Inteligência Analítica Aplicada ao
Processo de Participação Social

Extração Automática de Tópicos

InWeb/UFMG CEWEB/NIC.BR MJ

Agosto de 2015

Sumário

1 Contexto

1.1 Descrição

Durante a discussão de um anteprojeto de lei, indivíduos, sociedades e organizações podem comentar e sugerir mudanças em cada um dos artigos e parágrafos do texto do anteprojeto. No entanto, um mesmo tópico de interesse pode ser discutido e comentado em diversas partes do texto. Nesse contexto, identificar comentários referentes a um mesmo assunto amplo é uma tarefa interessante, uma vez que pode oferecer uma visão mais global do processo de discussão e uma forma de sumarização dos tópicos mais discutidos.

1.2 Motivação

Técnicas automáticas de extração de tópicos tem como objetivo agrupar informações de acordo com seu conteúdo semântico. No contexto do Anteprojeto de Lei de Proteção de Dados Pessoais, "consentimento de dados"e "proteção cibernética", por exemplo, poderiam ser considerados tópicos. Essas técnicas seriam especialmente úteis quando utilizadas em textos maiores, e também poderiam ser utilizadas para gerar estatísticas gerais sobre os assuntos mais comentados. No futuro, esses tópicos poderiam inclusive ajudar na organização do documento de anteprojetos de lei em si para exibição no site e posterior acesso pela população.

1.3 Entrada

Tanto os comentários quanto os documentos recebidos através do sítio de participação popular seriam dados como entrada para o método.

1.4 Saída

A saída da técnica permitiria uma visualização dos comentários agrupados de forma semântica, com estatísticas básicas dos assuntos mais comentados (por exemplo, número de usuários que ocmentaram sobre o tópico, porcentagem em relação ao total, entre outros). Cada tópico é representado por um conjunto de termos, que podem passar por uma avaliação qualitativa de um especialista para validação.

1.5 Estratégia

O processo de identificação de tópicos pode ser realizado de diversas maneiras. A abordagem utilizada neste projeto será baseada em métodos probabilísticos, como o *Latent Dirichlet Allocation* (LDA) [1]. Para cada documento, o método retorna a probabilidade dele pertencer a um determinado tópico, onde o número de tópicos é previamente definido. Inicialmente, criaremos uma base de dados a partir dos comentários submetidos no site. Já os documentos submetidos em formato pdf poderão passar por um processo de filtragem anterior, que pode ser feito manualmente ou usando os próprios métodos de tópicos.

2 Solução Proposta

Esta seção descreve os principais conceitos do *Latent Dirichlet Allocation* (LDA), método utilizado para identificação de tópicos neste projeto.

O LDA é um modelo estatístico que descreve coleções de documentos através de um conjunto de tópicos. A forma mais fácil de entendê-lo é por meio de seu processo generativo, i.e., o processo imaginário por meio do qual o modelo assume que os documentos são criados.

Um tópico é formalmente definido como uma distribuição de probabilidades sobre um vocabulário fixo. Por exemplo, o tópico *esportes* pode ser descrito por uma distribuição de probabilidade concentrada em palavras como *jogo, futebol, tênis, campo, bola, partida.* Assuma que estes tópicos são especificados antes dos documentos serem gerados. Assim, para cada documento na coleção, suas palavras são geradas em duas fases:

1. Escolha uma distribuição aleatória sobre os tópicos.
2. Para cada palavra no documento:
    a. Escolha um tópico a partir da distribuição de tópicos da etapa 1.
    b. Escolha uma palavra a partir do tópico escolhido na etapa 2.a.

Este modelo estatístico reflete a intuição de que documentos exibem múltiplos tópicos. Por exemplo, um documento pode pertencer aos tópicos *Esporte* e *Política.* Cada documento possui tópicos em diferentes proporções (etapa 1); cada palavra de cada documento é escolhida a partir de um dos tópicos (etapa 2b), sendo que o tópico selecionado é escolhido a partir da distribuição de tópicos para o documento que está sendo gerado (etapa 2a).

O objetivo de modelos de tópicos probabilísticos é descrever os tópicos de uma coleção de documentos de forma automática. Os documentos em si são observados, enquanto a estrutura de tópicos – os tópicos, as distribuições de tópicos por documento e a atribuição de tópicos por palavra e por tópico – são as variáveis latentes (i.e., variáveis não-observadas). O problema central da modelagem em tópicos é utilizar os documentos observados para inferir as estrutura de tópicos latentes.

Cabe ressaltar que os modelos não tem informação sobre o assunto da coleção, e os documentos não possuem anotações com tópicos e/ou palavras-chave. A distribuição interpretável dos tópicos surge por meio da computação da estrutura latente que, com maior probabilidade, gerou a coleção de documentos observada.

Formalmente, sejam K o número de tópicos, D o número de documentos e N o número de palavras do vocabulário. Os tópicos são representados por $\beta_{1:K}$, onde cada β_k é uma distribuição de probabilidade sobre o vocabulário da coleção. As distribuições de tópicos para o documento d são Θ_d, onde $\Theta_{d,k}$ é a proporção do tópico k no documento d. As atribuições de tópicos para o documento d são representadas pela variável aleatória z_d, onde $z_{d,n}$ é a atribuição de tópico para a palavra n no documento d. Finalmente, as palavras observadas no documento d são representadas como w_d, onde $w_{d,n}$ refere-se à palavra n no documento d, que representa um elemento a partir do vocabulário fixo da coleção.

A partir desta notação, o processo generativo do LDA corresponde à seguinte distribuição conjunta de variáveis observadas e latentes (i.e., não-observadas):

$$p(\beta_{1:K}, \Theta_{1:D}, z_{1:D}, w_{1:D}) = \prod_{i=1}^{K} p(\beta_i) \prod_{d=1}^{D} p(\Theta_d) \left(\prod_{n=1}^{N} p(z_{d,n}|\Theta_d) p(w_{d,n}|\beta_{1:K}, z_{d,n}) \right) \quad (1)$$

Ressalta-se que esta distribuição conjunta especifica um conjunto de dependências. Por exemplo, a atribuição de um tópico, $z_{d,n}$, depende das proporções dos tópicos em um documento Θ_d. Estas dependências definem o LDA, ou seja, elas materializam as suposições

Tabela 1: Termos que descrevem os 13 tópicos extraídos e o número de documentos que foi associado a cada tópico.

Id	Docs.	Palavras que descrevem os tópicos
1	136	deveria,liberdade,termo,empresa,acredito,**jornalisticos**,pesquisa,exclusivamente
2	80	artigo,operador,**responsabilidade**,somente,civil,atividade,devem,terceiros
3	79	dado,protecao,regras,**privacidade**,internacional,lei,incluir,garantir,atual,direitos
4	75	dado,acesso,direito,publico,sido,cancelamento,determinado,formato,inclusive
5	64	competente,artigo,sugerimos,casos,caso,titulares,medidas,importante,danos
6	64	**consentimento**,pessoais,titular,livre,coleta,controle,devem,fornecimento
7	62	dado,tratamento,titular,interesse,finalidade,caso,direitos,base
8	61	pessoa,seja,pais,uso,brasil,melhor,entidade,natural,conceito
9	51	lei,sendo,anteprojeto,empresas,autoridade,dispositivo,cumprimento,rede,legais
10	48	pessoas,the,data,tipo,importante,estado,concordo,qualquer,tecnologia
11	43	inciso,**europeia**,qualquer,proposta,possibilidade,disposto,contrato,lei
12	39	dado,pessoais,protecao,tempo,base,processo,contexto,sentido
13	34	**privacidade,internet,protecao,pessoais,brasil,direito,mecanismos,brasileiro,dado**

estatísticas por trás do processo generativo. Especificamente, as dependências são representadas na forma matemática por meio da distribuição de probabilidade conjunta (Equação 1).

3 Metodologia Experimental

O principal objetivo desse projeto é extrair tópicos automaticamente dos comentários submetidos à página de discussão do Anteprojeto de Lei de Proteção de Dados Pessoais. Para isso, a primeira decisão a ser tomada é qual o número de tópicos que se deve extrair dos documentos, que é um parâmetro do sistema. Inicialmente, utilizamos duas abordagens para definir o número de tópicos: (i) o número de eixos temáticos utilizados na página, que é igual a 13; (ii) o número de artigos descritos no texto da lei, que são 52.

Para essas duas configurações, utilizamos o LDA para extrair tópicos dos 836 comentários submetidos via Web, que foram descritos por um conjunto de 1,136 palavras. Um grande esforço foi dispendido na limpeza inicial dos dados.

4 Resultados

A seguir apresentamos os resultados obtidos considerando tanto 13 quanto 52 tópicos. Ao utilizar 13 tópicos, número equivalente a quantidade de eixos temáticos definidos manualmente, os tópicos descritos na Tabela 1 foram identificados.

Podemos observar, por exemplo, que os 136 documentos classificados como pertencentes ao tópico descrito na primeira linha da Tabela 1 discutem o ponto da lei não se aplicar a tratamentos de dados *realizados para fins exclusivamente jornalísticos*, o que vai de encontro ao eixo temático *Escopo e Aplicação*. Embora não exista um mapeamento de um para um entre tópicos e eixos, os tópicos também abordagem os eixos temáticos de consentimento, na definição da lei para o Brasil, na questão da privacidade, entre outros pontos relevantes. É importante ressaltar que, para técnicas baseadas em tópicos, um número maior de tópicos pode gerar uma distribuição de tópicos por documentos mais interessante.

Tabela 2: Termos que descrevem os 13 tópicos extraídos e o número de documentos que foi associado a cada tópico.

Id	Docs.	Palavras que descrevem os tópicos
1	50	**jornalisticos**,liberdade,exclusivamente,imprensa,amplo,jornais,expressamente
2	45	**consentimento**,expresso,livre,informado,termos,fornecido,coleta,casos,escrito
3	42	tratamento,finalidade,**consentimento**,contexto,garantir,relevantes,individuo
4	23	direito,pessoas,privado,liberdade,**consumidor**,titular,preciso,incisos,empresas
5	22	privacidade,protecao,lei,intimidade,dado,marco,civil,nacional,liberdade,sistemas
6	21	responsabilidade,atividade,dano,civil,culpa,risco,agente
7	21	natural,pessoal,direitos,**banco**,privada,proporcionalidade,identificadores,internet
8	21	pais,nacional,**territorio,brasileiros**,forum,brasil,brasileira,regra,diretiva
9	18	dado,**terceiros**,sentido,**bancos**,dispensado,expressa,suficiente,busca
10	17	regras,**internacional**,privacidade,mecanismos,globais,modelo,contratuais
11	17	dado,pessoais,protecao,acordo,deixar,governo,futura,dispor
12	16	internet,privacidade,brasil,mecanismos,direitos,**internacional,brasileiro,europa**
13	16	**sugerimos**,seguinte,inciso,feita,ampla,tratar,**sugiro**,realizados,multa

Por esse motivo, a seguir mostramos 13 exemplos de tópicos encontrados quando utilizamos o LDA para gerar 52 tópicos. Note que um mesmo documento pode estar associado a mais de um tópico, e que os tópicos podem ser redundantes (exemplo: os tópicos 2 e 3 falam sobre consentimento).

Dentre as descrições de tópicos listadas na Tabela 2, observamos que muitos dos tópicos descobertos quando utilizamos 13 tópicos ao invés de 52 aparecem novamente na lista de tópicos, mas novos tópicos também são descobertos.

É interessante ressaltar também que um tópico não necessariamente precisa descrever um assunto, mas pode encontrar outros tipos de estruturas "escondidas" nos dados. Por exemplo, o tópico 13 agrupa comentários que trazem sugestões de mudanças no texto da lei, tais como: "*Sugerimos a seguinte redação: IV - realização de pesquisa histórica, genealógica, artística, científica, cultural ou acadêmica, estatística ou de interesse público, garantidas as medidas de segurança aplicáveis*"ou "*Sugestão: inserção de princípio que venha a prestigiar a liberdade de iniciativa no tratamento de dados, desde que respeitadas as condições impostas pela lei, como forma de fazer valer a menção a direitos fundamentais de liberdade previstos no art. 1º*".

Abaixo apresentamos alguns dos comentários listados nos tópicos da Tabela 2. Para o Tópico 1, percebe-se que a maior polêmica está no que seriam *fins exclusivamente jornalísticos*. Para o Tópico 2, a idade em que se considera a pessoa apta a dar consentimento sobre os dados causa polêmica. Por fim, o tópico 4 discutido a proteção da informação pessoal em geral.

Tópico 1

Por qual motivo textos para fins exclusivamente jornalísticos não se encaixariam nessa lei? Discordo desse inciso.

Deve haver uma melhor definição sobre o que venha a ser fins jornalisticos. Carece de uma

melhor definição. Da forma como está abre espaço para diversas interpretações subjetivas

Gostei! A liberdade de expressão da imprensa requer responsabilidades. Precisa funcionar principalmente na prática e na clara interpretação do que seja fins exclusivamente jornalisticos

Dados colhidos para fins jornalisticos também devem estar submetidos à lei. Na atual conjuntura das comunicaçoes sociais do Brasil, a midia representa forte expressao poitica, demonstrando tendencias de pensamento politico e partidarismo. Imprescindivel que esteja submetida aos rigores da lei no que se trata de divulgaçao e tratamento de dados pessoais.

Há de se detalhar mais o fim jornalistico, pois no tempo de hoje com várias formas de se divulgar informações, pode haver abusos que afeta os Direitos previstos na Constituição federal.

Tópico 2

Concordo com os colegas que afirmaram que 12 anos não é uma idade apropriada para tomar esse tipo de decisão, de consentimento para tratamento de dados. Assim como se lê no referido artigo, trata-se de uma pessoa em desenvolvimento, portanto ainda não tem a maturidade suficiente para tal. E no mais, o artigo não deixa claro como deve ser esse respeito por parte de quem recebe os dados com essas pessoas fornecedoras. Dito isso, também acho interessante elevar essa idade para a partir dos 16 anos.

Deve ser corrigido para entre 12 e 16 anos com o consentimento dos pais e de 16 à 17 anos sem o consentimento dos pais. Deve ser levado em consideração a capacidade civil do indivíduo, ou seja, o artigo 3º do Código Civil, os absolutamente e relativamente incapazes

Não concordo com esse tipo de exposição sem o consentimento, mesmo que por profissionais da área específica. O indivíduo deve escolher como será o seu tratamento e à quem de confiança apresentar seus ditos dados sensíveis.

Uma exceção para interesse legítimo para o tratamento é importante principalmente no contexto do aumento de digitalização de processos comerciais e da sociedade e em conexão com a Internet das Coisas e análise de big data, em que o consentimento expresso e específico nem sempre pode ser obtido na prática. Nesses casos, deve haver outros motivos legítimos de tratamento para facilitar os usos de dados de forma responsável e transparente, que sejam benéficos para os indivíduos e para a sociedade e que permitam práticas comerciais legítimas e inovação, evitando danos e respeitando a privacidade dos indivíduos. A aplicação rígida de uma exigência de consentimento em casos que seria impraticável ou inadequado obter o consentimento válido, resultaria em consentimentos ilusórios, desinformados e sem sentido, prejudicando a eficaz proteção de privacidade. Assim,para garantir que as regras de privacidade de dados permaneçam tecnologicamente neutras e possam ser aplicadas contextualmente no futuro, é necessário fornecer motivos adicionais e mais flexíveis para o tratamento, como interesse legítimo (além das outras exceções ao consentimento previsto no anteprojeto).

Tópico 4

O artigo é desnecessário já que repete a definição e conceitos já dispostos na Lei. Caso seja mantido, sugerimos que a redação seja alterada para: "Art. 23. A comunicação ou interconexão de dados pessoais entre pessoas de direito privado dependerá de consentimento, ressalvadas as hipóteses de dispensa do consentimento previstas nesta Lei".

A legislação canadense "Personal Information Protection and Electronic Documents Act" permite, em sua Divisão 1 - Proteção da Informação Pessoal, parágrafo 7(f), a transmissão de dados sem o conhecimento do titular para fins estatísticos, acadêmicos ou de pesquisa, nos casos em que não é possível obter o consentimento do titular, devendo a organização informar ao órgão competente que irá transferir os dados previamente. É autorizado também, no parágrafo 7(g), a transmissão de dados feita por instituições cujas funções incluem a conservação de registros de importância histórica, e a transmissão é feita com o propósito de conservação destes registros. Sugerimos que o Anteprojeto, com o objetivo de não prejudicar a conservação de registros históricos, bem como facilitar a transferência de dados entre institutos de pesquisa, tenha dispositivo expresso autorizando a comunicação ou interconexão de dados pessoais entre pessoas de direito privado sem o consentimento do titular. Assim, sugerimos a criação de um § único ao artigo, com a seguinte redação: Parágrafo único. A comunicação ou interconexão de dados pessoais com fins de compartilhamento de pesquisa acadêmica, genealógica, estatística, histórica, científica ou com fins de conservação de arquivos ou registros com estas características, dispensa o consentimento prévio do titular.

As sanções não devem ser aplicadas somente à pessoas jurídicas de direito privado.

No caso, a lei deixa algumas definições em aberto, pois não especifica os reais responsáveis pelo tratamento das informações, se é a empresa como um todo ou o profissional da área de tecnologia que trabalha para tal, contratado ou terceirizado. Assim, questiona-se a possibilidade de não responsabilização da pessoa jurídica caso o responsável seja outro. Da mesma forma, caso a comunicação que se refere o art. 44 não se proceda, qual seria a sanção aplicada? Finalmente, com relação a proteção efetiva, sabe-se que a rede de informações é suscetível a diversos tipos de ataques, o que forçaria a manter-se um aparato destinado a coibir tais ataques, tornando o serviço mais custoso ao consumidor.

---

5 Limitações

Métodos como o LDA baseiam-se em padrões de co-ocorrência de palavras para encontrar tópicos. Assim, quando o número de documentos disponível é pequeno ou o conteúdo dos documentos é formado por textos curtos, esses métodos podem encontrar dificuldades de executar a tarefa.

6 Próximos passos

No momento, estamos trabalhando na extração dos tópicos dos pdf. A ideia é investigar se os assuntos tratados na Web são os mesmos enfatizados no pdfs. Estamos também aplicando os métodos de extração de tópicos nos textos da lei, afim de investigar como eles poderiam

casar com aqueles extraídos dos comentários. Por último, estamos analisando a relação entre os documetnos associados aos tópicos e os artigo aos quais eles estão associados.

Referências

[1] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.