Fundação Getúlio Vargas
Escola de Matemática Aplicada
Programa de Pós-Graduação em Matemática Aplicada

Técnicas de Processamento de Linguagem Natural Aplicadas às Ciências Sociais

Alice Duarte Scarpa

Rio de Janeiro - RJ, Agosto de 2017

Alice Duarte Scarpa

Técnicas de Processamento de Linguagem Natural Aplicadas às Ciências Sociais

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Matemática Aplicada da FGV, como parte dos requisitos necessários para a obtenção do Título de Mestre em Modelagem Matemática da Informação.

Fundação Getúlio Vargas

Escola de Matemática Aplicada

Programa de Pós-Graduação em Matemática Aplicada

Orientador: Renato Rocha Souza

Rio de Janeiro - RJ

Agosto de 2017

Ficha catalográfica elaborada pela Biblioteca Mario Henrique Simonsen/FGV

Scarpa, Alice Duarte
Técnicas de processamento de linguagem natural aplicadas às ciências sociais / Alice Duarte Scarpa. – 2017.
86 f.

Dissertação (mestrado) – Fundação Getulio Vargas, Escola de Matemática Aplicada.
Orientador: Renato Rocha Souza.
Inclui bibliografia.

1. Processamento da linguagem natural (Computação). 2. Linguística – Processamento de dados. I. Souza, Renato Rocha. II. Fundação Getulio Vargas. Escola de Matemática Aplicada. III. Título.

CDD – 006.35

ALICE DUARTE SCARPA

TÉCNICAS DE PROCESSAMENTO DE LINGUAGEM NATURAL APLICADAS ÀS CIÊNCIAS POLÍTICAS.

Dissertação apresentada ao Curso de Mestrado em Modelagem Matemática da Informação da Escola de Matemática Aplicada da Fundação Getulio Vargas para obtenção do grau de Mestra em Modelagem Matemática da Informação.

Data da defesa: 24/08/2017.

ASSINATURA DOS MEMBROS DA BANCA EXAMINADORA

Renato Rocha Souza
Orientador (a)

Moacyr Alvim Horta Barbosa da Silva

Flávio Codeço Coelho

Alexandre Luís Moreli Rocha

Agradecimentos

Família, Maurício, EMAp, Renato, pessoas que responderam à pesquisa.

Resumo

A enorme quantidade de documentos disponíveis atualmente representa um grande potencial de avanço para as Ciências Políticas. Ao mesmo tempo essa riqueza de informações gera um problema, pois não é mais possível que pesquisadores analisem todos os documentos manualmente. Técnicas modernas de processamento de linguagem natural têm um papel essencial a cumprir para auxiliar tal avanço. O objetivo desse trabalho é criar uma ferramenta baseada em processamento de linguagem de natural que ajude pesquisadores a navegar uma base de dados muito importante, o *Cablegate*, que é um conjunto de mais de 250 mil cabos diplomáticos de embaixadas dos Estados Unidos que foi publicado como parte do *WikiLeaks*. Essa é uma base muito importante que pode trazer uma nova luz sobre vários momentos-chave do início do século XXI.

Palavras-chave: documentos diplomáticos, processamento de linguagem natural.

Abstract

The vast amount of documents available nowadays presents a great opportunity for advancing Political Sciences. At the same time, this deluge of information poses a problem, because it is no longer feasible for researchers to analyze every document manually. Modern natural language processing techniques have an essential role in helping with this process. The goal of this work is to create a tool based on natural language processing techniques that helps researchers to navigate an important database, *Cablegate*, which is a corpus of over 250 thousand diplomatic cables sent between US embassies that was published as part of *WikiLeaks*. This is a very important database that can shed new light at key historical moments of the twenty-first century.

Keywords: natural language processing, diplomatic documents.

Lista de ilustrações

Figura 1 – Distribuição dos documentos ao longo dos anos 24
Figura 2 – Confidencialidade dos documentos 24
Figura 3 – Entre que cidades os documentos eram enviados 26
Figura 4 – A página inicial da ferramenta de busca 30
Figura 5 – A página de resultados da ferramenta de busca 31
Figura 6 – Um documento aberto na ferramenta 31
Figura 7 – Resultados para “brazil bolivia” 35
Figura 8 – Resultados para “oil” 36
Figura 9 – Resultados para “nuclear submarine” 37
Figura 10 – Formulário enviado a especialistas em Ciências Politicas 38

Lista de tabelas

Tabela 1 – Tópicos latentes encontrados pelo LDA 53
Tabela 2 – Tópicos latentes encontrados pelo LSI 59

Sumário

1 INTRODUÇÃO 1

Introdução 1
Motivação 1
Descrição do problema 1
Documentos 2
Organização deste trabalho 2

I ESTADO DA ARTE 3

2 RECUPERAÇÃO DA INFORMAÇÃO 5
2.1 Modelo vetorial 6
2.1.1 Problemas comuns 6
2.1.2 Pré-processamento 7
2.2 Representação matricial de um corpus 7
2.2.1 Notação 7
2.2.2 Construindo uma matriz de documentos 8
2.2.3 Tipos de representação matricial 9
2.3 Conclusão 10

3 LATENT SEMANTIC INDEXING (LSI) 11
3.1 PCA e SVD 11
3.2 Exemplos de Aplicações 13

4 LATENT DIRICHLET ALLOCATION (LDA) 15
4.1 Definições 16
4.2 Modelo 17
4.3 Aplicações do LDA 18
4.4 Conclusão 19

II O TRABALHO 21

5 BASE DE DOCUMENTOS 23
5.1 História 23
5.2 Visão geral 23
5.3 Distribuição geográfica 25

5.4 Importância 25

6 PROCESSAMENTO DOS DADOS 27
6.1 Pré-processamento 27
6.2 Ferramentas 27
6.2.1 Gensim 28
6.2.2 Mallet 28
6.3 Processamento 28
6.3.1 LSI 28
6.3.2 LDA 28
6.3.3 Definindo a quantidade de tópicos 29
6.4 Consultas 29
6.5 Interface 29

III PARTE FINAL 33

7 RESULTADOS E DISCUSSÃO 35
7.1 Exemplos 35
7.1.1 Brasil Bolívia 35
7.1.2 Petróleo 36
7.1.3 Submarinos nucleares 37
7.2 Análise 37

Conclusão e Trabalhos Futuros 41

REFERÊNCIAS 43

Glossary 47

ANEXOS 51

ANEXO A – TÓPICOS DO LDA 53

ANEXO B – TÓPICOS DO LSI 59

ANEXO C – PROGRAMA PARA GERAR O MAPA DE CONEXÕES POR CABOS 65

ANEXO D – PROGRAMA DE PROCESSAMENTO DE DADOS . 67
D.1 documentos.py 67

1 Introdução

Um sistema de busca de documentos inteligente pode ser uma ferramenta muito poderosa para pesquisadores na área de Ciências Políticas. Atualmente existe uma quantidade gigantesca de documentos disponíveis, e encontrar quais documentos são relevantes para uma determinada linha de pesquisa é uma tarefa muito complicada. Neste trabalho, propomos o uso de *LSI* e *LDA*, técnicas de processamento de linguagem natural, para resolver este problema.

Motivação

Encontrar exatamente quais são os documentos relevantes em um banco de dados com centenas de milhares de documentos é uma tarefa humanamente impossível. Para atacar esse problema pesquisadores podem utilizar heurísticas, como por exemplo filtrar por data ou local, fazer uso da busca por *palavras-chave*, recurso disponível para alguns bancos de dados ou utilizar ferramentas similares a que está sendo proposta neste trabalho.

Dois grandes problemas em buscar por palavras-chave são:

1. Sinônimos não são considerados, então documentos que contêm um termo equivalente não irão aparecer;
2. Se muitos documentos contêm as palavras-chave, ainda é difícil encontrar quais são os mais relevantes.

Latent Semantic Indexing (LSI) e *Latent Dirichlet Allocation* (LDA) são técnicas de processamento de linguagem natural que resolvem os dois problemas listados acima. Neste trabalho vamos criar um sistema de busca que gera resultados baseado nas duas técnicas, e iremos verificar como cada técnica satisfaz os nossos usuários. LSI é uma técnica muito simples que tem resultados empíricos bons (GRAESSER et al., 2000), enquanto que LDA é uma técnica mais moderna, baseada em métodos sólidos de estatística Bayesiana que tem tido muito sucesso recentemente (BLEI; NG; JORDAN, 2003).

Descrição do Problema

Este é um trabalho empírico que visa construir um sistema de busca de documentos diplomáticos que suporte encontrar documentos relevantes de maneira eficiente. Nessa dissertação nós explicamos como criar tal sistema, e verificamos se tal sistema beneficia seus usuários.

Documentos

Para este trabalho vamos construir um sistema de busca de documentos diplomáticos para a base de dados do *Cablegate* (WIKILEAKS, 2017a), que foi publicada como parte do projeto *WikiLeaks* (WIKILEAKS, 2017b). É uma base com mais de 250 mil cabos diplomáticos que foram enviados entre embaixadas dos Estados Unidos. Essa base de dados foi escolhida por seu grande interesse histórico, por ainda não ter sido estudada profundamente e pela grande quantidade de documentos que contém, sendo assim um exemplo perfeito para o sistema.

Organização

Na primeira parte do trabalho vamos olhar para o estado da arte. Primeiramente vamos introduzir o conceito de Recuperação da Informação, e daremos uma visão geral sobre o assunto. Depois iremos nos aprofundar nas duas técnicas utilizadas no trabalho, o LSI e o LDA. Na segunda parte do trabalho iremos explicar como foi realizado esse processamento, e como foi construído o sistema de busca. Na terceira parte do trabalho iremos avaliar a qualidade do sistema de busca, e verificar qual das duas técnicas funciona melhor para o problema deste trabalho, baseado na avaliação dos usuários.

Parte I

Estado da Arte

2 Recuperação da Informação

Recuperação da Informação é uma área da ciência da computação que lida com a obtenção automática de informação associada a um conjunto de documentos. O primeiro sistema computadorizado de recuperação da informação foi descrito em 1948, por Holmstrom (HOLMSTROM, 1948). Desde então houve grandes avanços na área.

Entender linguagem natural é um dos problemas mais antigos e difíceis da computação, desde as pesquisas iniciais em inteligência artificial dos anos 60. Computadores entendem apenas linguagens precisas e extremamente técnicas. Em cima delas, é possível construir outras linguagens mais fáceis de usar, mas que ainda requerem o mesmo nível de precisão. A fala humana é repleta de ambiguidades e depende de conhecimento prévio compartilhado pelos comunicadores.

O processo de classificar bases de dados manualmente é muito suscetível a erros, além de tomar muito tempo: Mesmo que um humano demore apenas dez segundos por documento, a quantidade de tempo necessária para processar bases de documentos atuais (ou mesmo considerando o tamanho médio de uma base há 30 anos) é proibitivamente alta. Mesmo quando há tempo para uma classificação humana rápida, a precisão da classificação sofre consideravelmente. No caso específico dessa dissertação, a base tem mais de 250 mil documentos. Mesmo restringindo-nos aos documentos originados em ou destinados ao Brasil, analisar os 5 mil documentos disponíveis não seria economicamente viável.

A ideia por trás de recuperação da informação é indexação eletrônica de documentos, combinada com uma forma rápida de encontrar os documentos mais relevantes para termos específicos. O objetivo é construir um sistema que encontre todos os documentos relevantes e a menor quantidade de documentos não-relacionados possível. Uma visão geral pode ser vista em (BAEZA-YATES; RIBEIRO-NETO et al., 1999).

Em seu artigo, van Rijsbergen (RIJSBERGEN, 2000) descreve quatro famílias de técnicas de recuperação da informação: as baseadas em espaços vetoriais de documentos/palavras, as técnicas probabilísticas, as que utilizam gramáticas formais e as baseadas em redes Bayesianas.

Neste trabalho iremos utilizar duas técnicas que se baseiam na representação vetorial de textos, *Latent Semantic Indexing* (LSI) e *Latent Dirichlet Allocation* (LDA). A última é uma técnica mais moderna que combina também métodos probalísticos. Vamos agora explicar mais detalhadamente como o modelo vetorial funciona.

2.1 Modelo vetorial

Modelos vetoriais são amplamente utilizados para representar documentos em sistemas de recuperação de informação. Nesse modelo, documentos são representados por vetores com uma entrada para cada termo que aparece nele. Nessa representação, um documento é visto como um saco de palavras, isto é, as posições das palavras nos documentos não são consideradas, apenas a quantidade de vezes que cada palavra aparece. Em bases com uma grande quantidade de documentos, como a que foi utilizada para esse projeto, a matriz de documentos tem dimensões muito grandes, o que poderia ter um custo computacional muito alto. Mas essa matriz é extremamente esparsa por natureza, pois apenas uma fração pequena das palavras do corpus aparece em cada documento. A matriz ser esparsa possibilita o uso técnicas de armazenamento que utilizam consideravelmente menos memória, como por exemplo um dicionário de chaves, com apenas os termos não-nulos. Matrizes esparsas também possibilitam cálculos mais rápidos.

Ao representar o conjunto de dados como uma matriz, é possível usar muitas técnicas de Álgebra Linear Aplicada vindas de subáreas como Processamento de Imagens e Aprendizado por Máquinas. No século XXI, houve um rápido crescimento nesta área de pesquisa, que tem impressionado pesquisadores e praticantes de todo mundo pela sua simplicidade e eficácia.

2.1.1 Problemas comuns

Alguns dos problemas comuns encontrados ao representar documentos como sacos de palavras são descritos abaixo.

- Sinônimos: A princípio, o algoritmo não sabe que duas palavras representam o mesmo conceito. Apesar de ser possível fornecer um dicionário de sinônimos pagando um custo computacional razoável, é possível que a relação entre duas palavras seja dependente de contexto.

- Polissemia: Uma única palavra pode ter diferentes significados dependendo do contexto. Por exemplo, se procuramos por “manga camisa”, não queremos que a palavra “manga” faça com que a busca retorne resultados envolvendo a fruta manga.

Em geral, resolver tais imprecisões da linguagem natural de modo automático é difícil. Apesar disso, alguns algoritmos, como o LSI e o LDA, são capazes de determinar que duas palavras são sinônimas se treinados em um corpus apropriadamente grande.

2.1.2 Pré-processamento

Existem técnicas que podem ser aplicadas no corpus, antes de qualquer tratamento algorítmico profundo, para aumentar a eficiência de algoritmos de processamento natural de linguagens. Tais técnicas têm como objetivo remover parte do conteúdo original que não contém informação. Duas dessas técnicas utilizadas neste trabalho estão descritas abaixo.

Stopwords são palavras extremamente comuns, como artigos, preposições, conjunções e afins. Tais palavras comumente aparecem em todos os textos, e portanto não servem para diferenciar quais documentos devem ser retornados. Além disso, como tais palavras adicionam um ruído que pode obscurecer a informação relevante de outras palavras mais especializadas, é de praxe removê-las antes de rodar os algoritmos de aprendizado.

Para determinar quais palavras são *stopwords*, existem duas opções: Comparar uma lista de stopwords de uma dada língua, ou fixar uma frequência acima da qual palavras devem ser consideradas *stopwords*.

Além disso, uma mesma palavra pode ser modificada com sufixos que não são relevantes para buscas (por exemplo, verbos conjugados). O processo de *stemming* consiste em encontrar um prefixo de uma dada palavra comum a todas as suas variações, similar ao conceito de radical de um verbo.

2.2 Representação matricial de um corpus

Nesta seção iremos definir a notação que será utilizada para tratar matricialmente de um *corpus* ao longo de todo o trabalho. Também iremos exemplificar a construção de uma matriz para representar um *corpus* e mostrar algumas possibilidades diferentes de processamento que podem ser utilizadas na construção de tal matriz.

2.2.1 Notação

O objeto a ser estudado é um conjunto de documentos, denominado *corpus*, que será denotado por $\mathcal{D}$. Os documentos de $\mathcal{D}$ são ordenados, de modo que

$$\mathcal{D} = \{d_0, \ldots, d_{m-1}\}.$$

Cada documento é uma sequência de palavras de um dicionário $\mathcal{W}$. Novamente temos uma ordem para as as palavras, de modo que

$$\mathcal{W} = \{w_0, \ldots, w_{n-1}\}.$$

2.2.2 Construindo uma matriz de documentos

O objetivo dessa subseção é exemplificar, num corpus concreto, a construção de uma matriz de documentos X. O elemento x_{ij} dessa matriz X representa a relação entre o documento d_i e o termo w_j. A dimensão da matriz é $m \times n$, em que m é a quantidade de documentos e n é a quantidade de termos.

O corpus a ser considerado neste exemplo irá consistir dos seguintes três documentos.

Número	Texto
0	Comprei um jogo ontem de manhã.
1	João assistiu ao jogo do Cruzeiro ontem.
2	Ontem e hoje eu trabalhei, ontem não dormi.

O primeiro passo para construir a matriz é gerar um dicionário a partir das frases. No caso acima, o dicionário é mostrado abaixo.

0	comprei	6	joão	12	e
1	um	7	assistiu,	13	hoje
2	jogo	8	ao	14	eu
3	ontem	9	jogo,	15	trabalhei
4	de	10	do	16	não
5	manhã	11	cruzeiro	17	dormi

Iremos agora construir as linhas da matriz. Lembrando que cada coluna corresponde a uma das palavras do dicionário, o documento 0 é então representado por

$$\begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix},$$

o documento 1 é representado por

$$\begin{pmatrix} 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix},$$

e o documento 2 tem a seguinte representação:

$$\begin{pmatrix} 0 & 0 & 0 & 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 \end{pmatrix}.$$

A matriz X é, portanto,

$$X = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 \end{pmatrix}.$$

Note que no caso consideramos a posição $x_{i,j}$ como sendo a quantidade de vezes em que a palavra i apareceu no documento j. Na próxima seção iremos apresentar outras formas de construir essa matriz.

2.2.3 Tipos de representação matricial

A matriz acima foi gerada utilizando através de uma simples contagem de palavras. Uma ideia útil é a de aplicar uma função à contagem obtida antes de formar a matriz. Para formalizar esse conceito, definamos a função tf, que consiste em considerar a quantidade de vezes que uma palavra aparece num texto. Em outras palavras,

$$\mathrm{tf}(t, d) = \text{número de ocorrências do termo } t \text{ no documento } d. \tag{2.1}$$

Temos então as seguintes representações matriciais para um corpus.

Contagem

Nesse tipo, a matriz é dada por uma contagem simples, como vimos na seção anterior.

$$x_{i,j} = \mathrm{tf}(w_j, d_i).$$

Binário

Podemos construir a matriz de modo a considerar apenas se a palavra aparece ou não no documento, isto é,

$$x_{i,j} = \begin{cases} 1 & \text{se o documento } d_i \text{ contém a palavra } w_j \\ 0 & \text{caso contrário} \end{cases}.$$

Logarítimico

Outra função comumente utilizada é $\log x$, o que nos dá

$$x_{i,j} = \log \mathrm{tf}(w_j, d_i). \tag{2.2}$$

TF-IDF

Também podemos levar em conta quantas vezes um termo aparece no conjunto de documentos como um todo. Se o corpus $\mathcal{D}$ tem m documentos um termo t aparece em em exatamente x desses documentos, definiremos

$$\mathrm{idf}(t, \mathcal{D}) = \log \frac{m}{x}.$$

A função tf-idf, que relaciona o documento d com o termo t, pode então ser definida por

$$\text{tf-idf}(t, d, \mathcal{D}) = \mathrm{tf}(t, d) \cdot \mathrm{idf}(t, \mathcal{D}),$$

onde $\mathrm{tf}(t, d)$ foi definida na equação (2.1).

A representação matricial em questão é então dada por obtida é dada por

$$x_{i,j} = \text{tf-idf}(w_j, d_i, \mathcal{D}).$$

2.3 Conclusão

Neste capítulo introduzimos o conceito de recuperação da informação e definimos o modelo vetorial que será utilizado neste trabalho. Nos capítulos seguintes apresentaremos duas técnicas de recuperação da informação que foram utilizadas neste trabalho: *Latent Semantic Indexing* e *Latent Dirichlet Allocation*.

3 Latent Semantic Indexing (LSI)

A LSI é um conjunto de procedimentos estatísticos automatizados para medir quantitativamente a semelhança de significado entre duas palavras ou grupos de palavras. De maneira um pouco mais específica, é uma técnica de recuperação da informação baseada no modelo vetorial introduzido no Capítulo 2. A técnica foi introduzida no contexto de recuperação da informação em 1988 (DEERWESTER, 1988) e, apesar de sua simplicidade, obteve bons resultados (GRAESSER et al., 2000). A LSI usa decomposição de valor singular (SVD), explicado a seguir, que é uma forma geral de análise fatorial, para condensar uma grande matriz de dados do tipo *word-by-context* em uma muito menor, mas que ainda contenha informações importantes sobre os dados. A LSI fornece uma maneira através da qual associações desconhecidas entre as palavras podem ser induzidas a partir de uma grande análise de como essas palavras co-ocorrem com todas as outras palavras na língua. LSI também pode ser usado para determinar a semelhança de palavras ou documentos com documentos externos a ele (MARTIN; BERRY, 2007).

A técnica consiste em 4 passos:

1. Construir uma matriz de documentos a partir do *corpus*
2. Fazer a decomposição SVD da matriz obtida
3. Escolher n componentes principais
4. Utilizar uma métrica de semelhança, como por exemplo o cosseno, para encontrar o documento mais semelhante

O passo 1 já foi abordado no capítulo anterior. Nesse capítulo vamos ver os passos 2, 3 e 4.

3.1 PCA e SVD

A *análise de componentes principais* (PCA) é um método de redução de dimensionalidade. Dados pontos num espaço vetorial de dimensão alta[1], o PCA permite encontrar uma projeção dos pontos num subespaço de dimensão k que preserve as características geométricas e de *clustering* dos pontos originais.

[1] No nosso caso, o espaço vetorial tem dimensão igual ao número de palavras; os pontos são as linhas da matriz descrita na seção 2.2 e portanto representam documentos.

A PCA está intimamente relacionada com o conceito de *decomposição em valores singulares* (SVD). Um *valor singular* de uma matriz real M é um número real σ tal que existe um par de vetores unitários (u, v) satisfazendo

$$Mu = \sigma v \qquad \text{e} \qquad M^T v = \sigma u.$$

A decomposição em valores singulares de uma matriz $m \times n$ é uma maneira de escrever a matriz M na forma

$$M = U \Sigma V^T$$

onde U é uma matriz $m \times m$ cujas colunas são ortogonais, V é uma matriz $n \times n$ cujas colunas são ortogonais, e Σ é uma matriz diagonal com entradas $\sigma_1 \geq \ldots \geq \sigma_n \geq 0$, os valores singulares de M.

Em termos mais intuitivos, escolhemos eixos $\{u_1, \ldots, u_m\}$ na esfera unitária do domínio de M; essa esfera será levada num elipsóide $\{\sigma_1 v_1, \ldots \sigma_n v_n\}$ no contradomínio de M, de modo que M leva o i-ésimo eixo que fixamos na esfera no i-ésimo eixo da elipse (se $m < n$, então alguns dos eixos são levados no vetor nulo). É conveniente ordenar os eixos do elipsóide em ordem decrescente de tamanho.

Quando a matriz é quadrada, um conceito relacionado é o de autovetor. No caso em que a matriz é (quadrada e) simétrica, a decomposição em autovetores sempre existe e tem forte significado geométrico. A decomposição em valores singulares preserva muita das propriedades da decomposição em autovetores, mas se aplica a toda matriz, o que garante grande aplicabilidade do método.

Ela possibilita implementar o PCA de um jeito simples: Se chamarmos de Σ_k a matriz que é igual a Σ nos primeiros k elementos da diagonal e igual a zero nos demais elementos, a matriz $M_k = U \Sigma_k V^T$ é a "melhor aproximação de posto k" para M. Em outras palavras, os pontos representados por M_k estão num subespaço de dimensão k e, sob essa restrição, minimizam a soma dos quadrados das distâncias aos pontos correspondentes em M.

Um tópico recente de pesquisa é o uso de algoritmos randomizados para obter uma aproximação, usando muito menos memória e tempo que os algoritmos determinísticos convencionais, para a decomposição SVD. A biblioteca *gensim*, utilizada nos experimentos dessa dissertação, usa o algoritmo randomizado descrito em (HALKO; MARTINSSON; TROPP, 2009) para realizar a decomposição SVD eficientemente.

De acordo com (LANDAUER; FOLTZ; LAHAM, 1998), "o número de dimensões retidas [no LSI] é uma consideração empírica". A quantidade de dimensões ótima depende da distribuição de palavras e da complexidade dos documentos, isto é, as características

estilísticas do corpus devem ser levada em conta. Assim, o trabalho de escolher a dimensão ótima é frequentemente feito com ajuste humano.

Uma vez feito o processamento do LSI, podemos comparar a similaridade de dois documentos encontrando os vetores correspondentes aos mesmos e calculando o cosseno do ângulo feito entre os vetores (também chamado de coeficiente de correlação linear de Pearson).

Dado um corpus $\mathcal{D}$ e um documento D (não necessariamente do corpus) qualquer, podemos então ordenar os documentos de $\mathcal{D}$ de acordo com sua similaridade com D. Para buscar os documentos mais relevantes para um termo de busca, basta considerar que a sequência de termos de busca é o documento D e realizar o procedimento acima.

3.2 Exemplos de Aplicações

Apresentamos a seguir alguns exemplos onde o LSI foi aplicado com sucesso em diversas áreas.

Um primeiro exemplo de aplicação é o trabalho de (LANDAUER; FOLTZ; LAHAM, 1998), avaliando a LSI, em que os autores mostraram que o ângulo entre sinônimo e os antônimos possuem uma média de cosseno entre os vetores 12 vezes maior do que a mesma medida de similaridade entre palavras não relacionadas.

Como uma avaliação da capacidade do modelo de aprender representações adequadas de significado de palavras, (DUMAIS; LANDAUER; LITTMAN, 1997) testaram o quão bem o modelo se comportou em um experimento realizado em um questionário com 80 questões do tipo sinônimos, onde, dada uma palavra de teste, o modelo teve que escolher a resposta mais altamente associada de um grupo de quatro opções. As decisões foram tomadas selecionando a escolha de resposta que apresentou o maior valor de cosseno entre ele e a palavra avaliada.

O algoritmo computacional empregado pela LSI também foi usado em uma série de tarefas de recuperação da informação. Para a tarefa de recuperar documentos relevantes com base na pesquisa com termos inseridos pelo usuário, a LSI oferece uma melhoria de 16% em relação aos algoritmos que retornam documentos com base apenas em correspondência de palavras com o termo de pesquisa (DUMAIS et al., 1994). A LSI também foi utilizada com sucesso para recomendar artigos relevantes para pesquisadores com base em uma seleção de artigos que o pesquisador leu e gostou (FOLTZ; DUMAIS, 1992).

Finalmente, a LSI pode auxiliar na seleção de documentos educacionais que correspondem ao nível de leitura do estudante com o intuito de maximizar a aprendizagem (KINTSCH, 1994).

4 Latent Dirichlet Allocation (LDA)

No capítulo anterior vimos uma técnica de redução de dimensionalidade, o LSI, que alcançou desempenho significativo na classificação de documentos e modelagem de tópicos. O método, baseado em redução de dimensionalidade, tem garantias teóricas de preservar as características mais importantes da base de texto, descartando apenas conteúdo relativamente supérfluo.

Na prática, no entanto, os resultados obtidos indicavam uma *melhora* significativa (e não apenas desempenho similar) de algoritmos de classificação após tal redução de dimensionalidade. Não havia fundamentação teórica para essa melhora de qualidade dos resultados obtidos pelo LSI. Em seu artigo, Papadimitriou, Raghavan, Tamaki e Vempala (PAPADIMITRIOU et al., 1998) propuseram um modelo probabilístico que gerasse corpus (*modelo probabilístico generativo*), com a intenção de estudar sob quais condições era possível garantir teoricamente que o LSI tivesse o bom desempenho observado na prática. Esse modelo generativo inspirou o uso de técnicas Bayesianas rigorosas na área de modelagem de tópicos, e nesse capítulo iremos estudar uma das mais bem-sucedidas técnicas dessa categoria.

Neste capítulo, descrevemos o modelo Latent Dirichlet Allocation (LDA), um modelo probabilístico generativo. O modelo pode ser aplicado para redução de dimensionalidade e classificação de documentos, com fundamentação estatística rigorosa e modelagem probabilística feita especificamente para o problema de análise de corpus.

Esse modelo foi apresentado em um artigo na revista Journal of Machine Learning Research por Blei, Ng e Jordan (BLEI; NG; JORDAN, 2003) e alcançou um grande sucesso tendo, até o momento, mais de 17.750 citações. As áreas de pesquisa em que o método está sendo usado incluem mineração de dados (HAN; PEI; KAMBER, 2011), aprendizado semi-supervisionado (ZHU, 2006), e análise de sentimento (PANG; LEE et al., 2008). Notáveis aplicações do método são categorização de cenas em fotos (LAZEBNIK; SCHMID; PONCE, 2006) e classificação de artigos científicos por tópicos (GRIFFITHS; STEYVERS, 2004), dois dentre muitos exemplos.

O LDA mostra como técnicas de inferência Bayesiana podem ser úteis em diversos domínios, especialmente aqueles que envolvem vários níveis de estrutura. No caso estudado, o corpus é um conjunto de documentos, onde cada documento é uma mistura de tópicos e cada tópico é uma mistura de palavras, e o LDA gera e classifica documentos e tópicos respeitando tal estrutura. A flexibilidade das técnicas usadas no LDA faz com que ele possa ser incorporado em modelos mais complexos, uma propriedade que não é compartilhada pelo LSI.

4.1 Definições

Como já vimos nos Capítulos 2 e 3, utilizamos modelos baseados em sacos de palavras, onde a ordem das palavras no documento pode ser desconsiderada. Na linguagem de teoria da probabilidade esta é uma suposição de permutabilidade das palavras no documento (ALDOUS, 1985), algo menos restritivo que supor que as palavras são independentes e identicamente distribuídas. Supor permutabilidade permite construir métodos computacionalmente eficientes. Ao considerar representações permutáveis para documentos e palavras são necessários modelos que capturam a permutabilidade de ambos, palavras e documentos, que é a proposta do LDA (BLEI; NG; JORDAN, 2003).

O modelo Latent Dirichlet Allocation (LDA) leva tudo isso em consideração da seguinte maneira:

1. Uma palavra é um elemento do dicionário $\mathcal{W}$ (ver capítulo 2). Para os propósitos do modelo vetorial, cada palavra é um vetor cuja única posição não-nula é a correspondente ao seu índice no dicionário (com valor 1). Assim, as palavras formam uma base do espaço vetorial com dimensão n. Formalmente, a i-ésima palavra do dicionário é associada ao vetor v que satisfaz

$$v_j = \begin{cases} 1 & \text{se } i = j \\ 0 & \text{caso contrário.} \end{cases}$$

2. Um documento é uma sequência de palavras denotado por $\mathbf{d} = (w_0, w_1, \ldots, w_{\ell_d})$.

3. Um corpus é uma coleção de m documentos denotado por $\mathcal{D} = (\mathbf{d}_0, \mathbf{d}_1, \ldots, \mathbf{d}_{m-1})$.

O LDA é baseado em inferência estatística. Especificamente, assumiremos que o corpus de entrada tem uma distribuição subjacente, e o objetivo do método é descobrir (inferir) os parâmetros dessa distribuição de modo que o modelo generativo consiga produzir documentos similares e estimar a probabilidade de um documento ter sido gerado através desse modelo.

Assim, o LDA não apenas designa alta probabilidade aos membros do corpus, mas também designa alta probabilidade a outros documentos "similares". A ideia básica é que documentos são representados como misturas aleatórias de tópicos latentes (desconhecidos), onde cada tópico é caracterizado por uma distribuição de probabilidades sobre as palavras. Isso é mais geral do que considerar que as palavras são independentes e identicamente distribuídas (iid), como comentado acima.

4.2 Modelo

O LDA considera que cada documento do corpus $\mathcal{D}$ foi gerado através do seguinte processo. As definições das distribuições de probabilidade das variáveis aleatórias utilizadas abaixo (distribuição de Dirichlet, distribuição multinomial) podem ser encontradas no glossário.

1. Escolha $N \sim \text{Poisson}(\xi)$, o número de palavras do documento. Nenhuma característica da distribuição de Poisson será usada na análise, de modo que é possível alterar esse aspecto do processo com facilidade.
2. Escolha $\theta \sim \text{Dirichlet}(\alpha)$, um vetor em que a i-ésima coordenada corresponde à densidade de ocorrência do i-ésimo tópico no documento.
3. Para cada uma das N palavras w_n:
   a) Escolha um tópico $z_n \sim \text{Multinomial}(\theta)$.
   b) Escolha uma palavra w_n de $p(w_n|z_n, \beta)$ uma distribuição multinomial condicionada no tópico z_n.

A dimensionalidade k da distribuição de Dirichlet corresponde ao número de tópicos, e é um parâmetro escolhido pelo usuário. O objetivo principal do algoritmo é estimar $\alpha \in \mathbb{R}^k$, vetor que rege as densidades dos tópicos nos documentos, e $\beta \in \mathbb{R}^{k \times n}$, matriz com a probabilidade de uma dada palavra pertencer a um dado tópico.

O primeiro passo para o uso do modelo acima é encontrar uma fórmula que, dados os parâmetros α e β, diz a probabildade de um corpus $\mathcal{D}$ ser gerado a partir do procedimento acima. Com ela em mãos, investigaremos o procedimento de encontrar a distribuição dos valores de α e β que melhor se adapta a um corpus já existente.

Dados os parâmetros α e β, a distribuição conjunta da mistura de tópicos θ, um conjunto de N tópicos $\mathbf{z}$ e um conjunto de N palavras $\mathbf{w}$ é dado por

$$p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{N} p(z_n|\theta) p(w_n|z_n, \beta), \tag{4.1}$$

onde $p(z_n|\theta)$ é simplesmente θ_i para o único i tal que $z_n^i = 1$. Integrando sobre θ e somando em z, obtemos a distribuição marginal do documento

$$p(\mathbf{w}|\alpha, \beta) = \int p(\theta|\alpha) \prod_{n=1}^{N} p(z_n|\theta) p(w_n|z_n, \beta) d\theta. \tag{4.2}$$

Finalmente, tomando o produto das probabilidades marginais de um único documento, obtemos a probabilidade de um corpus da seguinte maneira:

$$p(\mathbf{D}|\alpha, \beta) = \prod_{d=1}^{M} \int p(\theta_{\mathbf{d}}|\alpha) (\prod_{n=1}^{N} p(z_{dn}|\theta_{\mathbf{d}}) p(w_{dn}|z_{dn}, \beta) d\theta_{\mathbf{d}}. \tag{4.3}$$

Os parâmetros α e β são amostrados exatamente uma vez no processo de geração do corpus. As variáveis θ_d são amostradas uma vez por documento. Finalmente, as variáveis z_{dn} e w_{dn} são amostradas uma vez para cada palavra em cada documento. Portanto, o LDA envolve três níveis. Nesse modelo, os documentos podem ser associados a vários tópicos. O LDA postula que cada palavra dos documentos observados é gerada por um tópico escolhido aleatoriamente, com distribuição dada por um parâmetro θ escolhido aleatoriamente de acordo com o hiperparâmetro α.

Voltamos agora nossa atenção para procedimentos de inferência e estimação dos parâmetros em um modelo LDA. O problema inferencial mais importante que precisamos resolver para usar o LDA é o de computar a distribuição à posteriori das variáveis ocultas, dado um documento:

$$p(\theta, z|\mathbf{w}, \alpha, \beta) = \frac{p(\theta, z, \mathbf{w}|\alpha, \beta)}{p(\mathbf{w}|\alpha)} \tag{4.4}$$

Esta distribuição à posteriori é intratável para fazer inferência exata. Por isso, se faz necessário estimar os parâmetros das distribuições de forma aproximada, através de algoritmos. Há uma grande variedade de algoritmos que podem ser considerados para LDA, incluindo a aproximação de Laplace e Monte Carlo via Cadeias de Markov (JORDAN, 1998). O algoritmo utilizado nesse trabalho é baseado em Monte Carlo via Cadeias de Markov.

Então, dado um corpus de documentos $\mathbf{D} = \{\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_M\}$, queremos encontrar os parâmetros α e β que maximizam a log-verossimilhança marginal dos dados

$$l(\alpha, \beta) = \sum_{d=1}^{M} \log p(\mathbf{w}_d|\alpha, \beta) \tag{4.5}$$

Como descrito acima, a quantidade $p(\mathbf{w}|\alpha, \beta)$ não pode ser computada de modo tratável, mas uma vez que seja encontrada uma boa aproximação, existem algoritmos que permitem encontrar iterativamente estimadores de máxima verossimilhança para os vetores de parâmetros α e β.

4.3 Aplicações do LDA

A LDA foi originalmente proposta para documentos de texto, mas ela pode ser vista como uma metodologia de descoberta de tópicos mais geral. O modelo foi estendido com sucesso a inúmeras aplicações em outros domínios, incluindo: reconhecimento de objetos (CAO; FEI-FEI, 2007; FEI-FEI; PERONA, 2005; FERGUS; PERONA; ZISSERMAN, 2003; RUSSELL et al., 2006; WANG; GRIMSON, 2008), processamento de linguagem natural (BLEI; NG; JORDAN, 2003; BOYD-GRABER; BLEI, 2009), análise de vídeo (NIEBLES; WANG; FEI-FEI, 2008; WANG; MA; GRIMSON, 2007), filtragem colaborativa (MARLIN, 2004), filtragem de spam (BÍRÓ; SZABÓ; BENCZÚR, 2008),

web-mining (MEI et al., 2006), desambiguação de autoria (ROSEN-ZVI et al., 2004) e segmentação de diálogo (PURVER et al., 2006).

Uma aplicação na área de reconhecimento de imagens foi feita por Sivic et al (SIVIC et al., 2005), que aplicou a LDA da maneira mais direta à analise de imagens. Nesse artigo, as imagens desempenham o papel dos documentos, as palavras-chave desempenham o papel de palavras e as categorias de objeto desempenham o papel de tópicos. O modelo não supervisionado baseado em LDA utilizado por Sivic mostrou desempenho competitivo de categorização comparado a algoritmos que utilizavam 400 imagens marcadas à mão para cada classe de treinamento, sem utilizar dados rotulados.

A seguir, listamos vantagens da LDA no reconhecimento de imagens. Muitas dessas vantagens se estendem a outros domínios.

1. Representação de baixa dimensão. A LDA representa de forma concisa cada imagem como uma mistura aleatória de tópicos.

2. Não supervisionado. A LDA não requer dados rotulados, podendo ser utilizada quando rotular manualmente não é uma opção devido ao volume total.

3. Representação de tópicos intuitiva. A LDA descobre um conjunto de tópicos latentes de toda a coleção de imagens que é expressa como uma distribuição por meio de palavras-chave. Esses tópicos têm um apelo intuitivo e, em geral, correspondem a objetos reais no conjunto de treinamento. Isso facilita a compreensão do que o modelo aprendeu, além de resumir o conteúdo da coleção de imagens como um todo.

4. Categorização e segmentação simultânea. A LDA produz uma representação de mistura de tópicos para cada imagem, bem como cada componente de imagem dentro de uma imagem. Isso permite que a categorização seja feita no nível da imagem e a segmentação multi-objeto seja feita no nível do componente da imagem.

5. Leva em consideração a polissemia visual, a mesma palavra-chave encontrada em dois diferentes contextos será diferenciada.

4.4 Conclusão

Nesse capítulo, motivamos o uso do LDA e explicamos como ele funciona. Depois iremos utilizar o LDA como motor de uma ferramenta para encontrar documentos diplomáticos relevantes sobre um determinado tópico. Isso é possível porque as técnicas estatísticas do LDA permitem inferir, a partir dos documentos, tópicos relevantes que sumarizam os textos do corpus.

Parte II

O Trabalho

5 Base de documentos

A base de documentos para a qual será construída um sistema de busca é o *Cablegate*. São mais de 250 mil documentos diplomáticos que foram enviados entre embaixadas dos Estados Unidos, todos em língua inglesa. Esses documentos foram publicados como parte do projeto *WikiLeaks* (WIKILEAKS, 2017b).

5.1 História

Cablegate é o nome pelo qual ficou conhecido o conjunto de cabos diplomáticos dos Estados Unidos que foram vazados em 2010. O conjunto contém 251.287 documentos entre dezembro de 1966 e fevereiro de 2010. Quando lançada foi a maior base de documentos confidenciais publicada até então.

Esses documentos foram tornados públicos como parte do projeto *WikiLeaks* (WIKILEAKS, 2017b). O projeto *WikiLeaks* é uma organização jornalística internacional dedicada a publicar anonimamente documentos confidenciais.

5.2 Visão geral

Nessa seção faremos uma análise geral dos metadados da base de documentos para estabelecer um ponto de início para a sua análise.

É importante levar em conta que a base não representa todos os documentos diplomáticos entre embaixadas americanas, e que se trata de uma amostra enviesada pela sua forma de obtenção. Não podemos então tirar conclusões gerais sobre as comunicações entre embaixadas americanas a partir da análise desse banco de dados, mas podemos usar essas informações para saber o que está disponível no banco de dados.

Primeiramente observamos as datas dos documentos. A base contém cabos diplomáticos de 1963 a 2010, e podemos ver claramente na Figura 1 ilustra que a base está concentrada entre 2003 e 2010.

Cada documento da base tem uma confidencialidade associada, a distribuição das classificações dos documentos pode ser vista na Figura 2.

Na figura, vemos 3 tipos de confidencialidades diferentes, além de 2 variações sobre estes. As definições são baseadas em quanto risco à segurança nacional dos EUA seriam causados pela divulgação dos documentos, de acordo com o *National Industrial Security Program Operating Manual* (NISPOM, 2017) eles significam:

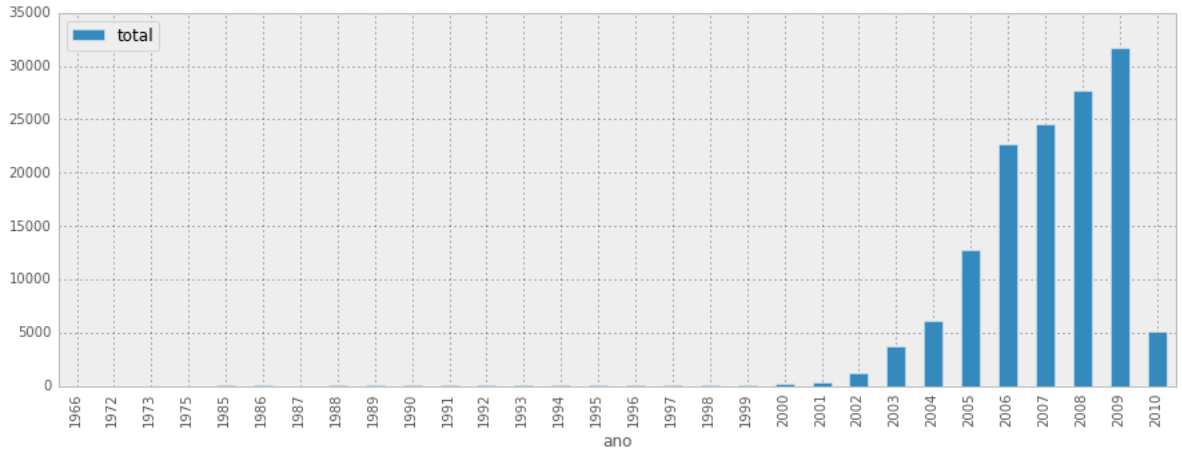


Figura 1 – Distribuição dos documentos ao longo dos anos

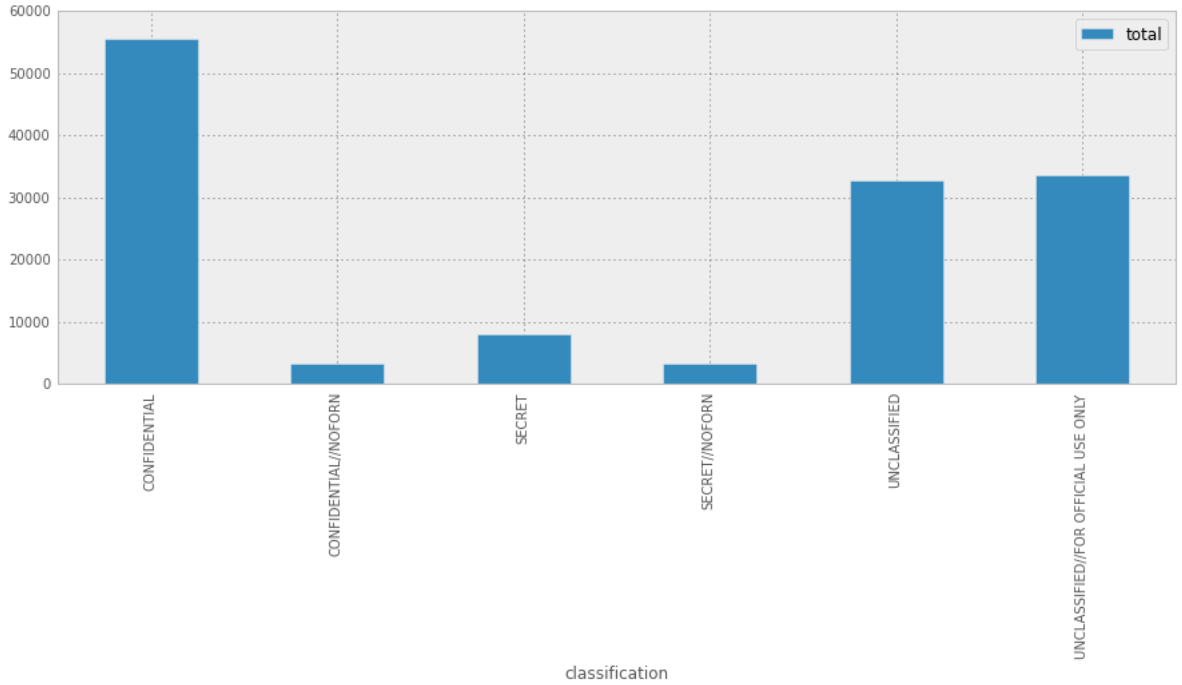


Figura 2 – Confidencialidade dos documentos

Top Secret

Top Secret é o nível mais alto de classificação, e significa que a divulgação de tal documento causaria danos excepcionalmente grandes à segurança dos EUA. A base de dados não contém nenhum documento Top Secret.

Secret

Um documento é classificado como *Secret* quando sua divulgação causaria sérios danos à segurança nacional dos EUA.

Confidential

Confidential significa que sua divulgação causaria algum dano ou seria prejudicial à segurança nacional norte-americana.

Unclassified

Unclassified não é uma classificação no sentido estrito, mas é usado no sistema de classificação para descrever documentos que não chegam a uma classificação específica ou que já foram desconfidencializados. Em geral esses documentos são de baixo impacto.

NOFORN

NOFORN é uma sigla para *No foreign*, que significa que o documento só pode ser divulgado para cidadãos dos EUA.

For official use only

É um termo técnico definido pelas leis de liberação de informação dos EUA. Documentos com essas denominação podem estar isentos de ser divulgados pela lei.

5.3 Distribuição geográfica

No mapa abaixo as linhas marcam embaixadas de origem e de destino de documentos, e a grossura de uma linha representa a quantidade de documentos entre os pontos. Foi construído com o programa no Anexo C.

5.4 Importância

Os documentos contêm análises de diplomatas quanto ao país que estão, seus líderes e sua atual situação política. Essas análises nunca foram publicadas antes e trazem à tona não só o ponto de vista dos Estados Unidos como fatos que não haviam sido divulgados antes. Com isso, essa base de dados pode ser utilizada por cientistas políticos brasileiros para complementar a análise de vários acontecimentos da década de 2000, principalmente os que ocorreram entre 2003 e 2009. O objetivo deste trabalho é auxiliar essa análise.

Um exemplo da importância dessa base de dados é ela ter sido considerada um dos catalizadores da revolução de 2010-2011 na Tunísia (ALJAZEERA, 2011), por ter revelado excessos do então presidente.

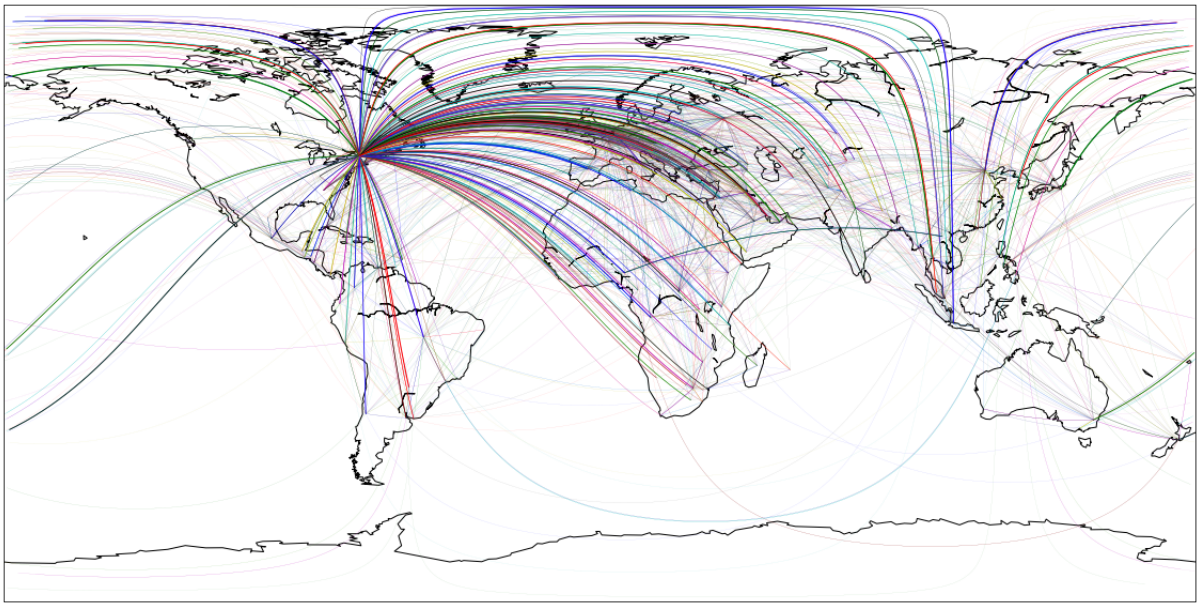


Figura 3 – Entre que cidades os documentos eram enviados

6 Processamento dos dados

Neste capítulo iremos descrever todo o processo realizado para criar uma ferramenta de busca de documentos diplomáticos a partir de um *corpus* de documentos.

6.1 Pré-processamento

Os documentos foram distribuídos na forma de uma tabela em PostgreSQL (GROUP, 2017) pelo site do WikiLeaks (WIKILEAKS, 2017b). O primeiro passo foi converter o banco de dados PostgreSQL num banco de dados SQLite, como modo conveniente de distribuir o banco de dados.

Após isso, o procedimento inicial de tratamento dos textos foi o seguinte:

1. Os documentos são tokenizados (convertidos em uma sequência de palavras através do tratamento de pontuação e espaçamento), e é feito *stemming* em cada um dos mesmos.

2. O dicionário é construído. Para isso, fazemos uma passada nos documentos, e o *Gensim* monta um dicionário do Python com um identificador numérico para cada palavra que aparece no corpus.

3. Elimina-se do dicionário palavras que aparecem menos de 5 vezes ou em mais de 50% dos textos. Essas palavras correspondem a palavras que não podem ser aprendidas devido a um baixo número de amostras, ou a *stopwords*, respectivamente.

4. Uma segunda passada é feita para converter os documentos em uma sequência de códigos numéricos (sacos de palavras) correspondentes às palavras que restaram no dicionário.

Ao final desse processo tempos uma matriz de documentos, que chamamos de *corpus*. O código para essa primeira parte, juntamente com lógica para salvar tais dados no disco e recuperá-los posteriormente, está no Anexo D.1.

6.2 Ferramentas

Neste trabalho utilizamos fortemente algumas ferramentas de *Software Livre*. Descrevemos abaixo as ferramentas de aprendizagem por máquinas que utilizamos.

6.2.1 Gensim

Gensim ("generate similar", objetivo da versão inicial do projeto) é uma ferramenta mantida por Radim Řehůřek. É uma implementação em Python de vários algoritmos de aprendizagem por máquinas, como LDA e LSI (ŘEHŮŘEK; SOJKA, 2010).

6.2.2 Mallet

Mallet (MCCALLUM, 2002) é uma das ferramentas mais estabelecidas de processamento natural de linguagens, incluindo modelagem de tópicos e classificação de documentos. Feita por Andrew McCallum e estudantes da Universidade de Massachusetts em Amherst, o Mallet foi pioneiro em implementações robustas do LDA (WALLACH; MIMNO; MCCALLUM, 2009).

6.3 Processamento

O processamento para a criação da ferramenta de busca é diferente para o *LSI* e para o *LDA*. Nessa seção detalharemos o processamento para as duas técnicas separadamente, bem como o processo utilizado para definir a quantidade de tópicos.

6.3.1 LSI

Como vimos no Capítulo 3, o LSI é uma aplicação da Decomposição em Valores Singulares à matriz de documentos. Dada a quantidade de documentos e termos, é importante explorar a esparsidade da matriz para obter algoritmos eficientes.

Como utilizamos a biblioteca *Gensim*, o algoritmo utilizado para esse propósito foi um algoritmo iterativo e randomizado de autoria de Halko, Martinsson e Tropp (HALKO; MARTINSSON; TROPP, 2009).

6.3.2 LDA

O algoritmo do LDA é muito mais sofisticado em termos matemáticos e estatísticos. Sendo assim, optamos por uma implementação robusta e estabelecida, a fornecida pelo projeto Mallet (MCCALLUM, 2002), que existe desde 2003 e fornece implementações em Java de vários algoritmos de processamento de linguagem natural.

Vários aperfeiçoamentos de engenharia implementados pelos autores melhoram significativamente o desempenho estatístico do algoritmo, como uma otimização judiciosa dos hiperparâmetros (WALLACH; MIMNO; MCCALLUM, 2009). No nosso caso, o treinamento ocorreu em 1.000 iterações, e os hiperparâmetros eram otimizados a cada 10 delas.

O Gensim oferece uma interface para chamar os algoritmos do Mallet através de sua própria API. Utilizamos tal interface para unificar o código das duas técnicas.

6.3.3 Definindo a quantidade de tópicos

Treinamos o modelo do LDA e LSI com várias escolhas de número de tópicos, entre $k = 50$ e $k = 300$. Com o valor de k próximo de 50, vários termos que gostaríamos de diferenciar eram considerados equivalentes, por exemplo, "Honduras" e "Cuba". Isso porque esses termos aparecem em vários documentos em comum, e com um número pequeno de tópicos não há informação o suficiente para diferenciá-los. Com valores de k acima de 200 sinônimos já não eram aprendidos tão bem, e a busca passava a depender muito do termo exato, sem extrapolar a partir do termo original. Portanto, empiricamente, a escolha $k = 100$ forneceu os melhores resultados, e esse número é o padrão oferecido na interface. Através do argumento numTopicos, é possível alterar o valor de k para 300, um valor que funciona bem para buscas textuais mais exatas, mas não aprende a relação entre "usina" e "energia" por exemplo.

6.4 Consultas

Para as consultas, utilizamos a classe MatrixSimilarity do pacote Gensim. Para usá-la, criamos um documento avulso a partir dos termos de busca, fazemos as mesmas normalizações descritas na Seção 6.1, e aplicamos a transformação dada pelo LSI ou LDA para obter uma representação vetorial v dos termos de busca no espaço vetorial correspondente.

A partir daí, os documentos do corpus são ordenados de acordo com a similaridade com os termos de busca. Mais precisamente, o cosseno de v (o vetor dos termos de busca) com cada um dos vetores correspondentes aos documentos do corpus é calculado, fornecendo a relevância de cada resultado, e os documentos são ordenados em ordem decrescente de cosseno.

6.5 Interface

Para expor os resultados aos usuários, uma ferramenta na web foi feita usando o projeto Flask, que fornece um modo simples de escrever páginas da Web dinâmicas em Python. A interface da ferramenta foi feita utilizando Bootstrap, que permite criar UIs (interfaces de usuário) funcionais com facilidade. A página está disponível em <http://cables.explique.me>, e está sendo servida a partir de um servidor remoto no Digital Ocean.

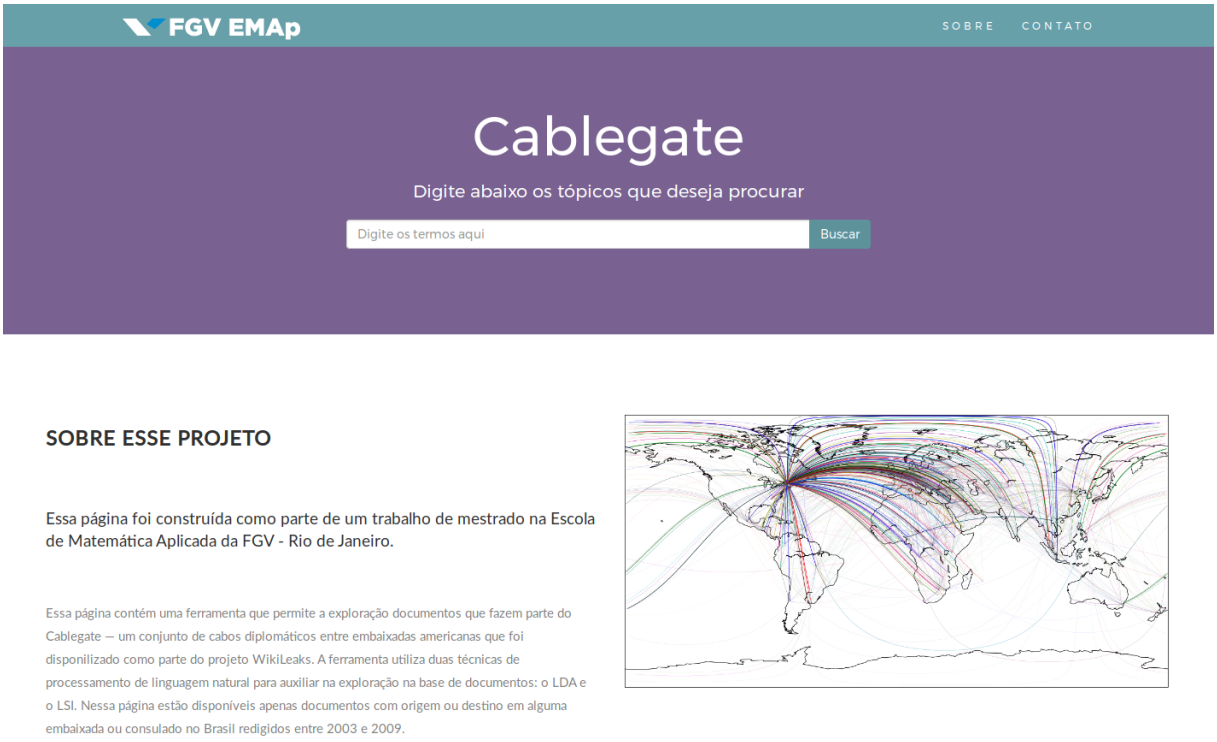


Figura 4 – A página inicial da ferramenta de busca

A interface web consiste em duas páginas. A primeira, página inicial, serve como ponto de entrada, por onde pesquisadores podem digitar quais termos procuram. Ela também contém informações sobre o projeto.

A segunda página é dinamicamente gerada a partir dos termos de busca, que chamamos de página de resultados. Ela contém os melhores documentos encontrados tanto pelo LSI quanto pelo LDA em ordem de relevância, bem como qual a relevância estimada por cada sistema para cada documento e a data do mesmo. Para transmitir ao usuário o grau de confiança que temos em cada resposta utilizamos um sistema de cores muito popular, em que verde significa que confiamos que o documento é relevante, amarelo significa que nossa informação é inconclusiva e vermelho significa que o documento provavelmente não é relevante.

Todos os documentos são clicáveis e ao clicar em um uma janela é aberta com todo o conteúdo do mesmo. Essa janela é um *Lightbox* que foi construída utilizando *Bootstrap*.

De modo a melhorar a experiência do usuários, treinamos em separado um modelo usando apenas os cabos diplomáticos que chegaram ou saíram do Brasil. Isso visa aumentar a relevância dos resultados para os especialistas em relações internacionais brasileiras.

Figura 5 – A página de resultados da ferramenta de busca

Figura 6 – Um documento aberto na ferramenta

Parte III

Parte Final

7 Resultados e Discussão

A utilidade da ferramenta foi estudada de duas formas: primeiro de maneira simples com experimentação direta, utilizando termos que foram mencionados em conversas com pesquisadores de ciências políticas da FGV, e através de um questionário de satisfação enviado diretamente a pesquisadores da fundação.

7.1 Exemplos

Nesta seção iremos observar os resultados obtidos pela ferramenta para alguns termos de busca. Tais termos foram escolhidos para exemplificar o desempenho das técnicas em contextos geopolíticos possivelmente relevantes para os pesquisadores.

7.1.1 Brasil Bolívia

Para a consulta "brazil bolivia" a ferramenta relata ter encontrado 5 documentos relevantes com o LSI enquanto o LDA reporta ter encontrado 6 documentos relevantes. Não há nenhuma interseção entre os documentos considerados relevantes por cada um dos sistemas.

Figura 7 – Resultados para "brazil bolivia"

Dos documentos considerados relevantes pelo LSI, 4 são sobre as relações entre os dois países com respeito a gás e petróleo, e um outro é a respeito da relação entre os dois países no contexto de tráfego de drogas.

Todos os documentos considerados relevantes para o LDA mencionam a Bolívia apenas no contexto de relações com a Venezuela. Isso se deve ao fato de um dos tópicos do LDA conter Bolívia e várias palavras relacionadas à Venezuela. Como existem vários documentos na base em que essa relação existe, o LDA aprendeu a relação entre os dois países.

7.1.2 Petróleo

Os resultados para a consulta "oil" incluem 13 documentos relevantes encontrados pelo LSI e 30 documentos relevantes encontrados pelo LDA. Todos os documentos considerados relevantes pelo LSI também foram considerados relevantes pelo LDA.

Figura 8 – Resultados para "oil"

Todos os resultados são sobre petróleo ou energia, muitos deles mencionando o pré-sal do ponto de vista de empresas americanas. Os 17 documentos que foram considerados relevantes pelo LDA e não para o LSI foram considerados inconclusivos pelo mesmo. Como "oil"é um termo de busca muito genérico e com muitos documentos relacionados, medidas de precisão acabam sofrendo com ruído.

7.1.3 Submarinos nucleares

Na consulta "nuclear submarine" o LSI encontra 11 documentos que considera relevantes, enquanto o LDA encontra 8. Há um documento que é considerado relevante por ambos.

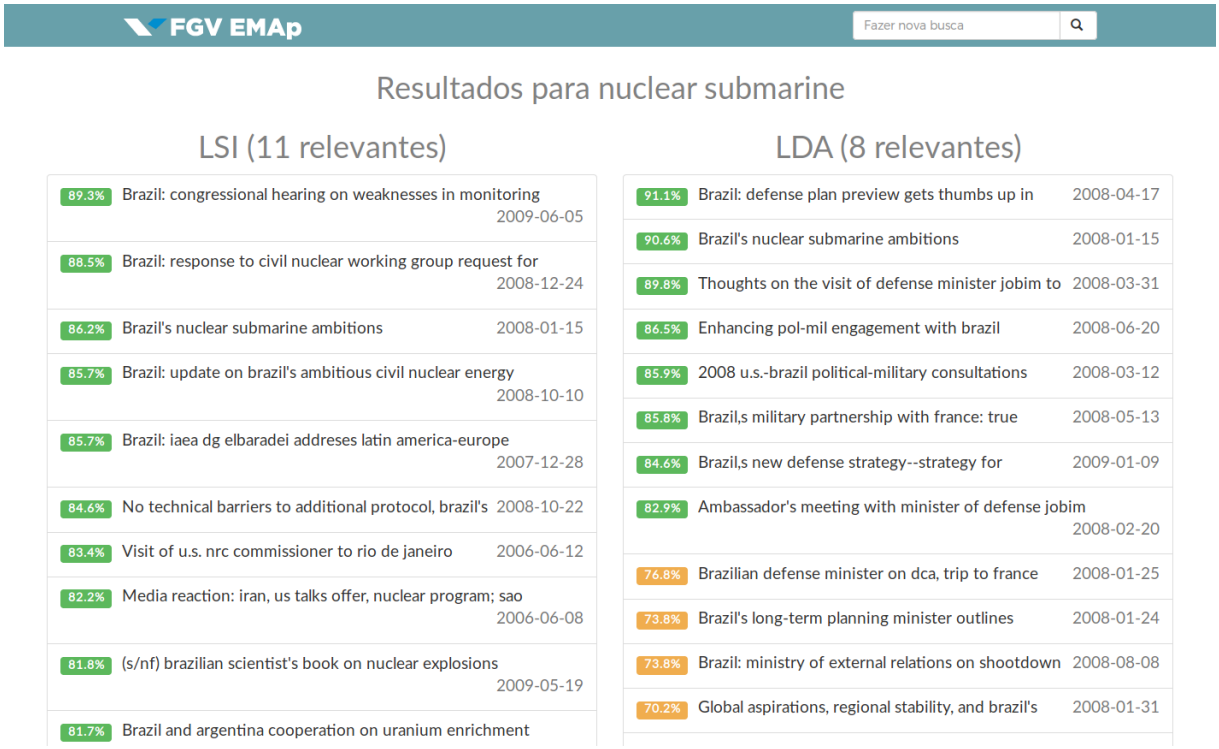


Figura 9 – Resultados para "nuclear submarine"

A resposta do LSI inclui documentos a respeito do programa nuclear civil brasileiro e sobre submarinos nucleares. A resposta do LDA também inclui documentos sobre submarinos nucleares, além de documentos sobre as forças armadas brasileiras.

7.2 Análise

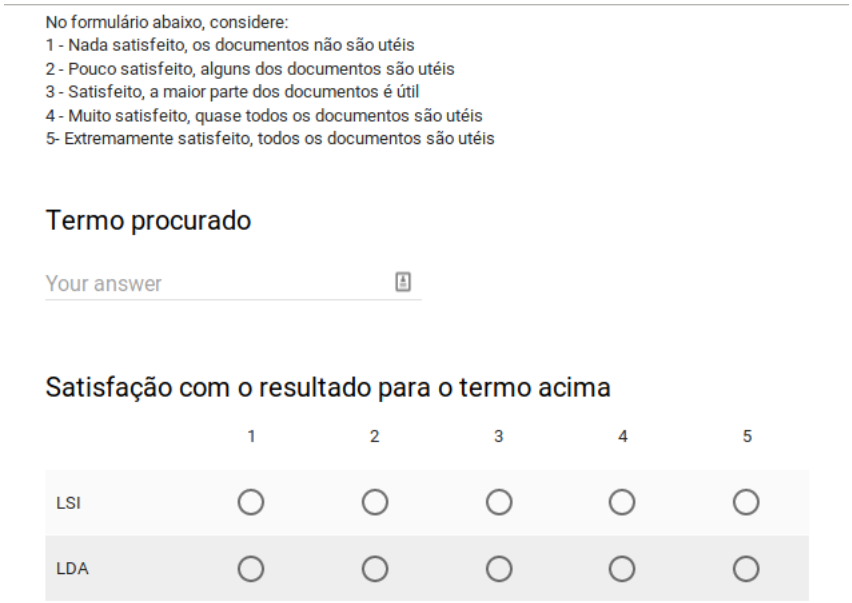
Para estudar a utilidade da ferramenta, um questionário sobre a satisfação do usuário com o algoritmo de recuperação da informação foi construído. Cada usuário foi instruído a realizar várias buscas de documentos em tópicos do seu interesse. Para cada busca o entrevistado deve avaliar a sua satisfação, em uma escala de 1 a 5 (ver Figura 10), com o resultado encontrado pelo algoritmo usando cada uma das técnicas: LSI e LDA.

Os objetivos dessa avaliação são dois: avaliar a satisfação geral do entrevistado com o algoritmo de recuperação de informação e comparar os resultados das duas técnicas entre si, se alguma delas se sairia melhor na opinião do entrevistado.

Devido a especificidade da ferramenta e do banco de dados, que é um banco de dados diplomáticos de embaixadas dos Estados Unidos que foram disponibilizados como

parte do *WikiLeaks*, optamos por uma avaliação qualitativa, feita por *experts* na área, em vez de uma avaliação quantitativa feita por leigos, que não teriam conhecimento para avaliar a qualidade dos resultados. Diante disso, alguns pesquisadores em Ciências Sociais da FGV foram convidados a participar da avaliação e dois deles gentilmente aceitaram responder o questionário, trazendo contribuições muito relevantes para esse trabalho. Esses pesquisadores possuem bacharelado, mestrado e doutorado em História ou Ciências Sociais e são pesquisadores ativos em Ciêmcias Sociais.

Normalmente, resultados desse tipo são avaliados através de um *Golden Set*, um conjunto de documentos ou buscas já avaliados. Como o banco de dados era muito extenso e até onde tenho conhecimento, não há nenhum trabalho prévio que o utilize, isso não foi possível para esse trabalho.


No formulário abaixo, considere:
1 - Nada satisfeito, os documentos não são utéis
2 - Pouco satisfeito, alguns dos documentos são utéis
3 - Satisfeito, a maior parte dos documentos é útil
4 - Muito satisfeito, quase todos os documentos são utéis
5- Extremamente satisfeito, todos os documentos são utéis

Termo procurado

Your answer

Satisfação com o resultado para o termo acima

	1	2	3	4	5
LSI	○	○	○	○	○
LDA	○	○	○	○	○

Figura 10 – Formulário enviado a especialistas em Ciências Politicas

A análise das avaliações feitas pelos cientistas políticos da FGV mostra que a ferramenta foi bem avaliada, em geral, pelos avaliadores com notas acima de 3 para a maioria da buscas. Não pareceu haver um consenso entre eles sobre a prevalência do LSI em relação ao LDA. Um deu notas maiores para o LSI e o outro deu notas maiores para o LDA. É interessante observar que em uma das buscas um dos avaliadores deu notas igualmente ruins (2), para as duas técnicas. No caso, o termo procurado pelo avaliador foi "Itaipu". Acreditamos que os pesquisadores provavelmente gostariam de encontrar resultados específicos sobre a usina Itaipu. No entanto, as técnicas LSI e LDA aprendem que a palavra Itaipu pertence ao tópico “energia”, e não distinguem documentos que falam de Itaipu de documentos que mencionam outros tópicos relacionados a energia.

No entanto, enfatizamos que aprender as relações entre termos é uma grande vantagem do LDA e do LSI e permite ao usuário obter uma visão geral do tópico: Outros

documentos relevantes sobre a situação energética do país provavelmente teriam mais informações relevantes para o usuário e viriam à tona que poderiam passar.

Um outro exemplo da mesma característica foi o termo de busca "Rio tourism", testado por um dos nossos entrevistados. Nele, o LSI teve desempenho melhor que o LDA, de acordo com o avaliador que deu nota 1 para o LDA e 3 para o LSI. Isso ocorreu porque, com a escolha de $k = 100$ tópicos, o LDA aglomerou a palavra "Rio" com os demais estados do Brasil num único tópico. Utilizando $k = 300$ tópicos, não ocorreu tal fenômeno e o LDA obteve resultados muito similares ao LSI.

Para ilustrar a situação, mostramos abaixo uma seleção dos tópicos encontrados pelo LDA. Cada linha representa um tópico. A lista completa de tópicos gerados pelo LDA e pelo LSI estão disponíveis no Anexos B e A.

research	technolog	scienc	innov	cooper	institut
0.051	0.046	0.039	0.029	0.021	0.020
anatel	internet	technolog	digit	standard	servic
0.033	0.023	0.022	0.021	0.020	0.020
environment	mercuri	environ	region	water	mine
0.053	0.035	0.032	0.019	0.018	0.018
trade	negoti	mercosul	ftaa	agreement	agricultur
0.064	0.059	0.029	0.028	0.025	0.024
gas	bolivia	petrobra	bolivian	natur	suppli
0.085	0.063	0.050	0.043	0.037	0.022
day	work	week	month	call	onli
0.016	0.012	0.011	0.011	0.009	0.009
china	chines	export	trade	market	vale
0.092	0.045	0.044	0.038	0.017	0.016

Assim, ao procurar por "ftaa" (Acordo de Livre Comércio das Américas, ALCA), por exemplo, o algoritmo de recuperação da informação também irá retornar documentos que tratem do Mercosul, pois foi identificado que essas palavras pertencem a um tópico comum. Se o usuário estivesse interessado apenas no tópico ALCA, isso é algo indesejado; por outro lado, se o usuário estivesse interessado em explorar informações sobre acordos de livre comércio como o Mercosul, obteria informação valiosa adicional.

No próximo capítulo apresentaremos algumas possíveis soluções para os problemas apontados pela avaliação dos entrevistados.

Conclusões e Trabalhos Futuros

Conclusões

A partir das informações que encontramos na literatura, em várias áreas de aplicação, sobre o sucesso da utilização da LSI e LDA como técnicas de recuperação da informação e também dos bons resultados na avaliação dos entrevistados, concluímos que o algoritmo de recuperação da informação aqui apresentado pode ser bastante útil como ferramenta para pesquisadores da área de Ciência Política.

Porém, os problemas apontados pelos avaliadores mostram uma discrepância entre as expectativas do usuário e o que a ferramenta pode fazer. Para resolver esse problema, existem algumas técnicas possíveis:

- Dar peso parcial à buscas exatas (que incluem a palavra procurada) e não só ao tópico da palavra.
- Explicar, de modo intuitivo, como os algoritmos utilizados pela ferramenta funcionam, de modo a enfatizar os casos de uso mais compatíveis com a técnica de extração de tópicos.
- Expor, na interface web, a funcionalidade (disponível através do argumento numTopicos, ver Subseção 6.5) de selecionar manualmente o número de tópicos a ser usado no LDA, de modo que o usuário possa escolher o nível de generalização apropriado.

Num trabalho futuro, iremos explorar tais técnicas.

Trabalhos Futuros

Além de explorar as técnicas mencionadas na seção anterior, uma outra vertente possível para trabalhos futuros seria lidar com dificuldades que tivemos nesse trabalho. A grande dificuldade desse trabalho foi a avaliação. Não havia nenhum *Golden Set* para a base de dados escolhida, o que impossibilitava avaliação mecânica. Devido ao nível de conhecimento técnico necessário para avaliar a ferramenta, apenas um conjunto muito pequeno de especialistas na área tinha condições de avaliá-la, o que dificultou muito a avaliação humana. Seria interessante reproduzir as mesmas técnicas em um banco de dados já estudado previamente, para o qual já exista um *Golden Set*. Também seria interessante reproduzir as mesmas técnicas com um banco de dados não tão técnico, que possa ser avaliado por um conjunto maior de pessoas.

A ferramenta criada é específica para o banco de dados do *Cablegate*. A criação de uma ferramenta genérica, na qual o usuário apenas entre com um banco de dados e um sistema de busca é criado no *Browser* também pode ser útil para pesquisadores que lidam com vastas bases de documentos.

Referências

ALDOUS, D. J. Exchangeability and related topics. In: *École d'Été de Probabilités de Saint-Flour XIII—1983*. [S.l.]: Springer, 1985. p. 1–198. Citado na página 16.

ALJAZEERA. *AlJazeera*. [S.l.], 2011. Disponível em: <http://www.aljazeera.com/news/africa/2011/01/201114101752467578.html>. Citado na página 25.

BAEZA-YATES, R.; RIBEIRO-NETO, B. et al. *Modern information retrieval*. [S.l.]: ACM press New York, 1999. v. 463. Citado na página 5.

BÍRÓ, I.; SZABÓ, J.; BENCZÚR, A. A. Latent dirichlet allocation in web spam filtering. In: ACM. *Proceedings of the 4th international workshop on Adversarial information retrieval on the web*. [S.l.], 2008. p. 29–32. Citado na página 18.

BLEI, D. M.; NG, A. Y.; JORDAN, M. I. Latent dirichlet allocation. *Journal of Machine Learning Research*, v. 3, n. Jan, p. 993–1022, 2003. Citado 4 vezes nas páginas 1, 15, 16 e 18.

BOYD-GRABER, J. L.; BLEI, D. M. Syntactic topic models. In: *Advances in neural information processing systems*. [S.l.: s.n.], 2009. p. 185–192. Citado na página 18.

CAO, L.; FEI-FEI, L. Spatially coherent latent topic model for concurrent segmentation and classification of objects and scenes. In: IEEE. *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*. [S.l.], 2007. p. 1–8. Citado na página 18.

DEERWESTER, S. Improving information retrieval with latent semantic indexing. 1988. Citado na página 11.

DUMAIS, S.; LANDAUER, T. K.; LITTMAN, M. L. Automatic cross-linguistic information retrieval using latent semantic indexing. In: . [S.l.: s.n.], 1997. p. 16–23. Citado na página 13.

DUMAIS, S. T. et al. Latent semantic indexing (lsi) and trec-2. *Nist Special Publication Sp*, NATIONAL INSTIUTE OF STANDARDS & TECHNOLOGY, p. 105–105, 1994. Citado na página 13.

FEI-FEI, L.; PERONA, P. A bayesian hierarchical model for learning natural scene categories. In: IEEE. *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*. [S.l.], 2005. v. 2, p. 524–531. Citado na página 18.

FERGUS, R.; PERONA, P.; ZISSERMAN, A. Object class recognition by unsupervised scale-invariant learning. In: IEEE. *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*. [S.l.], 2003. v. 2, p. II–II. Citado na página 18.

FOLTZ, P. W.; DUMAIS, S. T. Personalized information delivery: An analysis of information filtering methods. *Communications of the ACM*, ACM, v. 35, n. 12, p. 51–60, 1992. Citado na página 13.

GRAESSER, A. et al. Latent semantic analysis captures casual, goal-oriented, and taxanomic strucutres. In: *Proceedings of the Cognitive Science Society*. [S.l.: s.n.], 2000. v. 1, n. 1. Citado 2 vezes nas páginas 1 e 11.

GRIFFITHS, T. L.; STEYVERS, M. Finding scientific topics. *Proceedings of the National academy of Sciences*, National Acad Sciences, v. 101, n. suppl 1, p. 5228–5235, 2004. Citado na página 15.

GROUP, P. G. D. *PostgreSQL*. [S.l.], 2017. Disponível em: <https://www.postgresql.org/>. Citado na página 27.

HALKO, N.; MARTINSSON, P.-G.; TROPP, J. A. Finding structure with randomness: Stochastic algorithms for constructing approximate matrix decompositions. California Institute of Technology, 2009. Citado 2 vezes nas páginas 12 e 28.

HAN, J.; PEI, J.; KAMBER, M. *Data mining: concepts and techniques*. [S.l.]: Elsevier, 2011. Citado na página 15.

HOLMSTROM, J. E. *Indexing and other Library Services*. [S.l.]: Royal Society Scientific Information Conference, 1948. Citado na página 5.

JORDAN, M. I. *Learning in graphical models*. [S.l.]: Springer Science & Business Media, 1998. v. 89. Citado na página 18.

KINTSCH, W. Text comprehension, memory, and learning. *American psychologist*, American Psychological Association, v. 49, n. 4, p. 294, 1994. Citado na página 13.

LANDAUER, T. K.; FOLTZ, P. W.; LAHAM, D. An introduction to latent semantic analysis. *Discourse processes*, Taylor & Francis, v. 25, n. 2-3, p. 259–284, 1998. Citado 2 vezes nas páginas 12 e 13.

LAZEBNIK, S.; SCHMID, C.; PONCE, J. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: IEEE. *Computer vision and pattern recognition, 2006 IEEE computer society conference on*. [S.l.], 2006. v. 2, p. 2169–2178. Citado na página 15.

MARLIN, B. M. Modeling user rating profiles for collaborative filtering. In: *Advances in neural information processing systems*. [S.l.: s.n.], 2004. p. 627–634. Citado na página 18.

MARTIN, D. I.; BERRY, M. W. Mathematical foundations behind latent semantic analysis. *Handbook of latent semantic analysis*, Mahwah, NJ: Lawrence Erlbaum Associates, p. 35–56, 2007. Citado na página 11.

MCCALLUM, A. K. Mallet: A machine learning for language toolkit. Http://mallet.cs.umass.edu. 2002. Citado na página 28.

MEI, Q. et al. A probabilistic approach to spatiotemporal theme pattern mining on weblogs. In: ACM. *Proceedings of the 15th international conference on World Wide Web*. [S.l.], 2006. p. 533–542. Citado na página 19.

NIEBLES, J. C.; WANG, H.; FEI-FEI, L. Unsupervised learning of human action categories using spatial-temporal words. *International journal of computer vision*, Springer, v. 79, n. 3, p. 299–318, 2008. Citado na página 18.

NISPOM. *NISPOM*. [S.l.], 2017. Disponível em: <https://web.archive.org/web/20110807094949/http://www.dss.mil:80/isp/odaa/nispom06.html>. Citado na página 23.

PANG, B.; LEE, L. et al. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, Now Publishers, Inc., v. 2, n. 1–2, p. 1–135, 2008. Citado na página 15.

PAPADIMITRIOU, C. H. et al. Latent semantic indexing: A probabilistic analysis. In: ACM. *Proceedings of the seventeenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems*. [S.l.], 1998. p. 159–168. Citado na página 15.

PURVER, M. et al. Unsupervised topic modelling for multi-party spoken discourse. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*. [S.l.], 2006. p. 17–24. Citado na página 19.

ŘEHŮŘEK, R.; SOJKA, P. Software Framework for Topic Modelling with Large Corpora. In: *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Valletta, Malta: ELRA, 2010. p. 45–50. <http://is.muni.cz/publication/884893/en>. Citado na página 28.

RIJSBERGEN, C. K. van. Getting into information retrieval. In: *Lectures on information retrieval*. [S.l.]: Springer, 2000. p. 1–20. Citado na página 5.

ROSEN-ZVI, M. et al. The author-topic model for authors and documents. In: AUAI PRESS. *Proceedings of the 20th conference on Uncertainty in artificial intelligence*. [S.l.], 2004. p. 487–494. Citado na página 19.

RUSSELL, B. C. et al. Using multiple segmentations to discover objects and their extent in image collections. In: IEEE. *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*. [S.l.], 2006. v. 2, p. 1605–1614. Citado na página 18.

SIVIC, J. et al. Discovering objects and their location in images. In: IEEE. *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*. [S.l.], 2005. v. 1, p. 370–377. Citado na página 19.

WALLACH, H. M.; MIMNO, D. M.; MCCALLUM, A. Rethinking lda: Why priors matter. In: *Advances in neural information processing systems*. [S.l.: s.n.], 2009. p. 1973–1981. Citado na página 28.

WANG, X.; GRIMSON, E. Spatial latent dirichlet allocation. In: *Advances in neural information processing systems*. [S.l.: s.n.], 2008. p. 1577–1584. Citado na página 18.

WANG, X.; MA, X.; GRIMSON, E. Unsupervised activity perception by hierarchical bayesian models. In: IEEE. *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*. [S.l.], 2007. p. 1–8. Citado na página 18.

WIKILEAKS. *Cablegate*. [S.l.], 2017. Disponível em: <https://wikileaks.org/plusd/pressrelease/>. Citado na página 2.

WIKILEAKS. *WikiLeaks*. [S.l.], 2017. Disponível em: <https://wikileaks.org/>. Citado 3 vezes nas páginas 2, 23 e 27.

ZHU, X. Semi-supervised learning literature survey. *Computer Science, University of Wisconsin-Madison*, v. 2, n. 3, p. 4, 2006. Citado na página 15.

Glossário

análise fatorial É uma técnica para se reduzir o número de variáveis de uma base de dados, identificando o padrão de correlações ou de covariância entre elas e gerando um número menor de novas variáveis latentes, não observadas, calculadas a partir dos dados brutos. 11

autovetor Um escalar λ é valor próprio (ou autovalor) de um operador linear $A\colon V \to V$ se existir um vetor x diferente de zero tal que $A\mathbf{x} = \lambda\mathbf{x}$. Qualquer vetor x satisfazendo tal igualdade é dito ser um autovetor de A. 12

coeficiente de correlação linear de Pearson Ele quantifica a força de associação linear entre duas variáveis e, portanto, descreve quão bem uma linha reta se ajustaria através de nuvem de pontos. Se os pontos caem exatamente sobre uma linha crescente então $r = 1$, e se eles caem exatamente sobre uma linha decrescente, $r = -1$. Calcula-se o coeficiente de correlação de Pearson segundo a seguinte fórmula:

$$\rho = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{cov(X,Y)}{\sqrt{var(X) \cdot var(Y)}}$$

. 13

distribuição conjunta Considere duas variáveis aleatórias d X e Y com fdp ou fp $p_X(x)$ e $p_Y(y)$ oriundas dos espaços amostrais Ω_X e Ω_Y . Seja Ω_{XY} o espaço amostral de todos os possíveis pares observados (x, y), chamado espaço conjunto de X e Y. A função de densidade de probabilidade (de distribuição) conjunta de X e Y, $p_{X,Y}(x, y)$ é definida por:

$$p_{X,Y}(x, y) = P(X = x, Y = y) \qquad \text{para todo } (x, y) \in \Omega_{XY}.$$

. 17

distribuição de Dirichlet Uma variável aleatória θ com distribuição de Dirichlet tem como imagem o conjunto

$$\sum_{i=1}^{k} \theta_i = 1 \qquad \text{e} \qquad \theta_i \geq 0 \;\; \forall\, 1 \leq i \leq k,$$

tendo função de densidade de probabilidade dada por

$$p(\theta|\alpha) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\Pi_{i=1}^{k}\Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1}, \ldots, \theta_k^{\alpha_k - 1},$$

onde o parâmetro α é um vetor de dimensão k com componentes $\alpha_i > 0$, e $\Gamma(x)$ é a função Gama. 17

distribuição à posteriori A distribuição da probabilidade a posteriori de uma variável aleatória, dado o valor de outra, pode ser calculada com o teorema de Bayes, multiplicando-se a distribuição da probabilidade a priori pela função de verossimilhança, e depois dividindo-a pela constante de normalização, como se segue:

$$f_{X|Y=y}(x) = \frac{f_X(x)L_{X|Y=y}(x)}{\int_{-\infty}^{\infty} f_X(x)L_{X|Y=y}(x)\,dx}$$

. 18

distribução multinomial A distribuição multinomial com parâmetros n e $(\theta_1, \ldots, \theta_k)$ é uma distribuição discreta (seus valores são vetores de inteiros com k coordenadas) dada por

$$\mathbb{P}(X_1 = x_1, \ldots, X_k = x_k) = \frac{n!}{x_1! \cdots x_n!} \pi_{i=1}^{k} \theta_i^{x_i}$$

se $\sum_{i=1}^{k} x_i = n$ e 0 caso contrário. 17

espaço amostral Conjunto de resultados possíveis de um experimento aleatório $\mathcal{E}$. Frequentemente denotado por Ω. 50

estimador de máxima verossimilhança O estimador de máxima verossimilhança de θ é o valor que maximiza a função de verossimilhança $L(\theta; x_1, \ldots, x_n)$ e é um estimador com propriedades ótimas para famílias importantes de distribuições de probabilidade.. 18

experimento aleatório constituem situações onde os acontecimentos possuem variabilidade de ocorrência, isto é, o mesmo experimento pode ter vários resultados diferentes quando repetido sob as mesmas condições. Frequentemente denotado por $\mathcal{E}$. 48

função de densidade de probabilidade Uma função $f\colon \mathbb{R} \to [0, +\infty)$ é denominada função densidade de probabilidade se satisfaz às seguintes propriedades:

$$f(x) \geq 0 \text{ para todo } x \in \mathbb{R} \qquad \text{e} \qquad \int_{-\infty}^{\infty} f(x)dx = 1$$

Além disso, definimos para qualquer $c, d \in \mathbb{R}_x$, com $c < d$, a probabilidade do evento $c < X < d$, como

$$\mathbb{P}(c < X < d) = \int_c^d f(x)dx.$$

Se a função depender de algum valor $\alpha \in \mathbb{R}^K$, dizemos que α é um parâmetro da função. Se a variável aleatória X for discreta em vez da integral temos a soma e a função passa a ser chamada de função de distribuição.. 47

função Gama A função Gama (representada pela letra maiúscula grega Γ) é uma extensão da função fatorial para o conjunto dos números reais e complexos, com o argumento subtraído em 1. Se n é um inteiro positivo define-se da seguinte forma:

$$\Gamma(n+1) = n!$$

. 47

identicamente distribuídas Duas variáveis aleatórias X e Y são iguais em distribuição se $\mathbb{P}(X \leq x) = \mathbb{P}(Y \leq x)$ para todo x. 49

independentes e identicamente distribuídas (iid) ver independência e identicamente distribuídas. 16

independência Duas variáveis aleatórias X e Y são independentes se e só se satisfazem a seguinte propriedade:

$$\mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x)\mathbb{P}(Y = y),$$

ou seja, se a probabilidade de ocorrência simultânea de X e Y for igual ao produto das respectivas probabilidades individuais. . 49

inferência Bayesiana A inferência Bayesiana deriva a probabilidade posterior (*à posteriori*) de um evento a partir da combinação, através da Regra de Bayes. da informação dada pela probabilidade anterior (*à priori*) do evento e da informação dada pela amostra (*verossimilhança*), através de um modelo de probabilidade para os dados observados. 15

inferência estatística Tem como objetivo estudar generalizações sobre uma população através de evidências dadas pela amostra. 16

log-verossimilhança Consideremos uma população e uma variável aleatória X, relacionada a essa população, com função de probabilidade (se X é uma variável aleatória discreta) ou função densidade de probabilidade (se X é uma variável aleatória contínua) $f(x, \theta)$, sendo θ o parâmetro desconhecido. Retiremos uma amostra aleatória simples de X, de tamanho n, $X_1, \ldots, X_n$, e sejam $x_1, \ldots, x_n$ os valores efetivamente observados. A função de verossimilhança L é definida por

$$L(\theta; x_1, \ldots, x_n) = f(x_1; \theta) \times \ldots \times f(x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta).$$

que deve ser interpretada como uma função de θ. A função de log-verossimilhança é simplesmente o logaritmo, em geral, na base e, da função de verossimilhança.. 18

modelo probabilístico Um modelo probabilístico de um fenômeno engloba o experimento aleatório envolvendo o fenômeno em estudo, seu espaço amostral e os eventos definidos nesse espaço, e uma medida de probabilidade associada a esses eventos. 15

permutabilidade Um conjunto finito de variáveis aleatórias $X_1, \ldots, X_n$ é permutável se

$$f(X_1, \ldots, X_n) = f(X_{\pi(1)}, \ldots, X_{\pi(n)})$$

para qualquer permutação $\pi(1), \ldots, \pi(n)$ do conjunto $\{1, 2, \ldots, n\}$. 16

posto O posto de uma matriz A é o número de linhas não-nulas quando a mesma está escrita na forma reduzida escalonada por linhas ou, equivalentemente, o número de linhas ou colunas linearmente independentes de A. 12

probabilidade Número não negativo atribuído a cada um dos resultados possíveis de um espaço amostral, de modo que a soma seja 1. 16

probabilidades marginais A partir da distribuição conjunta de X e Y definimos as distribuições marginais de cada uma como:

$$p_X(x) = \sum_y p_{X,Y}(x, y)$$

$$p_Y(Y) = \sum_x p_{X,Y}(x, y)$$

. 17

técnicas Bayesianas Referem-se a métodos de probabilidade e estatística baseados no Teorema de Bayes, ou Regra de Bayes, (Thomas Bayes, 1702-61), que iniciou uma área dentro da estatística chamada de *Inferência Bayesiana*, enunciado a seguir:

$$P(H \mid E) = \frac{P(E \mid H) \cdot P(H)}{P(E)}$$

. 15

variável aleatória Uma varíavel aleatória representa as características numéricas de interesse em uma população. 17, 47

Anexos

ANEXO A – Tópicos do LDA

Na tabela abaixo, cada linha representa um tópico latente encontrado pelo LDA. Cada tópico é uma mistura de palavras, e os coeficientes de cada palavra estão listados embaixo de cada uma das palavras. Para fins de visualização, mostramos apenas as seis palavras mais importantes de cada tópico.

Tabela 1 – Tópicos latentes encontrados pelo LDA

energi	electr	generat	power	sector	plant
0.083	0.043	0.027	0.025	0.020	0.020
bank	financ	central	mantega	fund	imf
0.062	0.044	0.035	0.031	0.029	0.024
patent	licens	drug	pharmaceut	health	compulsori
0.053	0.032	0.025	0.025	0.021	0.020
foreign	invest	bank	tax	law	compani
0.031	0.021	0.019	0.018	0.016	0.014
secretari	furlan	trade	commerci	codel	gutierrez
0.037	0.019	0.017	0.017	0.016	0.016
privat	project	infrastructur	invest	sector	public
0.044	0.043	0.041	0.031	0.023	0.021
polit	ani	make	believ	onli	view
0.014	0.011	0.010	0.009	0.009	0.009
lula	trade	sbu	reform	remain	foreign
0.016	0.014	0.012	0.010	0.007	0.007
visit	offic	meet	discuss	repres	trip
0.072	0.021	0.016	0.014	0.014	0.013
human	resolut	poloff	council	vote	mre
0.067	0.041	0.025	0.025	0.022	0.020
lula	polit	administr	polici	reform	elect
0.149	0.031	0.028	0.014	0.014	0.014
cuba	cuban	castro	fidel	regim	trade
0.190	0.056	0.051	0.023	0.017	0.015
pt	parti	elect	lula	candid	deputi
0.109	0.071	0.022	0.021	0.013	0.012
port	governor	sul	grand	santo	parana
0.043	0.025	0.021	0.020	0.019	0.019
militari	defens	armi	forc	fleet	servic

0.077	0.030	0.017	0.016	0.016	0.015
research	technolog	scienc	innov	cooper	institut
0.051	0.046	0.039	0.029	0.021	0.020
anatel	internet	technolog	digit	standard	servic
0.033	0.023	0.022	0.021	0.020	0.020
environment	mercuri	environ	region	water	mine
0.053	0.035	0.032	0.019	0.018	0.018
trade	negoti	mercosul	ftaa	agreement	agricultur
0.064	0.059	0.029	0.028	0.025	0.024
gas	bolivia	petrobra	bolivian	natur	suppli
0.085	0.063	0.050	0.043	0.037	0.022
day	work	week	month	call	onli
0.016	0.012	0.011	0.011	0.009	0.009
china	chines	export	trade	market	vale
0.092	0.045	0.044	0.038	0.017	0.016
post	ministri	request	reftel	inform	respons
0.034	0.033	0.030	0.026	0.025	0.018
summit	mercosul	region	south	argentina	america
0.054	0.047	0.033	0.028	0.028	0.028
mre	varga	ministri	machado	environ	meet
0.059	0.039	0.033	0.029	0.024	0.022
terror	terrorist	activ	financi	money	launder
0.039	0.030	0.025	0.018	0.018	0.018
congress	bill	legisl	senat	law	deputi
0.052	0.049	0.037	0.031	0.026	0.025
colombia	farc	colombian	region	ecuador	dca
0.072	0.033	0.029	0.025	0.024	0.022
serra	psdb	governor	mayor	parti	citi
0.065	0.044	0.042	0.038	0.035	0.026
industri	busi	sector	export	fiesp	product
0.063	0.034	0.030	0.027	0.027	0.023
river	region	transport	project	environment	road
0.025	0.024	0.024	0.021	0.021	0.019
haiti	assist	minustah	effort	mission	afghanistan
0.109	0.025	0.025	0.016	0.014	0.012
increas	larg	problem	system	mani	accord
0.016	0.011	0.011	0.010	0.010	0.009
ethanol	product	sugar	produc	price	agricultur

0.068	0.043	0.027	0.027	0.021	0.020
program	train	cooper	technic	assist	activ
0.049	0.031	0.021	0.018	0.018	0.017
tax	invest	treati	mre	ceo	forum
0.032	0.022	0.022	0.021	0.020	0.017
growth	rate	econom	gdp	interest	economi
0.039	0.028	0.026	0.018	0.018	0.017
number	travel	visa	consul	employe	day
0.048	0.044	0.043	0.039	0.032	0.030
aircraft	transfer	technolog	fighter	boe	french
0.024	0.023	0.022	0.021	0.020	0.017
polic	law	feder	crime	enforc	public
0.045	0.045	0.037	0.024	0.023	0.023
cattl	amazon	acr	cane	peru	rainforest
0.027	0.018	0.017	0.015	0.014	0.014
health	case	flu	ministri	report	diseas
0.071	0.054	0.030	0.022	0.019	0.015
fire	forest	climat	chang	water	peru
0.050	0.037	0.034	0.033	0.029	0.021
polic	pcc	crime	attack	secur	report
0.044	0.023	0.020	0.019	0.018	0.016
souza	train	depart	unit	preto	date
0.046	0.036	0.031	0.030	0.027	0.025
rio	janeiro	favela	polic	citi	secur
0.127	0.059	0.045	0.022	0.018	0.016
ambassador	meet	minist	discuss	sobel	secretari
0.083	0.045	0.028	0.025	0.020	0.017
bank	financi	credit	crisi	percent	econom
0.040	0.029	0.028	0.026	0.019	0.017
bush	chvez	iraq	daili	editori	circul
0.021	0.015	0.014	0.013	0.013	0.013
report	press	media	public	day	made
0.040	0.034	0.028	0.028	0.013	0.012
secur	felix	dhs	abin	system	intellig
0.067	0.040	0.017	0.017	0.016	0.016
drug	polic	traffick	border	feder	narcot
0.051	0.023	0.023	0.017	0.016	0.015
gob	usg	ministri	sbu	meet	concern

0.117	0.021	0.020	0.016	0.012	0.010
agenc	author	requir	regul	approv	inform
0.035	0.028	0.027	0.024	0.017	0.013
court	case	justic	judg	suprem	decis
0.067	0.040	0.029	0.028	0.026	0.024
anac	airport	aviat	airlin	carrier	flight
0.044	0.042	0.038	0.027	0.020	0.019
exchang	market	foreign	debt	real	rate
0.044	0.039	0.027	0.026	0.024	0.023
petrobra	oil	compani	salt	rio	pre
0.070	0.055	0.033	0.023	0.022	0.021
parti	senat	pmdb	deputi	coalit	sbu
0.069	0.042	0.041	0.026	0.021	0.017
obama	reaction	media	crisi	global	editori
0.043	0.023	0.023	0.022	0.019	0.019
foreign	polici	itamarati	barbosa	south	diplomat
0.074	0.061	0.045	0.022	0.019	0.016
biofuel	energi	ethanol	develop	cooper	secretari
0.090	0.071	0.021	0.018	0.016	0.014
develop	particip	meet	group	discuss	confer
0.022	0.018	0.017	0.016	0.016	0.015
alckmin	lula	elect	parti	candid	campaign
0.058	0.044	0.036	0.032	0.025	0.023
usd	million	compani	billion	busi	invest
0.046	0.046	0.043	0.025	0.017	0.016
chavez	venezuela	venezuelan	bolivia	region	moral
0.087	0.072	0.026	0.025	0.024	0.019
muslim	communiti	lebanes	church	lebanon	arab
0.046	0.037	0.035	0.024	0.021	0.017
tax	gob	fiscal	budget	revenu	feder
0.050	0.032	0.020	0.020	0.020	0.019
sourc	chile	environment	energi	argentina	region
0.024	0.015	0.012	0.012	0.009	0.009
minist	palocci	lula	pt	cabinet	ministri
0.054	0.054	0.031	0.024	0.023	0.017
visitor	embassi	servic	post	agenc	icass
0.020	0.017	0.017	0.014	0.013	0.012
jobim	defens	militari	unger	minist	secur

0.078	0.076	0.032	0.024	0.021	0.019
research	genet	approv	project	nih	foreign
0.048	0.031	0.028	0.026	0.020	0.020
sbu	sensit	unclassifi	protect	accord	told
0.167	0.012	0.011	0.011	0.011	0.011
air	control	forc	traffic	procedur	shootdown
0.082	0.047	0.031	0.029	0.019	0.018
post	offic	gov	applic	br	www
0.029	0.025	0.022	0.021	0.019	0.014
nuclear	iaea	mourao	technolog	energi	addit
0.099	0.025	0.022	0.020	0.016	0.016
varig	debt	compani	leas	million	sharehold
0.056	0.037	0.021	0.015	0.013	0.012
iran	iranian	middl	east	israel	visit
0.079	0.030	0.026	0.026	0.023	0.021
afro	racial	black	action	quilombo	communiti
0.054	0.035	0.027	0.025	0.022	0.021
cooper	unit	interest	relat	relationship	region
0.031	0.021	0.018	0.017	0.015	0.014
prison	pcc	inmat	facil	system	human
0.124	0.037	0.023	0.021	0.019	0.017
plan	accord	report	agreement	ministri	announc
0.030	0.023	0.017	0.015	0.014	0.014
da	silva	director	costa	luiz	gome
0.074	0.066	0.032	0.027	0.025	0.023
amorim	garcia	lula	biato	gob	patriota
0.060	0.046	0.028	0.020	0.020	0.018
climat	chang	emiss	develop	deforest	reduct
0.056	0.050	0.032	0.028	0.018	0.017
percent	rate	increas	month	product	industri
0.114	0.027	0.026	0.018	0.016	0.015
labor	traffick	tip	child	victim	forc
0.061	0.051	0.029	0.024	0.022	0.021
space	agreement	embassi	zelaya	russian	russia
0.060	0.040	0.029	0.027	0.026	0.026
amazon	forest	deforest	land	sustain	environment
0.054	0.046	0.030	0.026	0.018	0.017
piraci	ipr	enforc	pirat	good	feder

0.054	0.019	0.018	0.017	0.012	0.012
neve	candid	rousseff	serra	pt	dilma
0.039	0.033	0.032	0.028	0.027	0.026
indigen	indian	land	demarc	communiti	reserv
0.096	0.049	0.044	0.023	0.020	0.020
mst	land	worker	labor	movement	union
0.054	0.038	0.037	0.033	0.027	0.025
africa	develop	south	initi	india	foreign
0.059	0.038	0.034	0.021	0.019	0.018
gob	sbu	biofuel	percent	lula	energi
0.010	0.010	0.009	0.009	0.009	0.008
polic	feder	investig	murder	arrest	case
0.046	0.038	0.031	0.023	0.022	0.021
program	educ	school	student	social	univers
0.060	0.037	0.028	0.022	0.015	0.014
articl	parti	text	paragraph	decre	agreement
0.040	0.026	0.017	0.017	0.016	0.016
pt	scandal	dirceu	cpi	corrupt	investig
0.027	0.023	0.022	0.021	0.020	0.018

ANEXO B – Tópicos do LSI

Na tabela abaixo, cada linha representa um tópico latente encontrado pelo LSI. Cada tópico é uma mistura de palavras, e os coeficientes de cada palavra estão listados embaixo de cada uma das palavras. Para fins de visualização, mostramos apenas as seis palavras mais importantes de cada tópico.

Tabela 2 – Tópicos latentes encontrados pelo LSI

visitor	travel	icass	tdi	servic	embassi
0.228	0.169	0.163	0.120	0.110	0.099
visitor	icass	travel	tdi	medevac	lula
-0.249	-0.187	-0.167	-0.136	-0.108	0.108
pt	parti	alckmin	psdb	serra	pmdb
0.316	0.282	0.242	0.230	0.188	0.180
iran	obama	chavez	editori	percent	reaction
-0.252	-0.186	-0.176	-0.143	0.129	-0.126
nuclear	polic	mourao	growth	mre	iran
-0.147	-0.144	-0.136	0.132	-0.123	-0.121
forest	polic	amazon	iran	mourao	deforest
0.170	0.162	0.155	-0.134	-0.132	0.123
piraci	energi	polic	petrobra	forest	nuclear
-0.228	0.175	-0.169	0.153	0.153	0.151
iran	mercosul	nuclear	trade	ftaa	bank
0.329	-0.263	0.248	-0.170	-0.150	0.148
petrobra	gas	forest	deforest	oil	bolivia
-0.424	-0.233	0.198	0.184	-0.179	-0.162
piraci	ethanol	iran	chavez	alckmin	serra
0.302	0.190	0.185	-0.159	0.157	0.144
nih	research	dr	project	approv	deforest
0.354	0.238	0.153	0.148	0.135	-0.128
piraci	serra	alckmin	iran	mercosul	dirceu
-0.219	0.202	0.183	-0.178	-0.150	-0.141
coaf	intellig	freita	lope	search	list
-0.363	-0.209	-0.201	-0.162	-0.159	-0.156
piraci	prison	mercosul	obama	pcc	nih
-0.301	0.252	0.247	-0.220	0.194	0.193
ethanol	alckmin	biofuel	piraci	chavez	serra

-0.405	0.241	-0.221	0.197	0.172	0.163
nuclear	chvez	anac	mercosul	prison	mourao
-0.186	-0.173	-0.169	-0.146	0.131	-0.127
n	flu	h	chavez	nuclear	ethanol
-0.244	-0.241	-0.235	0.188	0.181	0.167
iran	mourao	prison	pcc	iranian	ahmadinejad
0.349	-0.267	-0.231	-0.182	0.173	0.131
flu	h	n	chavez	case	health
-0.329	-0.316	-0.315	-0.210	-0.156	-0.135
cuba	cuban	chvez	castro	bush	china
0.663	0.261	-0.194	0.174	-0.134	0.103
cuba	chvez	obama	ethanol	anac	nuclear
-0.264	-0.261	0.206	-0.180	-0.170	0.139
labor	traffick	prison	pcc	child	tip
0.331	0.286	-0.271	-0.245	0.174	0.160
chavez	chvez	anac	iran	iraq	bush
-0.304	0.269	-0.216	-0.211	0.171	0.165
deforest	iran	chvez	climat	petrobra	emiss
-0.291	-0.151	-0.139	-0.139	-0.135	-0.133
demarch	labor	energi	chvez	ethanol	bolivian
0.177	0.153	-0.151	-0.146	0.143	0.140
alckmin	anac	ethanol	palocci	china	bolivia
-0.211	-0.173	-0.166	-0.152	0.148	-0.146
china	chines	obama	alckmin	percent	pt
-0.419	-0.310	0.192	0.156	-0.134	-0.134
palocci	alckmin	summit	chvez	chavez	mst
0.198	0.196	-0.182	0.182	-0.180	-0.175
china	chines	nuclear	mercosul	garcia	ethanol
-0.277	-0.201	-0.188	0.162	-0.140	0.133
zfr	summit	labor	energi	felix	dirceu
-0.808	-0.134	-0.116	-0.100	0.096	-0.079
zfr	summit	dirceu	felix	palocci	pmdb
0.497	-0.204	-0.136	0.136	-0.134	0.134
mst	land	prison	palocci	indigen	stang
-0.281	-0.220	0.184	-0.180	-0.167	-0.140
alckmin	serra	chavez	china	pt	neve
0.440	-0.222	0.185	0.147	-0.133	-0.119
energi	serra	mourao	senat	pt	petrobra

-0.207	-0.178	0.168	-0.147	0.141	0.141
human	adorno	ambassador	piraci	chavez	indigen
0.176	0.175	-0.152	0.142	-0.140	0.138
mst	mourao	dirceu	piraci	pmdb	space
0.216	-0.200	0.200	-0.144	-0.141	0.136
summit	iraq	chvez	chavez	favela	alckmin
0.249	-0.206	0.175	-0.153	-0.153	-0.151
obama	chvez	palocci	mourao	labor	afro
-0.276	-0.257	0.169	-0.167	0.149	-0.145
mst	chvez	chavez	summit	palocci	energi
-0.272	-0.210	0.185	-0.148	0.146	-0.133
palocci	obama	dirceu	crisi	mourao	china
-0.368	-0.251	0.176	0.160	0.154	-0.139
patent	drug	licens	mercosul	compulsori	senat
-0.209	-0.198	-0.162	0.162	-0.143	0.143
chvez	iraq	palocci	gaza	copenhagen	israel
0.197	-0.190	0.160	0.147	-0.142	0.141
hondura	zelaya	mst	chvez	iraq	chavez
0.251	0.243	-0.195	0.164	-0.148	-0.148
zelaya	hondura	coup	mourao	honduran	dirceu
-0.342	-0.334	-0.195	-0.194	-0.185	0.182
felix	mst	obama	terror	favela	labor
-0.279	-0.242	-0.182	-0.165	0.154	-0.152
felix	garcia	summit	obama	energi	copenhagen
-0.384	0.179	-0.161	0.160	0.142	-0.133
palocci	felix	anac	dca	varig	iraq
-0.226	0.202	-0.180	-0.157	0.156	-0.140
varig	bank	haiti	reform	meirell	mourao
0.214	0.157	-0.132	-0.126	0.125	-0.119
summit	mercosul	palocci	doha	iraq	she
-0.197	-0.173	0.142	0.137	-0.136	0.128
felix	jobim	mre	dca	air	shootdown
0.198	-0.158	-0.153	-0.152	0.145	0.141
summit	ftaa	garcia	amorim	indigen	space
0.218	-0.212	-0.201	-0.165	-0.156	-0.139
palocci	felix	shootdown	air	space	garcia
0.260	0.248	-0.229	-0.207	0.186	-0.158
garcia	dirceu	anatel	palocci	gaza	summit

-0.362	0.175	-0.147	-0.146	0.118	-0.118
garcia	stang	anac	felix	varig	shootdown
0.243	0.220	0.195	0.186	-0.157	-0.147
ftaa	garcia	mercosul	cpi	mourao	mst
-0.232	0.225	0.187	0.144	-0.135	0.131
indigen	indian	mst	demarc	demarch	funai
0.464	0.248	-0.242	0.199	0.141	0.137
mourao	nuclear	demarch	afro	amorim	mercosul
-0.246	0.172	0.164	0.137	0.126	-0.123
stang	varig	ftaa	mst	mre	dirceu
0.245	0.217	-0.215	-0.193	0.182	-0.157
stang	anatel	dump	summit	indigen	energi
-0.274	-0.236	0.148	-0.131	0.130	0.129
mercosul	usd	muslim	biofuel	indigen	gsp
-0.188	-0.176	0.165	-0.146	-0.135	0.116
ftaa	felix	anatel	stang	garcia	amorim
-0.220	0.177	0.173	-0.165	0.149	-0.144
amorim	favela	anatel	garcia	tfca	haiti
-0.245	0.223	-0.212	0.151	0.144	-0.139
anatel	muslim	varig	port	space	amorim
-0.310	-0.209	-0.181	0.127	0.118	0.111
demarch	afro	felix	mercosul	amorim	tfca
0.229	0.179	0.168	0.160	-0.157	-0.154
muslim	biofuel	felix	afro	summit	ambassador
-0.259	0.204	0.194	0.185	0.146	-0.139
amorim	muslim	doha	dirceu	gsp	pt
-0.213	0.155	0.150	0.139	0.135	-0.135
muslim	mercuri	kosovo	lebanes	amorim	usd
0.257	0.169	0.132	0.125	0.120	0.115
port	russian	anatel	russia	muslim	varga
0.211	-0.191	-0.186	-0.180	0.155	-0.139
mercuri	varga	rousseff	sarney	dirceu	trq
-0.233	0.180	-0.166	-0.129	-0.124	-0.118
mercuri	amorim	dirceu	garcia	varga	trq
-0.208	0.192	-0.189	-0.171	-0.135	-0.132
muslim	mre	rousseff	mercosul	amorim	trq
-0.245	0.145	0.137	-0.124	-0.119	-0.119
rousseff	ftaa	stang	gsp	fiesp	felix

-0.165	0.153	0.149	0.144	0.135	0.130
mercuri	kosovo	trq	labor	varga	agricultur
0.280	-0.141	-0.131	0.127	0.125	-0.117
mercuri	trq	dirceu	raw	quarter	korea
-0.162	-0.149	0.136	-0.128	0.128	0.125
dirceu	cdc	cpi	ftaa	ppp	hhs
-0.249	0.232	0.166	0.139	0.122	0.121
korea	port	she	traffick	labor	cdc
-0.166	0.139	0.138	-0.134	0.129	-0.128
mercuri	gsp	dirceu	mma	tax	mre
-0.377	-0.152	0.129	-0.123	0.111	0.101
imf	jobim	cpi	dirceu	haiti	forest
0.218	0.171	0.151	-0.131	-0.128	-0.125
korea	north	dprk	labor	traffick	haiti
-0.216	-0.163	-0.161	-0.154	0.148	0.133
drug	imf	environment	mercuri	korea	gsp
0.140	-0.134	-0.133	0.129	0.128	0.119
kassab	dirceu	serra	mantega	debt	gsp
-0.183	-0.165	0.147	0.139	-0.131	0.128
visa	muslim	environment	student	kosovo	mercuri
-0.177	0.138	0.123	-0.120	-0.116	-0.115
varga	dirceu	gsp	cavalcanti	polic	neve
0.204	0.180	0.163	-0.137	-0.124	0.113
pinta	haiti	gama	kassab	mercuri	gsp
-0.138	0.130	-0.129	-0.123	-0.122	0.120
gama	pinta	dirceu	licens	bush	amorim
0.222	0.199	-0.156	0.126	0.125	-0.125
haiti	amorim	pmdb	correa	cdc	kassab
-0.181	0.174	-0.169	-0.138	0.133	-0.133
gama	pinta	afghanistan	tfca	adorno	cdc
-0.279	-0.258	-0.216	0.214	-0.136	0.133
afghanistan	jobim	dirceu	gama	haiti	biato
-0.212	0.165	0.141	0.135	0.134	-0.130
amorim	kassab	imf	varga	meirell	cavalcanti
0.158	0.146	-0.137	-0.135	0.134	-0.118
she	serra	haiti	tfca	human	sarney
0.178	-0.176	0.156	-0.133	-0.131	-0.130
globe	kassab	neve	jobim	gsp	student

-0.191	0.140	-0.129	-0.123	0.118	-0.115
port	sarney	food	bush	pt	biofuel
0.154	0.133	0.121	-0.115	0.111	0.105
imf	meirell	gsp	cpi	sarney	prison
-0.183	-0.149	-0.132	-0.129	0.127	-0.120
afghanistan	netto	bush	galvao	jaguarib	deforest
0.256	0.139	-0.133	0.128	0.116	-0.116
varga	neve	bush	imf	prison	port
0.315	0.152	0.138	-0.137	0.117	-0.114
imf	galvao	mercuri	ibama	esth	ambassador
-0.306	-0.145	0.142	-0.132	-0.120	0.110
pmdb	fort	gsp	correa	polic	imf
-0.163	0.136	-0.125	0.124	-0.124	0.124
pmdb	patent	she	cabinet	ambassador	mantega
0.141	0.133	0.126	0.120	-0.118	-0.114
globe	patriota	bush	pt	moral	ppp
-0.167	0.157	0.124	0.118	-0.112	-0.105
imf	ppp	cavalcanti	kassab	meirell	tax
0.270	-0.169	0.158	-0.145	0.118	0.117

ANEXO C – Programa para gerar o mapa de conexões por cabos

```
import json
import numpy as np
import matplotlib.pyplot
import math
from mpl_toolkits.basemap import Basemap
from geopy.geocoders import Nominatim

def acha_coordenadas(lugar):
        with open("coordenadas.json", "r") as f:
        coord = json.load(f)
        if lugar in coord:
                return coord[lugar]

        geolocator = Nominatim()
        loc = geolocator.geocode(lugar)
        try:
                coord[lugar] = loc.latitude, loc.longitude
        except:
                coord[lugar] = None
        with open("coordenadas.json", "w") as f:
                json.dump(coord, f, sort_keys=True,
                          indent=4, separators=(",",": "))
        return coord[lugar]

coordenadas = {}
for codigo, cidade in codigo_para_nome.iteritems():
        coordenadas[codigo] = acha_coordenadas(cidade)
coordenadas = {key: value for key,value in coordenadas.iteritems()
                           if value is not None}

plt.rcParams["figure.figsize"] = (20, 20)
m = Basemap(projection="cyl", lon_0=0, resolution="c")
m.drawcoastlines()
```

```
for origem, destino in grafo.keys():
        if origem not in coordenadas or destino not in coordenadas:
                continue
        lat1, lon1 = coordenadas[origem]
        lat2, lon2 = coordenadas[destino]
        line, = m.drawgreatcircle(lon2, lat2, lon1, lat1,
                                lw=0.04*math.sqrt(grafo[(origem, destino)
        p = line.get_path()
        cut_point = np.where(np.abs(np.diff(p.vertices[:, 0])) > 200)[0]
        if cut_point:
                cut_point = cut_point[0]
                new_verts = np.concatenate(
                                        [p.vertices[:cut_point, :],
                                        [[np.nan, np.nan]],
                                        p.vertices[cut_point+1:, :]]
                                        )
                p.codes = None
                p.vertices = new_verts
plt.show()
```

ANEXO D – Programa de processamento de dados

D.1 documentos.py

```
import cPickle as pickle

from gensim import corpora
from cablegate import banco, tokenizador

ORIGENS_POR_PAIS = banco.ORIGENS_POR_PAIS

class Documentos(object):
    def __init__(self, pais):
        self.pais = pais

        self.dicionario = None
        self.corpus = None
        self.corpus2db = None
        self.carrega_modelo()

    def computa_vazios(self):
        pais = self.pais

        if not self.dicionario:
            self.dicionario = corpora.Dictionary()

            for _, doc in banco.Banco(pais):
                self.dicionario.doc2bow(tokenizador.processa(doc),
                                        allow_update=True)

            self.dicionario.filter_extremes(no_below=5, no_above=0.5,
                                            keep_n=None)
            self.dicionario.save("data/binary_dict_%s" % pais)

        if not self.corpus or not self.corpus2db:
```

```
        dicionario = self.dicionario
        corpus2db = {}

        class CorpusStream(object):
            def __iter__(self):
                for docid, (dbid, doc) in enumerate(banco.Banco(pais)):
                    corpus2db[docid] = dbid
                    yield dicionario.doc2bow(tokenizador.processa(doc))

        corpora.MmCorpus.serialize('data/corpus_%s.mm' % pais,
                                   CorpusStream())
        self.corpus = corpora.MmCorpus('data/corpus_%s.mm' % pais)

        self.corpus2db = corpus2db
        with open('data/corpus2db_%s.p' % pais, 'wb') as fd:
            pickle.dump(corpus2db, fd)

    def carrega_modelo(self):
        try:
            self.dicionario = \
                corpora.Dictionary.load("data/binary_dict_%s" % self.pais)
            self.corpus = corpora.MmCorpus("data/corpus_%s.mm" % self.pais)
            with open('data/corpus2db_%s.p' % self.pais, 'rb') as fd:
                self.corpus2db = pickle.load(fd)
        except IOError:
            self.computa_vazios()
```