7
Estudo de Caso

Se os fatos não se encaixam na teoria, modifique os fatos.

Albert Einstein

Com o objetivo de propor uma aplicação para o sistema e avaliar alguns resultados de forma sistemática um estudo de caso em menor escala foi montado. O escopo reduzido serviu de facilitador para observação dos efeitos de determinadas melhorias e obtenção de algumas conclusões. O principal objetivo buscado aqui foi o de fazer uma prova de conceito, mostrando que técnicas de agrupamento, aplicadas sobre um modelo enxuto de representação textual, seria capaz de resolver uma dada ambigüidade lexical. O modelo utilizado foi o vetorial montado pelo sistema proposto no capítulo 6.

7.1
Montagem do Corpus

O *corpus* foi obtido a partir de consultas em máquinas de busca na *Web* sobre a palavra chave "LULA" e a partir de então se buscou dentro dos resultados documentos que versassem sobre o Presidente do Brasil e sobre o molusco marinho. Por meio de classificação humana os textos são classificados da seguinte forma:

- 1 – 10: textos onde o presidente LULA cita termos associados à pesca, podendo incluir também o molusco marinho LULA, como no caso específico do documento 10. Esses textos foram propositalmente inseridos para testar o sistema devido a sua característica dúbia com relação à classe, apesar de, por intermédio humano, serem considerados textos sobre política por conterem em todos os casos o nome do presidente LULA;
- 11 – 30: textos exclusivamente referentes ao molusco marinho LULA. Nessa categoria encontram-se receitas, dicas sobre pesca, relatos de pescarias do molusco e ainda um artigo sobre pesquisas

com a LULA gigante, animal muito raramente encontrado e nunca observado com vida em seu habitat natural até hoje[26];

- 31 – 50: textos diversos sobre o presidente LULA associados à política, economia, ao país em geral ou até ao mundo.

7.2
Resultados

A fase de pré-processamento contou basicamente com quatro ações, além da retirada de *stopwords* baseada em uma *stoplist* contendo 214 palavras:

1) Eliminação de caracteres especiais, pontuação e acentuação;
2) Eliminação de caracteres numéricos;
3) Normalização de termos para a formatação minúscula;
4) Normalização de termos para o singular;
5) Normalização de termos realizada por um algoritmo de lematização baseado no *Porter-stemmer*.

Dessa forma, ao final do pré-processamento, apenas termos com símbolos entre [a-z], com a restrição adicional de terem comprimento maior que 3 caracteres em sua cadeia, foram incorporados ao léxico. Os termos com cadeia de caracteres menor que 4 foram eliminados por serem considerados possuidores de pouca representatividade semântica e assim contribuiu-se para redução da dimensão do modelo vetorial.

Conforme (CARRILHO, 2007), para o processamento da mineração existem duas abordagens diferentes: a análise semântica, na qual o foco é a funcionalidade dos termos dos textos; e a análise estatística, que se preocupa com a freqüência de aparição de cada termo. Dentro do estudo de caso proposto ambas as abordagens são empregadas em momentos distintos. Durante a etapa de pré-processamento (veja seção 2.2 – Pré-Processmento) a análise semântica é utilizada para normalização e conseqüente enxugamento do léxico e num segundo

[26] Data de publicação do artigo (27/10/2007), que falava sobre a autópsia de uma dessas criaturas, que havia sido encontrada com excelentes condições, ou seja, sem nenhum pedaço arrancado por seu predador preferido, o cachalote. Posteriormente assisti a um programa de televisão dedicado à criatura, onde um pesquisador japonês conseguiu algumas fotografias dela viva em seu habitat por meio de uma câmera fixada junto a uma isca.

momento, durante a montagem do modelo vetorial (veja seção 3.3 – Modelo Vetorial) a análise estatística é usada sobre a saída da etapa de pré-processamento analisada semanticamente.

No primeiro experimento os itens 4 e 5 da fase de pré-processamento foram suprimidos e o léxico ficou com 2672 termos distintos, ou seja, construiu-se um espaço vetorial 2672-dimensional.

De certa forma pode-se dizer que o agrupamento visa a simular a construção de uma taxonomia, porém com isso sendo feito sem a interação humana e sem nomeação semântica dos rótulos da taxonomia. As taxonomias refletem uma caracterização coletiva ou individual de como os itens podem ser hierarquicamente classificados (ADAMO, 2000) e de certa maneira o agrupamento também reflete essa caracterização. No caso específico do estudo de caso uma taxonomia possível poderia ser "POLÍTICA" e "PESCA" ou "PRESIDÊNCIA" e "CULINÁRIA", por exemplo.

Os agrupamentos hierárquicos (veja seção 4.2. - Métodos de Formação dos Grupos) apresentados a seguir foram todos executados com o uso de *group average link* (veja seção 4.3.2 – Métricas de Similaridade).

A montagem do léxico 1 não contou com nenhuma estratégia de normalização de termos e por esse motivo indexou muitos termos distintos representantes de um mesmo conceito semântico demonstrando a importância de algum processo de redução ao radical. Isso é facilmente exemplificado por meio de um extrato do léxico (figura 22), que apresenta 11 vocábulos que traduzem conceitos fortemente relacionados ao termo *pesca*. Isso faz com que o espaço vetorial do modelo que representa o documento se torne muito grande e dificulte a tarefa de agrupamento, pois dispersa um mesmo conceito semântico em diversas dimensões do modelo vetorial.

```
.
.
.
1756. pesca
1757. pescada
1758. pescado
1759. pescador
1760. pescadores
1761. pescando
1762. pescar
1763. pescaria
1764. pesque
1765. pesqueira
1766. pesqueiro
.
.
.
```

Figura 22 – Extrato do léxico 1 indexado

O agrupamento hierárquico do experimento 1 pode ser visto na figura 23. Podem-se observar dois grupos principais que se juntam no valor 9 da escala de distância entre grupos. O primeiro grupo (A) é formado em sua grande maioria por textos sobre o presidente LULA. Esse grupo também abrangeu alguns textos da categoria dúbia, os documentos 2, 4, 6 e 9, não sendo, este fato, considerado um problema, visto que o termo LULA nestes documentos se refere ao presidente do Brasil.

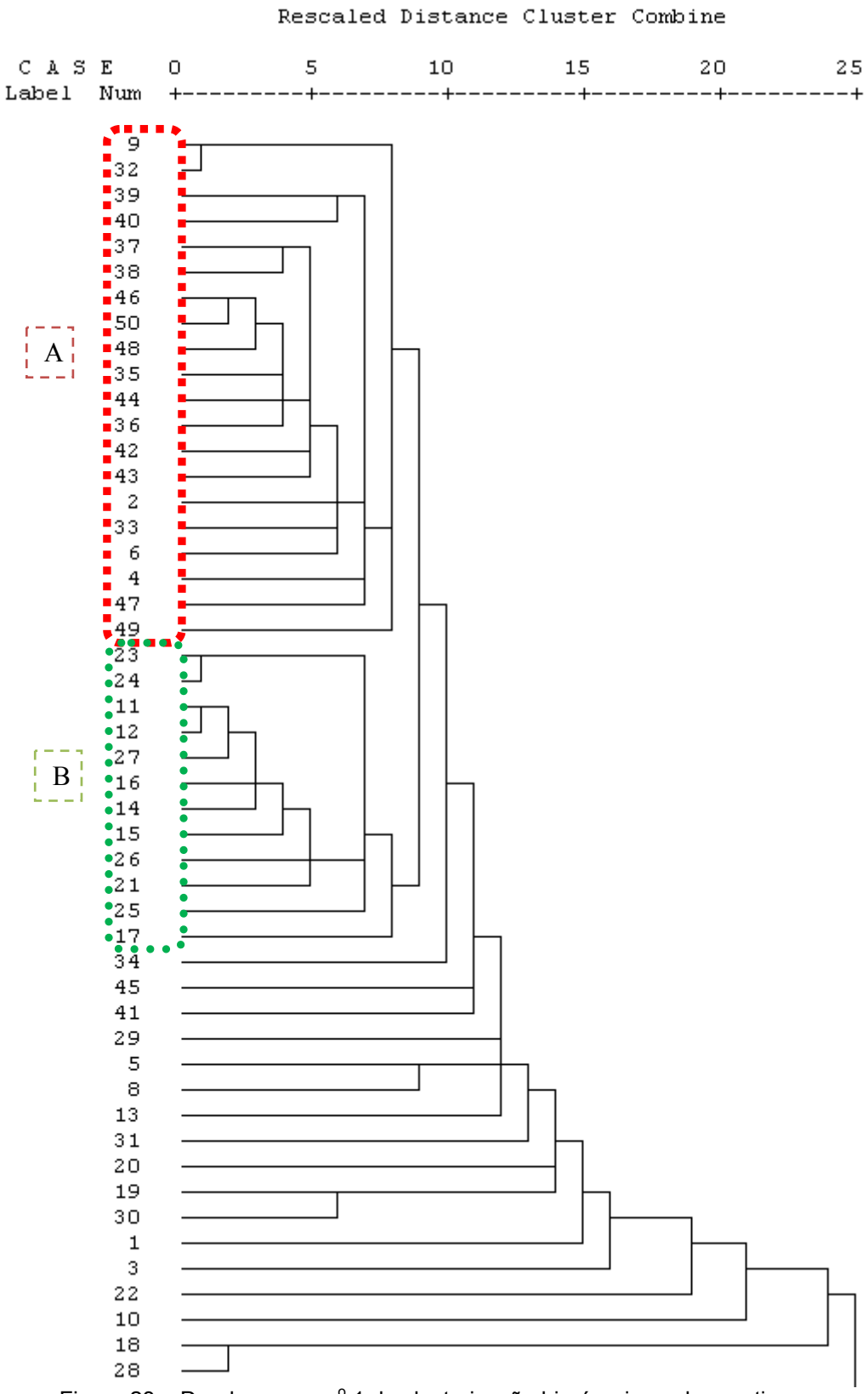

Figura 23 – Dendrograma n°.1 de clusterização hierárquica aglomerativa

Pode ser considerado problemático o fato do grupo A não ter sido capaz de abranger os documentos 31, 34, 41 e 45, que só foram agrupados após a reunião dos grupos A e B em um mesmo grupo. O documento 34 é incorporado no passo seguinte (escala de distância 10) e logo depois, na escala 11, são incorporados o 41 e o 45. Porém a melhor escala de corte para formação dos grupos para obtenção de uma divisão mais precisa das duas classes seria na escala 9, o que demonstra ainda certa confusão na separação dos grupos, principalmente com respeito ao grupo B (que representaria a categoria relacionada ao molusco), que foi bem menos abrangente incluindo apenas 12 dos 20 documentos da categoria. Já sob o ponto de vista de precisão os dois grupos foram muito bem formados, com 100% de precisão. Uma boa abordagem para essa situação seria oferecer 3 grupos para decisão do usuário, o A, o B e outro grupo formado pelos demais.

O segundo experimento contou com a normalização de termos para o singular e, prontamente, reduziu o léxico em 11%, deixando-o com 2462 termos. O extrato do léxico, apresentado na figura 24, seccionado do mesmo segmento semântico do exemplo anterior do léxico 1 comprova a redução e possui agora apenas 10 termos relacionados ao conceito semântico *pesca*. Uma redução tímida ainda, porém capacitou o agrupamento a se tornar mais abrangente. A figura 25 apresenta o dendrograma pelo agrupamento hierárquico aglomerativo sendo, diferentemente do experimento 1, a classe A representante do molusco marinho e a classe B do presidente.


```
.
.
.
1592. pesca
1593. pescada
1594. pescado
1595. pescador
1596. pescando
1597. pescar
1593. pescaria
1599. pesque
1600. pesqueira
1601. pesqueiro
.
.
.
```

Figura 24 – Extrato do léxico 2 indexado

A redução da dimensão do modelo vetorial e a conseqüente maior coesão entre termos (conteúdo gráfico da palavra) e conceitos (conteúdo semântico da palavra) contribuíram de sobremaneira para a melhor formação dos grupos. Vê-se pelo trecho extraído do léxico (figura 24) que o termo *pescadores* foi fundido ao termo *pescador* por eles se tratarem de termos que representam o mesmo conceito semântico e a indexação de ambos conjuntamente facilita qualquer tipo de análise baseada na idéia ou conceito a ser expresso pelas palavras.

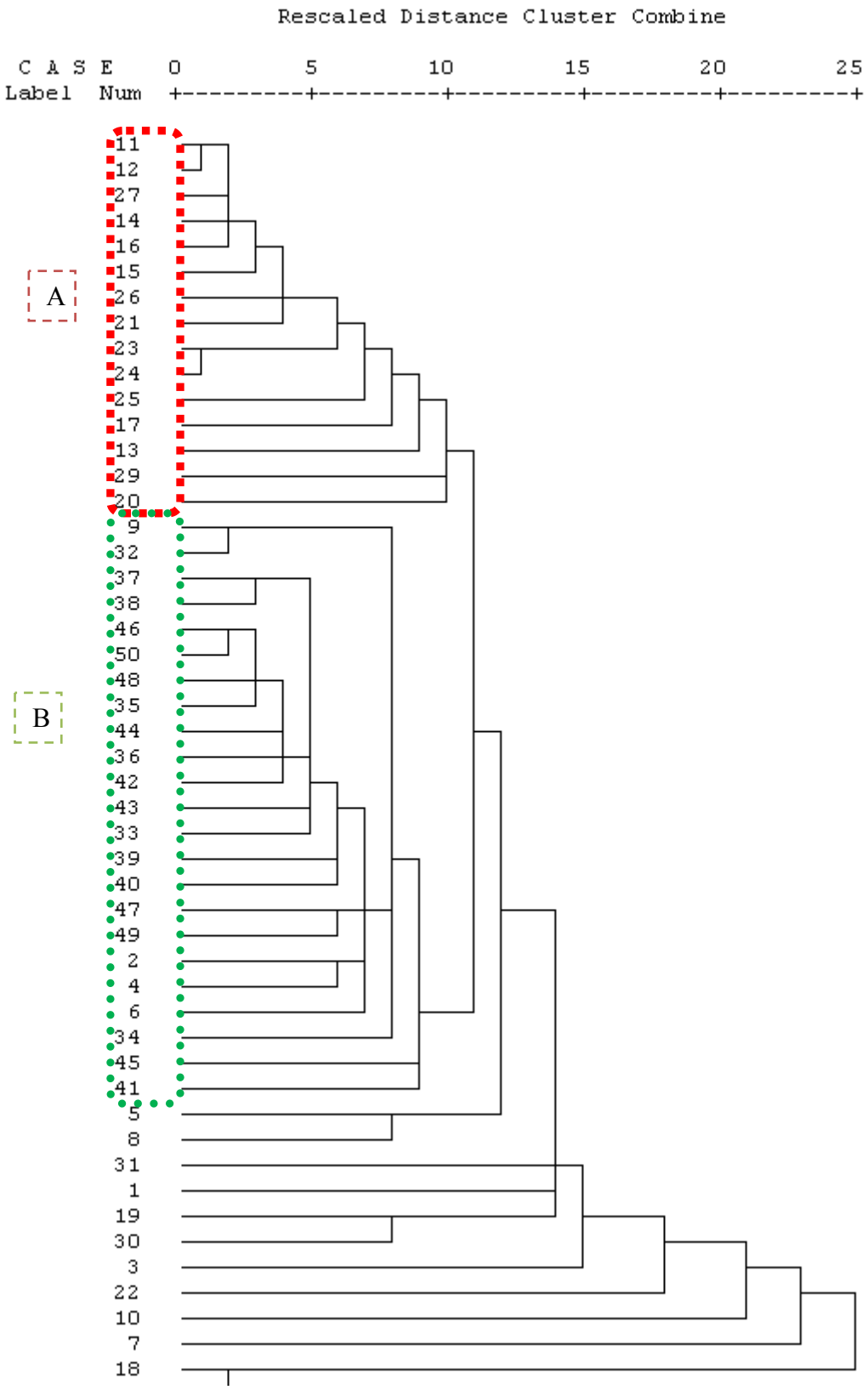

Figura 25 – Dendrograma n°.2 de clusterização hierárquica aglomerativa

Pode-se verificar que o grupo A (molusco) obteve abrangência de 75%, maior que os 60% do experimento anterior. Já o grupo do presidente deixou de fora apenas um documento e incluiu dessa vez 4 dos 10 documentos da categoria dúbia.Mais uma vez a precisão foi de 100% em ambos os casos, considerando a escala 11 como corte para formação dos grupos, pois em 12 ocorre a fundição dos dois grandes grupos.

O experimento 3 contou com uma estratégia mais complexa de normalização. As idéias utilizadas no algoritmo Porter-stemmer foram adaptadas para o português, tornando o processo de normalização mais robusto. A abordagem é calcada na eliminação de prefixos (*ad*, *in*, etc.) e sufixos (*ação*, *ável*, *mente*, *ava*, *asse*, *or*, etc.). Percebe-se facilmente, por meio da figura 26, que o mesmo trecho extraído dos léxicos dos experimentos anteriores está agora muito mais enxuto.

```
.
.
.
1002. pesc
1003. pescad
1004. pesqu
.
.
.
```

Figura 26 – Extrato do léxico 3 indexado

Nesse experimento o léxico 3 teve redução de dimensão de quase 29% em relação ao léxico sem normalização, ficando com um espaço 1922–dimensional. A redução do tamanho do léxico com a maior coesão entre os termos e os conceitos por ele transmitidos proporcionou uma formação de grupos mais rápida. No grupo A 12 dos 15 documentos foram agrupados na escala 6 contra apenas 10 na escala 6 no experimento anterior, o que demonstra o surgimento de similaridades mais fortes como conseqüência da estratégia de normalização empregada. Por outro lado, sem se importar com a escala de distância, as métricas de eficiência no agrupamento permaneceram as mesmas do experimento 2, com precisão de 100% para ambas as categorias e 75% na categoria relativa ao molusco e 77% na categoria do presidente. Apenas 24% de indecisão foi registrada, com 12 documentos pertencendo à categoria ambígua.

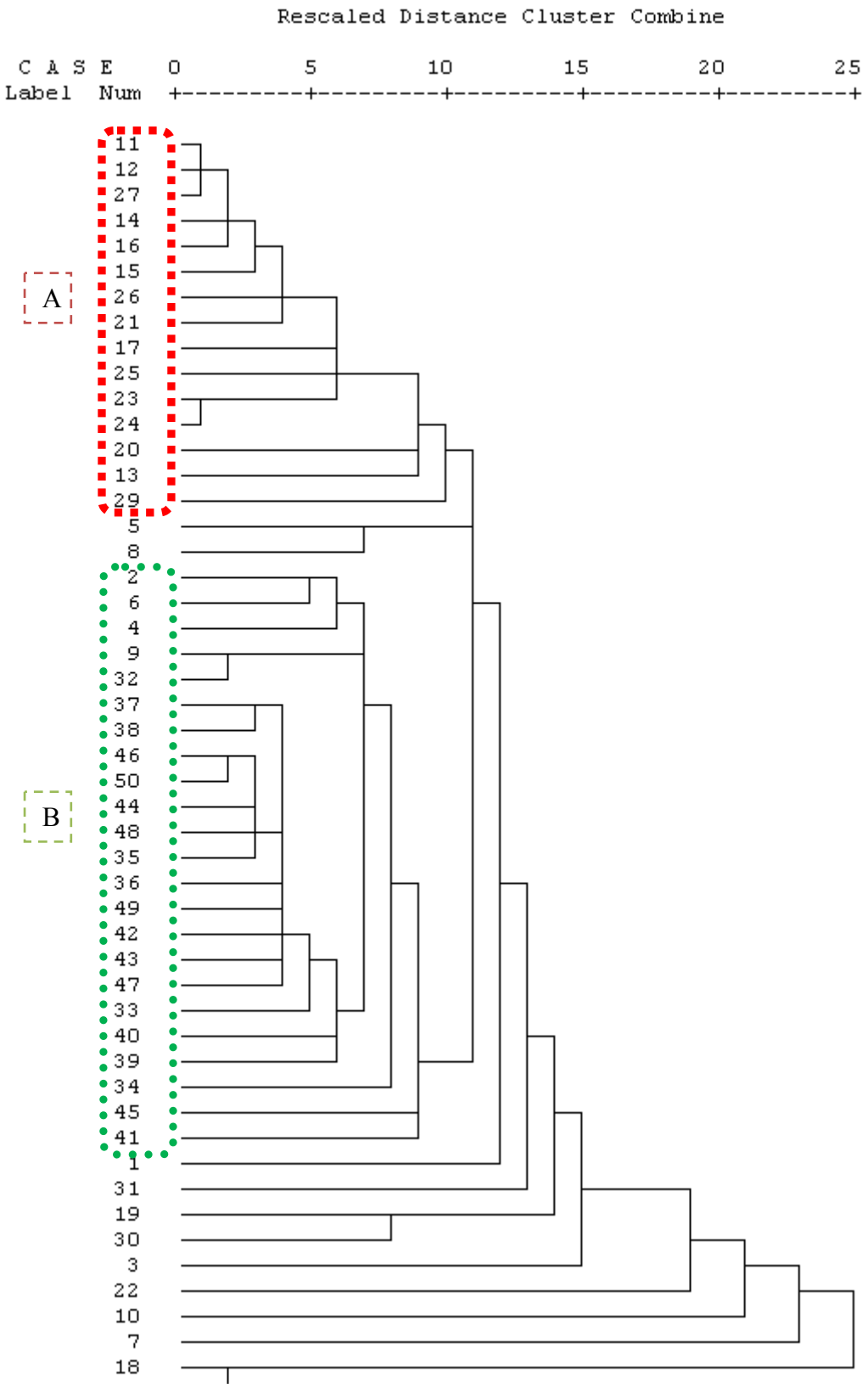

Figura 27 – Dendrograma n°.3 de clusterização hierárquica aglomerativa

Como resumo dos experimentos, no que tange a tarefa de agrupamento hierárquico com corte na escala mais adequada à caracterização das duas classes (presidente LULA e molusco LULA), tem-se o seguinte:

Experimento	Estratégia	Precisão (molusco)	Precisão (presidente)	Abrangência (molusco)	Abrangência (presidente)	Ambigüidade
1	Nenhuma normalização	100%	100%	60%	67%	34%
2	Normalização de plurais	100%	100%	75%	77%	24%
3	Lematização baseada em Porter adaptada à língua portuguesa	100%	100%	75%	77%	24%

Tabela 2 – Resumo dos resultados do agrupamento hierárquico

Com a utilização do algoritmo *k-means* para agrupamento dos documentos o resultado foi ainda melhor. Porém esse método tira proveito de, devido ao ambiente controlado e universo reduzido do estudo de caso, sabermos *a priori* que os documentos devem se dividir em dois grupos. Em uma situação prática esse valor deveria ser obtido de forma dinâmica. A abordagem utilizada para isso parte da obtenção do somatório das distâncias de cada item ao centróide de seu respectivo grupo. É intuitivo que a medida que o número de grupos se tona maior, os mesmos tendem a ser mais coesos e em conseqüência essa soma diminui. Por isso motivo um aumento dessa soma associado ao aumento do número de grupos pode ser indício de que o agrupamento começou a se perder e, portanto, ser usado na conclusão do número mais correto para a quantidade de grupos. O presente trabalho irá demonstrar essa abordagem através do gráfico de decaimento do somatório das distâncias dos itens aos respectivos centróides de seus grupos.

Para enxergar quão bem separados estão os grupos utilizar-se-á um gráfico de *silhouette* sobre os grupos de saída do algoritmo *k-means*. Esse gráfico apresenta um medida de quão próximo cada item do grupo está dos itens dos outros grupos. Valores próximos de +1 indicam itens que estão muito distantes dos grupos vizinhos; 0 indica itens que não se distingue ao certo qual o grupo; e -1 indica itens que, provavelmente, foram atribuídos ao grupo errado.

Em todos os 3 exemplos (sem normalização de termos, normalização de termos plurais e lematização) a divisão em dois grupos obteve 100% de precisão e 100% de abrangência. O primeiro exemplo, que não contou com nenhuma estratégia de normalização, apresentou muita confusão, segundo o gráfico de *silhouette*, no que tange a coesão interna e dispersão externa do grupo 1, grupo do molusco lula. Observam-se muitos documentos com valores negativos de *silhouette* nesse grupo. Já no grupo do presidente Lula apenas dois documentos apresentam essa característica, apesar dos valores de *silhouette* ficarem em torno de 0,5, não indicando uma certeza tão grande assim. Tudo isso pode ser visto na figura 28.

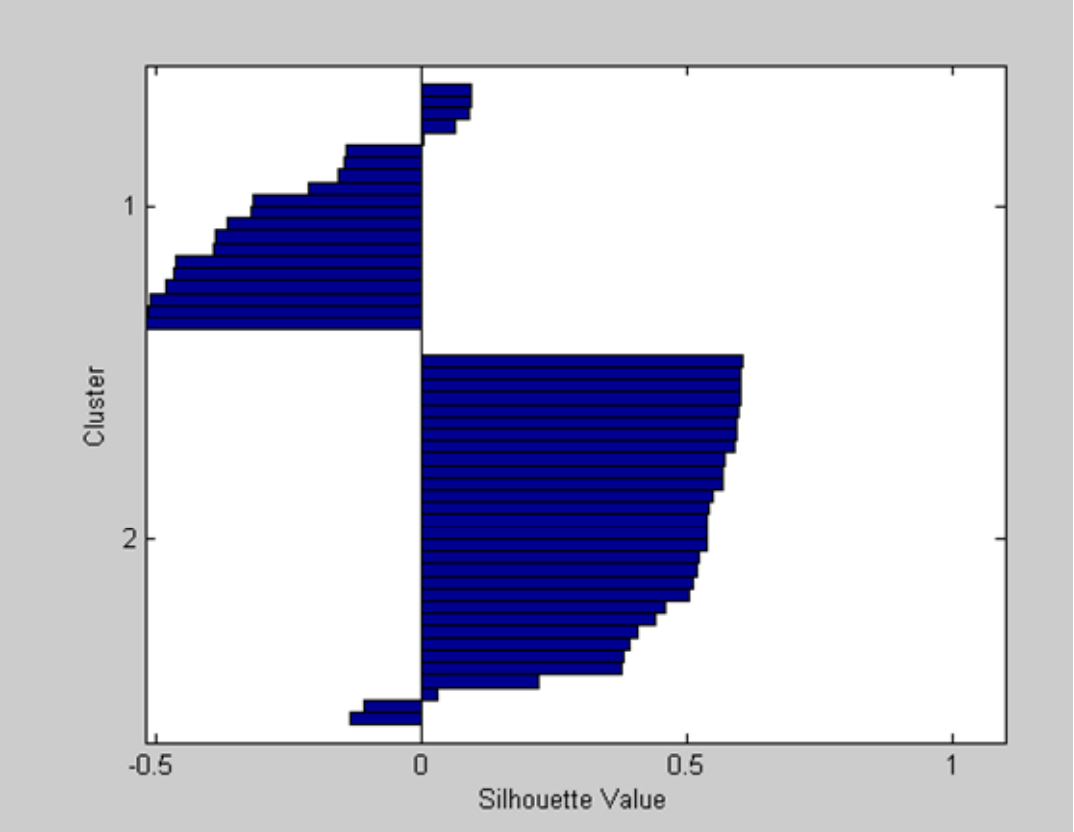

Figura 28 – Gráfico silhouette do experimento 1 com 2 grupos

Avaliando-se o gráfico do somatório das distâncias entre os itens e os respectivos centróides de seus grupos observa-se que para 11 grupos o valor de soma das distâncias apresenta um "vale" interessante indicando coesão interna dos grupos para a divisão em 11 grupos, conforme a figura 29.

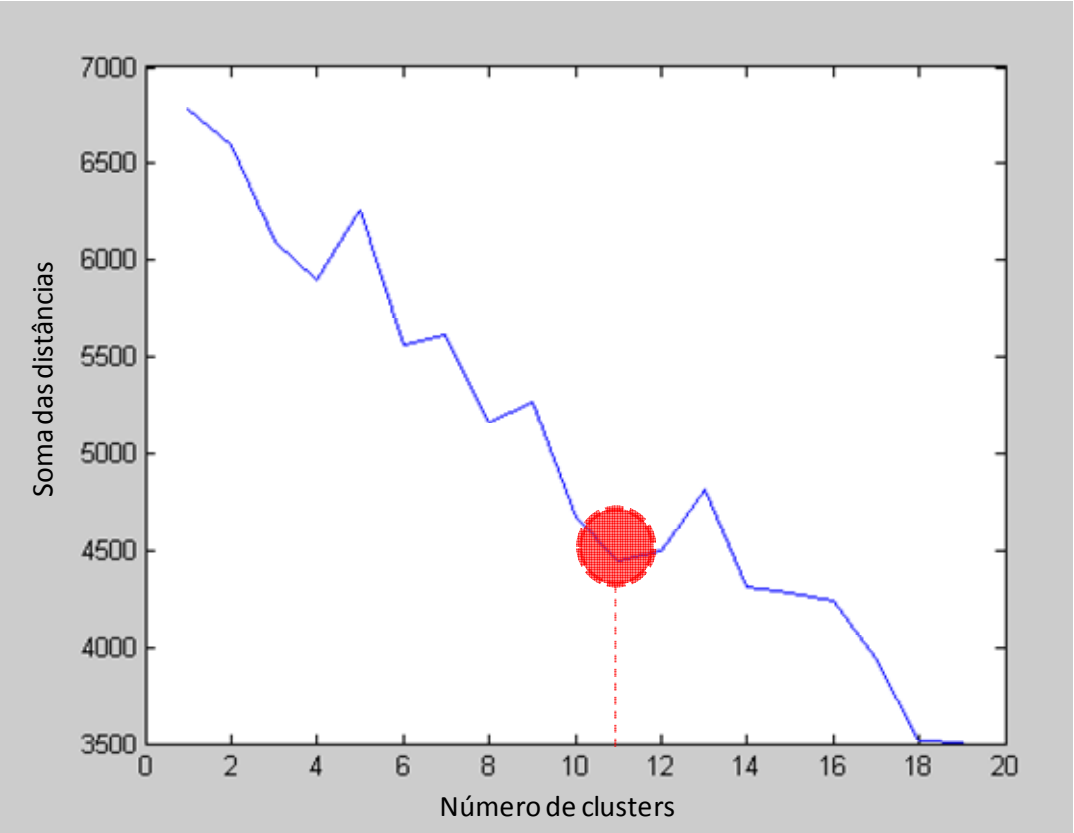

Figura 29 – Gráfico de soma das distâncias para o experimento 1

Porém, uma divisão em 11 grupos resultou em grupos muito bons com poucos documentos e grupos com mais documentos e com dispersão externa, em relação a outros grupos, muito ruim. Pode-se observar pela figura 30 que o grupo 8 apresentou valores de *silhouette* que chegam a ser menores que -0,8, demonstrando extrema imprecisão na formação desse grupo. Algo semelhante ocorre no grupo 11, formado por apenas 3 documento, que se apresentaram muito dispersos entre si. Na prática esse agrupamento seria descartado devido à formação de grupos menos coesos.

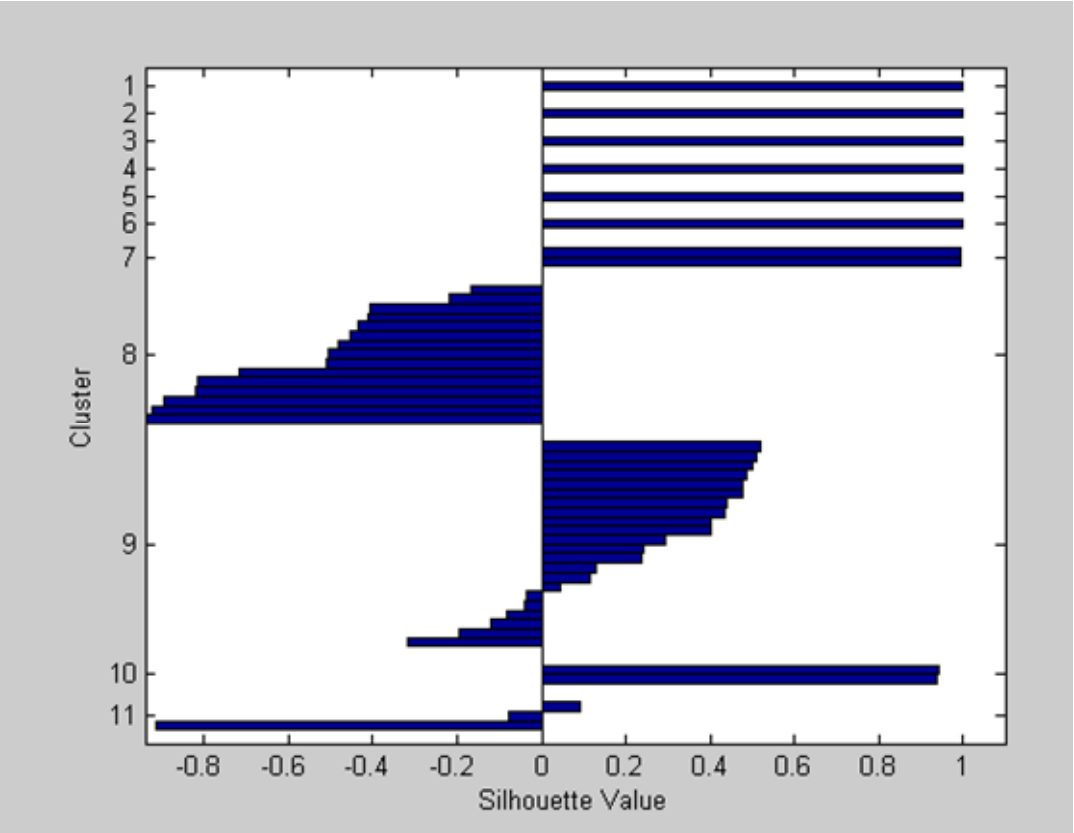

Figura 30 – Gráfico silhouette do experimento 1 com 11 grupos

O experimento 2, com normalização de termos baseada na supressão das palavras no plural e incorporação delas na respectiva dimensão do termo correspondente no singular, demonstrou, conforme a figura 31, melhoria de resultados no agrupamento. Assim como já havia acontecido para a mesma estratégia com o agrupamento hierárquico o grupo referente ao molusco projetou algumas incertezas. Já o grupo do presidente ficou com todos os valores de *silhouette* positivos, com exceção do documento 10, onde coexistia a palavra lula tanto se referindo ao presidente como também ao molusco marinho. Foi bastante interessante perceber como essas características foram capturadas pelo processo de agrupamento.

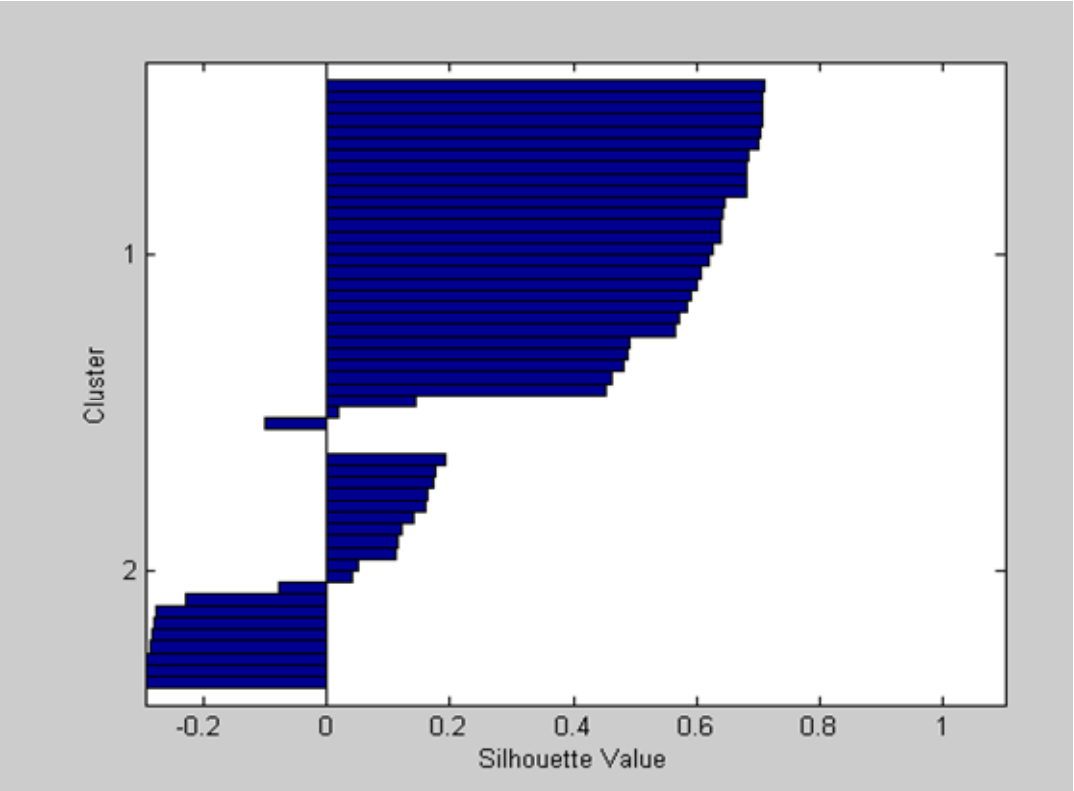

Figura 31 – Gráfico silhouette do experimento 2 com 2 grupos

Nessa abordagem número 2 de normalização de termos a curva de decaimento dos somatórios das distâncias não teve um "vale" muito expressivo que se caracteriza com grande probabilidade uma subdivisão em grupos melhor. Para efeito de avaliação foi tentada uma divisão em 12 grupos seguindo o indício destacado pelo círculo sobre o gráfico da figura 32. Porém pela figura 33 é possível observar que os grupos com mais documentos continuaram apresentando grande grau de confusão com valores de *silhouette* extremamente pequenos, o que refuta a hipótese de melhor ajuste com a subdivisão em mais grupos.

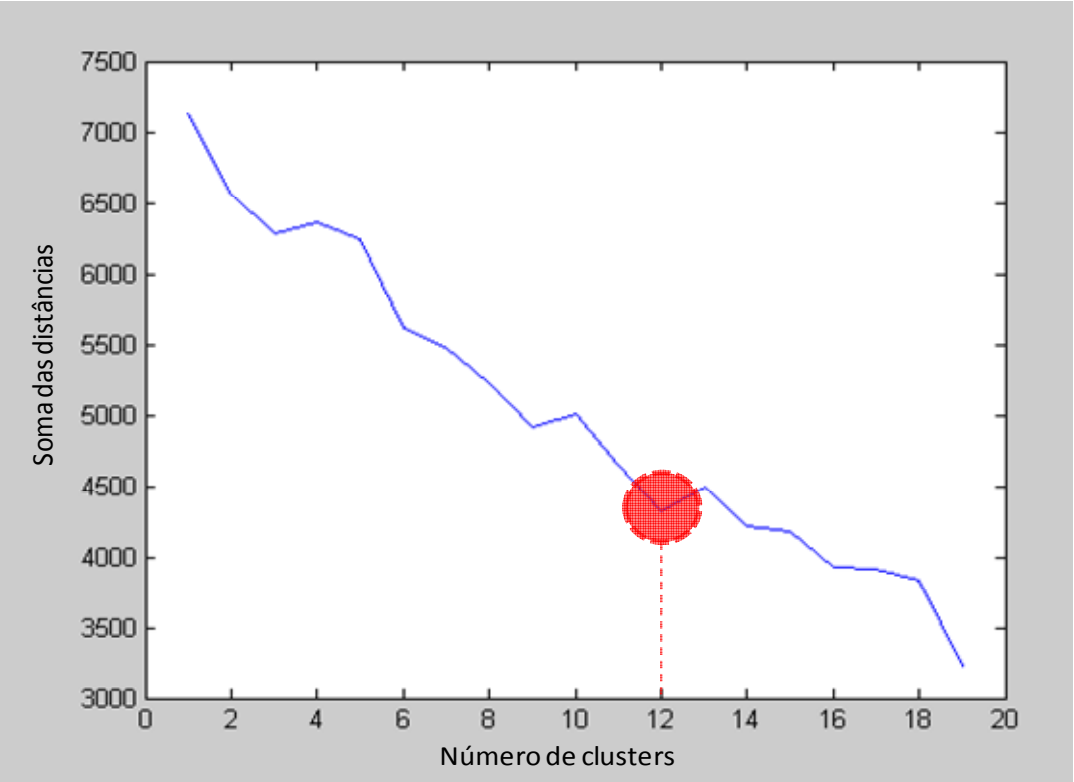

Figura 32 – Gráfico de soma das distâncias para o experimento 2

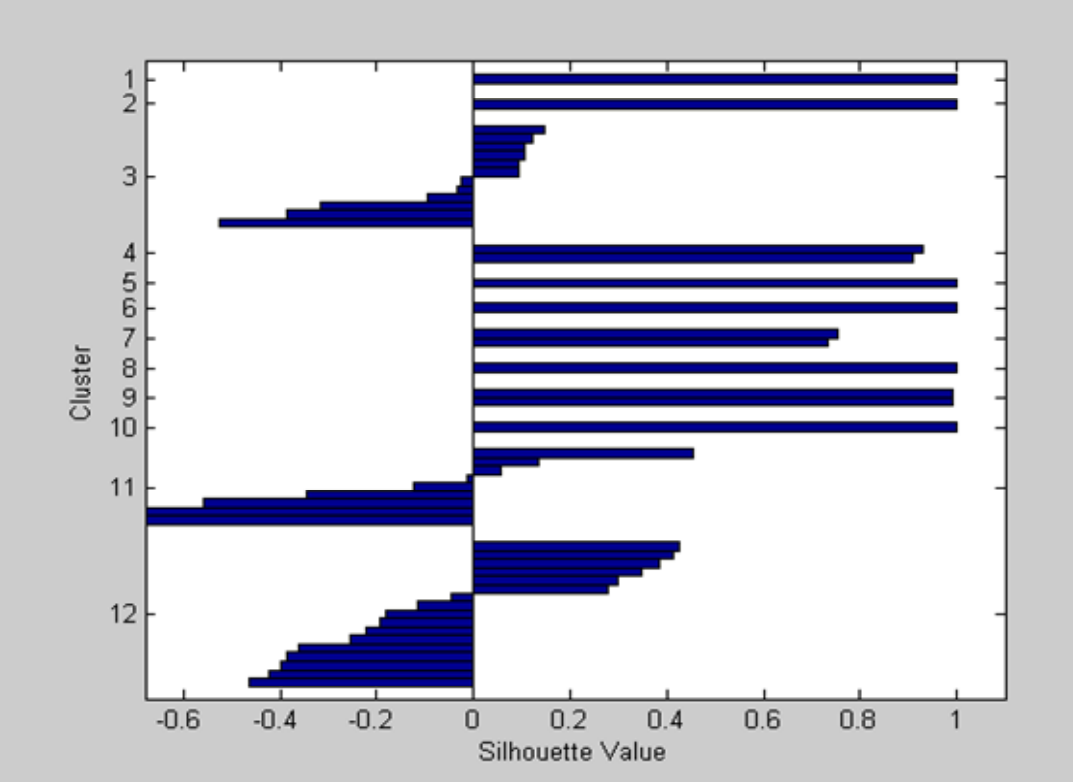

Figura 33 – Gráfico silhouette do experimento 2 com 12 grupos

Os ganhos de resultado do experiment 3 em relação ao 2, observados com o agrupamento *k-means* foram muito sutis. A lematização mais completa e maior redução na dimensão do léxico fortaleceram a coesão interna dos grupos e dispersão externa entre eles. Isso pode ser verificado pela figura 34, que apresenta valores de *silhouette* superiores aos do experimento 2 apesar de um desenho extremamente similar. Os valores positivos máximo se aproximaram mais de 0,8 (no grupo 1, do presidente) enquanto que os valores mínimos (grupo 2, do molusco) se aproximaram de -0,2, demonstrando ligeira diminuição da incerteza de alguns documentos dentro desse grupo.

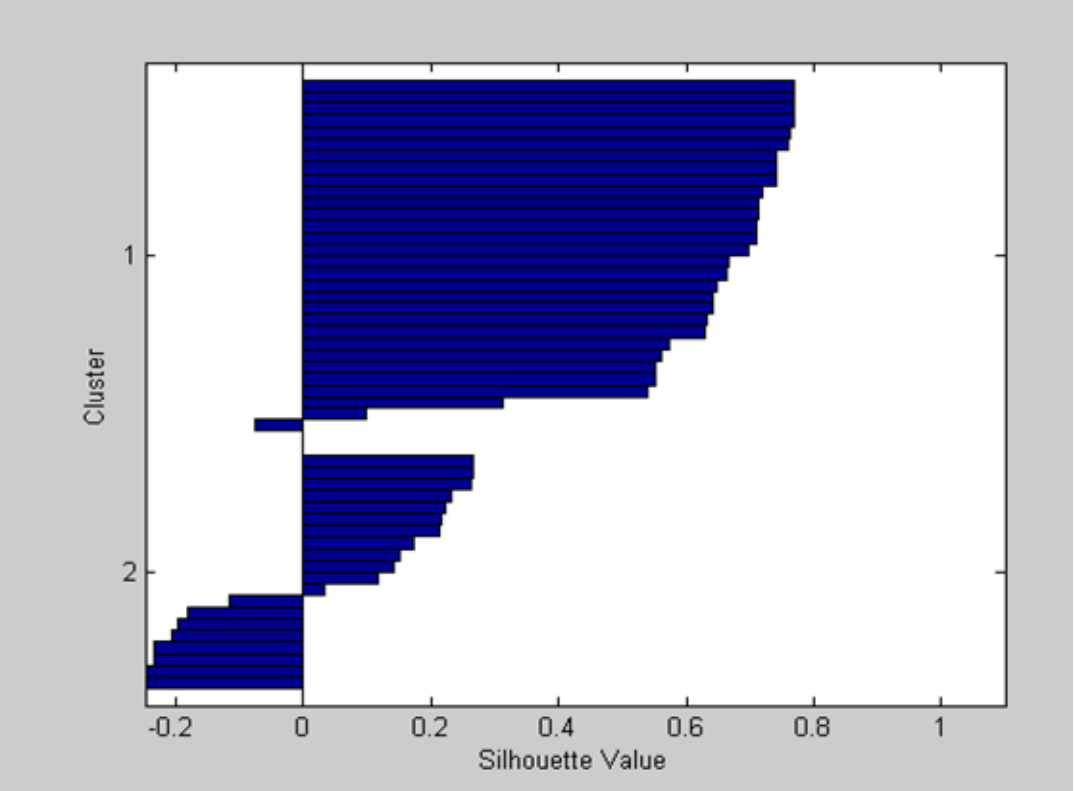

Figura 34 – Gráfico silhouette do experimento 3 com 2 grupos

Para efeito de teste e confirmação de que a melhor formação dos grupos seria baseada na divisão em 2, buscou-se um novo valor para o número de grupos e, conforme destacado na figura 35, pelo gráfico de decaimento do somatório das distâncias do itens aos respectivos centróides de seus grupos, optou-se por 7. Porém a figura 36 apresenta os mesmos problemas das divisões em mais de dois grupos anteriores, ou seja, grande grua de incerteza nos grupos que reúnem mais documentos. É importante frisar que esse procedimento, de teste de outros valores

para o número de grupos, é útil considerando que em um caso prático não se saberia *a priori* o número de classes dos documentos.

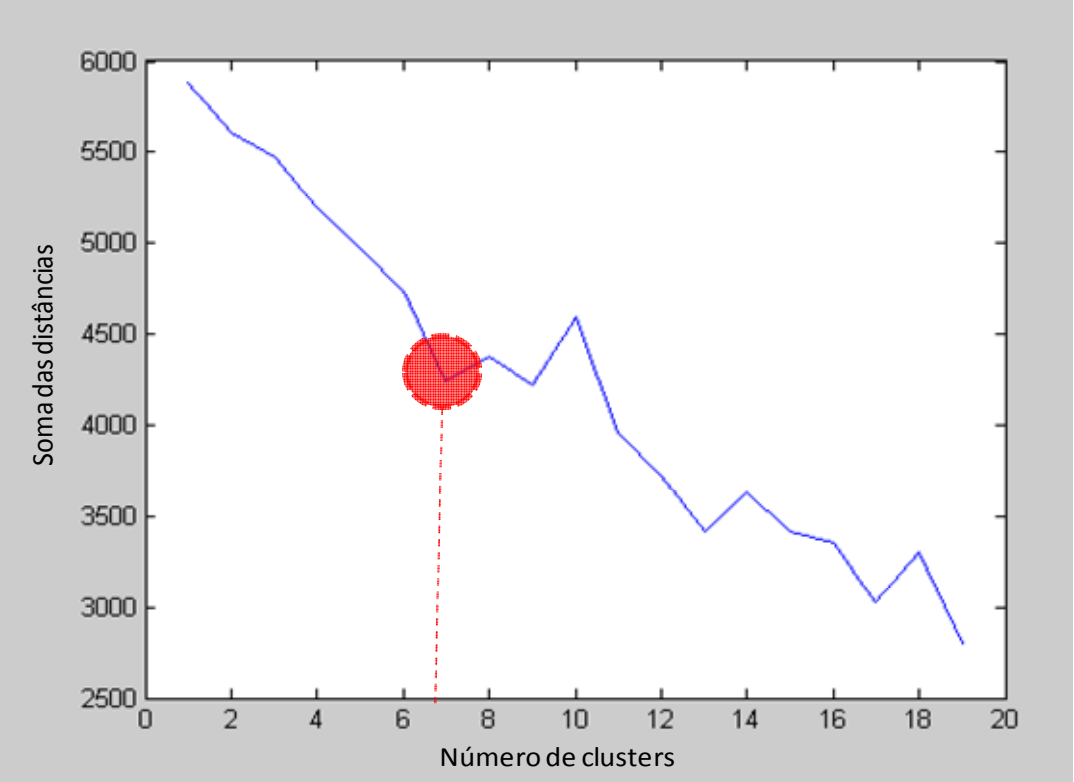

Figura 35 – Gráfico de soma das distâncias para o experimento 3

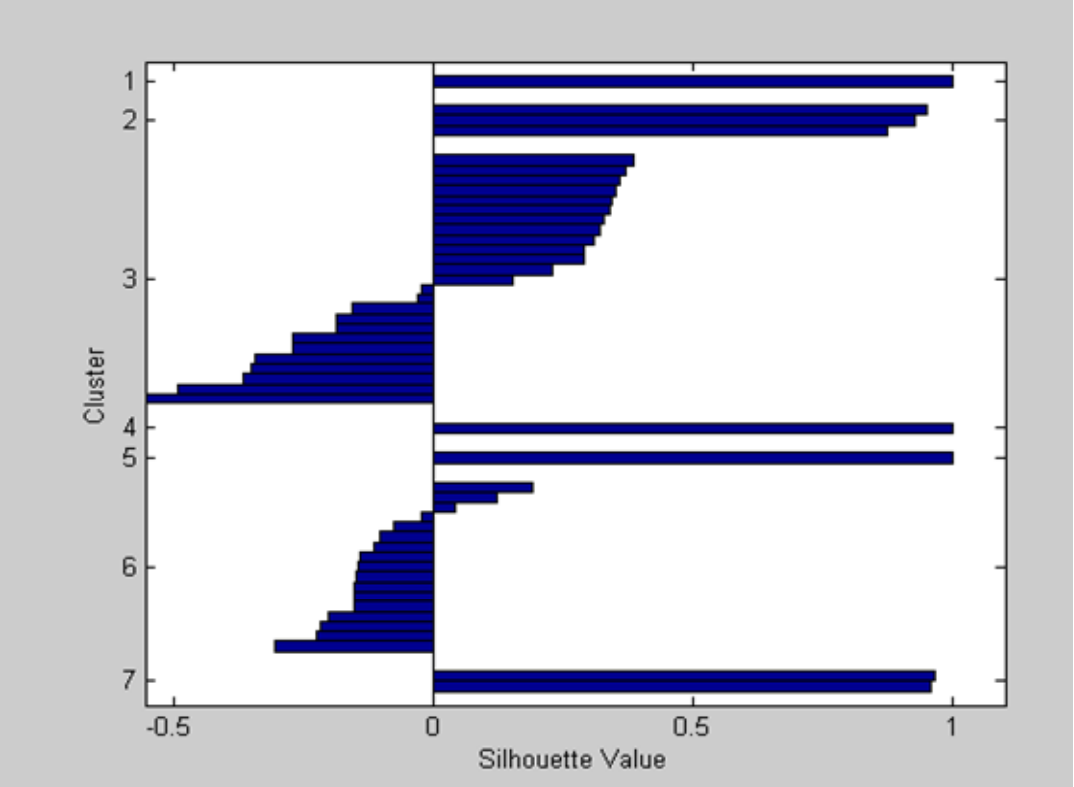

Figura 36 – Gráfico silhouette do experimento 3 com 7 grupos

7.3
Trabalhos futuros

O uso de informações lingüísticas para identificação de termos compostos (veja seção 2.2.2.1 – Identificação de colocações) ou ainda melhoria do procedimento de redução ao radical (veja seção 2.2.2.3 – Lematização) durante o processo de *tokenização* deve ser capaz de melhorar ainda mais os resultados do agrupamento e conseqüentemente da recuperação de informações. A primeira técnica contribui para melhoria na tradução das palavras em conceitos, com estes tendendo a serem mais específicos por serem constituídos por mais de um termo. A segunda técnica contribui para redução da dimensionalidade do léxico, facilitando a tarefa de agrupamento, que realiza cálculos de similaridade. Para essa redução podem ser aplicadas outras técnicas para extração de características como **Análise de Componentes Principais**[27] (JOLLIFFE, 1986) e a **Decomposição em Valores Singulares**[28] (WALL, 2002) dentre outras. Porém a utilização dessas regras gera novas características, no caso termos, não possuem significado representativo no universo do léxico por serem resultantes de uma transformação do conjunto inicial. Isso dificulta a interpretação dos grupos formados, porém para o caso prático se prestaria muito bem ao objetivo de fornecer ao índice a informação do grupo ao qual o documento mais se enquadra e contribuir assim para maior precisão da recuperação da informação.

A utilização de uma rede semântica de termos como faz a WordNet pode auxiliar de sobremaneira o processo. O estabelecimento de um relacionamento do tipo hipernímia-hiponímia, por exemplo, pode ser usado para redução de dimensionalidade do espaço vetorial que modela os documentos, por meio do uso de conceitos mais gerais, tentando dessa forma reduzir o *gap* entre o conhecimento humano (dos conceitos associados às palavras) e as decisões de agrupamento dirigidas pelos dados (palavras). Outra estratégia para alcançar os mesmo objetivos seria o uso de um tesauro que contivesse conceitos sobre os relacionamentos existentes entre categorias estabelecidas de forma hierárquica, conforme (BANG, 2006) ou ainda algum outro tipo de ontologia que auxiliasse

[27] Do inglês Principal Component Analysis.
[28] Do inglês Singular Value Decomposition.

todo o processo. Esse dicionário de termos poderia ser constantemente realimentado com novos termos até então desconhecidos e assim se tornar cada vez mais robusto e "sábio".

Outro ganho a ser obtido com o uso de um dicionário é a possibilidade de correção de termos. O algoritmo de distância de Levenshtein (LEVENSHTEIN, 1966), popularmente conhecido como distância de edição, poderia ser usado para detectar e corrigir pequenos erros de grafia transformando uma palavra estranha em outra contida no dicionário e que tenha pequena distância de edição. Por exemplo, no estudo de caso construído, o léxico indexou a palavra "trabalhano", o que foi um erro óbivio de construção do morfema. A avaliação humana diria prontamente que a palavra correta é "trabalhando" ou até mesmo sequer perceberia o erro lendo automaticamente "trabalhando". Essa característica poderia ser simulada por meio do uso de um algoritmo de distância de edição e confronto de termos indexados com outros pertencentes a um dicionário. Dessa maneira, evita-se a adição de mais uma dimensão ao espaço vetorial, que de forma errada deve contribuir para piora nos resultados.

Outras abordagens um pouco mais complexas de agrupamento poderiam ser tentadas como o *bisection k-means* (STEINBACH, 2000) ou ainda algo específico para agrupamento de documentos textuais. Os algoritmos baseados em modelo de árvore de morfemas poderiam ser tentados no intuito de alcançar melhores resultados.

Sob o aspecto de desempenho do sistema uma alternativa para se conseguir aumentar a capacidade de textos processados poderia ser o uso de *swap* mais freqüente entre memória volátil e memória persistente.

O principal trabalho a ser realizado deveria ser a implementação do *pipe-line* proposto de forma mais automatizada. O presente trabalho executa diversas tarefas, em módulos separados, com o objetivo de compor o fluxo proposto nas figuras 19 e 20, porém ainda demanda de muito trabalho manual para início de procedimentos e alimentação de procedimentos subseqüentes. Um trabalho como esse permitiria que o sistema fosse testado ao extremo e novas abordagens de pré-processamento e também de agrupamento pudessem ser empregadas com objetivo de se realizar um ajuste mais fino e se obter melhores resultados.

7.4
Conclusões

Conforme apresentado durante a exposição teórica, várias abordagens para tratamento de ambigüidades textuais, no sentido de palavras, se utilizam de informações do contexto para tomar suas decisões. Porém a metodologia de busca por informações do contexto pode variar e o presente trabalho propôs a utilização de técnicas de agrupamento para realizar a tarefa de detecção de ambigüidades. Aliado a isso foram utilizadas técnicas de indexação especiais que carregam as informações de contexto obtidas pelos agrupamentos e dessa forma possibilitam que as ambigüidades sejam efetivamente localizadas e utilizadas para tomada de decisões sobre as informações a serem recuperadas oferecendo opções ao usuário e contribuindo assim para aumento de precisão na recuperação da informação buscada.

O primeiro ponto de destaque deve ser dado à importância do trabalho de pré-processamento. Isso ficou muito claro e exemplificado pela melhoria de resultados obtida com a indexação de termos com alguma estratégia de normalização em contraste com a indexação dos termos sem nenhuma estratégia de normalização. A lematização empregada contribuiu para redução do léxico e conseqüentemente para melhoria de qualidade do agrupamento por meio do aumento das similaridades entre documentos após a tradução das palavras para seus respectivos conceitos semânticos. A simples normalização das palavras para sua forma singular alavancou os resultados, pois obviamente o valor semântico de uma palavra dentro de um documento, no sentido de indicar, por exemplo, o assunto ao qual o documento se refere, independe do número que a palavra está grafada. A normalização, nesse caso, pode ser descrita como uma função N que mapeia o universo das palavras no universo de termos, ou seja, $M: P \rightarrow T$, conforme a figura 37.

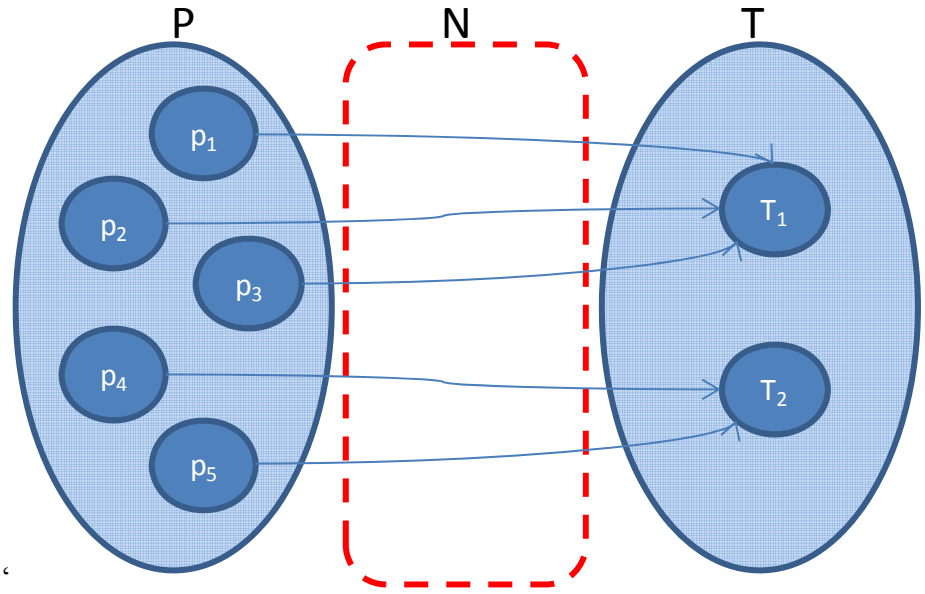

Figura 37 – Função de normalização, que mapeia palavras em termos

Em geral, não ocorreram muitas diferenças nos resultados com respeito à variação das técnicas de agrupamento utilizadas. O agrupamento hierárquico apresentou confusão para integração dos documentos dúbios na categoria mais adequada, ou seja, a do presidente Lula. Usando *k-means* foram obtidos resultados 100% corretos, tanto em precisão como em abrangência, porém deve se levar em consideração que em um ambiente real não seria sabido *a priori* a quantidade de grupos e esse valor teria que ser estimado através de algum procedimento. Foi dada uma sugestão para obtenção do melhor valor do número de grupos, o de avaliação da curva de decaimento do somatório das distâncias dos itens aos centróides de seus grupos. Porém deve se buscar uma forma automatizada de capturar o melhor valor. Para ambos os caso de agrupamento, hierárquico e heurístico, a mudança das métricas de similaridade e distância não incutiu em diferencial sensível nos resultados. Conforme apresentado no capítulo 4 (principalmente no item 4.2 – método de formação dos grupos) técnicas de agrupamento aplicadas especificamente a documentos textuais ainda estão sendo muito estudadas e uma grande variedade de novas técnicas têm surgido, porém os grandes paradigmas não apenas de formas de agrupamento e métricas de similaridade ou distância como também de modelos de representação dos documentos para que os agrupamentos possam ser aplicados não vêm apresentando grandes mudanças. Talvez alguma grande ruptura nos padrões atuais

possa ser capaz de revolucionar esse ambiente, porém enquanto isso, o que se vê, são apenas algumas pequenas mudanças e ajustes sutís objetivando resolver problemas pontuais e atuar em ramo mais específico do amplo universo da mineração de textos. Cabe ressaltar que não foi utilizada nenhuma técnica de agrupamento aplicável sobre modelagem na forma de árvore de morfemas conforme alguns trabalhos citadas também no item 4.2 e estas foram sugeridas como trabalhos futuros para confronto de resultados.

Em relação à indexação concluiu-se que melhorias nesse aspecto respondem por ganhos de desempenho em velocidade tanto na indexação quanto na recuperação da informação e ainda sob o aspecto de utilização de recursos de memória volátil. Um grande esforço foi gasto para construção de uma estrutura de dados enxuta capaz de suportar o carregamento de uma grande massa de dados textuais em memória para realização dos cálculos. A maior parte do tempo de execução do sistema é gasta com entrada e saída, que por sinal poderia ter sido ainda mais utilizada para permitir aumento na capacidade de processamento da massa de dados textuais, porém com conseqüente ônus de velocidade. A flexibilidade da estrutura de indexação também foi bastante importante para que fosse possível carregar dados dos grupos obtidos para dentro dos índices e proceder de forma rápida com a recuperação das informações já com a detecção das ambigüidades, oferecendo assim opções para que o usuário decidisse mais precisamente qual a informação a ser buscada.