Publication_Date	2426	0.14%	0.29
Reporting_Justice	18825	1.06%	2.22
Author	3107	0.17%	3.22
Co-author	493	0.03%	0.51
Collection_Title	569	0.03%	0.59
Publisher	1060	0.06%	1.10
Title	5969	0.33%	6.19
Year_of_Publication	802	0.04%	0.83
Clause	608	0.03%	0.07
Institution	79	0.00%	0.01
Legal_Act	19549	1.10%	2.32
Origin	305	0.02%	0.04
Paragraph	2836	0.16%	0.34
Section	7735	0.43%	0.92
Subsection	2813	0.16%	0.33

To better support further research using this corpus to test algorithms for legal element extraction, we have also added part-of-speech (POS) information. Every token from every sentence was tagged with its corresponding POS tag by its context in the sentence. The POS tagging process was performed with the spaCy [7] library for text processing with Python. The corpus generated during this study is available under request[8], where every annotated excerpt is in CSV format. Each line in the file is made of five columns separated by space, which identify in order: (1) the sentence ID, (2) the token, (3) the POS tag, (4) the coarser-grained tag, and (5) the fine-grained tag. Table 4.6 present an example of tagged text with a Precedent.

For the coarse and fine-grained entities, the tokens tagged with O are outside of named entities, the B_* prefix tag indicates the first token of a named entity, and I_* prefix indicates all following tokens of the same entity. Nevertheless, in cases where coarse entities have no fine-grained annotation, the last column is filled with X, such as in Table 4.7:

[7] `https://spacy.io/`

[8] The dataset generated during the current study is available from the corresponding author on reasonable request.

Table 4.6: Example of tagged content with a Precedent.

Sent. ID	Token	POS	Coarser-grained	Fine-grained
9b4f40	;	PUNCT	O	O
9b4f40	Inq	PROPN	B_Precedent	B_Legal_Procedure_Class
9b4f40	2126	NUM	I_Precedent	B_Legal_Procedure_Number
9b4f40	,	PUNCT	I_Precedent	O
9b4f40	Relator	NOUN	I_Precedent	B_Reporting_Justice
9b4f40	Sepúlveda	NOUN	I_Precedent	B_Reporting_Justice
9b4f40	Pertence	NOUN	I_Precedent	I_Reporting_Justice
9b4f40	,	PUNCT	I_Precedent	O
9b4f40	Pleno	NOUN	I_Precedent	B_Court
9b4f40	,	PUNCT	I_Precedent	O
9b4f40	DJe	NOUN	I_Precedent	B_Publication_Date
9b4f40	26.04.2007	NOUN	I_Precedent	B_Publication_Date
9b4f40	.	PUNCT	O	O

Table 4.7: Example of tagged content without fine-grained annotation.

Sent. ID	Token	POS	Coarser-grained	Fine-grained
62ef86	da	NOUN	O	O
62ef86	Resolução	NOUN	B_Legislative_Reference	**X**
62ef86	nº	NOUN	I_Legislative_Reference	**X**
62ef86	80	NUM	I_Legislative_Reference	**X**
62ef86	do	PROPN	I_Legislative_Reference	**X**
62ef86	CNJ	NOUN	I_Legislative_Reference	**X**
62ef86	–	PROPN	O	O

4.4.3 Results

As a result, we built a manually annotated corpus based on those annotations that exceeded our criteria of quality: excerpts annotated by at least two annotators and those with an average inter-annotator agreement score greater than 0.7. Therefore, we present the largest corpus of Brazilian legal decisions manually annotated for the NER task with nested legal entities — to the best of our knowledge. The given corpus contains excerpts from 594 decisions (62,933 sentences; 1,782,395 tokens; 33,055 coarser-grained annotations, and 57,573 fine-grained annotations). We also experimented with the presented corpus with two different NER strategies: Conditional Random Fields (CRF) and bidirectional Long-Short Term Memory Networks with CRF (BiLSTM-CRF) for both levels of NE.

Unlike the results presented in Correia et al. (2019), for the coarser-grained entities, the BiLSTM-CRF model outperformed the CRF for the NER task where the F1 score exceeds 0.9, but for the fine-grained entities recognition, the CRF outperformed the BiLSTM-CRF. The performance improvement observed with BiLSTM-CRF was due to the improvements made on the pre-trained word embedding. We trained a *word2vec* with all decisions in our collection, while in Correia et al. (2019) the model was trained with 200 thousand decisions. We also carefully evaluated textual content used in the training, resulting in a dataset with less noise, improving the model performance.

Due to the quality of the BiLSTM-CRF observed when in the "out word", we opt to use the same approach for fine-grained entities. The study still lacks a conclusive significance test[9] even so, the observed results outperform the results presented in Correia et al. (2019) for this type of model. One of the issues observed in Correia et al. (2019) regarding the BiLSTM-CRF performance was that much of the words present in the corpus could not be found in the vocabulary of the pre-trained word2vec. Since our embedding is trained based on the whole collection, this scenario is less likely to happen.

4.5 Building the Extractor

As stated in Section 2.1, due to the STF's plurality of different kinds of jurisdictions, a massive amount of new cases are filed at the court every day, and most of them are appeals that seek to reverse rulings from lower courts on constitutional grounds. The court has developed institutional changes such as the general repercussion requirement to cope with this demand. As a result, since 2017, the productivity score (the percent ratio of the number of cases ended in a year by the number of new cases filed) was above 100% (Supremo Tribunal Federal, 2021). However, there is a long way down untill the court finally meets with its demand. Despite the effort, the average time for a case between gets its final decision is close to a year – in 2020 it was 314 days – and there is an enormous amount of old processes pending for a final judgment; in 2020 10% of the open cases were more than five years old.

As first observed by Hartmann and Chada (2015), the strategy to copy and past decisions content is also a strategy to leverage the court's production. It is reasonable that similar cases must have similar decisions also, much of the appeals that reach the court are about issues already decided by the court. So much so that the court started a project to develop artificial intelligence to help them to select the appeals regarding issues already decided[10].

Regarding the production of similar decisions, we have analyzed the content of the decision in order to propose a system for the NER in the

[9] To verify the significance, in Correia et al. (2019) we have performed a the *5x2cv* combined F test procedure to compare both algorithm's performance

[10] A Project called Victor that is an ongoing adoption process, as noticed by the court in `https://portal.stf.jus.br/noticias/verNoticiaDetalhe.asp?idConteudo=471331&ori=1,Lastaccessedon03/20/2022`

whole collection. We employed much of this proposal and now present our observations.

The first graphic in Figure 4.3 presents the total amount of monocratic decisions that we used in our study, distributed by the semester of publication. The two stacked areas represent the proportion between unique decisions versus copied decisions. By copied decisions, we mean a decision whose content can also be found in at least one other decision. To compare decisions, we removed content such as headers and footnotes, and then we computed a hash code for each decision's content. The comparison was made using the hash code[11]. The third graphic present the average size of a decision in words throughout time.

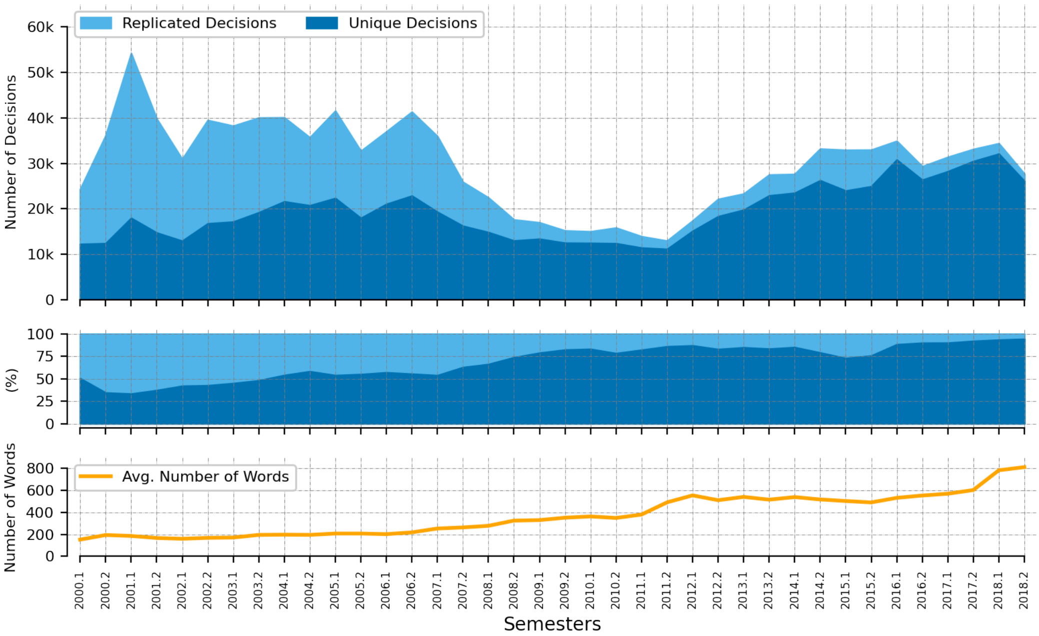


Figure 4.3: Decision uniqueness over the time.

The initial idea was to process only the "original" decisions and copy the results to its replicas. Observing Figure 4.3, it seams clear that the court has increased the production of original content and also that the average size of monocratic decisions increased over time. In 2018, we have much more unique content and substantially longer decisions when compared to 2006.2, for instance. Nevertheless, taking a step further, investigating the decisions' content by comparing sentences (phrases) among decisions reveals a more interesting scenario.

[11] To compute the hash code, we used the *blacke2b* algorithm with *hashlib* package for Python. More in: `https://docs.python.org/3/library/hashlib.html`

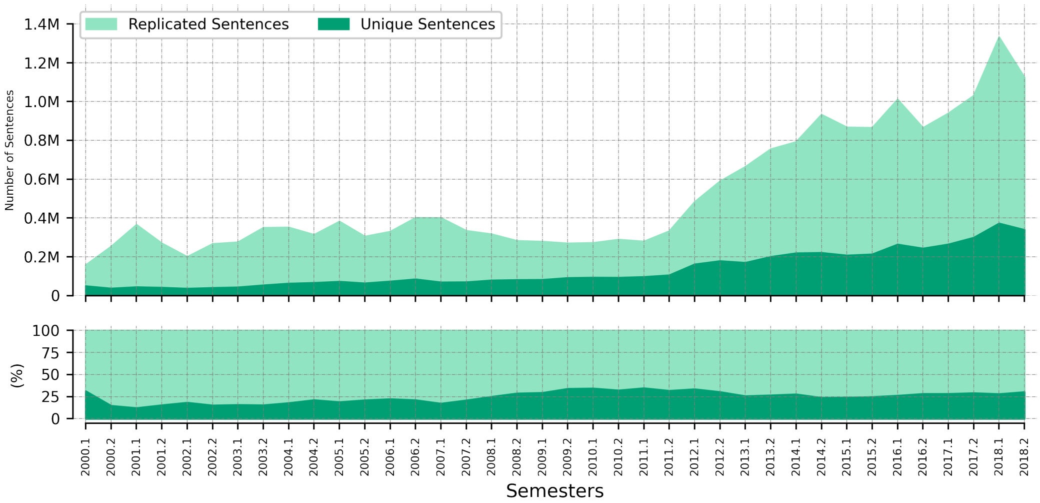


Figure 4.4: Sentence uniqueness over the time.

We broke down the content of the decisions into a set of sentences. We normalized the sentences by lowering the case and removing non-alphanumeric characters. Then we computed the hash code for each sentence, as we did with the content of the decisions, to verify its uniqueness in the whole collection of decisions. Figure 4.4 presents the distribution of sentences throughout the time, distinguishing unique sentences from the copies.

Despite the increasing production of new decisions, there are many similarities among them. In 2018, more than 90% of the decisions produced were unique content. However, only 25% of the content represents something really unique.

We have less than 5 million unique sentences for the period presented in both pictures. By processing only the unique sentences, we reduce by 80% the processing cost, even considering the data management challenge to keep tracking the decisions and sentences.

We have employed this strategy to prepare the data for some of the experiments presented in Chapter 6, by running the NER over the unique sentences. We also broke the NER task into steps: first extracting the coarser-grained entities and then extracting the fine-grained entities. So we could distribute the processing, making it even faster. To store and compare hash codes, the Redis[12] was employed.

[12] https://redis.io/

Table 4.8: Extraction Summary.

Year	Decisions	Extracted entities					
		Coarser-grained				Total	Avg. per decision
		Precedent	Legislative Ref.	Academic	Person		
2000	60,395	98,692	169,198	1,987	8,921	278,798	4.616
2001	93,915	146,411	258,340	5,698	15,373	425,822	4.534
2002	70,360	131,886	182,440	5,000	10,162	329,488	4.683
2003	78,214	199,612	238,037	4,950	11,944	454,543	5.812
2004	75,709	189,743	241,502	2,817	12,491	446,553	5.898
2005	74,208	227,983	245,306	2,975	16,096	492,360	6.635
2006	78,240	229,690	269,070	5,162	20,119	524,041	6.698
2007	61,808	242,780	252,215	4,347	10,464	509,806	8.248
2008	40,072	163,583	170,083	1,911	9,006	344,583	8.599
2009	32,197	143,970	148,926	1,159	10,204	304,259	9.450
2010	30,857	143,367	147,290	1,350	11,530	303,537	9.837
2011	26,912	154,425	149,647	3,419	13,383	320,874	11.923
2012	39,510	265,592	261,765	6,215	26,338	559,910	14.171
2013	50,791	338,189	349,360	5,366	30,289	723,204	14.239
2014	60,792	422,735	417,058	6,083	35,691	881,567	14.501
2015	65,857	434,862	410,761	7,901	36,095	889,619	13.508
2016	64,192	438,452	407,035	10,190	37,163	892,840	13.909
2017	64,461	471,504	441,019	11,688	46,708	970,919	15.062
2018	62,054	560,052	550,463	16,084	70,855	1,197,454	19.297

Table 4.8 presents an extraction summary with the quantities of extracted coarser-grained entities from the whole collection of decisions used in our study. In total, 10.8 million citations extracted: 48.93% are legislative references, 46.11% are precedents, 3.99% are references to persons (neither related to precedent nor academic citation), and only 0.96% are academics citations.

An important remark is to observe the rising usage of those entities throughout the time. In 2018, there were five times more citations than in 2000. This is a possible illustration of the effort put into the argumentation process to enrich the basis for the decision. Also, is remarkable the differences between the numbers from 2011 and 2012. In 2012 were published 46% more decisions than in 2011, but were 74% more citations. The difference is also pertinent when comparing any year after 2012 with any other year before 2012. In Chapter 6 we explore some of these differences.

5 Extracting Reliable Information from a Lager Collection of Documents

The tensor decomposition reveals patterns of events (latent topics) without an upfront classification. Also, a tensor analysis works in an arbitrary number of dimensions and thus can detect complicated relationships between several data attributes simultaneously (Henretty et al., 2018). For the dynamic topic modeling application with monocratic decisions from the STF, we propose two different approaches for tensor structure representation: the decision-oriented structure and the origin-oriented structure. Each reveals different information over the same object of study, the reasoning embedded in the monocratic decisions.

The differences between structures are regarding the first and the second mode. For both, the third mode is related to time, which might be related to annual or semiannual time slices. The first two following section presents both structures (Sections 5.1 and 5.1.1), and the Section 5.3 presents strategies for the second mode that represents the textual features.

5.1 Decision-Oriented Structure

The document-based structure is a proposal adapted from an experiment presented by Ahn et al. (2021) (which also inspired Haddock et al. (2020); Kassab et al. (2021)). In the experiment, Ahn et al. (2021) used a collection of texts grouped into six subjects to compare the associations of the latent topic with the subject variation. They used four groups to construct a tensor representation (“*for sale*”, “*baseball*”, “*atheism*” and “*space*”), and randomly select a fixed number of documents from each group. The text was preprocessed and computed the TF-IDF weights across documents to reflect how important a word is to a document in a collection of documents. The first mode of the third-order tensor represents documents, the second represents words, and the third represents time.

Figure 5.1 presents the proposed structure (Ahn et al., 2021). The top red part consists of “*for sale*” documents evenly spread across ten time periods (or time slices). Similarly, the blue part consists of “*space*” documents evenly spread across ten time slices. Lastly, the green part consists of "*atheism*" documents evenly spread across time slices 1-5, and the purple part consists of "*baseball*" documents evenly spread across time slices 6-10.

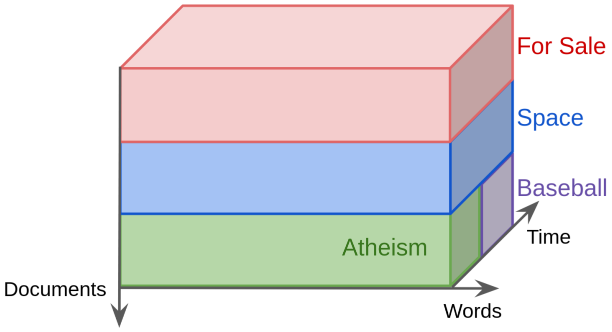


Figure 5.1: Visualization of the construction of the 4 Newsgroups tensor. Adapted from Ahn et al. (2021).

Ahn et al. (2021) computed the NNCPD for the entire tensor and analyzed the factor matrices A, B, and C resultant. Correlating topics (ranks), the words, document group, and temporal slices, they concluded that: *not only the keywords are meaningfully associated with the latent topics in the dataset, but also the NNCPD captures the topic variation across documents and exhibits the temporal topic information.*

Based on Ahn et al. (2021), on our proposal for the document-based structures, the second mode represents the document groups. For every time slice, we have a fixed number of decisions randomly selected within a group of decisions and published in the same time interval. For each group, an index interval is defined to set boundaries as in Figure 5.1.

The tensor structure resultant is a tensor, $\mathcal{X} \in \mathbb{R}^{d x w \times t}$, defined by $decision \times word \times time$, where $\mathcal{X}_{ijk}$ shows relevance score of the word j for the decision i in the time slot k, where the index of the decision i identifies the group witch it is related to.

5.1.1 Grouping Monocratic Decisions

There are different ways to group legal cases and, as a consequence, to group decisions based on the legal cases related to them[1].Every legal case filed in the STF is identified by a procedural class. The Court has 72 different procedural classes, and, as presented in Falcao et al. (2012); Correia et al. (2019), those classes can also be grouped into five main categories related to the different competencies of the STF:

[1] Decisions can also be grouped by type: if the decision is related to an internal appeal or a motion.

- **Appeal**: appeals that seek to reverse rulings from lower courts on constitutional grounds;
- **Writ**: decisions on writs, including those judged directly and on appeal which do not require constitutional review arguments;
- **Constitutional**: proceedings that deal with the constitutionality of laws in abstract;
- **Criminal**: when the Court acts as a regular fact finding case court for some federal public officials (*e.g.*, Senator, and Minister);
- **Others**: All remaining classes that did not fit the previous classifications.

Besides classes and categories, legal cases are also related to legal subjects (or "*Ramo do Direito*" in Portuguese) that indicate the main topics related to the case. The STF follows the CNJ norms which define 22 main topics[2]. As we present in Chapter 6, exploring the dynamic topic modeling based on any of those groupings, we can explore a different point of view over the decision reasoning throughout the latent topics. For instance, Figure 5.2 represents the proposed structure based on the legal case categories.

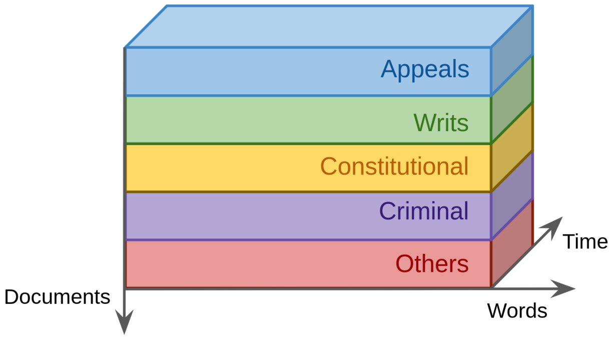


Figure 5.2: Visualization for a possible decision-based structure.

As we present in Chapter 6, exploring the dynamic topic modeling based on the document-based structure with different groupings strategies, we can explore different points of view over the decision patterns throughout the latent topics.

[2] The CNJ defines a set of 22 main topics that are further subdivided in sub-topics organized in a tree structure. The complete list is available in `https://www.cnj.jus.br/sgt/consulta_publica_assuntos.php`.

5.2 Origin-Oriented Structure

The origin-oriented structure takes into account the origins of each decision — who wrote them. The proposal is to explore the patterns-of-writings as latent topics. In other words, to explore the patterns of words, terms, and legal named entities from the authors' point of view and to compare the relationship between latent topics and authorship. Nevertheless, to think of the Justice as the sole author might be misleading. As observed in Section 2.1.1, monocratic decisions are not necessarily written by a single Justice. They have the help and advice of several clerks who assist in the decision-making and writing processes.

Furthermore, as mentioned in Section 2.1, for the period from 2000 to 2018, there were at least 16 Justices substitutions throughout Cabinets, and for every transition we expect that the new Justice will have kept part of the old Cabinet personnel. This suggests a path through which new Justices might be influenced by their predecessors. We could not prove this hypothesis, but we found elements that corroborate it (further in Chapter 6).

Therefore, we found that it would be more accurate to think of the entire Cabinet as the decisions' author instead of the Justice itself. Moreover, due to the many transitions, it would be hard to find significant periods where all the same eleven Justices were acting. Since the Chief Justice has a different caseload than a regular Justice, we removed the decisions published by the Justice's Cabinet when the Justice assumed this role. We also found gaps resulting from the long periods of vacancy, gaps that last less than a year.

Figure 5.3 presents the monocratic productions per Cabinet, and the average. In the first chart, we observe some discontinuities (when a line suddenly has no value) due to vacancy or when a Justice assumes the role as Chief Justice. This figure also reveals the differences in the production level among Justices. Even though all have a similar number of new cases (Falcao et al., 2014), some of the Justices are faster while reviewing a case, and some of them are more likely to send the case to one collegiate body, avoiding issuing a monocratic decision.

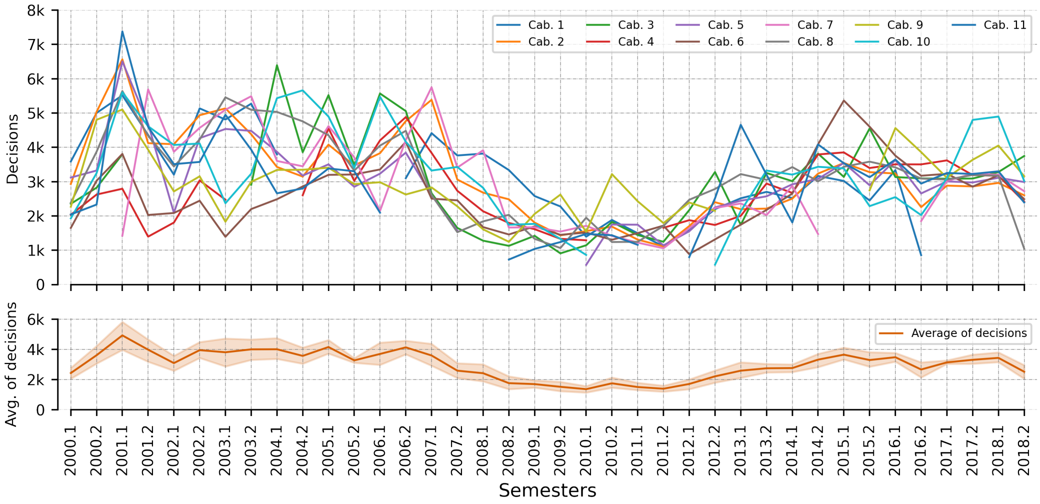


Figure 5.3: Production of monocratic decisions per Cabinet.

The result of this strategy is a tensor, $\mathcal{X} \in \mathbb{R}^{o \times w \times t}$, defined by $origin \times word \times time$, where $\mathcal{X}_{ijk}$ shows relevance score of the word (terms or legal named entities) j for the collection of decisions related to the origin i in the time slot k. For instance, Figure 5.4 presents a proposition of document-based structure based on the legal case categories.

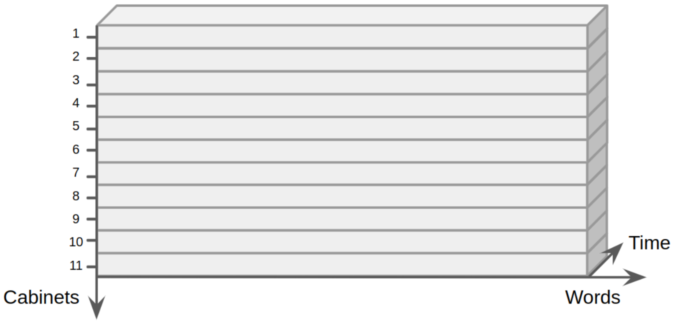


Figure 5.4: Visualization for a possible group-based structure.

5.3 The Textual Features Representation

We tried three different approaches for textual features selection based on words, terms, and legal named entities. For all proposed structures, the textual features are represented in the second mode of the third-order tensor. But, regardless of the approach, all monocratic decisions passed through the same preprocessing procedure for text enhancement and noise removal.

5.3.1 Text Processing

As the majority of the legal documents, legal decisions are official documents and, to testify their official origin, an extra-textual information is added to its content (*e.g.*, Justice's digital signature). As mentioned in Section 2.1 most of the monocratic decisions are PDF files, and for those types of files, the first step handles the header, footer, and digital signature removal. The header usually identifies the legal case to which the decision is related and, in some cases, the litigants and attorneys involved in the legal case. The footer of every page presents the page number and the document's digital signature. Decisions in RTF files have a clean text containing only the decision's content.

Figure 5.5 represents the text preprocessing steps. The second step is needed for both RTF and PDF files. By the end of the last page, we may have Justice's digital signature and the date/location of when and where the decision has been made. Thus, the second step consists of signatures removal. This information was also removed. In the third step, we have the textual content broken down into a collection of sentences.

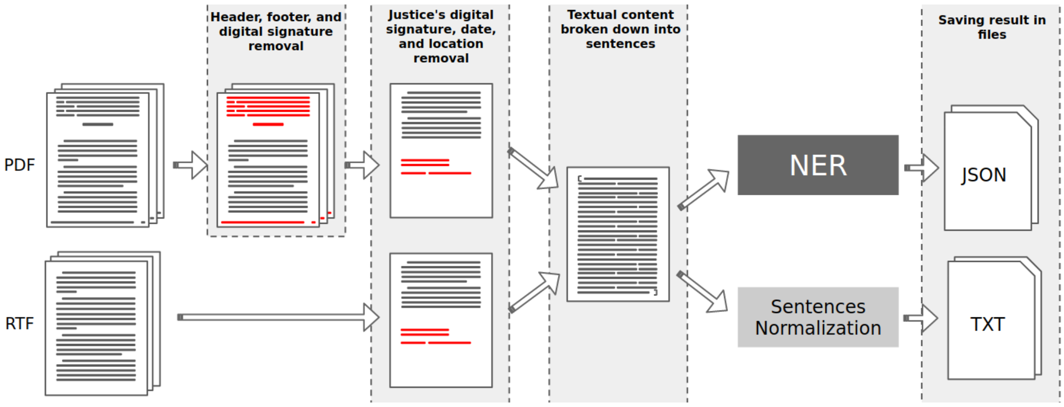

Header, footer, and digital signature removal
Justice's digital signature, date, and location removal
Textual content broken down into sentences
Saving result in files
PDF
RTF
NER
Sentences Normalization
JSON
TXT

Figure 5.5: Text processing in summary.

In the fourth step, legal named entities are extracted, structured, and saved in a JSON file for every one of the decision's sentences. This file is linked to the decision where all named entities are saved with their relative location in the original file. A second file is generated only with a normalized version of the decision's sentences. This normalization consists of diacritics removal and changing the words to lowercase.

5.3.2 Textual Feature Selection

The two files generated in the text preprocessing step can be used for the feature extraction. Usually, most of the studies for dynamic topic modeling use relevant words as textual features where the relevance is measured based on the word's frequency — the most relevant words define the second mode. However, a word might not be enough to promote an understanding of a certain latent topic in the legal domain. Most of the meaningful terms in the legal domain are similar to open compound words (expressions comprising a sequence of word), with at least two words. For example, while the term "*repercussao geral*" (*general repercussion*) is related to an essential instrument in the STF's decision-making processing, the two words separately might be related to very different contexts. Another example is the term "*Código de Processo Civil*" (*Code of Civil Procedure*), which refers to the *Law 13,105/15*. Meanwhile, the term "*Processo Civil*" (*Civil Procedure*) refers to a legal subject, and the individual words "*Código*", "*Processo*" , and "*Civil*" may be related to many different contexts.

Also, as presented in Section 2.2, named entities are defined by a sequence of words with more than one word, except for "person", which may be defined by one single word (a rare case). A "precedent" is comprised of at least two words: a legal case class acronym and a legal case number (*e.g.* "*HC 122*"). This two-word pattern is a common pattern for precedent citation.

As a compromise solution, our proposal relies on bi-grams as textual features. A feature selection based on the most frequent sequence of two words found in the corpus, ignoring stop-words [3]. This strategy is not enough to represent complex terms as "*Código de Processo Civil*", but a sequence of bi-grams can give a much more relevant context to interpret a latent topic than individual words.

For the document-oriented structure, decisions are converted to term frequency-inverse document frequency (TF-IDF) vector representations using sklearn[4] *TFIDFVectorizer*. In this conversion process, stop-words were ignored by using the selection available in *nltk* for Portuguese[5], and only the most frequent bi-grams were considered. In a very similar way, for the origin-oriented structure, decisions are converted to term frequency vector representation, where stop-words were also ignored, and only the most frequent bi-grams were considered.

[3] Those word with high frequencies in Portuguese that does not add much information, such as articles, prepositions, pronouns, conjunctions.

[4] `https://scikit-learn.org`

[5] `https://www.nltk.org`

5.3.3 Tuning Textual Feature Selection with Named Entities

Every coarser-grained named entity was converted into a textual representation embedding its fine-grained entities. The idea is to add a suffix related to its fine-grained entities (if there is one) to every word in the coarser-grained named entity. Then we converted the words to lowercase and removed diacritics. With the textual versions of the named entities within a decision, we can use this textual representation as part of the decision's content or select some of them for a specific analysis (*e.g.*, an analysis focused on person citation only).

However, we added an extra textual transformation step for precedent citations. The precedent has the most standardized citation format among the coarser-grained entities. For instance, there is a finite number of legal procedure classes. Between the two highest judicial bodies (STF and the STJ), there are 93 classes. Also, these classes are usually cited by their acronyms (*e.g.*, *HC* is the acronym for the class *Habeas Corpus*). Since it is common for the citation of a precedent in a form *class + number* (*e.g.*, *HC 11.23*), we mapped and converted every legal procedure class to its acronym. For those classes not present in the mapping, no transformations were made. The legal procedure number citations were also normalized to a numeric sequence without ' . ' (used as the thousands separator in Portuguese). Furthermore, we removed the role identification for the reporting justice (*e.g.*, "*Min.*", "*Ministro*", "*Rel.*", "*Relator*") keeping only the Justice's name, and, also, we joined the reporter's full name, with "_" between names.

For instance, taken the first citation presented in Figure 2.4:

> "*Habeas Corpus 232.618 / MS, Rel Min. Jorge Mussi, DJe de 20.06.2012*"

after the transformation, became:

> "*lpc_hc lpn_232618 lpo_ms rj_jorge_mussi pd_20.06.2012*".

Where the prefixes *lpc*, *lpn*, *lpo*, *rj*, and *pd* stand for legal procedure class, number, origin, reporting justice, and publication date, respectively.

For legislative reference, we also have an analogous scenario. There is a finite number of legal acts, institutions, and origins, but there are multiple ways of citing these elements. For instance, the legal act *Law 13,105/15* may also be cited as "*Código de Processo Civil*" or its acronym "*CPC*", or even as "*Novo Código de Processo Civil*" (*New Code of Civil Procedure*). Since we could not build a reliable mapping, we decided to keep those entities without further transformation. We only removed the thousands separator from numbers.

5.4 Tensor Decomposition for Latent Topic Identification

To build the tensor representation and the tensor decomposition, we used Tensorly[6], a Python library that also implements the NNCPD algorithm[7]. For the tensor creation, every time slice was individually computed. For every time slice, we have a matrix $document \times terms$ or $origin \times terms$, and a list of time slices is converted into a tensor representation with Tensorly.

Tensorly uses the Tensorflow library in the background to perform the NNCPD decomposition. The function for the decomposition is a distributed version optimized for its usage in computer architecture with multiple CPUs, GPUs, or TPUs. However, it presented an exponential growth in memory as a limitation, although the decomposition is fast in processing time.

5.5 Quality Assessment

When working with topic modeling, usually the quality of the topics is measured based on a *topic coherence score*, and there is a significant number of measurements in the literature that try to measure the coherence. Most of these measurements are based on the co-occurrence frequencies of terms within a reference corpus and distributional semantics. Since topics are interpreted by the terms related to them, the idea behind these metrics is that pairs of terms that co-occur frequently or are close to each other within a semantic space are likely to contribute to higher levels of coherence (O'Callaghan et al., 2015). However, due to the nature of the NNCPD, where words and topic relevance evolve along time — word and topic relevance can not be seen as a static value —, those metrics are not suitable for measuring the quality of the result. As presented in Ahn et al. (2021), the quality of the NNCPD results for dynamic topic modeling can be assessed via a quantitative and qualitative evaluation.

Regarding the quantitative evaluation of the tensor decomposition, as presented in Section 2.3.2.3, the reconstruction error $E(\mathcal{X} : A, B, C)$ (Equation 2-8) measures how much from the original information (the tensor $\mathcal{X}$) is still in the resulting factors (the matrices A, B, and C). The reconstruction error is closely related to the number of ranks (latent topics), where smaller numbers of topics are associated with higher reconstruction error. Nevertheless, defining the smallest number of topics for a perfect decomposition ($E(\mathcal{X} : A, B, C) = 0$)

[6] http://tensorly.org

[7] http://tensorly.org/stable/modules/generated/tensorly.decomposition.non_negative_parafac.html

is an NP-hard problem, and there is no straightforward algorithm to determine the rank of a specific given tensor (Kolda and Brett, 2009).

In the experiments presented in the following Chapter, we discuss the quality assessment of the proposed method in practical scenarios.

5.6 The Process in Summary

In this Chapter, we presented our approach for dynamic topic modeling, focusing on exploring the monocratic decisions issued by the STF over a relevant period of time. Despite that flows, we believe that our proposal can be adapted to different scenarios. The easiest one would be the usage in different Brazilian Courts (*e.g.*, a State Court) where few adaptations would be needed regarding decisions grouping. Since the STF, as a supreme court, has influences on all other courts in Brazil, it is very likely that the annotated corpus presented in Chapter 4 will produce good results. Thus, in principle, the whole method could be replicated.

Table 5.1 presents the proposed method for dynamic topic modeling with NNCPD for tensor decomposition.

Table 5.1: The method in summary.

Structure	First Mode	Second mode		Third Mode
		Features	Weight	
Document-Oriented	Document	Bi-grams	TF-IDF	Year or Semester
		Legal NE		
Origin-Oriented	Origin	Bi-grams	Frequency	
		Legal NE		

The idea of the origin-oriented structure could be easily adapted to a tensor representation of a large collection of documents produced individually by a few authors, such as news and social web messages. Even adapted to a more complex scenario, for instance, documents could be companies' annual reports and a company could be seen as the author. So, we could investigate how the writing in these reports evolves and what it could reveal about the company concerns along the time.

6
Experiments and Case Studies

In the following sections of this chapter, we present three sequences of experiments where we applied the tensor decomposition with NNCPD for dynamic topic modeling on the legal context by exploring the temporal evolution of the patterns-of-writings through latent topics. Furthermore, we explore the different strategies to structure the tensor and the method's capability to "explain" the result through visualizations and tables, correlating latent topics with decisions, terms, and time.

With these experiments, we intend to demonstrate the process's capability to produce coherent and reliable results. Nevertheless, it is important to state that the discussions of this experiments' results are presented from a Computer Science perspective, but, as far as possible, we also tried to present a legal interpretation. Indeed, a legal expert would present a richer discussion of the results' implications.

In the first Section, we present three experiments focusing on the decision-oriented tensors and the textual content of the decisions. In Section 6.2, we explore the legal named entities and the origin-oriented tensors. In Section 2.3 we present a case study related to the evolution of the STF's vocabulary. Finally, in Section 6.4 we present summary of the evaluation of these experiments.

6.1
Exploring Decisions' Content from Different Categories of Legal Case

As mentioned in Section 2.1, the STF acts on a wide variety of cases due to its broad jurisdiction. In this Section, we explore the proposed decision-oriented structure over the textual content for latent topic modeling three different scenarios:

- *EXP1.1*: Comparing the different categories. We compare decisions related to different categories of legal cases through the dynamic latent topics generated by the proposed process.
- *EXP1.2*: Restricting the analysis to decisions from the *Writ* category.
- *EXP1.3*: Comparing decisions in different time intervals. We zoom-in on the results of *EXP1.2* exploring two different intervals.

The following Subsections describe these experiments and their results.

6.1.1 EXP1.1: Comparing the Different Categories

In the first experiment with legal decisions, our goal is to verify the process capability to produce coherent dynamic topics over a sample of decisions, in a design that is inspired by Ahn et al. (2021). For this experiment, we distributed the collection of decisions on the 5 categories of the legal case related to them: *Appeal*, *Writs*, *Constitutional*, *Criminal*, and *Others*. Those are the same categories presented in Subsection 5.1.1.

Table 6.1: Distribution of decisions grouped according with the case's type.

Year	Appeal	Writs	Criminal	Constitutional	Others
2000	58,848	1,038	185	68	256
2001	91,868	1,349	199	154	348
2002	68,232	1,401	136	164	449
2003	76,226	1,329	160	130	385
2004	73,222	1,729	86	142	560
2005	70,041	3,108	126	153	796
2006	73,935	3,404	97	106	708
2007	59,807	1,702	22	37	245
2008	37,857	1,981	10	38	186
2009	29,14	2,805	8	59	187
2010	28,203	2,352	4	69	229
2011	24,116	2,524	18	102	159
2012	31,693	7,060	100	297	360
2013	40,446	9,063	141	482	660
2014	49,733	9,730	150	508	671
2015	54,219	9,858	264	520	998
2016	51,813	10,965	175	608	632
2017	47,565	14,739	345	786	1,027
2018	42,595	17,466	481	796	717

As shown in Table 6.1, the number of decisions per category varies greatly; however, for this experiment, we sampled uniform amounts of decisions from each group and we structured a decision-oriented tensor $\mathcal{X} \in \mathbb{R}^{d\times b\times t}$ $(\mathit{decision} \times \mathit{bi\text{-}gram} \times \mathit{time})$. Based on the selected data, we computed the TF-IDF weight of every bi-gram across all decisions and selected the 3000 bi-grams with the highest frequency. For this experiment, we defined a slot of 150 decisions per group per year based on the average of publications per year of the less representative group, the *Criminal* group. For the period in which less than 150 decisions were published, we added mock decisions (documents without content). For instance, in 2006, we had only 97 decisions related to *Criminal* cases, so we added 53 mock decisions in the *Criminal* slot for the time slice that represents decisions published in 2006.

As a result, we ended up with a third-order tensor $\mathcal{X}$ of shape $(750 \times 3000 \times 19)$, where the first mode represents documents, the second represents bi-grams, and the third represents the publication year. The structure of the tensor is shown in Figure 6.1. The top blue part consists of the first sample of 150 decisions on appeal cases published in the 18 time slices. Similarly, each colored part represents one of the groups, as mentioned earlier.

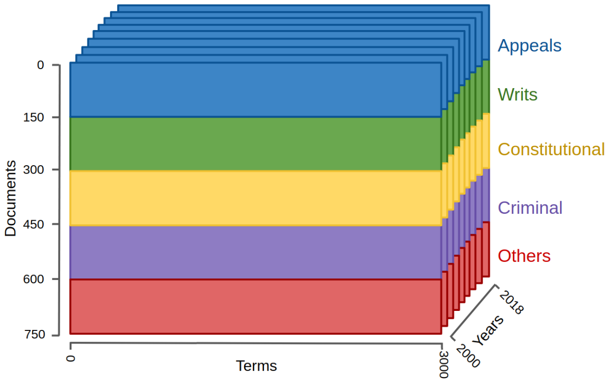


Figure 6.1: The tensor structure (frontal slice).

We ran the NNCPD for the entire tensor for 5 topics, expecting to see topics strongly related to each of the 5 categories. Figure 6.2 and Table 6.2 present a representation of the factors resultant (matrices A, B, and C). The first heat map in Figure 6.2 represents the latent topic presence in a given year (factors A) on a scale from 0 to 1. The second heat map presents identical information but indexes of documents instead of years on the horizontal axis. Since the group is the only feature explicitly in common between two decisions with the same index but different time-slices, we added the group identification to the graphic representation.

Furthermore, to build the visualizations, the columns related to the topics of the matrices A and B were normalized to values between 0 and 1 (Figure 6.2 transposed the matrices; hence, the topics became rows). Also, the topics were ordered according to when each reached its highest value. Topic 0 in such figure was the last one to reach its maximum value.

On the first heat map in Figure 6.2 (in blue scale), we observed the latent topics emerging, evolving, and fading; latent topic 3 is a good example of this. An event worth mentioning in topics 1 and 2 is the fading and emerging of both topics that coincide with the period where the number of decisions for the category was below the average, and mock decisions were used. For the period between 2007 and 2011, the majority of the decisions related to *Criminal* were mock decisions with empty content. With the rise of new decisions in 2012, the topic 1 emerged again, reaching its highest value in 2015.

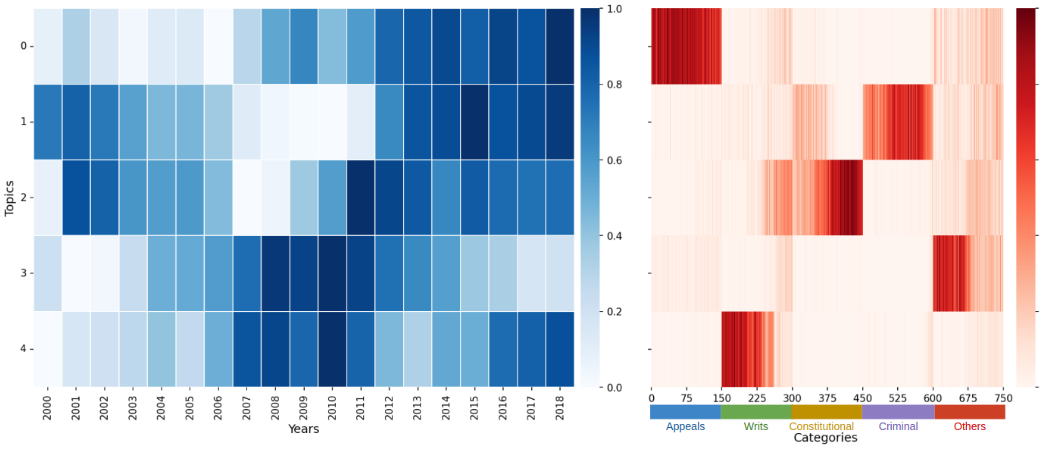


Figure 6.2: Comparing decisions from different types of legal case. NNCPD factors A and B.

Comparing both heat maps, we observe that, as expected, each topic is strongly correlated to one category, but none of these topics are exclusively associated with a single category. For instance, most of the occurrences of the latent topic 1 were in *Criminal* related decisions; however, there are significant occurrences in decisions related to categories *Constitutional* and *Others*, and less present in those related to *Appeals* and *Writs*.

Regarding the relationship between topics and categories, the second heat map also shows a peculiar scenario when topics seem to merge and fade among documents. This behavior can be explained in each time slice. The documents were sorted based on their legal class under the boundary set for each group. At first, it was not intentional — the collection was already sorted by the legal class before the sampling —, but it also indicates that the 5 categories were not enough to group these decisions. In the next experiment, presented in Section 6.1.2 we will explore this scenario for the *Writs* category.

The NNCPD reconstruction error $\left\|\mathcal{X} - \hat{\mathcal{X}}\right\|_F = 0.964$. As stated in Section 2.3, this value indicates the quantity of the original information which can be recovered from the factors A, B, and C. The smaller the error, the better the topic's ability to represent the nuances of the original information contained in the tensor. Hence, the larger the error, the less specific and detailed the topics will be. In Section 6.4 we present a brief discussion about the reconstruction error and topic modeling results, comparing the different values observed in the experiments and the case studies presented in this chapter.

Comparing the topics illustrated in Figure 6.2 with the bi-grams associated with each topic presented in table 6.2, we can observe that there is a coherence between latent topics and the bi-grams associated with them. For instance, as stated in Subsection 5.1.1 *Criminal* cases are related to cases filed in the STF evolving federal public officials. Usually, one of the litigants in this type of criminal action is the Federal Prosecution Office (or *Ministério Público Federal*), represented by the Attorney General (or *Procurador Geral*). Therefore, the term's association with the latent topic 1 shows a strong correlation with the category *Criminal*, reinforcing the argument that the process is capable of producing coherent dynamic topics.

Table 6.2: Correlation between topics, categories, and terms.

Topic	Correlated to	Top 5 terms (Portuguese / English)
0	Appel	*'recurso extraordinário'*, *'acórdão recorrido'*, *'nego seguimento'*, *'ss 1o'*, and *'art 557'*.
		'extraordinary appeal', *'contested judgment'*, *'deny continuation'*, *'1st paragraph'*, and *'Art 557'*.
1	Criminal	*'ministerio publico'*, *'deputado federal'*, *procurador geral'*, *'acao penal'*, and *'prisão preventiva'*.
		'public minister', *'federal deputy'*, *'attorney general'*, *'penal action'*, and *'preventive arrest'*.
2	Constitutional	*'presente acao'*, *'advogado geral'*, *'procurador geral'*, *'lei 868'*, and *'amicus curiae'*.
		'present action', *'advocate general'*, *'attorney general'*, *'law 868'*, and *'amicus-curiae'*.
3	Others	*'acao cautelar'*, *'recurso extraordinário'*, *'medida cautelar'*, *'suspensivo recurso'*, and *'medida liminar'*.
		'precautionary action', *'extraordinary appeal'*, *'precautionary measure'*, *'suspensive appeal'*, and *'preliminary injunction'*.
4	Writs	*'habeas corpus'*, *'medida liminar'*, *'impetrado contra'*, *'autoridade coatora'*, and *'prisão preventiva'*.
		'habeas-corpus', *'precautionary measure'*, *'filed against'*, *'coactor authority'*, and *'preventive arrest'*.

6.1.2 EXP1.2: Comparing Decisions from Writs' Legal Cases

As an extension of the experiment *EXP:1.1*, we narrowed down our scope to monocratic decisions related to the *Writs* category. Based on the distribution presented in Table 6.1, for this experiment we randomly selected 1,000 decisions related to *Writs* for every year. Also, as in the previous experiment, we computed the TF-IDF weight of every bi-gram across all decisions and selected the 3000 bi-grams with the highest frequency. Therefore, we structured a decision-oriented tensor $\mathcal{W} \in \mathbb{R}^{d \times b \times t}$ $(decision \times bi - grams \times time)$, as presented in Figure 6.3.

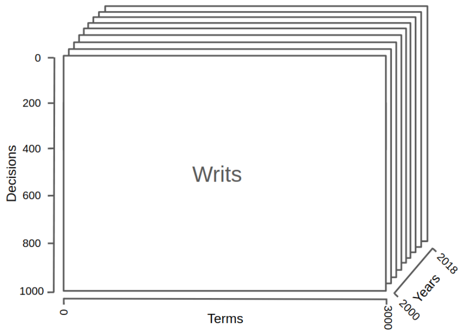


Figure 6.3: The structure of the tensor $\mathcal{W}$ (frontal slice).

With this experiment, the intention is to use the decision-oriented structure to explore one group only and to explore a little further the latent topics that emerge and fade on the resulting matrix A related to the decisions, as observed in the experiment *EXP:1.1*. Also, this design might shed some light on the evolution of the monocratic decisions in the *Writs* category. We ran the decomposition over the structured tensor asking for 10 latent topics. Figure 6.4, and Table 6.3 presents an overview regrading the decomposition results, factors A, B, and C. The NNCPD reconstruction error $\left\|\mathcal{X} - \hat{\mathcal{X}}\right\|_F = 0.958$.

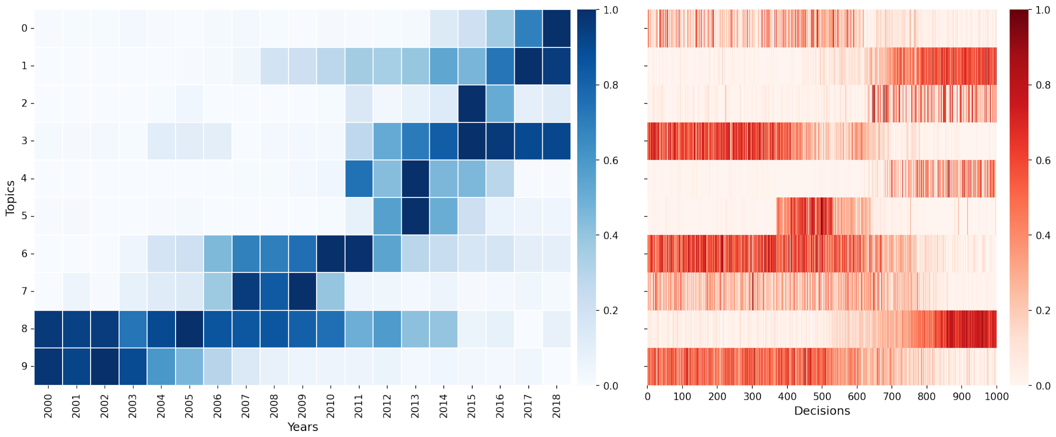


Figure 6.4: Comparing decisions on writs legal case.

We can draw some conclusions based on heat maps presented in Figure 6.4. But first, it is important to interpret the emerging and fading of topics on the second heat map. As presented in Section 5.1.1 the *Criminal* category is related to a group of 7 legal procedure classes, and, when structuring the tensor, we placed the decision sorting them by the legal class acronym of the case related to them, on lexicographical order. Figure 6.5 presents the ordering illustration per time slice[1]. Therefore, the emerging and fading of topics throughout decisions are a consequence of the decision ordering. For

[1] Instead of ordering by legal class, we could define a proportion per class and assure a slot for each one, but there is a great presence variance among these classes over time.

instance, topic 9 is closely related to *habeas corpus* (HC), meanwhile, topic 8 is closely related to *Rcl* (acronym for '*Reclamação*', or *Complaint*) cases.

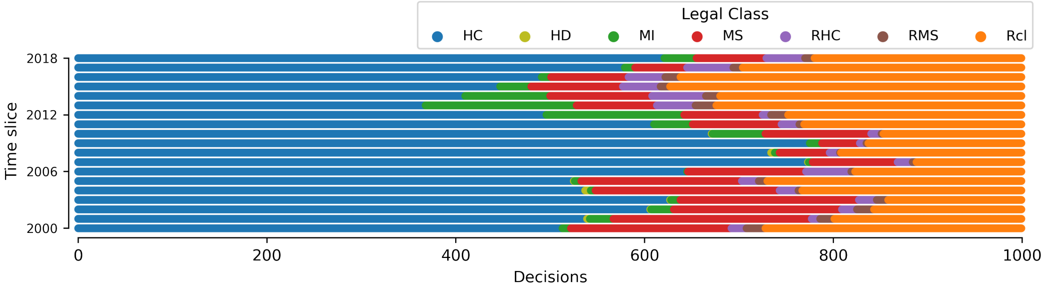


Figure 6.5: Decisions ordering per time slice, sorted by legal case classes.

In Figure 6.4, we observe at least two distinguishable moments: before and after 2012. Before 2012, we had a predominant latent topic (topic 8) related to a specific group of decisions that only began to fade around 2010, and there are three topics related to a certain group of decisions that seems to be aligned: topic 6 and 7 began to merge when topic 9 began to fade. In 2012, we had a scenario where six latent topics are fading or merging, and there is no clear predominance of a particular topic. After 2012, two topics were predominant, each related to a certain group of decisions. Meanwhile, topic 5 differs from the others and relates to an entirely different group of decisions.

Table 6.3: Terms and topics.

Topic	Top terms (Portuguese / English)
0	*'SP Rel', 'segunda turma', 'HC 122', 'contra decisão'*, and *'Carmén Lúcia'*. *'SP Reporter', 'second panel', 'HC 122', 'against decision'*, and *'Carmén Lúcia'*.
1	*'súmula vinculante', 'agravo regimental', 'repercussão geral', 'ato reclamado'*, and *'recurso extraordinário'*. *'binding ruling', 'regimental grievance', 'general repercussion', 'claimed act'*, and *'extraordinary appeal'*.
2	*'repercussão geral', 'recurso extraordinário', 'Tema 246', 'Art 543'*, and *'processo civil'*. *'general repercussion', 'extraordinary appeal', 'Theme 246', 'Art. 543'*, and *'civil procedure'*.
3	*'prisão preventiva', 'processo penal', 'ordem pública', 'agravo regimental'*, and *'codigo penal'*. *'preventive detention', 'penal procedure', 'public order', 'regimental grievance'*, and *'penal code'*.
4	*'Art 71', 'ADC 16', 'responsabilidade subsidiária', 'ss 1o'*, and *'encargos trabalhistas'*. *'Art. 71', 'ADC 16', 'subsidiary responsibility', '1st Paragraph'*, and *'labor benefits'*.
5	*'aposentadoria especial', 'ss 4o', 'Lei 8213', 'art 57'*, and *'Lei Complementar'*. *'special retirement', '4th Paragraph', 'Law 8213', 'Art. 57'*, and *'Complementary Law'*.
6	*'súmula 691', 'medida liminar', 'liberdade provisoria', 'ação penal'*and *'contra decisão'*. *'binding ruling 691', 'preliminary injunction', 'provisional freedom', 'penal action'*, and *'against decision'*.
7	*'contra ato', 'corpus impetrado', 'determino remessa', 'originária deste'*, and *'desta decisao'*. *'against act', 'corpus filed', 'determine shipment', 'originating from'*, and *'this decision'*.
8	*'presente reclamação', 'medida liminar', 'ação direta', 'desta corte'*, and *'decisão proferida'*. *'present complaint", 'preliminary injunction', 'direct action', 'of this court'*, and *'decision rendered'*.
9	*'despacho vistos', 'medida liminar', 'indefiro pedido', 'vistos etc'*, and *'determino remessa'*. *'dispatch visas', 'preliminary injunction', 'I reject the request', 'visas etc'*, and *'determine shipment'*.

Nevertheless, our focus was to explore the category *Writs*, not legal classes, so we decided to keep this distribution.

When associating topics and bi-grams (Table 6.3), we observe that topic 5 is closely related to the citation of Article 57 of Law 8,313 regarding the especial retirement in decisions related to a writ of mandamus (or '*Mandado de Segurança*', MS). Topic 4, in turn, is closely related to Article 71 of Law 5,452 and, also, related to the citation of the legal procedure *ADC 16* (ADC is an acronym for *Declaratory Action of Constitutionality*, or '*Ação Declaratória de Constitucionalidade*'). Both legislation and legal case are related to the subject *labor rights*, and, in November 2010, the STF's Plenary judged the *ADC 16*, and published a collegiate decision. The publication date is coherent with the topic's emergence, indicating that the decision was used as a precedent in different decisions. Three years after the publication of the *ADC 16* the topic 4 reaches its top, and then, the topic losses relevance. We observed an identical result in Correia et al. (2019) when exploring precedent relevance over the time where we found that writs precedents tend to reach their top three years after their publication and then begin to lose their relevance. Also, in 2017, paragraph 4 of Article 71 was edited, which tracks with the behavior of topic 4, which began to fade in 2014, but disappeared in 2017. These temporal facts are indications of coherence between terms, topics, and time.

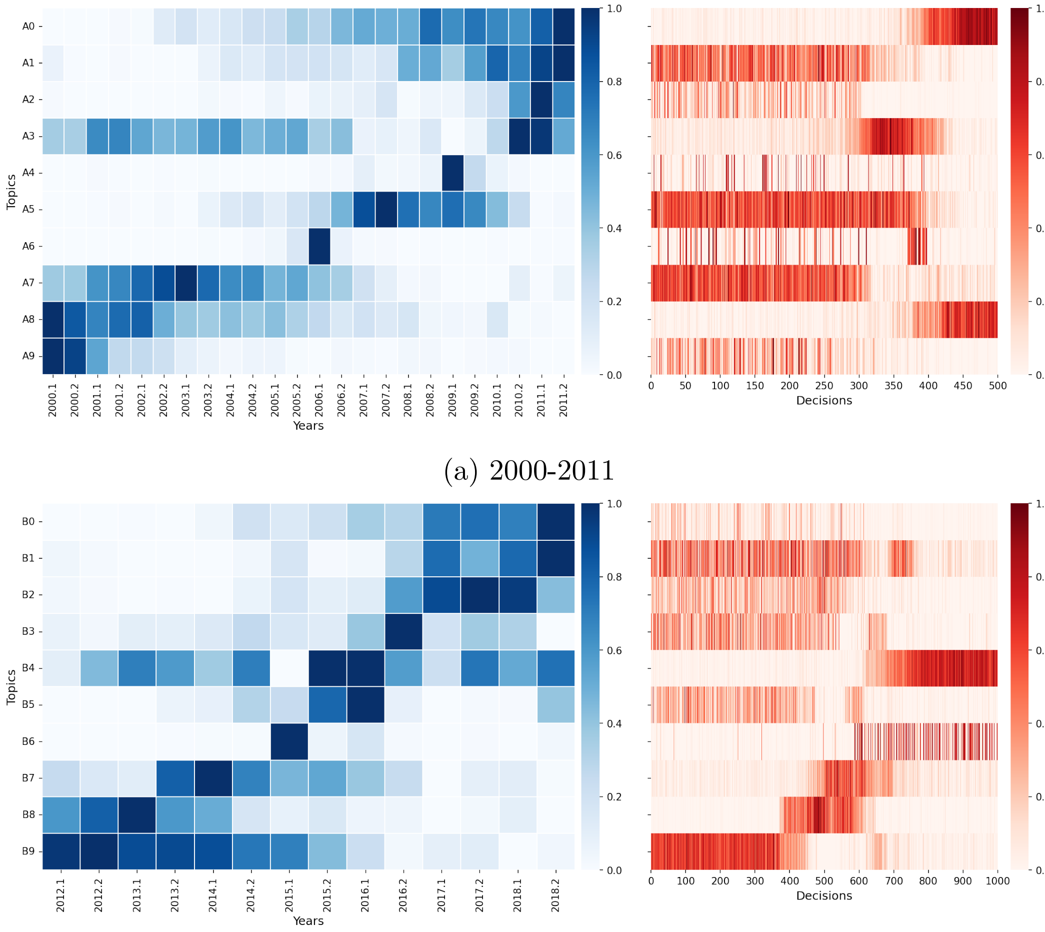
Overall, two different sets of topics present two different histories about decisions related to *HC* and *Rcl* cases, respectively. The topics 0, 3, 6, 7, and 9 present the evolution of the *HC* cases, meanwhile the topics 1, 2, 4, and 8 present the evolution of the *Rcl* cases. But still, a general overview of a complex scenario.

6.1.3 EXP1.3: Comparing Decisions in different time intervals

As observed in the experiment *EXP1.2*, significant changes seemed to happen around 2012. Narrowing the scope to two specific moments, before and after 2012, we can use the proposed process to capture even more information by specifying a higher number of topics.

Figure 6.6 presents a result comparison of the two tensors, $\mathcal{W}_A$ and $\mathcal{W}_B$ built based on the decisions related to *Writs* cases published before 2012 (2000-2011) and decisions published after 2012 (2012-2018), respectively. To facilitate the discussion, we added prefixes to the topics: the prefix '*A*' indicates topics generated from the first collection of decisions (200-2011), and the prefix '*B*' for the second collection (2012-2018).

Each collection of decisions was treated as a separated corpus for both tensor building processes, following the same decision-oriented structure as the previous experiment. Again, we used a semester as the window size for these tensors' temporal modes. Furthermore, since there were fewer writs decisions published over the years between 2000 and 2012, compared to the second time interval, we used only 500 decisions per time slice to structure the first tenser, and 1000 were used per time slice in the second one.



(a) 2000-2011

(b) 2012-2018

Figure 6.6: zooming in over decisions on *Writs* case.

Tables 6.4 and 6.5 present the list of terms associated to each topic in both decompositions. To compare the latent topics produced based on the tensors $\mathcal{W}$, $\mathcal{W}_A$, and $\mathcal{W}_B$, we used the Jaccard's distance, where the distance between two topics, defined as sets of terms A and B, is computed as:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}. \tag{6-1}$$

Since the terms related to a topic is a ranked list of terms, we limited the set to the top 20 terms.

Comparing topic modeling results from the one based on the whole and the one based on the first period is like a zoom-in over the collection. For instance, topic 8 (from *EXP1.1*) seems to have broken into at least two topics, topics A0 and A8 (Jaccard distances of 0.21 and 0.48, respectively). Also, topic

9 has a close distance to topics A7 and A9 (0.54 and 0.33), a statement also supported when comparing the topics' evolution and the decisions they are related to. Topic 9 began to fade in 2004, almost the same moment where topic A7 also began to fade, and topic 9 was already close to none. All these topics are related to *HC* cases. Thus, comparing the bi-grams associated with each topic, we have an interpretation (beyond the metric) that is coherent with the association between these topics — a similar scenario for comparing 8 and the topics A0 and A8.

Table 6.4: Terms and topics for the period 2000-2011.

Topics	Top terms (Portuguese / English)
A0	*‘súmula vinculante’*, *‘presente reclamação’*, *‘decisão proferida’*, *‘Art 71’*, and *‘agravo regimental’*.
	‘binding ruling’, *‘present complaint’*, *‘rendered decision’*, *‘Art 71’*, and *‘regimental grievance’*.
A1	*‘prisão preventiva’*, *‘liberdade provisoria’*, *‘ação penal’*, *‘cádigo penal’*, and *‘medida liminar’*.
	‘preventive arrest’, *‘provisional release’*, *‘penal action’*, *‘penal code’*, and *‘preliminary measure’*.
A2	*‘súmula 691’*, *‘turma unânime’*, *‘HC 79’*, *‘contra decisão’*, and *‘turma maioria’*.
	‘súmula 691’, *‘panel unanimous’*, *‘HC 79’*, *‘against decision’*, and *‘panel majority’*.
A3	*‘contra ato’*, *‘procurador geral’*, *‘concurso público’*, *‘medida liminar’*, and *‘liquido certo’*.
	‘against act’, *‘attorney general’*, *‘public tender’*, *‘preliminary measure’*, and *‘right liquid’*.
A4	*‘originaria deste’*, *‘federal motivo’*, *‘desta decisão’*, *‘determino remessa’*, and *‘decisão vistos’*.
	‘originate from this’, *‘federal motive’*, *‘from this decision’*, *‘determine remittance’*, and *‘visa decision’*.
A5	*‘corpus impetrado’*, *‘contra ato’*, ‘ §1^o ’, *‘Art 21*, and *‘contra decisao’*.
	‘corpus filled’, *‘against act’*, *‘1st Paragraph’*, *‘Art 21’*, and *‘against decision’*.
A6	*‘Art 2º’*, *‘Lei 072’*, *‘integralmente fechado’*, *‘exame criminologico’*, and *‘regime integralmente’*.
	‘2nd Art’, *‘Law 072’*, *‘completely closed’*, *‘criminological examination’*, and *‘completely regime’*.
A7	*‘despacho vistos’*, *‘indefiro pedido’*, *‘medida liminar’*, *‘hc 79’*, and *‘determino remessa’*.
	‘dispatch visas’, *‘deny application’*, *‘preliminary measure’*, *‘HC 79’*, and *‘determine remittance’*.
A8	*‘presente reclamação’*, *‘decisao reclamada’*, *‘fazenda pública’*, *‘ação direta’*, and *‘medida liminar’*.
	‘present complaint’, *‘complained decision’*, *‘public treasury’*, *‘direct action’*, and *‘preliminary measure’*.
A9	*‘emenda constitucional’*, *‘originariamente habeas’*, *‘julgar originariamente’*, *‘justiça competente’*, and *‘união remetam’*.
	‘constitutional amendment’, *‘originally habeas’*, *‘originally judge’*, *‘competent justice’*, and *‘union send’*.

We got a similar observation of the results obtained based on the second interval (2012-18). Topic 5 (from *EXP1.1*) was the first to reach its maximum value in the complete analysis in 2013. This topic is closer to the topic B8 (distance of 0.74), and, due to windows size, B8 gives us more detail about the topic's evolution. Also, both topics are related to the same type of decisions. The topics B1, B3, B5, and B9 are closely related to topic 3 (0.54, 0.48, 0.33, and 0.29). Those topics seem like a detailed version of topic 3.

Another result worth pointing out is identifying the short-lasting topics presented in all three representations. Topic 2 is the shortest in the complete

Table 6.5: Terms and topics for the period 2012-2018.

Topics	Top terms
B0	*‘SP Rel’*, *‘RHC 114’*, *‘HC 117’*, *‘hc 122’*, and *‘AGR SP’*. *‘SP Reporter’*, *‘RHC 114’*, *‘HC 117’*, *‘hc 122’*, and *‘AGR SP’*.
B1	*‘prisão preventiva’*, *‘agravo regimental’*, *‘ordem pública’*, *‘regime inicial’*, and *‘documento eletrônico’*. *‘preventive arrest’* *‘regimental grievance’*, *‘public order’*, *‘initial regime’*, and *‘electronic document’*.
B2	*‘segunda turma’*, *‘prisão preventiva’*, *‘SP primeira’*, *‘corpus impetrado’*, *‘contra decisao’*. *‘second panel’*, *‘preventive arrest’*, *‘SP first’*, *‘corpus filled’*, and *‘against decision’*.
B3	*‘agravo regimental’*, *‘súmula 691’*, *‘contra decisao’*, *‘regime inicial’*, *‘código penal’*. *‘regimental grievance’*, *‘súmula 691’*, *‘against decision’*, *‘initial regime’*, and *‘penal code’*.
B4	*‘súmula vinculante’*, *‘repercussão geral’*, *‘recurso extraordinário’*, *‘presente reclamacao’*, *‘agravo regimental’*. *‘biding precedent’*, *‘general repercussion’*, *‘extraordinary appeal’*, *‘present complaint’*, and *‘regimental grievance’*.
B5	*‘prisão preventiva’*, *‘segunda turma’*, *‘ordem pública’*, *‘corpus impetrado’*, and *‘processo penal’*. *‘preventive arrest’*, *‘second panel’*, *‘public order’*, *‘corpus filled’*, and *‘penal case’*.
B6	*‘repercussão geral’*, *‘recurso extraordinário’*, *‘tema 246’*, *‘comprovada culpa’*, *‘órgão especial’*. *‘general repercussion’*, *‘extraordinary appeal’*, *‘theme 246’*, *‘proven guilt’*, and *‘special body’*.
B7	*‘concurso público’*, *‘agravo regimental’*, *‘contra ato’*, *‘líquido certo’*, and *‘Art 236’*. *‘public tender’*, *‘regimental grievance’*, *‘against act’*, *‘right liquid’*, and *‘Art 236’*.
B8	*‘aposentadoria especial’*, *‘ §4º ’*, *‘Lei Complementar’*, *‘Lei 8213’*, *‘Art 57’*. *‘especial retirement’*, *‘4th paragraph’*, *‘Complementary Law’*, *‘Law 8,213’*, and *‘Art 57’*.
B9	*‘prisão preventiva’*, *‘medida liminar’*, *‘súmula 691’*, *‘processo penal’*, and *‘codigo penal’*. *‘preventive arrest’*, *‘preliminary measure’*, *‘súmula 691’*, *‘penal case’*, and *‘penal code’*.

analysis lasts two years, but it is closely related to topic B6 (distance of 0.53), which means that the topic had lasted for two and half years. The topics A4 and A6 show great relevance for a semester. However, as more topics are asked for the decomposition process, more short-lasting topics will appear, associating small events (patterns) to topics.

6.2 The Legal Named Entities Usage over the Time

Section 4.5 presents the NER process employed on the collection and Section 5.3.2, how to use the legal named entities as text features for dynamic topic modeling through tensor decomposition. In this Section, we explore the proposed process with a focus on the legal named entities extracted from the collection of monocratic decisions in two different scenarios:

- *EXP2.1*: Exploring the usage of legal NE on different categories of legal cases.
- *EXP2.2*: Exploring the precedents citation per Cabinets on *Writs*, but limiting precedent citation scope to the fine-grained entities: *legal class procedure*, *legal number procedure*, and *Reporting Justice*.

Our goal with these experiments is to test the capability of the NNCPD over origin-oriented tensor to produce meaningfully and interpretable associations between latent topics and the decisions' origins. The following Subsections describe these experiments and their results.

6.2.1 EXP2.1: Legal NE Usage on Different Categories of Legal Cases.

Since legal cases are related to one of the five categories, as presented in the Subsection 5.1.1, we can use the categories as origins and investigate the relationship between entities and these categories over time. Therefore, we structured a origin-oriented tensor $\mathcal{C} \in \mathbb{R}^{c \times b \times t}$, defined by $category \times bi\text{-}gram \times time$, where c_{ijk} represents the relevance score of the bi-gram j on decisions related to the category i and published on k. The i category is one of the five presented in Subsection 5.1.1.

The bi-grams are related to coarser-grained entities extracted from each decision within the collection. For every decision's content, we kept only the coarser-grained citations. Also, for every category, a particular corpus was defined. Furthermore, only the bi-grams that appear in at least two of the five corpus were considered, leaving us with 2,654 bi-grams, and computed the TF-IDF score. We did this as a strategy to balance the difference in the number of decisions between categories. The resulting tensor is presented in Figure 6.7.

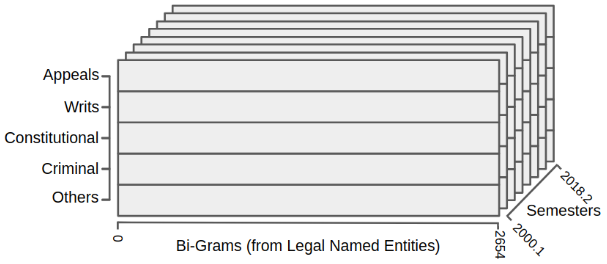


Figure 6.7: Frontal slice of the tensor $\mathcal{X}_{Category} \in \mathbb{R}^{c \times b \times t}$.

Figure 6.8 and Table 6.6 present an overview of the resulting topic modeling. On Table 6.6, the bi-grams were colored identifying the coarser-grained entity which its came from: in blue, fine-grained entities related to precedents, in orange, those related to legislative references, and in green, those related to person citations (a person neither related to a precedent nor academic citation).

The topic modeling revealed topics closely related to distinct categories. Only topic 1, 12, and 13 presented a significant correlation with more them one category. Also, we observed that the topics related to the same category are complementary in a temporal perspective. For instance, for each of the

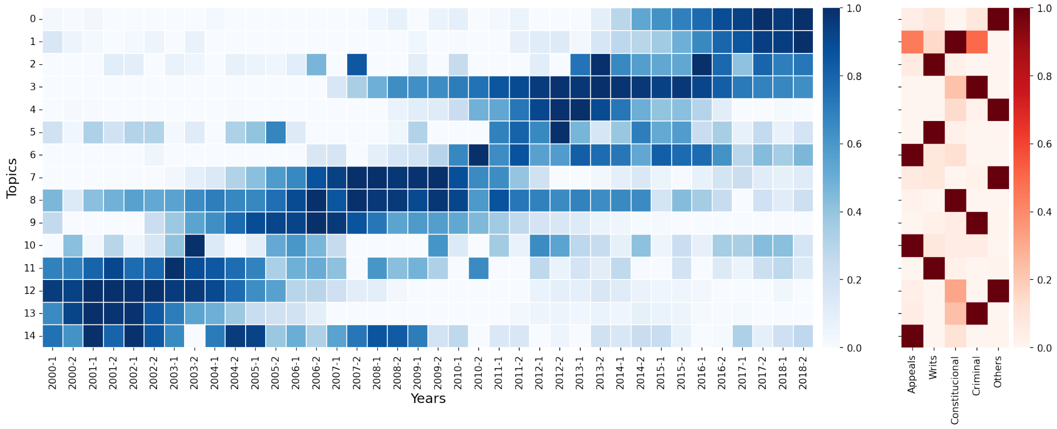


Figure 6.8: Legal named entities usage for different types of legal case.

topics related to *Criminal* cases (topics 1, 3, 9, and 13), when one topic begins to fade, a new one emerges.

Table 6.6: Topics defined by Legal Named Entities.

Topic	Top 10 terms
0	‘*súmula vinculante*’, ‘*primeira turma*’, ‘*segunda turma*’, ‘*dias toffoli*’, ‘*luiz fux*’, ‘*teori zavascki*’, ‘*lei 11*’, ‘*codigo processo*’, ‘*rosa weber*’, and ‘*carmen lucia*’
1	‘*cpc 2015*’, ‘*art 102*’, ‘*constituicao federal*’, ‘*tribunal pleno*’, ‘*luiz fux*’, ‘*teori zavascki*’, ‘*dias toffoli*’, ‘*n constituicao*’, ‘*primeira turma*’, and ‘*art 1*’
2	‘*codigo penal*’, ‘*sepulveda pertence*’, ‘*processo penal*’, ‘*pertence plenario*’, ‘*codigo processo*’, ‘*penal art*’, ‘*plenario maioria*’, ‘*marco aurelio*’, ‘*tribunal pleno*’, and ‘*1a turma*’
3	‘*constituicao federal*’, ‘*súmula 279*’, ‘*segunda turma*’, ‘*primeira turma*’, ‘*processo civil*’, ‘*codigo processo*’, ‘*art 543*’, ‘*constituicao republica*’, ‘*supremo tribunal*’, and ‘*carmen lucia*’
4	‘*adc 16*’, ‘*art 40*’, ‘*súmula 691*’, ‘*súmula vinculante*’, ‘*carmen lucia*’, ‘*art 71*’, ‘*vinculante 10*’, ‘*art 57*’, ‘*4o constituicao*’, and ‘*constituicao federal*’
5	‘*celso mello*’, ‘*codigo penal*’, ‘*penal art*’, ‘*cf art*’, ‘*gurgel santos*’, ‘*monteiro gurgel*’, ‘*art 85*’, ‘*roberto barroso*’, ‘*processo penal*’, and ‘*mello rtj*’
6	‘*aco 1*’, ‘*art 102*’, ‘*lei 9*’, ‘*f constituicao*’, ‘*celso mello*’, ‘*tribunal pleno*’, ‘*constituicao federal*’, ‘*constituicao republica*’, ‘*art 12*’, and ‘*marco aurelio*’
7	‘*súmula 691*’, ‘*lei 8*’, ‘*art 21*’, ‘*art 102*’, ‘*codigo penal*’, ‘*celso mello*’, ‘*hc 79*’, ‘*cezar peluso*’, ‘*habeas corpus*’, and ‘*1o ristf*’
8	‘*celso mello*’, ‘*art 542*’, ‘*processo civil*’, ‘*cc 7*’, ‘*codigo processo*’, ‘*art 102*’, ‘*art 21*’, ‘*lei 9*’, ‘*cpc art*’, and ‘*carlos velloso*’
9	‘*art 557*’, ‘*constituicao federal*’, ‘*art 102*’, ‘*celso mello*’, ‘*sepulveda pertence*’, ‘*lei 9*’, ‘*lei 8*’, ‘*constituicao republica*’, ‘*carlos velloso*’, and ‘*processo civil*’
10	‘*lei 9*’, ‘*lei complementar*’, ‘*constituicao federal*’, ‘*2o lei*’, ‘*lei estadual*’, ‘*art 12*’, ‘*artigo 12*’, ‘*art 7o*’, ‘*f constituicao*’, and ‘*paragrafo unico*’
11	‘*lei 8*’, ‘*codigo penal*’, ‘*processo penal*’, ‘*lei 5*’, ‘*art 102*’, ‘*art 21*’, ‘*codigo processo*’, ‘*constituicao federal*’, ‘*art 84*’, and ‘*penal art*’
12	‘*celso mello*’, ‘*art 102*’, ‘*hc 79*’, ‘*constituicao federal*’, ‘*lei 8*’, ‘*art 21*’, ‘*lei 9*’, ‘*moreira alves*’, ‘*art 1o*’, and ‘*adc 4*’
13	‘*re 226*’, ‘*art 557*’, ‘*cpc art*’, ‘*constituicao federal*’, ‘*lei 8*’, ‘*5o xxxvi*’, ‘*moreira alves*’, ‘*art 21*’, ‘*art 544*’, and ‘*art 5o*’
14	‘*celso mello*’, ‘*constituicao federal*’, ‘*adi 2*’, ‘*adi 1*’, ‘*art 103*’, ‘*mello adi*’, ‘*art 21*’, ‘*lei 9*’, ‘*lei complementar*’, and ‘*moreira alves*’

In general, the vast majority of top 100 terms related to the 15 latent topics (presented in Table 6.7) are related to the citation of legislative references and precedents (57% and 40%, respectively). However, we observed on topic 5 the presence of person citations, where 10% of the top 100 terms were related to it. In Table 6.6, two bi-grams emerged for the topic 5: '*gurgel santos*' and '*monteiro gurgel*'; both seem related to the name *Roberto Monteiro Gurgel Santos*, a Brazilian Jurist who became Attorney General (or *Procurador-geral*) in 2009, staying until 2013. His tenure coincides with the presence of the topic.

Also, as observed in Section 4.5, the citation of precedent in the early 2000's was less frequent than the citations to legislative references (in 2000, the legislative reference were cited 70% more them precedents), and topics 11 and 13 illustrate that scenario, as the majority of the top 100 terms related to these topics (87% and 75%, respectively) are related to legislative references citation. Also, these topics are closely related to different categories, *Writs* and *Criminal*. Even so, in topics 12 and 14, 45% and 42% (respectively) of the top 100 terms are related to citation of precedents.

Table 6.7: Distribution of the Top 100 per entity.

Topic	Legislative Ref.	Precedent	Academic	Person
0	36	64	0	0
1	64	34	0	2
2	52	46	0	2
3	54	46	0	0
4	42	58	0	0
5	59	30	0	11
6	50	48	0	2
7	40	60	0	0
8	55	41	4	0
9	64	36	0	0
10	74	24	0	2
11	88	6	0	6
12	53	45	0	2
13	75	25	0	0
14	58	42	0	0

6.2.2
EXP2.2: Exploring the precedents citation per Cabinets on Writs cases.

Despite the number of topics presented in the previous experiment, we can go further. For the *EXP2.2* we narrowed down the scope to the *Writs* case, as in *EXP1.2*, but limiting to only three fine-grained entities from precedent citations: *legal class procedure*, *legal number procedure*, and *reporting justice*. As presented in Section 2.2, the legal class and the number identifies a legal case cited within the precedent, so we transformed this two entities into one compound term, as one single token by join these two information with "_" (*e.g.*, "*lpc_hc lpn_232618*", became "hc_1234"). The *reporting justice* was already a single token, as presented in Subsection 5.3.3.

For instance, the precedent in its textual form "*lpc_hc lpn_232618 lpo_ms rj_jorge_mussi pd_20.06.2012*"

became: "*hc_232618 jorge_mussi*".

Regarding the legal case transformation, an especial case is the citation to binding rulings ("*súmulas*" and "*súmulas vinculantes*") which represents not a legal case but a common understanding of the court regarding a specific issue. Also, the "*súmulas vinculantes*" represent a special type of biding precedent that only the STF is allowed to issue and which must be respected by all lower courts. The binding precedent was transformed in the following way: "*súmulas*" became "*sum*", and "*súmulas vinculantes*" became "*sum_vinc*".

Our main goal with this experiment is to explore the relationship between each of these entities with decisions issued by each Cabinet on Writs. To do so, we propose two different analyses: (I) using only the *legal case*, and (II) the second one using only the *Reporting Justice*. In the scenario I, we seek to correlate latent topics with patterns of legal case citations and Cabinets. With the second one, scenario II, we explore how each Justice (represented by the Cabinet) cites precedents reported by their pairs and him/herself. For both scenarios, we structured an origin-oriented tensor: the tensor $\mathcal{C}$ for the first scenario, and $\mathcal{R}$ for the second one.

6.2.2.1 Scenario I: Using only the legal case citations

Figure 6.9 present the tensor structure for the first scenario $\mathcal{C} \in \mathbb{R}^{c \times e \times t}$, defined by *cabinet* × *entity* × *time*, where c_{ijk} represents the relevance score of the entity (legal case) i on the decision related to *Writs* cases, issued by the Cabinet j, and published on k. The entity score was measured by the entity frequency in the corpus divided by the entity frequency in the corpora, a simplified version of the TF-IDF. Also, we chose to include only citations that appeared in at least two different corpus in the second mode, and we limited the temporal interval to 2012-18 due to the lower number of citations per year before 2012, as presented in Table 4.8.

We ran the NNCPD for the entire tensor for 15 topics, with a reconstruction error $\left\|\mathcal{C} - \hat{\mathcal{C}}\right\|_F = 0.697$. in Figure 6.10, a representation of the factors A and B, and in Table 6.8 the top 5 entities according factor C. Observing Figure 6.10 we see that most of the latent topics are closely related to one Cabinet only. For instance, topics 8 and 9 are complementary (topic 9 emerged when topic 8 began to fade), and both are strongly associated with Cabinet 10. Therefore, these topics might be related to a transition of a citation pattern on decisions issued by that Cabinet.

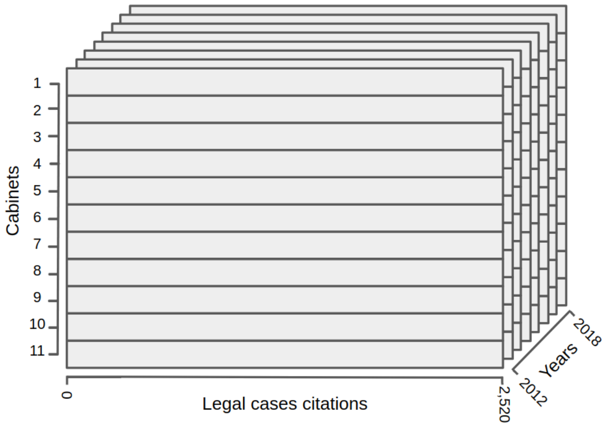


Figure 6.9: The tensor structure (frontal slice).

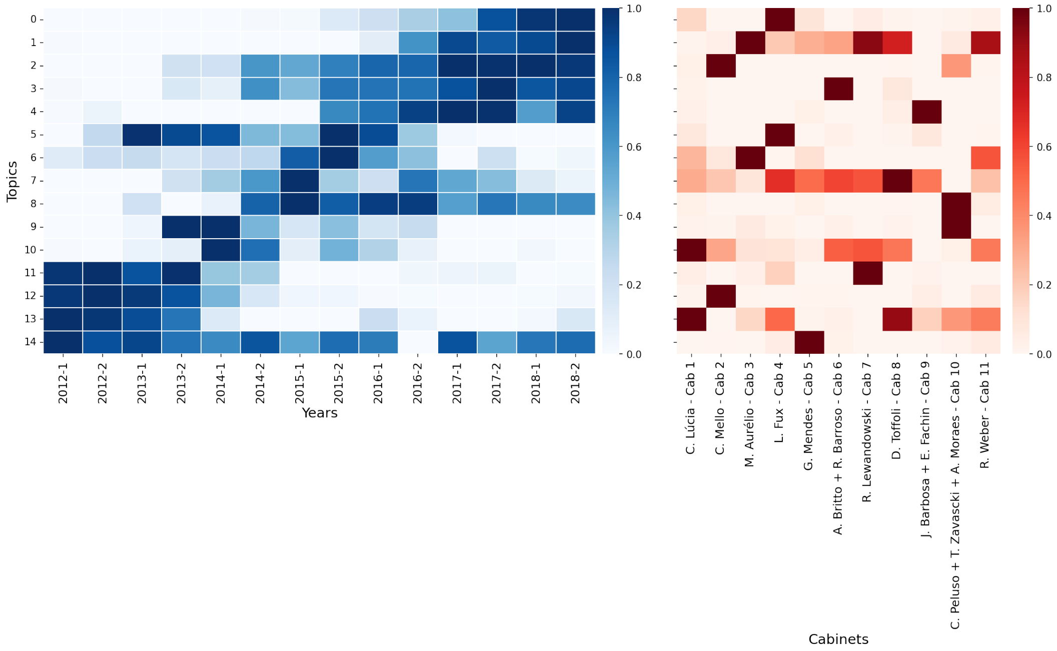


Figure 6.10: Legal named entities usage for writs legal cases.

Some topics, however, are closely related to many Cabinets, such as topics 1, 7, 10, and 13, representing a common pattern of citation among Cabinets. These topics are also complementary and might be related to a transition of an understanding regarding common issues in *Writs*' cases. Since precedents can be seen as elements of the argumentation process presented in a decision, this transition in topics might be related to changes on what was once seen as a common-sense among the Justices. For instance, topic 7 represents a common citation pattern among the majority of the Cabinets but began to fade when another common pattern (topic 1) emerged.

Observing Table 6.8, we see the presence of biding precedent citations on most of the long-lasting latent topics. For instance, topic 1, in which two of the five most relevant citations are binding rulings. The biding precedent is

Table 6.8: Topics represented by legal cases and bind precedents citations.

Topic	Top 5 legal case and binding rulings citations
0	*Sum. vinc. 26, HC 130439, ADI 4451, RHC 108877,* and *ADI 2135.*
1	*HC 126292, Sum. 339, Sum. vinc. 26, Sum. vinc. 33,* and *RE 964246.*
2	*RHC 114961, RHC 114737, HC 122718, Sum. 339,* and *Sum. vinc. 26.*
3	*RHC 115983, RHC 114890, RHC 117491, HC 100994,* and *HC 86656.*
4	*HC 79775, RCL 17867, RCL 20026, RHC 135560,* and *RCL 11985.*
5	*RCL 11985, AC 2177, MI 1194, RE 561836,* and *MI 940.*
6	*RE 760931, Sum. vinc. 26, RCL 2, RCL 1,* and *Sum. vinc. 11.*
7	*ADI 4357, RE 760391, RCL 16492, RCL 19907,* and *Sum. 279.*
8	*HC 97009, ADI 4357, RCL 31, ADI 2602,* and *HC 117090.*
9	*MS 28371, ADI 2602, ADI 4357, RHC 108877,* and *HC 97009.*
10	*Sum. vinc. 33, RCL 7547, RCL 4940, AC 2177,* and *ADI 4167.*
11	*HC 84349, HC 99031, RCL 14419, RCL 12994,* and *ADIN 3395.*
12	*MI 715, MI 670, RTJ 110 555, RTJ 129 1199,* and *HC 79775.*
13	*MI 1194, MI 375, RCL 7547, RCL 8150,* and *RE 591797.*
14	*HC 85185, HC 79775, HC 79776, RHC 108877,* and *RCL 9545.*

very often applied to dismiss cases, either because the issue doesn't fall within the court's purview or because it deals with something previously settled by the court. Even so, regarding issues already settled, topic 1 began to emerge in 2016, the same year where the STF's Plenary issued collegiate decisions on case *HC 126292*[2] and case *RE 964246*[3], and both collegiate decisions are related to the general repercussion requirement[4]. These cases illustrate two different scenarios where the court decided to apply the general repercussion requirement, becoming an example for a future similar case and becoming a valued precedent to justify the dismissal of a case.

6.2.2.2 Scenario II: Using only the Reporting Justice citations

Similarly to the structure used in Scenario I, for Scenario II, we structured a tensor $\mathcal{J} \in \mathbb{R}^{c \times e \times t}$, defined by $cabinet \times entity \times time$, where j_{ijk} represents the relevance score of the reporting Justice citation i on the decision related to *Writs* cases, issued by the Cabinet j, and published on k. The entity score was measured by the entity frequency in corpus divided by the entity frequency in the corpora. Also, we selected to compose the second mode only the citations that appear in at least two different corpus, and we limited the temporal interval to 2012-18 due to the lower number of citations per year before 2012.

[2] HC 126292. Decision available in: `https://redir.stf.jus.br/paginadorpub/paginador.jsp?docTP=TP&docID=10964246`

[3] RE 964246. Decision available in `https://redir.stf.jus.br/paginadorpub/paginador.jsp?docTP=TP&docID=12095503`

[4] A definition of general repercussion is presented in Subsection 2.1.1, it is a filter created by constitutional amendment in order to enable the court to select the concrete constitutional review cases it deems relevant.

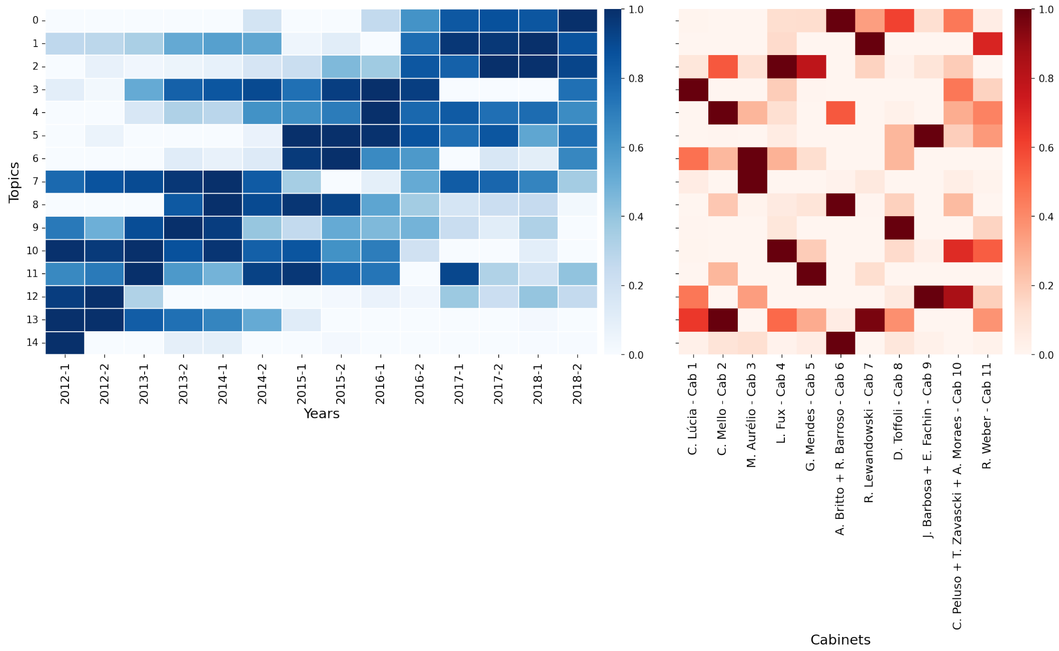


Figure 6.11: Legal named entities usage for writs legal cases.

Similarly to the structure used in the Scenario I, we also ran the NNCPD for the entire tensor for 15 topics, with a reconstruction error $\left\|\mathcal{J} - \hat{\mathcal{J}}\right\|_F = 0.393$ (the lowest reconstruction error among the experiments). In Figure 6.11, a representation of the factors A and B, and in Table 6.9 the top 5 reporting Justices according factor C.

Similar to Scenario I, most of the topics are closely related to one Cabinet only. However, this relationship is not as strong as in Scenario I. For instance, topic 0 is closely related to Cabinet 6, but it is also significantly close to other Cabinets. The Cabinets, in turn, are often related to more than the two topics. Cabinet 6 is the one worth mentioning to be closely related to four different and complementary topics (topics 0, 4, 8, and 14). The patterns that we observe in topics 8 and 14 identify most of the decisions issued by Cabinet 6 until 2015.1. After this date, the Cabinet's citation pattern became closer to what was issued by the other Cabinets.

Another observation worth pointing out is the effect of the absence of the Chief Justice's decisions. As stated in Section 2.1, we omitted the decisions issued by the Justice when they are the Chief Justice. This absence is presented in the Figure 6.11 as gaps in the topic's history. For instance, between September 2016 and September 2018, Justice Cármen Lúcia was the Chief in Justice, and the absence of her decisions, while in this role, impacted the topics related to Cabinet 1 (her Cabinet). Topic 3 has almost no value for the semesters 2017.1, 2017.2, and 2018.1, and a similar effect can be seen on topic 6, but with less impact. When Justice Cármen Lúcia returned, as regular Justice, both topics "appeared" again. Also, similar effects are noticed for the period when other Justices assumed that role: Ricardo Lewandowski (2014-16), Joaquim Barbosa (2012-14), and Ayres Britto (2012).

Table 6.9: Topics represented by Reporting Justice citations.

Topic	Closely related to Cabinet	Top 5 Reporting Justice citations.
0	6	*Teori Zavascki (Cab. 10), Ricardo Lewandowski (Cab. 7), Luiz Fux (Cab. 4), Rosa Weber (Cab. 11),* and *Cezar Peluso (Cab. 10).*
1	7	*Gilmar Mendes (Cab. 5), Cármen Lúcia (Cab. 1), Dias Toffoli (Cab. 8), Luiz Fux (Cab. 4),* and *Roberto Barroso (Cab. 6).*
2	4	*Rosa Weber (Cab. 11), Dias Toffoli (Cab. 8), Roberto Barroso (Cab. 6), Joaquim Barbosa (Cab. 9),* and *Teori Zavascki (Cab. 10).*
3	1	*Ellen Gracie (Cab. 11), Gilmar Mendes (Cab. 5), Joaquim Barbosa (Cab. 9), Marco Aurélio (Cab. 3),* and *Luiz Fux (Cab. 4).*
4	2	*Gilmar Mendes (Cab. 5), Teori Zavascki (Cab. 10), Ricardo Lewandowski (Cab. 7), Dias Toffoli (Cab. 8),* and *Maurício Corrêa (Cab. 4).*
5	9	*Roberto Barroso (Cab. 6), Luiz Fux (Cab. 4), Teori Zavascki (Cab. 10), Rosa Weber (Cab. 11),* and *Dias Toffoli (Cab. 8).*
6	3	*Cezar Peluso (Cab. 10), Ayres Britto (Cab. 6), Sepúlveda Pertence (Cab. 8), Luiz Fux (Cab. 4),* and *Gilmar Mendes (Cab. 5).*
7	3	*Ayres Britto (Cab. 6), Ellen Gracie (Cab. 11), Luiz Fux (Cab. 4), Dias Toffoli (Cab. 8),* and *Marco Aurélio (Cab. 3)* .
8	6	*Luiz Fux (Cab. 4), Rosa Weber (Cab. 11), Cármen Lúcia (Cab. 1), Teori Zavascki (Cab. 10),* and *Cezar Peluso (Cab. 10).*
9	8	*Cármen Lúcia (Cab. 1), Menezes Direito (Cab. 8), Ricardo Lewandowski (Cab. 7), Sepúlveda Pertence (Cab. 8),* and *Carlos Velloso (Cab. 7).*
10	4	*Cármen Lúcia (Cab. 1), Ricardo Lewandowski (Cab. 7), Joaquim Barbosa (Cab. 9), Cezar Peluso (Cab. 10),* and *Dias Toffoli (Cab. 8).*
11	5	*Moreira Alves (Cab. 9), Cezar Peluso (Cab. 10), Marco Aurélio (Cab. 3), Maurício Corrêa (Cab. 4),* and *Rosa Weber (Cab. 11).*
12	9	*Gilmar Mendes (Cab. 5), Marco Aurélio (Cab. 3), Cármen Lúcia (Cab. 1), Sepúlveda Pertence (Cab. 8),* and *Cezar Peluso (Cab. 10).*
13	2 and 7	*Cezar Peluso (Cab. 10), Sepúlveda Pertence (Cab. 8), Gilmar Mendes (Cab. 5), Carlos Velloso (Cab. 7),* and *Maurício Corrêa (Cab. 4).*
14	6	*Célio Borja (Cab. 11), Marco Aurélio (Cab. 3), Moreira Alves (Cab. 9), Maurício Corrêa (Cab. 4),* and *Cármen Lúcia (Cab. 1).*

In Table 6.9 we also indicate to what Cabinet the topic has a strong relationship. This table allows us to investigate how often a Justice cites precedent from another Justice. For instance, it is not common the self-citation among Cabinets on decisions related to *Writs* case. Nevertheless, topic 9 illustrates a situation where Cabinet 8 cites previous Justices from the same Cabinet. Another observation, for the period from 2012-18, only Justice Dias Toffoli was in Cabinet 8 (and still is nowadays), and topic 9 is closely related to his Cabinet, and two of the five top citations are related to his predecessors. Furthermore, topic 9 represents patterns that last for most of the period analyzed but began to fade in 2016, when topic 0 emerged.

6.3 The STF's Vocabulary Study Case

For this final experiment, we adapted the proposed process to explore the court's vocabulary changes throughout the time, where the latent topics are used to highlight the changes in the patterns of the word's usage. We grouped decisions by Cabinets (similar to EXP2.2) and by semester. We computed the set of unique words and their frequencies per group, considering only the unique sentences. To avoid the terms less relevant and an oversized tensor representation, we ignored from these sets words that appear in less than 1% or with a frequency higher than the third quartile. With the criterion, most of the words ignored are typos and foreign words; with the second criterion, very common words were ignored, such as "*stf*", "*processo*" (*law case*) and "*tribunal*" (*court*).

As presented in Section 4.5, the STF has been increasing its production of unique content (unique monocratic decisions), and it's also producing longer decisions — on average, decisions published in 2018 are four times longer than the one published in 2000. However, as also presented in that section, when inspecting the sentences within a decision, only 25% are indeed unique. With the experiment presented in this section, we took another step forward by exploring the court's vocabulary in time.

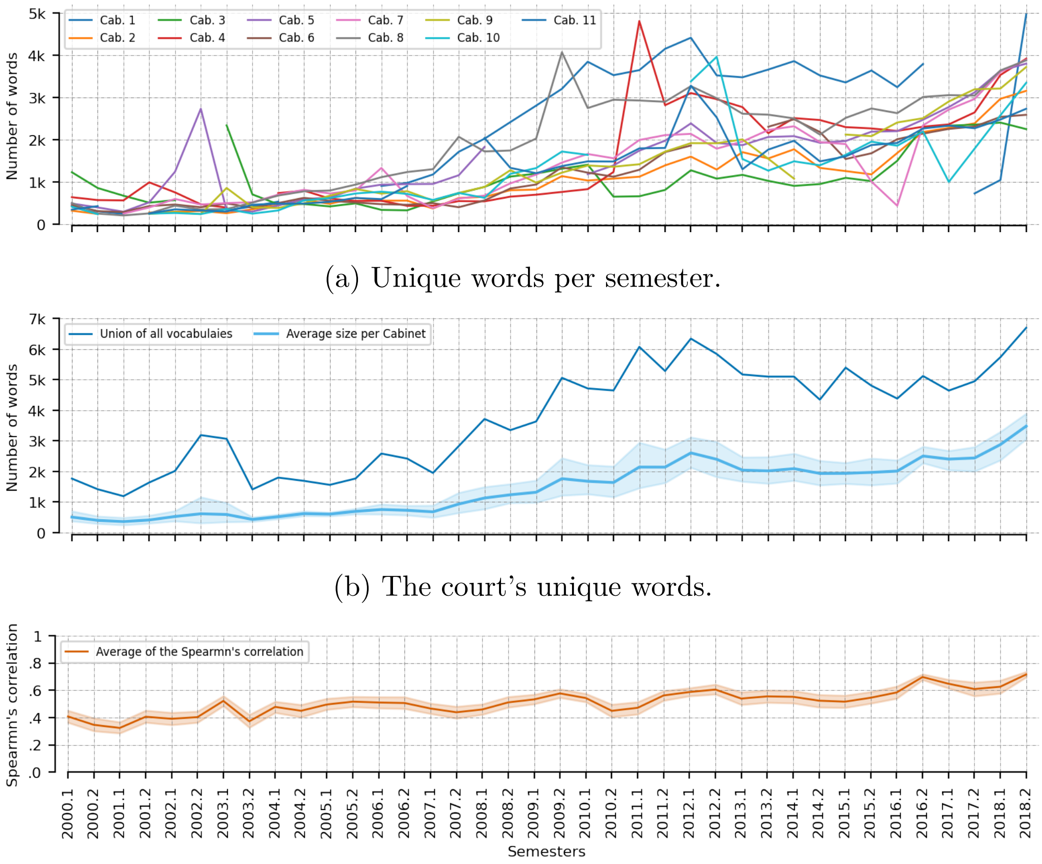
First, before presenting the tensor representation, we present some analysis of the court's vocabulary. Each Cabinet contributes differently to form the court's vocabulary — we call the "court's vocabulary" the aggregation of all Cabinet's vocabulary. As presented in Figure 6.12.a, each Cabinet's vocabulary evolved in distinct ways[5] varying its contribution, in time, to form the court's vocabulary (Figure 6.12.b).

With the rising of the word embedding techniques, such as the *word2vec* Mikolov et al. (2013), many of the studies that explore semantic textual features, or systems that learn from a collection of text (*e.g.*, text classifiers) ignore the temporal aspects and text origins (*i.e.* who wrote it). However, when dealing with legal documents, we point out the relevance of the time and the origin. It's worth mentioning that the correlation between Cabinets' vocabularies was lower in the past (Figure 6.12.c) indicating that decisions from different Cabinets were dissimilar in this respect. With time, this correlation has increased.

Figure 6.13 depicts the correlation between Cabinets' vocabularies in more detail[6]. In the early 2000s (Figure 6.13.a), there was litt.e correlation between Cabinets and the total of words in the STF's vocabulary was only 1,765. The vocabulary variation does not necessarily indicate different content — two different texts can deliver the same message —, but the legal text is a special case of highly conventionalized language. In this case, the vocabulary

[5] We observe in Figure 6.12 some discontinuities (when a line suddenly has no value) this is due to vacancy or when a Justice assumes the role as Chief Justice.

[6] We observe in Figure 6.13 "X" and some values close to zero (Figure 6.13.b) this is also due to the vacancy in Cabinet or when the Justice assumes the role as Chief Justice.



(a) Unique words per semester.

(b) The court's unique words.

(c) Correlation between Cabinet's vocabularies.

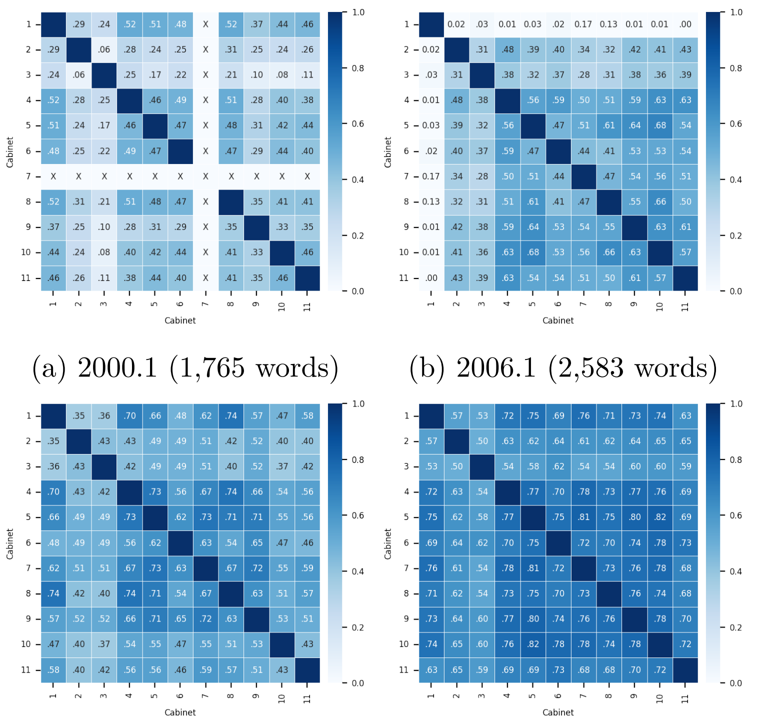
Figure 6.12: The STF's Vocabulary Evolution.

difference between documents might indicate semantics dissimilarity. In 2018 (Figure 6.13.a, we observe a completely different scenario, the court's vocabulary has 3.8 times more words, and the vocabularies employed in each Cabinet are much more similar than in the past.

6.3.1 Exploring the Vocabulary over Time

With topic modeling, we can also explore the dissimilarity among Cabinet's vocabularies. For the experiment, we structured an origin-oriented tensor $\mathcal{V} \in \mathbb{R}^{c \times w \times t}$, defined by *cabinet* × *word* × *time* — instead of bi-grams, words of the vocabulary —, where j_{ijk} represents the frequency of the word j on the decisions published by the Cabinet i, and published on k. The words in the second mode of the tensor are the result of the aggregation of all the words found in each group of decisions (per Cabinet per semester), resulting in a total of 14,145 words.

We also ran the NNCPD for the entire tensor for 15 topics, with a reconstruction error $\left\|\mathcal{V} - \hat{\mathcal{V}}\right\|_F = 0.570$. In Figure 6.14, a representation of



(a) 2000.1 (1,765 words) (b) 2006.1 (2,583 words)

(c) 2012.1 (6,337 words) (d) 2018.2 (6,690 words)

Figure 6.13: The STF's Vocabulary Evolution.

the factors A and B. The factor C, the words per topic, became too verbose and difficult to explore.

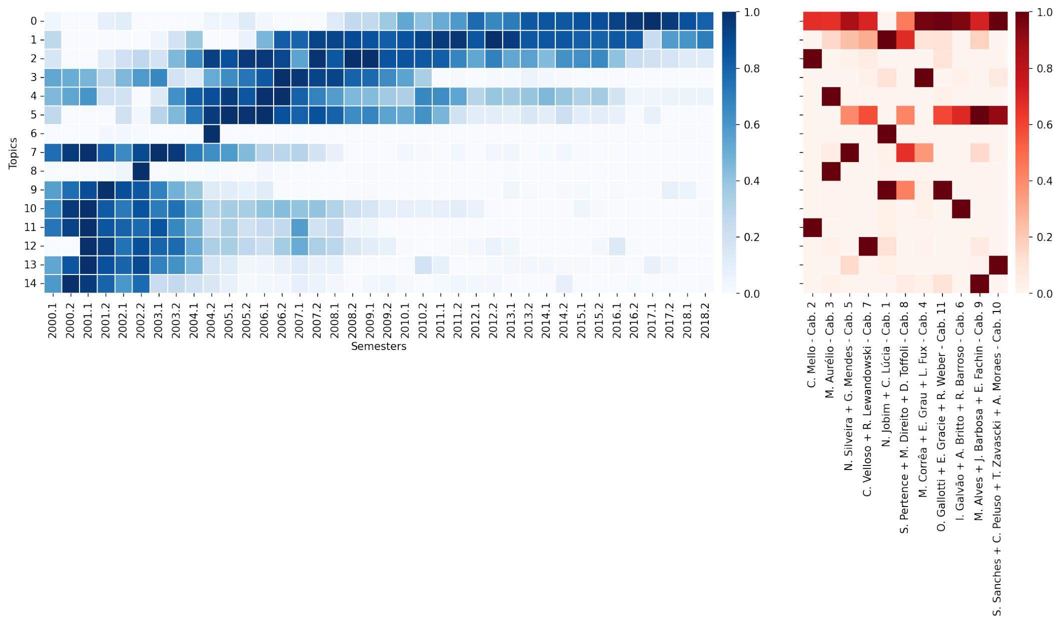


Figure 6.14: Exploring the vocabulary through latent topics.

Figure 6.14 gives us a visual representation of the dissimilarity between vocabularies. Before 2004, most of the predominant topics (8, 9, 10, 11, 12, 13 and 14) are closely related to one Cabinet only. These topics represent patterns of words usage that exclusively identify a Cabinet. After 2004, Cabinets began to share similar patterns and, in 2016, only two patterns were predominant in court, topics 0 and 1. Therefore, the vocabulary differences became less relevant, even with longer decisions.

Figure 6.14 also gives us insights regarding different moments in court history. We can split the data into three moments: (i) before 2004, (ii) between 2005 and 2011, and (iii) after 2012. Historic factors corroborate this idea. In December 2004, was approved the Constitutional Amendment 45 known as "Reform of the Judiciary", aimed to give more speed and efficiency to the judicial system. This Amendment created the "general repercussion" requirement, as presented in section 2.1.1, it is a filter created in order to enable the court to select the concrete constitutional review cases it deems relevant Mendes (2017). Also, it allowed the STF to issue binding rulings (*Súmulas Vinculantes*).

The second moment coincides with the increase in precedents citations on the decisions, a tendency confirmed, and enforced, with the approval of the New Code of Civil Procedure (March 2015) by Congress (Law 13,105) (Donizetti, 2015). We also observed the impact of the changes in 2012 in the experiments *EXP1.3* and *EXP2.2*.

6.4 Concluding Remarks

In summary, Table 6.10 presents an overview of the parameters and results of the tensor NNCPD decompositions. In this table, we also present the tensor density per experiment. The density is related to the number of zeros in the tensor: the more zeros, the sparser the tensor. We saw a clear correlation between the tensor sparsity and the Reconstruction error: the sparser the tensor, the greater the error. Consequently, with smaller the errors, the better is the latent topic modeling precision.

Table 6.10: Summary of the NNCPD decomposition.

Experiment	Structure	Tensor shape	Tensor density	Num. of topics	Rec. error	Running time (sec)
EXP1.1	decision-oriented	$19 \times 750 \times 3,000$	0.008	5	0.956	34.4
EXP1.2	decision-oriented	$19 \times 1,000 \times 3,000$	0.012	10	0.958	191
EXP1.3 [2000-11]	decision-oriented	$24 \times 500 \times 3,000$	0.008	10	0.970	85
EXP1.3 [2012-18]	decision-oriented	$14 \times 1,000 \times 3,000$	0.017	10	0.947	100
EXP2.1	origin-oriented	$38 \times 5 \times 2,654$	0.37	15	0.509	13.3
EXP2.2 [Scen. I]	origin-oriented	$14 \times 11 \times 2,520$	0.046	15	0.697	7.27
EXP2.2 [Scen. II]	origin-oriented	$14 \times 11 \times 225$	0.196	15	0.393	0.4
EXP3	origin-oriented	$38 \times 11 \times 14,145$	0.098	15	0.570	34

A better precision does not invalidate the modeling results. As we observed with the experiments, they were coherent. Nevertheless, the origin-oriented strategy delivered better results. Moreover, it is important to observe the running time of the NNCPD decomposition. The method is fast and cost-effective, for the NNCPD decomposition were used a regular modern notebook with a processor *11th Gen Intel(R) Core(TM) i7-1185G7 @ 3.00GHz* with 16Gb of RAM, without a discrete graphics card. But the memory would be a limitation for larger numbers of topics or to decompose larger tensors.

7
Conclusion

This chapter presents our final considerations about our work presented in this thesis. In Section 1.2 at the Introduction, we posed the main research question (*MRQ*) we aimed to answer and, in order to do it, we defined a sequence of related questions. In the subsequent chapters, we presented the path we paved and followed, seeking to answer all of them.

Regarding the research question *RQ1*, related to the legal named entities that we can find in an STF ruling, we mapped two levels of nested legal entities: four coarser-grained entities related to twenty-four nested ones (fine-grained). Also, we built the largest annotated corpus in Portuguese with a focus on the legal domain. This work would not have been possible without the collaboration and support of the FGV Direito Rio Law School and the researchers engaged in this study.

The reliability of the annotation process testifies the quality of the annotated corpus. Even so, we observed that the workload could be lower, with a smaller set of text annotated. The excerpts could be even smaller, and we could achieve a better distribution of coarser annotations and fine-grained annotations.

The set of annotated entities can also be improved by reducing the number of fine-grained entities by ignoring very rare entities. For instance, the "subsection" entity for fine-grained annotation of legislative references could be removed since it is present in less than 4% of the annotated legislative references. Also, the fine-grained annotation could be improved with a label suggestion tool based on a pattern-matching strategy (*e.g.*, to suggest legal procedure numbers and class) where the annotator could accept or not the suggestion. The pattern-matching approach can produce a good result for extracting simple precedent citations (only the references to previous legal procedures and its number). Our results showed that 93% of the precedent citations have the legal procedure class followed by the legal procedure number. However, only 22% of the annotated precedents citations do not have other fine-grained information, 78% also have other fine-grained entities, and 43% have at least 4 fine-grained entities.

Regarding the relevance of the mapped entities (*RQ1.a*), we took into account findings in studies in the legal domain (Falcão et al., 2019; Pereira et al., 2020). The extraction results presented in Section 4.5 confirms the relevance of these entities, showing how often they are cited. Nevertheless, as mentioned earlier, the set of entities can be improved, especially the fine-grain level.

The whole process employed for the mapping, the annotation task, and the annotated corpus building are also detailed in our publication, entitled *Fine-grained legal entity annotation: A case study on Brazilian Supreme Court*, published by the Journal *Information Processing & Management* (Correia et al., 2022).

Regarding the NER task, the extraction of the legal named entities (related to the sub-question *RQ1.b*), we presented in Section 4.5 the path that we followed to extract the entities from the collection of monocratic decisions that we have at our disposal.

Regarding the research question *RQ2*, we presented in Chapter 5 a proposal for a process that uses the dynamic topic modeling as a tool for an interpretable exploration of large collections of documents in a temporal perspective. Topic modeling relies on the NNCPD method for tensor decomposition, a method first explored for dynamic topic modeling by Ahn et al. (2021). The NNCPD reveals patterns of events, through latent topics, without an upfront classification, working in an arbitrary number of dimensions, which means it can detect complicated relationships between several data attributes simultaneously. We limited the scope to its application on monocratic decisions issued by the STF, but the proposed process can be adapted to many other similar scenarios — any large collection of documents written by a few authors regardless of the domain.

Regarding the sub-questions *RQ2.a*, *RQ2.b*, and *RQ2.c*, the primary motivation for choosing NCPD was its capability to produce coherent and interpretable results. We also introduced two different strategies of tensor structuring: based on the monocratic textual features (decision-oriented structure) and another based on the origin of the decision (origin-oriented structure). We also showed how to use the legal named entities as a textual feature.

The experiments showed the potential of the bi-gram usage for textual feature representation. Even so, despite the robustness of NCPC in noisy data, still lacks an evaluation regarding the noise of the employment of this strategy. For instance, in *EXP 2.1* we observed the presence of the bi-grams '*gurgel santos*' and '*monteiro gurgel*' with a strong correlation to a specific latent topic, and both bi-grams are related to the name *Roberto Monteiro Gurgel Santos*. A more intelligent form of representation would be to represent a person's names and common textual expressions as a single symbol. By doing this, the name *Roberto Monteiro Gurgel Santos* would be represented as a single symbol instead of two, reducing the noise inserted by the use of bi-grams.

In Chapter 6, our experiments and illustration present, through figures and tables, how to interpret the produced results. We also explored their coherence and their reliability by correlating latent topics with historical events. Throughout latent topics, we explored the usage of legal entities in the decision argumentation process throughout time. Nevertheless, our experiments were not enough to state the process accuracy for short-lasting topic identifications. Kassab et al. (2021) presents a possible path to identify short-lasting topics that can be used to improve our proposed process.

The proposed process still lacks an accurate qualitative assessment by legal domain experts. Nonetheless, given the observed results, we were able to give an answer to our main research question. The combination of textual feature extraction with latent topic modeling through tensor decomposition with NNCPD can extract reliable information from a large collection of legal documents. For the textual feature extraction, we presented the usage of the bi-grams and the usage of legal entities for a fine-grained evaluation of the collection in time. Regarding the tensor representation, both strategies presented in this study could produce coherent results, being the origin-oriented structure the one that presented better detail quality.

In the following section, we state some future works and enhancements.

7.1 Future Work

As mentioned earlier, the proposed process still lacks an accurate and extensive qualitative assessment. A future work to be developed is a qualitative evaluation with legal specialists. As a result, we could propose a quality measure for a given latent topic model based on the qualitative evaluation, correlating the results with features from (i) the corpus (*e.g.*, vocabulary and average document size), (ii) the tensor representation (*e.g.*, density and structure), and (iii) the NNCPD parameters (*e.g.*, number of topics).

Another future work regarding the qualitative evaluation would be comparing terms related to topics to terms listed by legal specialists. A group of legal specialists could produce a list of relevant words regarding a specific legal subject (*e.g.*, preventive arrest). Then, the proposed process could be used to extract latent topics over a collection of decisions closely related to the same legal subject. Finally, a quality assessment measure could be done based on the presence of the terms listed by the specialists on the latent topics.

The NER task was the most expensive task in the whole process. Even so, it still lacks some essential improvements regarding the standardization of the extracted entities. As an illustration, as mentioned in Section 5.3.3, consider the legislative reference fine-grained entities. There is a finite number of legal acts, institutions, and origins, but there are multiple ways of citing any of these elements in a decision. A relevant improvement would be the standardization of the different citation formats of these elements. For instance, the legal act *Lei 13.105/15* (*Law 13,105/15*) might also be cited as "*Código de Processo Civil*" (*Code of Civil Procedure*) or by its acronym "*CPC*", or even as "*Novo Código de Processo Civil*" (*New Code of Civil Procedure*). All these representations could be translated to a unique code (*e.g.*, "*LAW13105*"), and the effect of this would reflect as an enhancement in quality for the latent topics modeling. A similar solution would be interesting to codify common legal textual expressions, such as legal jargon, converting them to a single code.

The last experiment (Section 6.3) presented how to employ latent topic modeling to explore the vocabulary in time. We observed the presence of at least three distinguishable moments in the court's vocabulary history. Therefore, a possible future work would be a study case of how Machine Learning (ML) Systems deal with these changes in time, especially those based on *Deep Learning*. For instance, we could measure how significant the performance changes are when comparing ML models trained based on these three different moments. A possible result would be developing a validation test for ML models that measures how much of the training collection stills represent the current scenario — for example, measuring the current reliability of an ML System for monocratic classification trained with decisions published between 2012 and 2018.

Furthermore, our first publication regarding legal entities was Correia et al. (2019), where we discussed the precedent relevance. Thus, another future work would be to explore the precedent relevance from a topic modeling perspective, correlating latent topics with precedent relevance and comparing the results with the results presented in our first publication. Also, in Correia et al. (2019) precedents were structured as a graph representation and were applied network-science techniques for relevance measurements. Hence, a future work could also explore the tensor decomposition to study the graph changes in time.

Finally, the topic modeling presented in this thesis could be transformed into a tool for legal data exploration, a web application. Thus, another future work would be the development of an iterative tool for dynamic topic modeling, combining it with information retrieval. The challenges to doing such work would be numerous, mainly in the Legal domain but also in Information Extraction, Computing Optimization, and Human-Computer Interaction fields. However, such a tool could be helpful to legal researchers.

Bibliography

AHN, M.; EIKMEIER, N.; HADDOCK, J.; KASSAB, L.; KRYSHCHENKO, A.; LEONARD, K.; NEEDELL, D.; MADUSHANI, R. W. M. A.; SIZIKOVA, E. ; WANG, C.. **On large-scale dynamic topic modeling with nonnegative cp tensor decomposition**. Advances in Data Science, p. 181–210, 2021.

ALMEIDA, G. F. C. F.; ALMEIDA, D. S.. **Presidência do stf em números: de gracie a toffoli**. 2020.

ALMEIDA, G. F. C. F.; CHRISMANN, P. H. V.. **Os paradoxos da deliberação judicial colegiada**. Revista de Investigações Constitucionais, 6:165–188, 4 2019.

ALMEIDA, G. F. C. F.; HARTMANN, I. A.. **Decisões monocráticas**. In: Arantes, R. B.; Arguelhes, D. W., editors, O ESTADO DA ARTE DA PESQUISA EMPÍRICA SOBRE O STF. São Paulo, SP, 2022. Unpublished.

AMBROSINO, A.; CEDRINI, M.; DAVIS, J. B.; FIORI, S.; GUERZONI, M. ; NUCCIO, M.. **What topic modeling could reveal about the evolution of economics**. Journal of Economic Methodology, 25(4):329–348, 2018.

ANTONIAK, M.; MIMNO, D.. **Evaluating the stability of embedding-based word similarities**. Transactions of the Association for Computational Linguistics, 6:107–119, 2018.

ARGUELHES, D. W.; RIBEIRO, L. M.. **'the court, it is i'? individual judicial powers in the brazilian supreme court and their implications for constitutional theory**. Global Constitutionalism, 7(2):236–262, 2018.

CARVALHO, A. G. P. D.; ROESLER, C. R.. **O argumento de autoridade no supremo tribunal federal: uma análise retórica em perspectiva histórica**. Revista Direito, Estado e Sociedade, 2019.

CHALKIDIS, I.; ANDROUTSOPOULOS, I. ; MICHOS, A.. **Extracting contract elements**. In: PROCEEDINGS OF THE INTERNATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE AND LAW, p. 19–28, New York, New York, USA, 6 2017. Association for Computing Machinery.

CHALKIDIS, I.; ANDROUTSOPOULOS, I. ; ALETRAS, N.. **Neural legal judgment prediction in English**. In: ACL 2019 - 57TH ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, PROCEEDINGS OF THE CONFERENCE, p. 4317–4323. Association for Computational Linguistics (ACL), 2020.

DE JUSTIÇA, C. N.. **Resolução nº 65 de 16/12/2008**, 1 2009.

DE JUSTIÇA, C. N.. **Justiça em números 2021**. Technical report, Conselho Nacional de Justiça, 2021.

CORREIA, F. A.; NUNES, J. L.; ALMEIDA, GUILHERME FRANCA COUTO FERNANDES, A. A. A. ; LOPES, H.. **An exploratory analysis of precedent relevance in the brazilian supreme court rulings**. In: PROCEEDINGS OF THE ACM SYMPOSIUM ON DOCUMENT ENGINEERING 2019, DocEng '19, New York, NY, USA, 2019. Association for Computing Machinery.

CORREIA, F. A.; NUNES, J. L.; ALMEIDA, G. F. C. F.; ALMEIDA, A. A. ; LOPES, H.. **An exploratory analysis of precedent relevance in the brazilian supreme court rulings**. In: PROCEEDINGS OF THE ACM SYMPOSIUM ON DOCUMENT ENGINEERING, DOCENG 2019, p. 1–4, New York, New York, USA, 9 2019. Association for Computing Machinery, Inc.

CORREIA, F. A.; ALMEIDA, A. A. A.; NUNES, J. L.; SANTOS, K. G.; HARTMANN, I. A.; SILVA, F. A. ; LOPES, H.. **Fine-grained legal entity annotation: A case study on the brazilian supreme court**. Information Processing Management, 59(1):102794, 2022.

DADGOSARI, F.; GUIM, M.; BELING, P. A.; LIVERMORE, M. A. ; ROCKMORE, D. N.. **Modeling law search as prediction**. Artificial Intelligence and Law, p. 1–32, 2 2021.

DIAS, M.; BONÉ, J.; FERREIRA, J. C.; RIBEIRO, R. ; MAIA, R.. **Named entity recognition for sensitive data discovery in portuguese**. Applied Sciences 2020, Vol. 10, Page 2303, 10:2303, 3 2020.

DONIZETTI, E.. **A força dos precedentes no novo código de processo civil**. Direito UNIFACS – Debate Virtual, 0(175), Jul 2015.

2018 reform of eu data protection rules. Technical report, 2018.

FALCÃO, J.; CERDEIRA, P. C. ; ARGUELHES, D. W.. **I Relatório Supremo em Números - o Múltiplo Supremo**. Escola de Direito do Rio de Janeiro da Fundação Getulio Vargas, Rio de Janeiro, Dec 2012.

FALCÃO, J.; HARTMANN, I. A. ; CHAVES, V. P.. **III Relatório Supremo em Números: o Supremo e o Tempo**. Escola de Direito do Rio de Janeiro da Fundação Getulio Vargas, Rio de Janeiro, Sep 2014.

FALCÃO, J.; ARGUELHES, D. W. ; RECONDO, F.. **Onze supremos: O supremo em 2016**. p. 51 – 54, 2017.

FALCÃO, J.; HARTMANN, I. A.; ALMEIDA, G. F. C. F. ; CHAVES, L.. **V Relatório Supremo em Números: O Foro Privilegiado e o Supremo**. Escola de Direito do Rio de Janeiro da Fundação Getulio Vargas, Rio de Janeiro, 3 2017.

FALCÃO, J.; GÓES, S. B. C.; HARTMANN, I. ; ALMEIDA, G. F. C. F.. **A Realidade do Supremo Criminal**. Technical report, FGV Direito Rio, Rio de Janeiro, 2019.

GALGANI, F.; COMPTON, P. ; HOFFMANN, A.. **LEXA: Building knowledge bases for automatic legal citation classification**. Expert Systems with Applications, 42(17-18):6391–6407, 5 2015.

GREENE, D.; CROSS, J. P.. **Exploring the political agenda of the european parliament using a dynamic topic modeling approach**. Political Analysis, 25(1):77–94, 2017.

HADDOCK, J.; KASSAB, L.; KRYSHCHENKO, A. ; NEEDELL, D.. **On nonnegative cp tensor decomposition robustness to noise**. In: 2020 INFORMATION THEORY AND APPLICATIONS WORKSHOP (ITA), p. 1–7, 2020.

HAMILTON, W. L.; LESKOVEC, J. ; JURAFSKY, D.. **Diachronic word embeddings reveal statistical laws of semantic change**. In: PROCEEDINGS OF THE 54TH ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS (VOLUME 1: LONG PAPERS), p. 1489–1501, Berlin, Germany, Aug. 2016. Association for Computational Linguistics.

HAMILTON, W. L.; LESKOVEC, J. ; JURAFSKY, D.. **Cultural shift or linguistic drift? comparing two computational measures of semantic change**. In: PROCEEDINGS OF THE 2016 CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING, p. 2116–2121, Austin, Texas, Nov. 2016. Association for Computational Linguistics.

HARTMANN, I.. **Crise dos precedentes no supremo: O caso dos precedentes sobre liberdade de expressÃo**. REI - REVISTA ESTUDOS INSTITUCIONAIS, 6:109–128, 4 2020.

HARTMANN, I.; CHADA, D.. **A razão sem condições de qualidade (reason without conditions for quality)**. 2015.

HARTMANN, I. A.; KELLER, C. I.; CHADA, D.; VASCONCELOS, G.; NUNES, J. L.; CARNEIRO, L.; CHAVES, L.; BARRETO, M.; CORREIA, F. ; ARAÚJO, F.. **O impacto no sistema prisional brasileiro da mudança de entendimento do supremo tribunal federal sobre execução da pena antes do trânsito em julgado no hc 126.292/sp - um estudo empírico quantitativo**. Revista de Direito Econômico e Socioambiental, 9:399–426, 7 2018.

HENRETTY, T. S.; LANGSTON, M. H.; BASKARAN, M.; EZICK, J. ; LETHIN, R.. **Topic modeling for analysis of big data tensor decompositions**. In: Blowers, M.; Hall, R. D. ; Dasari, V. R., editors, DISRUPTIVE TECHNOLOGIES IN INFORMATION SCIENCES, volumen 10652, p. 52 – 64. International Society for Optics and Photonics, SPIE, 2018.

HÖFLER, S.; PIOTROWSKI, M.. **Building corpora for the philological study of Swiss legal texts**. Journal for Language Technology and Computational Linguistics, 26(2):77–89, 2011.

JI, D.; TAO, P.; FEI, H. ; REN, Y.. **An end-to-end joint model for evidence information extraction from court record document**. Information Processing and Management, 57(6):102305, 11 2020.

KANAPALA, A.; PAL, S. ; PAMULA, R.. **Text summarization from legal documents: a survey**. Artificial Intelligence Review, 51(3):371–402, 3 2019.

KASSAB, L.; KRYSHCHENKO, A.; LYU, H.; MOLITOR, D.; NEEDELL, D. ; REBROVA, E.. **Detecting short-lasting topics using nonnegative tensor decomposition**, 2021.

KOLDA, T. G.; BADER, B. W.. **Tensor decompositions and applications**. http://dx.doi.org/10.1137/07070111X, 51:455–500, 8 2009.

KRIPPENDORFF, K.. **Reliability in Content Analysis**. Human Communication Research, 30(3):411–433, 7 2004.

KULKARNI, V.; AL-RFOU, R.; PEROZZI, B. ; SKIENA, S.. **Statistically significant detection of linguistic change**. In: PROCEEDINGS OF THE 24TH INTERNATIONAL CONFERENCE ON WORLD WIDE WEB, WWW '15, p. 625–635, Republic and Canton of Geneva, CHE, 2015. International World Wide Web Conferences Steering Committee.

LEIBON, G.; LIVERMORE, M.; HARDER, R.; RIDDELL, A. ; ROCKMORE, D.. **Bending the law: geometric tools for quantifying influence in the multinetwork of legal opinions**. Artificial Intelligence and Law, 26(2):145–167, 6 2018.

LEITNER, E.; REHM, G. ; MORENO-SCHNEIDER, J.. **Fine-grained named entity recognition in legal documents**. In: LECTURE NOTES IN COMPUTER SCIENCE (INCLUDING SUBSERIES LECTURE NOTES IN ARTIFICIAL INTELLIGENCE AND LECTURE NOTES IN BIOINFORMATICS), volumen 11702 LNCS, p. 272–287. Springer, 9 2019.

LI, R.; TIAN, P. ; WANG, S.. **Study concept drift in 150-year english literature**. CEUR workshop proceedings, 2871:153–163, 2021. Publisher Copyright: © 2021 CEUR-WS. All rights reserved.; 1st Workshop on AI + Informetrics, AII 2021 ; Conference date: 17-03-2021.

LORENZETTO, B. M.; KENICKE, P. H. ; GALLOTTI. **Relação dos doutrinadores brasileiros de direito constitucional mais citados pelo supremo tribunal federal nos casos de controle concentrado de constitucionalidade**.

LUZ DE ARAUJO, P. H.; CAMPOS, T. E.; OLIVEIRA, R. R.; STAUFFER, M.; COUTO, S. ; BERMEJO, P.. **Lener-br: A dataset for named entity recognition in brazilian legal text**. In: LECTURE NOTES IN COMPUTER SCIENCE (INCLUDING SUBSERIES LECTURE NOTES IN ARTIFICIAL INTELLIGENCE AND LECTURE NOTES IN BIOINFORMATICS), volumen 11122 LNAI, p. 313–323. Springer Verlag, 9 2018.

MEDVEDEVA, M.; VOLS, M. ; WIELING, M.. **Using machine learning to predict decisions of the European Court of Human Rights**. Artificial Intelligence and Law, 28(2):237–266, 6 2020.

MENDES, C. H.. **The Supreme Federal Tribunal of Brazil**, p. 115–153. Cambridge University Press, 2017.

MERCHANT, K.; PANDE, Y.. **NLP Based Latent Semantic Analysis for Legal Text Summarization**. In: 2018 INTERNATIONAL CONFERENCE ON ADVANCES IN COMPUTING, COMMUNICATIONS AND INFORMATICS, ICACCI 2018, p. 1803–1807. Institute of Electrical and Electronics Engineers Inc., 11 2018.

MIKOLOV, T.; CHEN, K.; CORRADO, G. ; DEAN, J.. **Efficient estimation of word representations in vector space**. 2013.

O'CALLAGHAN, D.; GREENE, D.; CARTHY, J. ; CUNNINGHAM, P.. **An analysis of the coherence of descriptors in topic modeling**. Expert Systems with Applications, 42(13):5645–5657, 2015.

PEREIRA, T.; ARGUELHES, D. W. ; ALMEIDA, G. F. C. F.. **Quem decide no Supremo?** Technical report, FGV Direito Rio, Rio de Janeiro, 2020.

RINGLAND, N.; DAI, X.; HACHEY, B.; KARIMI, S.; PARIS, C. ; CURRAN, J. R.. **NNE: A Dataset for Nested Named Entity Recognition in English Newswire**. ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, p. 5176–5181, 6 2019.

SUPREMO TRIBUNAL FEDERAL. **Relatório de atividades 2019**. Technical report, Supremo Tribunal Federal, Brasilia, 2020.

FEDERAL, S. T.. **Relatório de de atividades do supremo tribunal federal - 2020**. Technical report, Supremo Tribunal Federal, 2021.

WANG, Z.; WU, Y.; LEI, P. ; PENG, C.. **Named entity recognition method of brazilian legal text based on pre-training model**. Journal of Physics: Conference Series, 1550:032149, 5 2020.

WYNER, A.; PETERS, W. ; KATZ, D.. **A case study on legal case annotation**. In: FRONTIERS IN ARTIFICIAL INTELLIGENCE AND APPLICATIONS, volumen 259, p. 165–174, 1 2013.