João Pedro Khair Cunha

Aplicação de técnicas de NLP e clusterização para segmentação de prospectos de fundos UCITS

PROJETO FINAL

DEPARTAMENTO DE INFORMÁTICA
Programa de Graduação em Engenharia da Computação

Rio de Janeiro
Julho de 2024

Pontifícia Universidade Católica
do Rio de Janeiro

João Pedro Khair Cunha

Aplicação de técnicas de NLP e clusterização para segmentação de prospectos de fundos UCITS

Relatório de Projeto Final II

Relatório de Projeto Final, apresentado ao Programa de Engenharia da Computação, do Departamento de Informática da PUC-Rio como requisito parcial para a obtenção do titulo de Bacharel em Engenharia da Computação.

Orientador: Prof. Álvaro de Lima Veiga Filho

Rio de Janeiro
Julho de 2024

João Pedro Khair Cunha

Graduando em Engenharia da Computação na PUC - Rio

Ficha Catalográfica

Khair Cunha, João Pedro

Aplicação de técnicas de NLP e clusterização para segmentação de prospectos de fundos UCITS / João Pedro Khair Cunha; orientador: Álvaro de Lima Veiga Filho. – 2024.

37 f: il. color. ; 30 cm

Projeto Final - Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Informática, 2024.

Inclui bibliografia

1. Informática – Trabalho de Conclusão de Curso. 2. Machine Learning. 3. Processamento de Linguagem Natural. 4. Clusterização. 5. UCITS. 6. Fundos de Investimento. 7. Finanças. I. Veiga, Álvaro. II. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Informática. III. Título.

CDD: 004

Para meus pais, por todo amor e apoio.

Agradecimentos

Para meu orientador e mentor Álvaro Veiga, por todo o apoio no início da minha vida profissional e durante a elaboração dessa tese.

Para meus pais, Filipe e Claudia, que sempre se sacrificaram para me proporcionar essa e todas as oportunidades.

Para meu avô, meu maior ídolo, que despertou em mim a curiosidade e a paixão pela Engenharia.

Para toda minha família, em especial Giovanna, Cleusa, Eligier e Marina, por me amar e torcer por mim, mesmo nos momentos mais difíceis.

Para meus amigos, em especial Bruno, Breno(s), Daniel, Fernando, Marcos, Pedro e Rafael, que me acompanharam nessa jornada acadêmica.

Para a PUC-Rio e seu corpo docente, por me proporcionar a infraestrutura e o conhecimento necessários para meu aprendizado.

Resumo

Khair Cunha, João Pedro; Veiga, Álvaro. **Aplicação de técnicas de NLP e clusterização para segmentação de prospectos de fundos UCITS**. Rio de Janeiro, 2024. 37p. Projeto Final – Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

O Processamento de Linguagem Natural (NLP) revolucionou a leitura e processamento automático de textos, culminando na emergência de modelos *Large Language Models* que possibilitaram uma compreensão e extração de informações em níveis jamais vistos. Neste estudo, explorou-se o uso de técnicas de NLP e Machine Learning para segmentar prospectos de fundos de investimento do tipo UCITS (*Undertakings for Collective Investment in Transferable Securities*), no intuito de otimizar a coleta de dados não estruturados contidos neles. Utilizando algoritmos de frequência de n-gramas e clusterização, esta tese busca expandir os horizontes de aplicação de Inteligência Artificial no âmbito do mercado financeiro.

Palavras-chave

Machine Learning; Processamento de Linguagem Natural; Clusterização; UCITS; Fundos de Investimento; Finanças.

Abstract

Khair Cunha, João Pedro; Veiga, Álvaro (Advisor). **Application of NLP and Clustering Techniques for Segmentation of UCITS Fund Prospectuses**. Rio de Janeiro, 2024. 37p. Projeto Final – Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

Natural Language Processing (NLP) has revolutionized the automatic reading and processing of texts, culminating in the emergence of Large Language Models that have made it possible to understand and extract information at unprecedented levels. This study explored the use of NLP and Machine Learning techniques to segment UCITS (Undertakings for Collective Investment in Transferable Securities) investment fund prospectuses in order to optimize the collection of unstructured data contained therein. Using n-gram frequency and clustering algorithms, this thesis seeks to expand the application horizons of Artificial Intelligence in the scope of financial markets.

Keywords

Machine Learning; Natural Language Processing; Clustering; UCITS; Investment Funds; Finance.

Sumário

Lista de figuras

Lista de tabelas

Lista de algoritmos

Lista de Códigos

Lista de Abreviaturas

NLP – *Natural Language Processing*

IA – Inteligência Artificial

ML – *Machine Learning* ou Aprendizado de Máquina

1
Introdução

A Inteligência Artificial emergiu como uma das tecnologias mais promissoras e transformadoras do século XXI. Com a capacidade de simular processos de pensamento humano e aprender com dados, a IA está revolucionando uma ampla gama de setores, desde a medicina até a indústria, da educação à economia. Seu impacto é evidente em nossas vidas diárias, à medida que assistentes virtuais, carros autônomos, sistemas de recomendação e muito mais se tornam parte integrante do nosso cotidiano.

A capacidade da IA de realizar tarefas que anteriormente eram exclusivas de seres humanos é uma área de pesquisa e aplicação que tem atraído a atenção de cientistas, engenheiros, empresários e acadêmicos em todo o mundo. Nessa ótica, o campo do Processamento de Linguagem Natural (*Natural Language Processing*, ou NLP) ganha destaque, ao permitir que máquinas compreendam e interajam com a linguagem humana de maneira sem precedentes. O NLP está revolucionando a forma de lidar com dados textuais, abrindo caminho para a leitura e interpretação automáticas de textos.

Este Trabalho de Conclusão de Curso tem como objetivo explorar o avanço do NLP e sua aplicação específica na análise automática de textos financeiros, com foco na segmentação automática de documentos de fundos de investimento. Serão abordados princípios e técnicas fundamentais do NLP, bem como estudos de caso que demonstram como essa tecnologia está sendo aplicada com sucesso para analisar, classificar e extrair informações críticas de documentos financeiros complexos. Adeamis, visa-se buscar soluções alternativas, mais simples e baratas que as soluções em voga atualmente, sem prejudicar a assertividade da análise.

2
Situação Atual

A inteligência artificial no setor financeiro remonta ao início dos anos 1980 e é utilizada no auxílio das mais diversas atividades. Em 1982, a empresa norte-americana Apex lançou o *PlanPower*, um programa de IA focado no planejamento fiscal e de investimentos para clientes de alta renda (NIELSON; BROWN; PHILLIPS, 1990). Em 1990, iniciou-se o uso da IA em detecção de fraudes, e estima-se que, em apenas 2 anos, foi capaz de evitar mais de US$ 1 bilhão em danos (SENATOR et al., 1995). Atualmente, o emprego de técnicas mais avançadas de IA, como Deep Learning e NLP, tem se tornado cada vez mais frequente, em aplicações como reconhecimento de imagens de satélite para decisões de investimento, modelos preditivos de preços de ativos, análise de crédito, entre muitas outras.

Atualmente, o uso de NLP para leitura automática de textos financeiros, embora bastante incipiente, mostra-se incrivelmente promissora. Segundo um estudo publicado em 2023 pela empresa de inteligência de mercado MarketsandMarkets, o mercado global de NLP em finanças em 2023 é de US$ 5.5 bilhões, e deve ultrapassar US$ 18 bilhões em 2028 (MARKETS, 2023). Com o advento dos modelos de *Large Language Model* (LLM), os estudos de processamento, leitura e extração de informações na área foram enormemente impulsionadas.

Dentre suas aplicações mais usuais, pode-se destacar a análise de sentimento de notícias e seu impacto no desempenho dos agentes financeiros. O uso de técnicas de Big Data e modelos de *Deep Learning* em reportagens de veículos globais de notícias sobre os mercados de capitais, como *StockTwits* e *seekingAlpha*, permite a melhor previsibilidade do sentimento de mercado (SOHANGIR et al., 2018). Ademais, a utilização do Chat-GPT revolucionou esse nicho, permitindo um maior aprofundamento e qualidade de extração de

informações dessas fontes (FATOUROS et al., 2023).

Nesse contexto, o subnicho de fundos de investimento é ainda pouco explorado. Um fundo de investimento é um mecanismo de aplicação de recursos em conjunto com outros investidores, permitindo a contratação de gestores profissionais, a redução de custos de transação e a diversificação de ativos para minimizar riscos e maximizar retornos (BANK, Acessado em 30/06/20234). O mercado europeu de fundos totalizou mais de US$ 19 trilhões (EFAMA, Acessado em 30/10/2023), de modo que a extração de dados de documentos é bastante interessante acadêmica e financeiramente. No entanto, tal tarefa traz diversos desafios, como a grande extensão e falta de padronização dos documentos, que podem chegar a milhares de páginas. Isso dificulta o uso de modelos mais sofisticados de LLM, os quais possuem limitações de tamanho de entrada (LI et al., 2024).

Relacionado ao problema supracitado, tem-se os esforços para adequadamente segmentar documentos via NLP. A estrutura organizacional dos textos, como formatação espacial, fonte e cores, além do contexto do conteúdo, ajuda o cérebro humano a agrupar a informação em trechos (parágrafos, seções e capítulos. Isso, no entanto, não é trivial para o processamento computacional, e disso derivam-se diversos estudos em documentos longos, como textos jurídicos (AUMILLER et al., 2021). Uma boa segmentação textual em subpartes facilita a extração de informações financeiras relevantes via modelos de Deep Learning e LLM, como feito por Zhu et. al (ZHU et al., 2024).

Por fim, a análise da frequência de palavras se faz bastante valiosa para a correta segmentação de documentos longos. A utilização de técnicas de pré-processamento, combinado a métricas como TF-IDF (*term frequecy-inverse document frequency*) (RAJARAMAN; ULLMAN, 2011), medida estatística de relevância de termos em um documento, mostrou-se bastante efetiva para textos longos de conteúdo geral, e foi empregado na mineração de dados para o setor financeiro (BACH et al., 2019). Os estudos no ramo de fundos comprova a

eficiência de diferentes técnicas que, combinadas com outras ferramentas como modelos não-supervisionados, podem aprimorar ainda mais os usos existentes no processamento de documentos na área de finanças.

3
Objetivos do Trabalho

O presente estudo objetiva explorar as técnicas de NLP e Machine Learning na segmentação de documentos financeiros longos, mais especificamente prospectos de fundos de investimentos do tipo UCITS (Undertakings for Collective Investment in Transferable Securities). A descrição e especificação legal dos UCITS e dos prospectos será um dos pontos focais do capítulo seguinte.

O trabalho foi organizado segundo as etapas abaixo.

1. **Criação de base de documentos:** foi realizada a coleta dos prospectos a partir de um algoritmo de web-scraping em Python.

2. **Pré-processamento:** cada prospecto foi digitalizado e processamento dinamicamente, utilizando técnicas de NLP

3. **Análise de frequência de n-gramas:** foi criado um vetor de frequência por página/seção do texto, o qual será usado como entrada para o modelo de classificação

4. **Segmentação:** foram testados diferentes algoritmos de clusterização, no intuito de classificar cada página do prospecto como:

   (a) **Geral:** possui informações relativas à companhia gestora ou a múltiplos UCITS

   (b) **Específico:** contém dados individuais relacionados a um fundo específico

5. **Extração de segmentos e tabelas:** serão selecionadas as páginas classificadas como "Específico"de cada prospecto, e serão extraídas todas as tabelas das páginas do tipo "Geral", no intuito de coletar o máximo de informação dos UCITS.

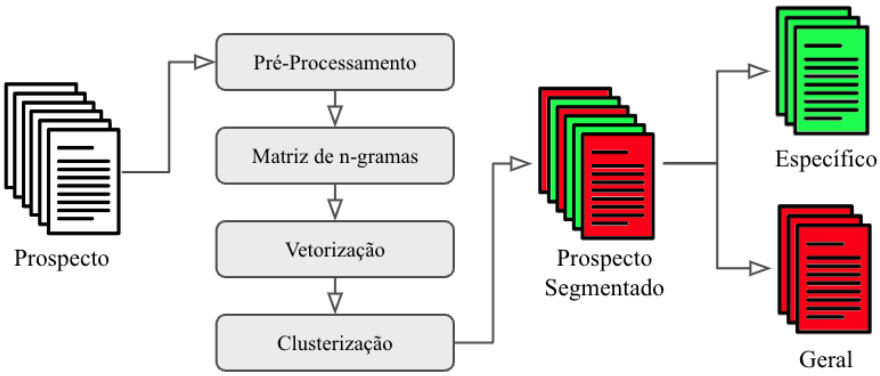


Figura 3.1: Proposta de Fluxograma do Projeto

4
Pesquisas Realizadas

Neste capítulo, serão abordadas as pesquisas realizadas para a elaboração desta tese. Os tópicos discutidos incluem desde o escopo financeiro, como UCITS e documentos referentes a eles, até temas mais técnicos, de pré-processamento, clusterização e métricas de avaliação do modelo.

4.1
UCITS

UCITS, acrônimo para *Undertakings for Collective Investment in Transferable Securities*, é o principal tipo de fundo de investimento estabelecido na Europa. É um arcabouço jurídico para o registro e negociação de veículos de investimento em valores mobiliários nos países europeus. (ESMA, 2024).

Atualmente, há mais de 30000 UCITS em atividade, representando um AUM (*Assets Under Management* - Ativos Sob Gestão) de mais de € 9 bilhões (ESMA, 2023). Segundo a Comissão Europeia, órgão legislativo supranacional responsável por essa regulamentação, os UCITS correspondem por 75% de todo o investimento da população europeia no mercado financeiro (European Commission, 2024).

Os UCITS são regulados conforme a diretiva 2009/65/EC da União Europeia, a qual estabelece que toda gestora de UCITS deve divulgar uma série de documentos acerca dos veículos de investimento, como o prospecto inicial, relatórios para cada ano fiscal, relatórios semestrais de desempenho, entre outros.

Embora as empresas administradoras de UCITS divulguem parte das informações em formato estruturado, como é o caso de demonstrações financeiras, grande parcela dos dados se encontram em fontes não-estruturadas, como textos, áudios ou vídeos. Estes configuram fonte de informação valiosa para atores do mercado financeiro, como bancos, gestoras ou empresas de in-

teligência e consultoria.

4.2 Prospecto de UCITS

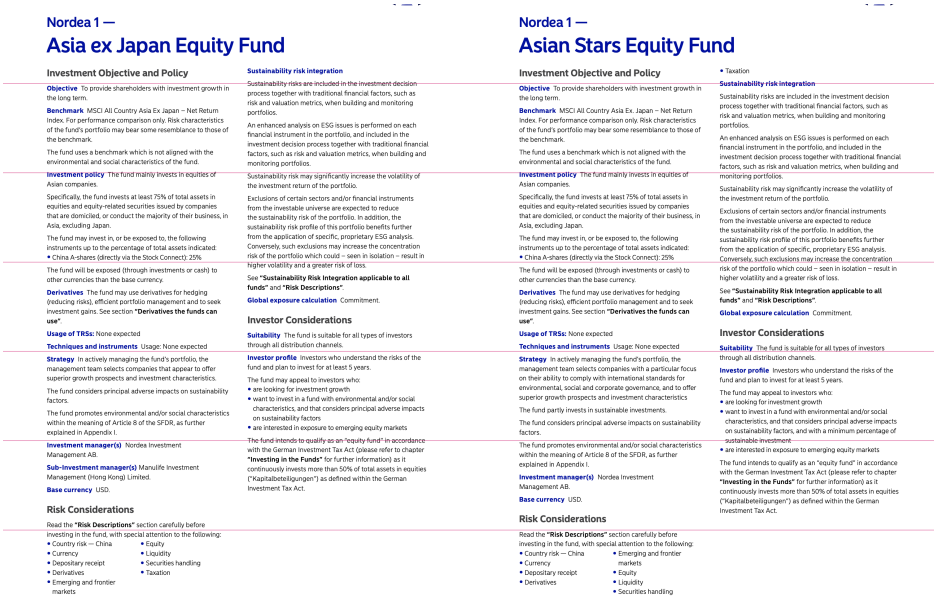
O prospecto é um documento financeiro que contém todas as informações sobre a companhia gestora e seus fundos de investimento. De acordo com a legislação europeia, é mandatória a divulgação do prospecto por empresas com AUM superior a € 1 milhão, ficando a critério de cada estado-membro elevar o limite mínimo para € 8 milhões (Directive 2009/65/EC, 2009).

Como já mencionado, o prospecto de fundos UCITS possui pouca padronização quanto à forma e organização dos dados, porém a diretiva da UE é bastante clara no tocante à informação obrigatório que deve estar contida nele, de acordo com o anexo I e o artigo 71 da diretiva 2009/65/EC. Dentre as principais, pode-se citar:

1. Informações sobre a companhia gestora, domiciliada ou não na União Europeia
    (a) Nome, natureza jurídica e endereços de registro e atividade
    (b) Data de incorporação e duração, caso seja de natureza limitada
    (c) Nome e posição de membros, gestores e supervisores
    (d) Dados sobre o capital gerido, estrutura acionária e mercados de valores mobiliários onde há listagem de ativos
2. Informações sobre os fundos de investimento
    (a) Nome, natureza jurídica
    (b) Data de criação
    (c) Dados operacionais, procedimentos, taxas e condições de recompra e distribuição de proventos
    (d) Políticas de recompra, distribuição de proventos e alocação de capital.

Além dos pontos acima, um dado relevante e presente na maioria dos prospectos é o SFDR (*Sustainable Finance Disclosures Regulation*), conjunto de informações acerca da atuação de cada UCITS no tocante à sustentabilidade, visando fomentar a economia social e ambientalmente sustentável (European Parliament, 2020).

Foi realizada a análise de mais de 4 mil prospectos e se concluiu que estes podem ser divididos em 2 grupos no tocante à organização dos dados. O primeiro diz respeito aos prospectos que incluem seções individuais de cada UCITS, contendo informações como objetivo de investimento, índice de referência, políticas de investimento e hedging, moeda, etc. Isso significa que é possível classificar as páginas e seções do texto nos 2 segmentos – "Geral" ou "Específico" –, conforme pontuado no capítulo anterior.



Nordea 1 —
Asia ex Japan Equity Fund

Investment Objective and Policy

Objective To provide shareholders with investment growth in the long term.

Benchmark MSCI All Country Asia Ex Japan – Net Return Index. For performance comparison only. Risk characteristics of the fund's portfolio may bear some resemblance to those of the benchmark.

The fund uses a benchmark which is not aligned with the environmental and social characteristics of the fund.

Investment policy The fund mainly invests in equities of Asian companies.

Specifically, the fund invests at least 75% of total assets in equities and equity-related securities issued by companies that are domiciled, or conduct the majority of their business, in Asia, excluding Japan.

The fund may invest in, or be exposed to, the following instruments up to the percentage of total assets indicated:
• China A-shares (directly via the Stock Connect): 25%

The fund will be exposed (through investments or cash) to other currencies than the base currency.

Derivatives The fund may use derivatives for hedging (reducing risks), efficient portfolio management and to seek investment gains. See section **"Derivatives the funds can use"**.

Usage of TRSs: None expected

Techniques and instruments Usage: None expected

Strategy In actively managing the fund's portfolio, the management team selects companies that appear to offer superior growth prospects and investment characteristics.

The fund considers principal adverse impacts on sustainability factors.

The fund promotes environmental and/or social characteristics within the meaning of Article 8 of the SFDR, as further explained in Appendix I.

Investment manager(s) Nordea Investment Management AB.

Sub-investment manager(s) Manulife Investment Management (Hong Kong) Limited.

Base currency USD.

Risk Considerations

Read the **"Risk Descriptions"** section carefully before investing in the fund, with special attention to the following:
• Country risk — China
• Currency
• Depositary receipt
• Derivatives
• Emerging and frontier markets
• Equity
• Liquidity
• Securities handling
• Taxation

Sustainability risk integration

Sustainability risks are included in the investment decision process together with traditional financial factors, such as risk and valuation metrics, when building and monitoring portfolios.

An enhanced analysis on ESG issues is performed on each financial instrument in the portfolio, and included in the investment decision process together with traditional financial factors, such as risk and valuation metrics, when building and monitoring portfolios.

Sustainability risk may significantly increase the volatility of the investment return of the portfolio.

Exclusions of certain sectors and/or financial instruments from the investable universe are expected to reduce the sustainability risk of the portfolio. In addition, the sustainability risk profile of this portfolio benefits further from the application of specific, proprietary ESG analysis. Conversely, such exclusions may increase the concentration risk of the portfolio which could – seen in isolation – result in higher volatility and a greater risk of loss.

See **"Sustainability Risk integration applicable to all funds"** and **"Risk Descriptions"**.

Global exposure calculation Commitment.

Investor Considerations

Suitability The fund is suitable for all types of investors through all distribution channels.

Investor profile Investors who understand the risks of the fund and plan to invest for at least 5 years.

The fund may appeal to investors who:
• are looking for investment growth
• want to invest in a fund with environmental and/or social characteristics, and that considers principal adverse impacts on sustainability factors
• are interested in exposure to emerging equity markets

The fund intends to qualify as an "equity fund" in accordance with the German Investment Tax Act (please refer to chapter **"Investing in the Funds"** for further information) as it continuously invests more than 50% of total assets in equities ("Kapitalbeteiligungen") as defined within the German Investment Tax Act.

Nordea 1 —
Asian Stars Equity Fund

Investment Objective and Policy

Objective To provide shareholders with investment growth in the long term.

Benchmark MSCI All Country Asia Ex. Japan – Net Return Index. For performance comparison only. Risk characteristics of the fund's portfolio may bear some resemblance to those of the benchmark.

The fund uses a benchmark which is not aligned with the environmental and social characteristics of the fund.

Investment policy The fund mainly invests in equities of Asian companies.

Specifically, the fund invests at least 75% of total assets in equities and equity-related securities issued by companies that are domiciled, or conduct the majority of their business, in Asia, excluding Japan.

The fund may invest in, or be exposed to, the following instruments up to the percentage of total assets indicated:
• China A-shares (directly via the Stock Connect): 25%

The fund will be exposed (through investments or cash) to other currencies than the base currency.

Derivatives The fund may use derivatives for hedging (reducing risks), efficient portfolio management and to seek investment gains. See section **"Derivatives the funds can use"**.

Usage of TRSs: None expected

Techniques and instruments Usage: None expected

Strategy In actively managing the fund's portfolio, the management team selects companies with a particular focus on their ability to comply with international standards for environmental, social and corporate governance, and to offer superior growth prospects and investment characteristics

The fund partly invests in sustainable investments.

The fund considers principal adverse impacts on sustainability factors.

The fund promotes environmental and/or social characteristics within the meaning of Article 8 of the SFDR, as further explained in Appendix I.

Investment manager(s) Nordea Investment Management AB.

Base currency USD.

Risk Considerations

Read the **"Risk Descriptions"** section carefully before investing in the fund, with special attention to the following:
• Country risk — China
• Currency
• Depositary receipt
• Derivatives
• Emerging and frontier markets
• Equity
• Liquidity
• Securities handling
• Taxation

Sustainability risk integration

Sustainability risks are included in the investment decision process together with traditional financial factors, such as risk and valuation metrics, when building and monitoring portfolios.

An enhanced analysis on ESG issues is performed on each financial instrument in the portfolio, and included in the investment decision process together with traditional financial factors, such as risk and valuation metrics, when building and monitoring portfolios.

Sustainability risk may significantly increase the volatility of the investment return of the portfolio.

Exclusions of certain sectors and/or financial instruments from the investable universe are expected to reduce the sustainability risk of the portfolio. In addition, the sustainability risk profile of this portfolio benefits further from the application of specific, proprietary ESG analysis. Conversely, such exclusions may increase the concentration risk of the portfolio which could – seen in isolation – result in higher volatility and a greater risk of loss.

See **"Sustainability Risk integration applicable to all funds"** and **"Risk Descriptions"**.

Global exposure calculation Commitment.

Investor Considerations

Suitability The fund is suitable for all types of investors through all distribution channels.

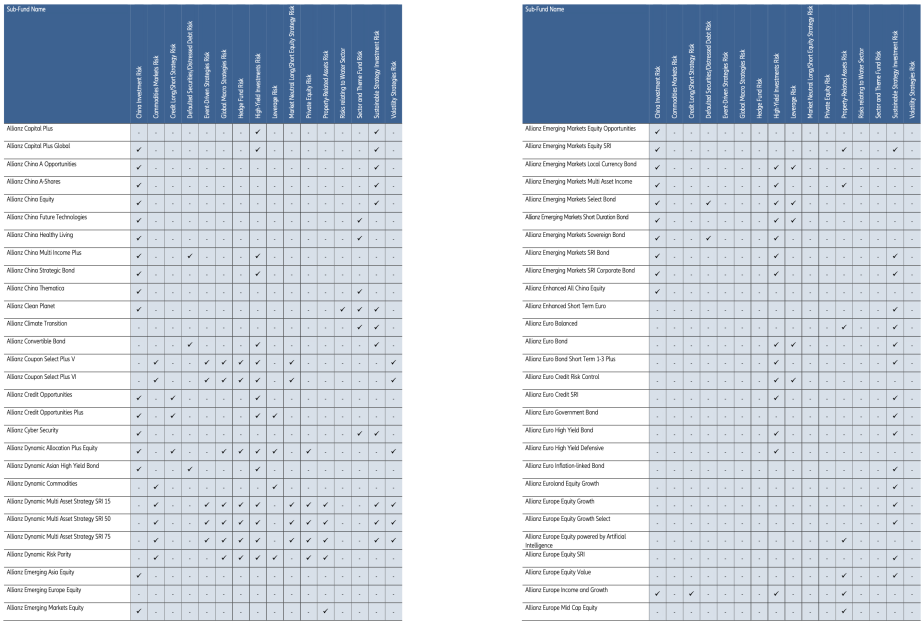
Investor profile Investors who understand the risks of the fund and plan to invest for at least 5 years.

The fund may appeal to investors who:
• are looking for investment growth
• want to invest in a fund with environmental and/or social characteristics, and that considers principal adverse impacts on sustainability factors, and with a minimum percentage of sustainable investment
• are interested in exposure to emerging equity markets

The fund intends to qualify as an "equity fund" in accordance with the German Investment Tax Act (please refer to chapter **"Investing in the Funds"** for further information) as it continuously invests more than 50% of total assets in equities ("Kapitalbeteiligungen") as defined within the German Investment Tax Act.

Figura 4.1: Exemplo de Prospecto Tipo 1 - Nordea Asset Management

Já o prospecto de tipo 2 não organiza os dados por fundo, mas em capítulos gerais, e conta com tabelas de dados específicos, onde cada linha corresponde a um UCITS. Nesses casos, não existe uma diferenciação de segmentos ao longo do documento, de modo que o tratamento será feito de maneira diferenciada, a partir de um método de extração de tabelas.



Sub-Fund Name	China Investment Risk	Commodities Market Risk	Credit Long/Short Strategy Risk	Defaulted Securities/Distressed Debt Risk	Event Driven Strategies Risk	Global Macro Strategies Risk	Hedge Fund Risk	High Yield Investments Risk	Leverage Risk	Market Neutral Long/Short Equity Strategy Risk	Private Equity Risk	Property-Related Assets Risk	Risks relating to Water Sector	Sector and Theme Fund Risk	Sustainable Strategy Investment Risk	Volatility Strategies Risk
Allianz Capital Plus	-	-	-	-	-	-	-	✓	-	-	-	-	-	-	✓	-
Allianz Capital Plus Global	✓	-	-	-	-	-	-	✓	-	-	-	-	-	-	✓	-
Allianz China A Opportunities	✓	-	-	-	-	-	-	-	-	-	-	-	-	-	✓	-
Allianz China A-Shares	✓	-	-	-	-	-	-	-	-	-	-	-	-	-	✓	-
Allianz China Equity	✓	-	-	-	-	-	-	-	-	-	-	-	-	-	✓	-
Allianz China Future Technologies	✓	-	-	-	-	-	-	-	-	-	-	-	-	✓	-	-
Allianz China Healthy Living	✓	-	-	-	-	-	-	-	-	-	-	-	-	✓	-	-
Allianz China Multi Income Plus	✓	-	-	✓	-	-	-	✓	-	-	-	-	-	-	-	-
Allianz China Strategic Bond	✓	-	-	-	-	-	-	✓	-	-	-	-	-	-	-	-
Allianz China Thematica	✓	-	-	-	-	-	-	-	-	-	-	-	-	✓	-	-
Allianz Clean Planet	✓	-	-	-	-	-	-	-	-	-	-	-	✓	✓	✓	-
Allianz Climate Transition	-	-	-	-	-	-	-	-	-	-	-	-	-	✓	✓	-
Allianz Convertible Bond	-	-	-	✓	-	-	-	✓	-	-	-	-	-	-	✓	-
Allianz Coupon Select Plus V	-	✓	-	-	✓	✓	✓	✓	-	✓	-	-	-	-	-	✓
Allianz Coupon Select Plus VI	-	✓	-	-	✓	✓	✓	✓	-	✓	-	-	-	-	-	✓
Allianz Credit Opportunities	✓	-	✓	-	-	-	-	✓	-	-	-	-	-	-	-	-
Allianz Credit Opportunities Plus	✓	-	✓	-	-	-	-	✓	✓	-	-	-	-	-	-	-
Allianz Cyber Security	✓	-	-	-	-	-	-	-	-	-	-	-	-	✓	✓	-
Allianz Dynamic Allocation Plus Equity	✓	-	✓	-	-	✓	✓	✓	✓	-	✓	-	-	-	-	✓
Allianz Dynamic Asian High Yield Bond	✓	-	-	✓	-	-	-	✓	-	-	-	-	-	-	-	-
Allianz Dynamic Commodities	-	✓	-	-	-	-	-	-	✓	-	-	-	-	-	-	-
Allianz Dynamic Multi Asset Strategy SRI 15	-	✓	-	-	✓	✓	✓	✓	-	✓	✓	✓	-	-	✓	✓
Allianz Dynamic Multi Asset Strategy SRI 50	-	✓	-	-	✓	✓	✓	✓	-	✓	✓	✓	-	-	✓	✓
Allianz Dynamic Multi Asset Strategy SRI 75	-	✓	-	-	✓	✓	✓	✓	-	✓	✓	✓	-	-	✓	✓
Allianz Dynamic Risk Parity	-	✓	-	-	-	✓	✓	✓	✓	-	✓	✓	-	-	-	-
Allianz Emerging Asia Equity	✓	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Allianz Emerging Europe Equity	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Allianz Emerging Markets Equity	✓	-	-	-	-	-	-	-	-	-	✓	-	-	-	-	-

Sub-Fund Name	China Investment Risk	Commodities Market Risk	Credit Long/Short Strategy Risk	Defaulted Securities/Distressed Debt Risk	Event Driven Strategies Risk	Global Macro Strategies Risk	Hedge Fund Risk	High Yield Investments Risk	Leverage Risk	Market Neutral Long/Short Equity Strategy Risk	Private Equity Risk	Property-Related Assets Risk	Risks relating to Water Sector	Sector and Theme Fund Risk	Sustainable Strategy Investment Risk	Volatility Strategies Risk
Allianz Emerging Markets Equity Opportunities	✓	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Allianz Emerging Markets Equity SRI	✓	-	-	-	-	-	-	-	-	-	-	✓	-	-	✓	-
Allianz Emerging Markets Local Currency Bond	✓	-	-	-	-	-	-	✓	✓	-	-	-	-	-	-	-
Allianz Emerging Markets Multi Asset Income	✓	-	-	-	-	-	-	✓	-	-	-	✓	-	-	-	-
Allianz Emerging Markets Select Bond	✓	-	-	✓	-	-	-	✓	✓	-	-	-	-	-	-	-
Allianz Emerging Markets Short Duration Bond	✓	-	-	-	-	-	-	✓	✓	-	-	-	-	-	-	-
Allianz Emerging Markets Sovereign Bond	✓	-	-	✓	-	-	-	✓	-	-	-	-	-	-	-	-
Allianz Emerging Markets SRI Bond	✓	-	-	-	-	-	-	✓	-	-	-	-	-	-	✓	-
Allianz Emerging Markets SRI Corporate Bond	✓	-	-	-	-	-	-	✓	-	-	-	-	-	-	✓	-
Allianz Enhanced All China Equity	✓	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Allianz Enhanced Short Term Euro	-	-	-	-	-	-	-	-	-	-	-	-	-	-	✓	-
Allianz Euro Balanced	-	-	-	-	-	-	-	-	-	-	-	✓	-	-	✓	-
Allianz Euro Bond	-	-	-	-	-	-	-	✓	✓	-	-	-	-	-	✓	-
Allianz Euro Bond Short Term 1-3 Plus	-	-	-	-	-	-	-	✓	-	-	-	-	-	-	✓	-
Allianz Euro Credit Risk Control	-	-	-	-	-	-	-	✓	✓	-	-	-	-	-	-	-
Allianz Euro Credit SRI	-	-	-	-	-	-	-	✓	-	-	-	-	-	-	✓	-
Allianz Euro Government Bond	-	-	-	-	-	-	-	-	-	-	-	-	-	-	✓	-
Allianz Euro High Yield Bond	-	-	-	-	-	-	-	✓	-	-	-	-	-	-	✓	-
Allianz Euro High Yield Defensive	-	-	-	-	-	-	-	✓	-	-	-	-	-	-	-	-
Allianz Euro Inflation-linked Bond	-	-	-	-	-	-	-	-	-	-	-	-	-	-	✓	-
Allianz Euroland Equity Growth	-	-	-	-	-	-	-	-	-	-	-	-	-	-	✓	-
Allianz Europe Equity Growth	-	-	-	-	-	-	-	-	-	-	-	-	-	-	✓	-
Allianz Europe Equity Growth Select	-	-	-	-	-	-	-	-	-	-	-	-	-	-	✓	-
Allianz Europe Equity powered by Artificial Intelligence	-	-	-	-	-	-	-	-	-	-	-	✓	-	-	-	-
Allianz Europe Equity SRI	-	-	-	-	-	-	-	-	-	-	-	-	-	-	✓	-
Allianz Europe Equity Value	-	-	-	-	-	-	-	-	-	-	-	✓	-	-	✓	-
Allianz Europe Income and Growth	✓	-	✓	-	-	-	-	✓	-	-	-	✓	-	-	-	-
Allianz Europe Mid Cap Equity	-	-	-	-	-	-	-	-	-	-	-	✓	-	-	-	-

Figura 4.2: Exemplo de Prospecto Tipo 2 - Allianz Global Investors Fund

4.3
Processamento de Linguagem Natural

O Processamento de Linguagem Natural (*Natural Language Processing*) é um subcampo interdisciplinar da ciência da computação e inteligência artificial, focado em capacitar computadores a processar dados em linguagem natural, relacionando-se estreitamente com recuperação de informações, representação do conhecimento e linguística computacional. Os dados são coletados em corpora de texto e processados utilizando abordagens baseadas em regras, estatísticas ou redes neurais de aprendizado de máquina e aprendizado profundo (Deep Learning). As principais tarefas de NLP incluem reconhecimento de fala, classificação de texto, compreensão e geração de linguagem natural.

O campo de pesquisa começou na década de 1940, após a Segunda Guerra Mundial, com o objetivo de criar máquinas capazes de tradução automática e criptografia. Hoje, a ascensão de redes neurais e modelos de Deep Learning mais complexos, como a tecnologia Transformer, abriu espaço para uma compreensão textual automática mais completa, assim como para a geração de texto, como o ChatGPT (VASWANI et al., 2017).

Para o processamento de um corpus textual via NLP, é necessário transformá-lo em uma grandeza numérica, como vetores. Isso é feito por meio de uma série de etapas, algumas das quais serão utilizadas neste trabalho.

4.4 Tokenização

A tokenização é um processo fundamental no tratamento de corpora textuais no âmbito do NLP. Consiste na segmentação de um texto contínuo em unidades menores, conhecidas como tokens, que podem ser palavras, frases ou até caracteres individuais. Este processo é essencial para a análise subsequente, pois permite que algoritmos manipulem e compreendam o texto de maneira estruturada. Durante a tokenização, são aplicadas regras específicas para lidar com pontuação, contrações e outros aspectos linguísticos que variam de idioma para idioma. A precisão na tokenização é crucial, pois influencia diretamente a eficácia de tarefas como a classificação de texto, análise de sentimentos e extração de informações, facilitando uma representação adequada do texto para algoritmos de aprendizado de máquina e modelos de linguagem avançados (AHO et al., 2006).

4.5 Lematização

A lematização é uma técnica de pré-processamento em NLP que consiste na transformação de palavras em sua forma base ou lema, levando em consideração o contexto e a parte do discurso da palavra (substantivo, verbo, adjetivo, etc.). Diferente da stemização, técnica mais primitiva que simplesmente remove sufixos para reduzir a palavra à sua raiz, a lematização proporciona maior precisão ao garantir que a forma base resultante seja uma palavra válida no idioma (GREEN et al., 2009).

Esse processo traz inúmeros benefícios para o processamento de texto, uma vez de acarreta a redução da dimensionalidade: várias palavras distintas, como por exemplo diferentes tempos verbais de uma ação, são reduzidas a

uma (o verbo no infinitivo), de modo a reduzir significativamente o número de tokens no corpus. Isso, no entanto, aumenta a necessidade de processamento computacional, e a análise do contexto e parte do discurso requerem um modelo de lematização dependente do idioma do texto.

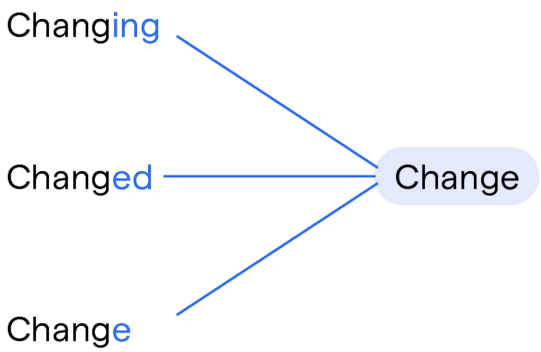


Figura 4.3: Processo de Lematização

4.6 N-Gramas

Um n-grama é uma sequência de n símbolos adjacentes em uma ordem específica, que podem ser letras, sílabas, palavras, fonemas ou pares de bases genômicas, coletados de um corpus de texto ou fala (BRODER et al., 1997). Eles permitem capturar contextos locais de palavras, proporcionando uma compreensão mais detalhada das relações entre termos, essencial para a análise de frequência de termos e detecção de padrões linguísticos.

N-gramas são fundamentais na construção de modelos de linguagem, utilizados para prever palavras em sequências, melhorar sistemas de tradução automática e reconhecimento de fala (FRANZ; BRANTS, 2006).

Assim, em uma sequência de k tokens, teremos $k - (n - 1)$ n-gramas, onde n representa o número de tokens aglutinados em cada n-grama.

$$x_1x_2x_3...x_{k-1}x_k \rightarrow \underbrace{\{(x_1x_2...x_n), (x_2x_3...x_{n+1}), ..., (x_{k-(n-1)}x_{k-(n-2)}...x_k)\}}_{\text{k-(n-1) elements}}$$

4.7
Clusterização K-means

O k-means clustering é uma técnica de aprendizado não supervisionado amplamente utilizada para particionar um conjunto de dados em k clusters distintos (LLOYD, 1982). No escopo de NLP e finanças, tal método é bem popular em estudos de sumarização e análise de sentimento (SUN et al., 2014).

O objetivo é minimizar a soma das distâncias quadráticas entre os pontos de dados e os centroides dos clusters a que pertencem. Inicialmente, selecionam-se aleatoriamente k centroides, onde cada centróide representa o centro de um cluster. Em seguida, cada ponto de dado é atribuído ao centróide mais próximo, utilizando a métrica de distância euclidiana. Formalmente, a distância euclidiana entre um ponto de dado x_i e um centróide μ_j é dada por:

$$(\mathbf{x}_i, \mu_j) = \sqrt{\sum_{l=1}^{n} (x_{il} - \mu_{jl})^2}$$

onde n é o número de características dos dados. Após a atribuição dos pontos aos centroides, recalculam-se os centroides como a média dos pontos de dados pertencentes a cada cluster:

$$\mu_j = \frac{1}{|C_j|} \sum_{\mathbf{x}_i \in C_j} \mathbf{x}_i$$

onde C_j é o conjunto de pontos de dados atribuídos ao cluster j e $|C_j|$ o número de pontos nesse cluster.

Este processo de atribuição de pontos e recalculação dos centroides é repetido iterativamente até que os centroides se estabilizem ou a mudança na soma das distâncias quadráticas entre iterações seja inferior a um valor de tolerância predefinido. A função objetivo a ser minimizada no k-means clustering é a soma das distâncias quadráticas dentro dos clusters, expressa por:

$$J = \sum_{j=1}^{k} \sum_{x_i \in C_J} \|\mathbf{x}_i - \mu_j\|^2$$

onde J representa a soma total das distâncias quadráticas entre os pontos de dados e os respectivos centroides dos clusters.

4.8 Clusterização Espectral

A clusterização espectral é uma técnica de Machine Learning que utiliza a informação espectral (autovalores) de uma matriz Laplaciana construída a partir dos dados de entrada para realizar um processo de redução de dimensionalidade antes de separar os dados em clusteres (LUXBURG, 2007).

A priori, dado um conjunto de m dados de entrada $(a_1, a_2, a_3, ..., a_m)$, constroi-se uma matriz ou grafo de similaridade $A_{m\times m}$, onde $A_{ij} \geq 0$ representa a medida de afinidade entre os dados a_i e a_j, que pode ser calculada de diversas formas. Um exemplo usual é o RBF (Radial Basis Function), ou função gaussiana de base radial (BUHMANN, 2000), dada por:

$$A_{ij} = \exp\left(-\frac{\|a_i - a_j\|^2}{2\sigma^2}\right), \quad \sigma \rightarrow \text{parâmetro livre}$$

Em seguida , calcula-se a matriz Laplaciana L do grafo como:

$$L = D - A$$

onde $D_{m\times m}$ é a matriz diagonal das valências, e $D_{ii} = \sum_j A_{ij}$.

Para o cálculo dos autovalores da matriz Laplaciana, utiliza-se algum método otimizado de resolução de equações matriciais, como o Multigrid Algébrico (AMG) (FALGOUT, 2006). Essa etapa visa encontrar os k ($k \ll m$) primeiros autovetores v (associados aos k menores autovalores não-nulos) que satisfazem

$$L\mathbf{v} = \lambda\mathbf{v}$$

Logo, obtidos os autovetores da matriz Laplaciana, os dados são mapeados em um espaço reduzido, o que tende a exibir agrupamentos mais facilmente distinguíveis (BELKIN; NIYOGI, 2003).

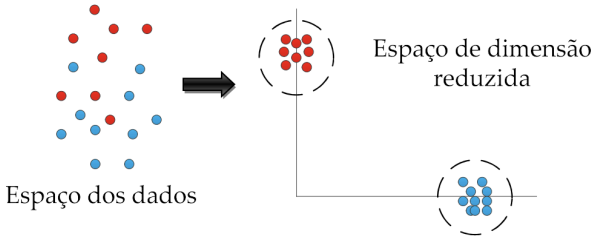


Figura 4.4: Redução de Dimensionalidade

Por fim, é feita a clusterização dos dados no espaço de dimensão reduzida. Esta pode ser por meio do K-means ou quaisquer outros modelos com mesmo objetivo. Duas alternativas, descritas abaixo, são a discretização e a decomposição QR.

4.8.1 Discretização

A discretização é um processo de segmentação de agrupamentos de dados por meio da conversão de valores contínuos em faixas discretas (YU; SHI, 2004). Alguns modelos de Machine Learning – supervisionados ou não –, como Árvore de Decisão (WINTERFELDT; EDWARDS, 1986) e Naive Bayes (BERRAR, 2001), requerem ou apresentam melhor funcionamento com atributos categóricos (WITTEN; FRANK; HALL, 2016). Ademais, variáveis de entrada numéricas podem possuir distribuição altamente assimétrica e dispersa, dificultando modelos tradicionais de clusterização.

4.8.2 Decomposição QR

A decomposição QR, também chamada de fatoração QR, é a decomposição de uma matriz $A = QR$, onde Q é uma matriz ortogonal e R uma matriz triangular superior (VELDE, 2005). Esse método é bastante comum na reso-

lução do problema de mínimos quadrados linear, e pode ser calculada a partir do processo de ortonormalização de Gram-Schmidt (GRAM, 1883).

Dados **v**, **u** vetores em $\mathbb{R}^n$, define-se o operador projeção por:

$$\text{proj}_u(v) = \frac{\langle v, u \rangle}{\langle u, u \rangle} u, \quad \langle v, u \rangle = \sum_{i=1}^{n} v_i \cdot u_i$$

Calcula-se a base otronormal de vetores e_k que satisfazem

$$e_k = \frac{u_k}{\|u_k\|}, \quad u_k = v_k - \sum_{j=1}^{k-1} \text{proj}_{u_j} v_k$$

Logo, tem-se a igualdade $A = QR$, sendo

$$Q = \begin{pmatrix} e_{11} & e_{21} & \dots \\ \vdots & \ddots & \\ e_{1k} & & e_{kk} \end{pmatrix}, \qquad R = \begin{pmatrix} \langle e_1, v_1 \rangle & \langle e_1, v_2 \rangle & \langle e_1, v_3 \rangle & \dots \\ 0 & \langle e_2, v_2 \rangle & \langle e_2, v_3 \rangle & \dots \\ 0 & 0 & \langle e_3, v_3 \rangle & \dots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

Essa técnica provê uma maneira simples e direta de atribuir clusteres ao conjunto de dados, visto que extrai as partições diretamente dos autovetores da matriz Laplaciana L. Embora não seja iterativo, pode muitas vezes superar as alternativas tanto em qualidade quanto em rapidez de processamento (DAMLE; MINDEN; YING, 2018).

4.9 Métricas de Avaliação

4.9.1 Matriz de Confusão

A matriz de confusão apresenta o resultado de um modelo de classificação binária, dividindo as amostras em verdadeiros positivos (VP), falsos positivos (FP), verdadeiros negativos (VN) e falsos negativos (FN):

		Previsto	
		Sim	Não
Real	Sim	VP	FN
	Não	FP	VN

Tabela 4.1: Matriz de Confusão

4.9.2 Acurácia

A Acurácia é uma métrica que avalia a proporção de previsões corretas feitas pelo modelo em relação ao total de previsões.

$$\text{Acurácia} = \frac{VP + VN}{VP + VN + FP + FN}$$

4.9.3 Precisão

A Precisão, ou valor preditivo positivo, mede a proporção de verdadeiros positivos entre todas as previsões positivas feitas pelo modelo. É particularmente útil quando o custo de um falso positivo é alto, pois indica a confiabilidade do modelo ao prever uma classe positiva.

$$\text{Precisão} = \frac{VP}{VP + FP}$$

4.9.4 Recall

A métrica de Recall, também conhecida como sensibilidade ou taxa de detecção, avalia a proporção de verdadeiros positivos identificados pelo modelo em relação ao total de casos positivos reais. Essa métrica é crucial quando a prioridade é minimizar falsos negativos, como é o caso, garantindo que a maioria dos casos positivos seja capturada pelo modelo.

$$\text{Recall} = \frac{VP}{VP + FN}$$

4.9.5 Especificidade

A Especificidade, ou taxa de verdadeiros negativos, mede a proporção de verdadeiros negativos entre todos os casos negativos reais.

$$\text{Especificidade} = \frac{VN}{VN + FP}$$

5
Projeto e Especificação do Sistema

Neste capítulo, serão discutidos as bibliotecas e funções necessárias para a criação do modelo de processamento, segmentação e extração dos prospectos. Será apresentado também uma descrição detalhada do dataset utilizado

5.1
Bibliotecas Utilizadas

Para a realização do projeto, foram empregadas diversas bibliotecas especializadas que, em conjunto, possibilitaram a execução eficiente das tarefas de pré-processamento de texto, geração de n-gramas e clusterização.

Inicialmente, a instalação e atualização das bibliotecas necessárias foi garantida pelo uso do `subprocess`, que automatizou o processo de instalação dos pacotes listados no arquivo `requirements.txt`. As bibliotecas do `nltk` (Natural Language Toolkit) foram utilizadas para o pré-processamento de texto, fornecendo ferramentas essenciais como a lematização de palavras e a remoção de stopwords. Especificamente, foram utilizados componentes como `stopwords`, `WordNetLemmatizer`, `word_tokenize` e `pos_tag`.

Para a manipulação dos documentos em PDF, a biblioteca `PyMuPDF` foi utilizada, permitindo a extração eficiente de textos e de tabelas. A biblioteca `pandas` foi usada para a manipulação de dados em formato tabular, permitindo a análise de n-gramas e a criação de matrizes de frequência. A visualização dos dados foi feita pela biblioteca `matplotlib`, e para a etapa de clustering, utilizou-se a biblioteca `scikit-learn`, particularmente as classes `KMeans` e `SpectralClustering`, as quais implementam os algoritmos de clusterização, permitindo a classificação e segmentação dos dados. A utilização do `ProcessPoolExecutor` da biblioteca `concurrent` também se mostrou essencial para a execução paralela das tarefas, garantindo a eficiência e rapidez do processamento.

Código 1: Bibliotecas utilizadas

```
import subprocess

# Install Requeirements
command = "pip3 install --upgrade pip & pip3 install -r ../
    requirements.txt"
try:
    subprocess.run(command, shell=True, check=True)
    print("Packages installed successfully!")
except subprocess.CalledProcessError as e:
    print(f"Error installing packages: {e}")

# Preprocessing
from time import time
import nltk
from string import punctuation
try :
    from nltk.corpus import stopwords
    from nltk.stem import WordNetLemmatizer
    from nltk import word_tokenize,pos_tag
except:
    print('Downloading nltk data')
    nltk.download()
    from nltk.corpus import stopwords
    from nltk.stem import WordNetLemmatizer
    from nltk import word_tokenize,pos_tag

# N Gram
import pymupdf
import pandas as pd
from concurrent.futures import ProcessPoolExecutor, as_completed

# Clustering
import numpy as np
import matplotlib.pyplot as plt
from sklearn.cluster import KMeans, SpectralClustering
```

5.2
Base de Dados

Os prospectos configuram-se como informação pública, existente nos sites das companhias gestoras que administram os fundos. Essas empresas geralmente possuem um único prospecto, compreendendo todos os seus fundos, porém há casos em que se divide o prospecto em alguns documentos, um por família (grupo) de fundos.

Foi realizada a coleta de mais de 5000 documentos. Após uma limpeza, excluiu-se cerca de 2000, entre eles arquivos corrompidos, documentos que não eram prospectos e alguns em outros idiomas que não o inglês. Assim, a base final contou com mais de 3000 prospectos de gestoras de países europeus, sobretudo Reino Unido, Luxemburgo, França e Alemanha.

Vale frisar, no entanto, que devido à carência de estudos de ciência de dados e Machine Learning voltados especificamente para a leitura e processamento de prospectos de UCITS, não havia documentos segmentados ou etiquetados, os quais serviriam de grupo teste para avaliação do modelo. Logo, foi feito um trabalho de etiquetamento (labeling) manual em 100 documentos, escolhidos aleatoriamente entre os arquivos coletados. Ao final, classificou-se mais de 20 mil páginas, as quais serão fundamentais para a análise da assertividade do modelo de clusterização e segmentação no capítulo seguinte.

6
Implementação e Avaliação

O capítulo atual visa descrever a implementação do código (Python) para o processamento discorrido no capítulo anterior e compilar os resultados do processo de pré-processamento, análise de n-grama, segmentação dos prospectos e extração de tabelas.

6.1
Pré-Processamento

O algoritmo a seguir descreve as etapas referentes ao pré-processamento do texto de um prospecto, desde a conversão para letras minúsculas até a lematização do texto já tokenizado. A entrada é uma cadeia de caracteres, ao passo que a saída é uma lista de tokens (palavras) lematizadas.

Algoritmo 1: Pré-Processamento

Entrada: texto
Saída: lista_tokens

```
texto_min ← converte_para_lowercase(texto)
texto_sem_url ← remove_url(texto_min)
texto_sem_pont ← remove_pontuacao(texto_sem_url)
texto_processado ← remove_stopwords(texto_sem_pont)
lista_tokens ← tokenize(texto_processado)
lista_tokens ← lemmatize(lista_tokens)

Retorna lista_tokens
```

6.2
Análise de Frequência de N-Gramas

No intuito de identificar padrões recorrentes no texto, permitindo uma compreensão mais profunda das temáticas e tendências presentes no documento, foi realizada a aglutinação dos tokens pré-processados em n-gramas.

A premissa por trás da escolha desse método, e o motivo pelo qual isso é útil para o processamento descrito nesse capítulo, é que dentro de um mesmo prospecto, cada seção individual de UCITS possui termos e informações

similares, como "objetivo de investimento", "taxa de administração", "política de sustentabilidade", entre outros. Logo, tais n-gramas possuirão uma alta frequência ao longo do texto.

Por outro lado, seções contendo informações gerais da companhia gestora possuirão termos e expressões que se repetirão menos no documento. Espera-se, portanto, que páginas e segmentos com dados individuais de fundos contenham um maior número de n-gramas com frequência alta do que partes de informação geral da gestora.

Após a criação dos n-gramas, foi construída uma matriz indicadora de frequência por página do texto. Para melhorar a qualidade da análise, foi aplicado um limiar de valor 15, removendo n-gramas pouco frequentes. A matriz também foi convertida para binária, passando a indicar a presença, ou não, de cada n-grama em cada página.

$$\mathbf{M_{m\times n}} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}$$

$$a_{ij} = \begin{cases} 1 & \text{se o n-grama } j \text{ existe na página} \\ 0 & \text{caso contrário} \end{cases}$$

$m =$ Número de páginas, $\quad n =$ Número de n-gramas com frequência > 15

O algoritmo a seguir ilustra o processo iterativo de criação da matriz de n-gramas, recebendo como entrada o prospecto, o número de páginas que compõem o segmento e o número de tokens do n-grama.

Algoritmo 2: Criação de N-gramas

Entrada: prospectos

```
para k = 1 até len(documentoProspecto)-numPaginas faça
    segmento ← páginas k até k+numPaginas-1
    texto ← extraiTexto(segmento)
    listaTokens ← preProcessa(texto)
    nGramas ← criaListaVazia()

    para n = 1 até len(listaTokens)-numNgramas faça
        incrementaLista(nGramas, listaTokens[n, n+nGramas-1])

matriz ← criaMatrizBinaria(
    dados = nGramas,
    eixoX = Ngrama,
    eixoY = Pagina
)

Retorna matriz
```

6.3 Segmentação do Prospecto

No intuito de facilitar a segmentação, converteu-se a matriz de n-gramas em um vetor numérico que será a entrada dos algoritmos utilizados. Considerando uma matriz binária $M_{m \times n}$, onde m é o número de páginas do prospecto e n é o número de n-gramas com frequência acima de 15, o vetor v é dado por:

$$\mathbf{v} = \left[\begin{array}{ccccc} v_1, & v_2, & v_3, & \ldots, & v_m \end{array} \right]$$

$$v_i = (-k + \sum_{p=1}^{n} a_{p,i})^c$$

$a_{p,i} \in \{0, 1\} \rightarrow$ presença do n-grama p na página i

$k, c =$ constantes

6.4 Clusterização

A segmentação foi realizada a partir dos modelos de clusterização K-means e Espectral, sendo este último testado com diferentes hiperparâmetros, cujos resultados serão expostos nas seções seguintes.

A implementação de ambos os modelos foi feita a partir da biblioteca de Machine Learning em Python `scikit-learn`. Para o K-means, foi utilizada a classe `KMeans`, com número de clusters igual a 2. Já para o espectral, por meio classe `SpectralClustering`, o número de clusters definido foi 2, a função de afinidade foi a RBF e o método de autovetores foi o Multigrid Algébrico (AMG).

O código abaixo descreve o funcionamento do modelo. Vale mencionar o uso do hiperparâmetro `random_state`, para garantir a replicabilidade do modelo, e testes adicionais no `SpectralClustering`, uma vez que o hiperparâmetro `assign_labels` aceita os modelos "kmeans", "discretize"(discretização) e "cluster_qr"(decomposição QR).

Como é possível observar no trecho de código abaixo, a função retorna o vetor `vector` etiquetado, ou seja, cada página possui agora uma classificação em "Geral"ou "Específico".

Código 2: Clusterização

```
import os
from pandas import DataFrame
import pandas as pd
import numpy as np
from sklearn.cluster import KMeans, SpectralClustering

def cluster(
        vector: DataFrame,
        filename: str,
        k: float = 0,
        c: float = 1.0,
        model_type: str = None,
        #['kmeans', 'discretize', 'cluster_qr']
):

    vector['transformed'] = np.maximum(0, vector['data']-k)**c

    if model_type is None:
        model = KMeans(
            n_clusters=2,
            random_state=0
        )
    else:
        model = SpectralClustering(
            n_clusters=2,
            random_state=0,
            eigen_solver='amg',
            assign_labels=model_type,
        )

    vector['cluster'] = model.fit_predict(vector[['transformed']])

    return vector
```

6.5 Extração de Tabelas

Por fim, resta extrair as tabelas dos segmentos etiquetados pelo modelo de clusterização como "Geral". Como previamente explicado, muitos prospectos contém toda ou parte da informação específica de cada UCITS em tabelas ao invés de capítulos do documento. Assim, faz-se necessário extrair as tabelas restantes.

Para isso, utilizou-se o pacote `PyMuPDF`, que possui métodos de reconhecimento e extração de tabelas. O algoritmo interno captura formas tabulares retangulares no prospecto, obtendo informações de borda, conteúdo, organização e localização dentro da página. Dada a grande variabilidade de formatação das tabelas, a extração em formato tabular (biblioteca `pandas`) não é adequada, e portanto decidiu-se por utilizar o atributo `bbox` das tabelas extraídas, o qual indica as coordenadas (x, y) da localização da tabela, para salvá-la como imagem.

6.6 Algoritmo Final

Ao final do processo descrito, tem-se o algoritmo abaixo, que perpassa desde o pré-processamento, separação de n-gramas, segmentação de texto e extração de tabelas.

Algoritmo 3: Processamento dos dados

Entrada: prospectos

```
1  para todo prospecto faça
2      listaTokens ← preProcessa(prospecto)
3      matriz ← criaMatrizNgramas(prospecto)
4      vetor ← somaColunas(matriz)
5      vetorComLabels ← cluster( vetor)
6
7      para todo segmento in vetorComLabels faça
8          se segmento = "Geral" então
9              tabelas ← extrailmagemTabelas(segmento)
10 Retorna vetorComLabels, tabelas
```

6.7 Resultados

Após o processamento e segmentação dos prospectos, avaliou-se o produto final de acordo com as métricas apresentadas no capítulo 4.

- **VP:** páginas corretamente classificadas como tipo "Específico".
- **VN:** páginas corretamente classificadas como "Geral".
- **FP:** páginas incorretamente classificadas como "Específico".
- **FN:** páginas incorretamente classificadas como "Geral".

Conforme previamente mencionado, foram realizados testes nas etapas de análise de n-grama e segmentação via clusterização.

6.7.1 Matriz de N-gramas

Em relação à codificação do texto em n-gramas, das 9 combinações de paginação e tokenização, a divisão em 3 páginas e trigramas mostrou-se mais promissora em capturar os padrões dentro de cada documento. Vale mencionar que tal análise é evidente nos prospectos de tipo 1 (ver item 4.2), nos quais cada fundo de investimento possui uma seção específica.

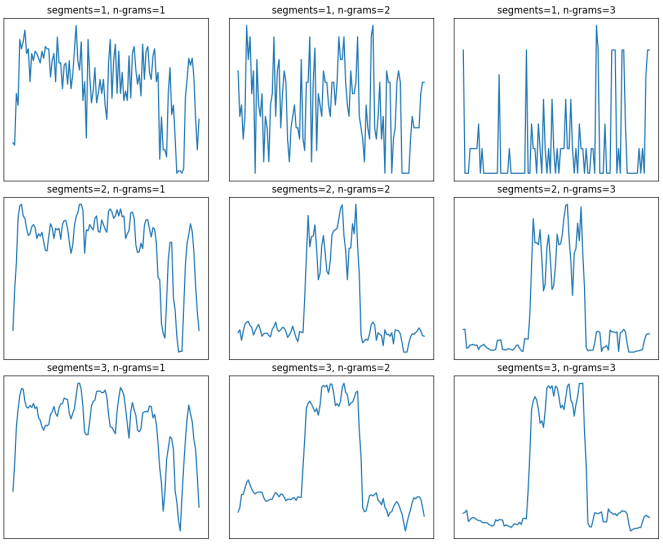


Figura 6.1: Comparação - Análise de N-gramas - Gestora LGIM

Como explicitado na figura acima, uma separação por página individual e aglutinação por unigrama impede a adequada identificação das diferentes seções internas do documento, uma vez que não captura o perfil de frequência de n-gramas ao longo do texto. Ao utilizar valores de 2 ou 3, tanto para paginação quanto para os n-gramas, é possível observar a formação de padrões mais bem definidos. Dessa maneira, a escolha final pode ser feita observando-se a tabela de métricas abaixo, que compila os resultados deste teste. Os valores das constantes k e c utilizadas foram de 50 e 0.5, respectivamente, e o modelo de clusterização foi o K-means.

Modelo	Acurácia	Precisão	Recall	Especificidade
segments=1, ngram=1	0.59	0.59	0.68	0.49
segments=1, ngram=2	0.80	0.94	0.66	0.96
segments=1, ngram=3	0.79	0.95	0.63	0.97
segments=2, ngram=1	0.47	0.49	0.66	0.26
segments=2, ngram=2	0.85	0.85	0.85	0.84
segments=2, ngram=3	0.89	0.94	0.85	0.94
segments=3, ngram=1	0.48	0.50	0.68	0.27
segments=3, ngram=2	0.78	0.73	0.89	0.66
segments=3, ngram=3	0.93	0.91	0.95	0.90

Tabela 6.1: Resultados - N-Gramas

O teste com segmentos de 3 páginas divididos em trigramas apresentou o melhor resultado, conforme a tabela acima. Vale frisar que, embora outros modelos tenham obtido melhores resultados em precisão ou especificidade, isso ocorreu às custas de um aumento significativo de falsos negativos, categoria que mais se quer evitar: ela significa que trechos com informação específica de UCITS foram classificados como "Geral", impedindo a correta segmentação e subsequente extração de dados.

Nessa ótica, a métrica mais relevante para observar a minimização de falsos negativos é o Recall: à medida que FN tende a zero, o Recall tende a 1 ($FN \ll VP \iff VP + FN \approx VP$). De fato, o modelo escolhido possui um valor dessa métrica bastante superior aos demais.

6.7.2 Clusterização

No tocante à segmentação do documento, testou-se primeiro as constantes de parametrização do vetor de n-gramas. A constante k terá valores de 0 ou 50, e servirá como um limiar para remover ruídos do vetor. Já a constante c, de valores 1 ou 0.5, é responsável por intensificar (ou não) a separação dos dados, visando facilitar a identificação dos clusters.

Ademais, avaliou-se a eficiência de 4 versões de modelos de clusterização: K-means simples, espectral com k-means, espectral com discretização e espectral com decomposição QR. Ao todo, portanto, foram testados 16 versões do modelo, combinando as técnicas de clusterização com as constantes k, c. Vale frisar que a separação dos n-gramas foi realizada com os valores provenientes do teste anterior: segmentos de 3 páginas e trigramas.

Modelo	Acurácia	Precisão	Recall	Especificidade
KMeans, k=0, c=1	0.82	0.83	0.64	0.81
KMeans, k=50, c=1	0.88	0.93	0.84	0.93
KMeans, k=0, c=0.5	0.83	0.80	0.93	0.73
KMeans, k=50, c=0.5	0.93	0.91	0.95	0.90
Spectral(KMeans), k=0, c=1	0.70	0.70	0.70	0.69
Spectral(KMeans), k=50, c=1	0.70	0.70	0.70	0.70
Spectral(KMeans), k=0, c=0.5	0.84	0.78	0.96	0.72
Spectral(KMeans), k=50, c=0.5	0.87	0.82	0.96	0.79
Spectral(Discretize), k=0, c=1	0.70	0.70	0.70	0.70
Spectral(Discretize), k=50, c=1	0.74	0.75	0.71	0.77
Spectral(Discretize), k=0, c=0.5	0.85	0.79	0.96	0.73
Spectral(Discretize), k=50, c=0.5	0.88	0.83	0.96	0.80
Spectral(ClusterQR), k=0, c=1	0.66	0.69	0.59	0.73
Spectral(ClusterQR), k=50, c=1	0.65	0.66	0.65	0.66
Spectral(ClusterQR), k=0, c=0.5	0.84	0.78	0.96	0.73
Spectral(ClusterQR), k=50, c=0.5	0.83	0.76	0.96	0.70

Tabela 6.2: Resultados - Clusterização

Observa-se que os resultados com $k = 50$ e $c = 0.5$ são melhores que as demais versão, visto que apresentam valores superiores das métricas analisadas. No que tange ao modelo de clusterização, a etapa extra de redução de dimensionalidade feita pelo `SpectralClustering` não teve sucesso em melhorar o modelo final. Embora este método tenha apresentado um ligeiro incremento no valor do Recall, as demais medidas pioraram significativamente, entre 10% e 20% em relação ao modelo K-means. Nesses casos, apesar de o número de falsos negativos diminuir, houve um aumento desproporcional na quantidade de falsos positivos, significando que o modelo classifica quase tudo como "Específico".

O modelo escolhido, portanto, foi a combinação da análise de n-gramas de valor 3 (trigramas), com divisão em segmentos de 3 páginas, e emprego do modelo de clusterização K-means, além de constantes k, c iguais a 50 e 0.5, respectivamente. A matriz de confusão resultante, construída a partir do grupo de teste da base de dados, é:

		Previsto	
		Específico	Geral
Real	Específico	9624 (90,3%)	442 (4,7%)
	Geral	1039 (9,7%)	8890 (95,3%)

Tabela 6.3: Matriz de Confusão Final

Por fim, os gráficos abaixo ilustram a segmentação final dos prospectos, com os parâmetros otimizados conforme supracitado. O primeiro é um típico exemplo de prospecto do tipo 1: observou-se a classificação de boa parte da primeira metade do texto como "Específico"e o restante como "Geral". Já o segundo gráfico segmentou adequadamente um prospecto do tipo 2, onde não há seções específicas para cada UCITS. Nesse caso, é possível ver que o modelo de clusterização K-means só encontrou 1 único cluster, como esperado.

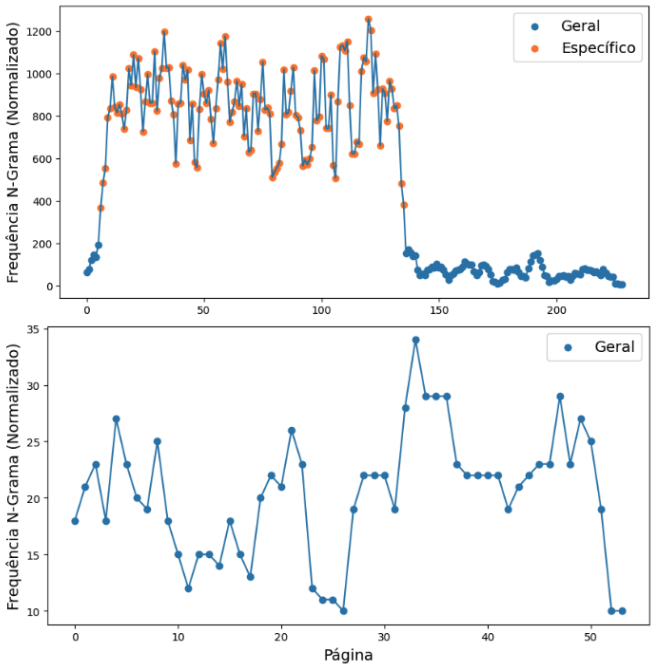


Figura 6.2: Segmentação de Prospectos do Tipo 1 e 2

7
Considerações Finais

Comforme se discorreu nos capítulos anteriores, o uso de técnicas de NLP e Machine Learning para a segmentação de documentos financeiros é um território ainda pouco explorado, mostrando-se economicamente interessante para todos os atores envolvidos no mercado financeiro.

A partir da análise minuciosa de um documento bastante específico – o prospecto de UCITS –, é possível:

1. Compreender mais profundamente a estrutura e informações contidas no texto.

2. Segmentar o documento, selecionando corretamente seções e páginas que contenham dados específicos de UCITS

3. Aumentar a assertividade e minimizar custos de extração de informações, uma vez que o modelo descarta boa parte do conteúdo "desnecessário" do prospecto (classificação "Geral")

Após a testagem de inúmeros cenários, concluiu-se que os modelos foram bem-sucedidos em capturar os padrões dentro de cada prospecto, embora haja pouca ou nenhuma padronização do mesmo. O melhor modelo foi a clusterização K-means com constantes $k = 50, c = 0.5$, segmentação por grupos de 3 páginas e em trigramas, com uma acurácia de 93% e recall de 96%. Fica evidente, portanto, o sucesso do modelo em classificar corretamente e, não somente isso, reduzir ao máximo o número de falsos negativos.

Para os próximos passos do projeto, seria interessante dar continuidade ao estudo, através do desenvolvimento de um algoritmo de extração de dados de UCITS. O modelo aqui descrito é bastante útil para segmentar e filtrar a informação dos prospectos, permitindo que terceiros utilizem somente seções que contenham informações específicas de fundos UCITS para alimentar,

por exemplo, uma aplicação de IA generativa, a qual será capaz de receber os segmentos de prospectos e extrair os dados dos UCITS. Dado que o processamento deste trabalho foi bem-sucedido na segmentação, menos de 50% do conteúdo dos prospectos analisados contém informação útil para o objetivo em questão (ou seja, classificado como "Específico"). Logo, a combinação do processamento via n-grama e clusterização com um modelo de extração de dados seria 2 vezes mais otimizada em termos de custo e tempo.

Em suma, este trabalho evidenciou a eficiência das técnicas de NLP e clusterização no processamento e segmentação de fundos europeus do tipo UCITS. A aplicação prática dessa tecnologia oferece uma imensa otimização de tempo e custos na coleta de dados financeiros de fontes não-estruturadas, fomentando a emergência de um ecossistema de informações mais amplo e transparente na esfera de finanças.

8
Referências bibliográficas

AHO, A. V. et al. **Compilers: Principles, Techniques, Tools**. 2nd. ed. Boston: Addison-Wesley, 2006. WorldCat. ISBN 978-0321486813.

AUMILLER, D. et al. Structural text segmentation of legal documents. **Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law (ICAIL '21)**, Association for Computing Machinery, p. 2–11, 2021. Disponível em: <https://dl.acm.org/doi/abs/10.1145/3462757.3466085>.

BACH, M. P. et al. Text mining for big data analysis in financial sector: A literature review. **Sustainability**, MDPI, 2019. Disponível em: <https://www.mdpi.com/2071-1050/11/5/1277>.

BANK, E. C. **Investment fund statistics**. European Central Bank, Acessado em 30/06/20234. Disponível em: <https://www.ecb.europa.eu/stats/financial_corporations/investment_funds/html/index.en.html>.

BELKIN, M.; NIYOGI, P. Laplacian eigenmaps for dimensionality reduction and data representation. **Neural Computation**, MIT Press, Chicago, IL, USA, v. 15, n. 6, p. 1373–1396, 2003. Disponível em: <https://doi.org/10.1162/089976603321780317>.

BERRAR, D. **Bayes' Theorem and Naive Bayes Classifier**. 2001. Accessed: 19-05-2024. Disponível em: <https://www.researchgate.net/profile/Daniel-Berrar/publication/324933572_Bayes'_Theorem_and_Naive_Bayes_Classifier/links/5d837aba92851ceb79143b04/Bayes-Theorem-and-Naive-Bayes-Classifier.pdf>.

BRODER, A. Z. et al. Syntactic clustering of the web. **Computer Networks and ISDN Systems**, v. 29, n. 8, p. 1157–1166, 1997.

BUHMANN, M. D. Radial basis functions. **Acta Numerica**, Cambridge University Press, v. 9, p. 1–38, January 2000. Disponível em: <https://doi.org/10.1017/S0962492900000015>.

DAMLE, A.; MINDEN, V.; YING, L. Simple, direct and efficient multi-way spectral clustering. **Information and Inference: A Journal of the IMA**, v. 8, n. 1, p. 181–203, 06 2018. ISSN 2049-8772. Disponível em: <https://doi.org/10.1093/imaiai/iay008>.

Directive 2009/65/EC. Directive 2009/65/ec of the european parliament. **Official Journal of the European Union**, European Union, 2009. Disponível em: <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:32009L0065>.

EFAMA. **Monthly Statistics May 2021**. EFAMA, Acessado em 30/10/2023. Disponível em: <https://www.efama.org/newsroom/news/monthly-statistics-may-2021-net-assets-ucits-and-aifs-reach-eur-20-trillion-first>.

ESMA. **ESG names and claims in the EU fund industry**. 2023. Disponível em: <https://www.esma.europa.eu/sites/default/files/2023-10/ESMA50-524821-2931_ESG_names_and_claims_in_the_EU_fund_industry.pdf>.

ESMA. **Fund Management**. European Securities and Markets Authority (ESMA), 2024. Disponível em: <https://www.esma.europa.eu/esmas-activities/investors-and-issuers/fund-management>.

European Commission. **Investment Funds: EU Laws and Initiatives Relating to Collective Investment Funds**. European Commission, 2024. Disponível em: <https://finance.ec.europa.eu/capital-markets-union-and-financial-markets/financial-markets/investment-funds_en>.

European Parliament. **REGULATION (EU) 2020/852 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 18 June 2020 on the establishment of a framework to facilitate sustainable investment, and amending Regulation (EU) 2019/2088**. 2020. Article 6, first paragraph. Disponível em: <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32020R0852&from=EN>.

FALGOUT, R. D. An introduction to algebraic multigrid. **Computing in Science and Engineering, vol. 8, no. 6, November 1, 2006, pp. 24-33**, 4 2006. Disponível em: <https://www.osti.gov/biblio/897960>.

FATOUROS, G. et al. Transforming sentiment analysis in the financial domain with chatgpt. **Machine Learning with Applications**, v. 14, 2023. Disponível em: <https://www.sciencedirect.com/science/article/pii/S2666827023000610>.

FRANZ, A.; BRANTS, T. **All Our N-gram are Belong to You**. 2006. Google Research Blog. Disponível em: <https://google-research.blogspot.com/2006/08/all-our-n-gram-are-belong-to-you.html>.

GRAM, J. Ueber die entwickelung reeller functionen in reihen mittelst der methode der kleinsten quadrate. **Journal für die reine und angewandte Mathematik**, v. 1883, n. 94, p. 41–73, 1883. Disponível em: <https://doi.org/10.1515/crll.1883.94.41>.

GREEN, N. D. et al. Webbanc: Building semantically-rich annotated corpora from web user annotations of minority languages. In: **2009 Nordic Conference of Computational Linguistics**. [S.l.: s.n.], 2009.

LI, T. et al. Long-context llms struggle with long in-context learning. 2024. Disponível em: <https://doi.org/10.48550/arXiv.2404.02060>.

LLOYD, S. P. Least squares quantization in pcm. **IEEE Transactions on Information Theory**, v. 28, n. 2, p. 129–137, 1982. Disponível em: <https://doi.org/10.1109/TIT.1982.1056489>.

LUXBURG, U. von. A tutorial on spectral clustering. **Statistics and Computing**, Springer, v. 17, p. 395–416, 2007. ISSN 0960-3174. Disponível em: <https://doi.org/10.1007/s11222-007-9033-z>.

MARKETS. **NLP in Finance Market**. 2023. 376 p. Disponível em: <https://www.marketsandmarkets.com/Market-Reports/nlp-in-finance-market-21737879.html>.

NIELSON, N.; BROWN, C. E.; PHILLIPS, M. E. Expert systems for personal financial planning. The Institute of Certified Financial Planners, 1990. Disponível em: <https://prism.ucalgary.ca/items/4d6643a5-ba25-469e-822e-85a2ac37ed59>.

RAJARAMAN, A.; ULLMAN, J. D. Data mining. In: _____. **Mining of Massive Datasets**. Cambridge University Press, 2011. p. 1–17. ISBN 978-1-139-05845-2. Disponível em: <https://www.cambridge.org/core/books/mining-of-massive-datasets/data-mining/9C89D09B5E0B7B987E0B1D0289F82B21>.

SENATOR, T. E. et al. The fincen artificial intelligence system: Identifying potential money laundering from reports of large cash transactions. IAAI-95 Proceedings, 1995.

SOHANGIR, S. et al. Big data: Deep learning for financial sentiment analysis. **Journal of Big Data**, v. 5, n. 3, 2018. Disponível em: <https://link.springer.com/article/10.1186/s40537-017-0111-6>.

SUN, F. et al. Pre-processing online financial text for sentiment classification: A natural language processing approach. In: **2014 IEEE Conference on Computational Intelligence for Financial Engineering Economics (CIFEr)**. [S.l.: s.n.], 2014. p. 122–129.

VASWANI, A. et al. Attention is all you need. **CoRR**, abs/1706.03762, 2017. Disponível em: <http://arxiv.org/abs/1706.03762>.

VELDE, E. F. V. de. QR Decomposition. In: **Concurrent Scientific Computing**. Springer, 2005, (Texts in Applied Mathematics, v. 16). p. 125–140. Disponível em: <https://www.springer.com/gp/book/9780387240264>.

WINTERFELDT, D. von; EDWARDS, W. Decision trees. In: **Decision Analysis and Behavioral Research**. [S.l.]: Cambridge University Press, 1986. p. 63–89. ISBN 0-521-27304-8.

WITTEN, I. H.; FRANK, E.; HALL, M. A. **Data Mining: Practical Machine Learning Tools and Techniques**. 4th. ed. [S.l.]: Morgan Kaufmann, 2016. 296 p.

YU, S. X.; SHI, J. Multiclass spectral clustering. **Technical Report**, 2004. Disponível em: <https://people.eecs.berkeley.edu/~jordan/courses/281B-spring04/readings/yu-shi.pdf>.

ZHU, L. et al. Llafs: When large language models meet few-shot segmentation. In: **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)**. [S.l.: s.n.], 2024. p. 3065–3075.