A modelagem de tópicos como proposta para classificação de banco de currículos de recrutamento organizacional

André Luiz Vidal Giampaolo

Departamento de Ciência da Computação – Universidade de Brasília (UnB)
Brasília – DF – Brasil

andregiampaolo@gmail.com

Abstract.
Given the large volume of information currently digitized in the organizational context, there are challenges related to the management of important these documents, especially in the classification of curricula in the recruitment process. The purpose of this study is to evaluate the use of topic modeling on an unclassified basis of curricula. To achieve the purpose, a web application was built using the algorithm latent Dirichlet allocation (LDA) in a set of curricula. To validate the coherence of the generated topics, the word intrusion technique was used with a specialist in the area. The result of the study was favorable to the use of topic modeling within a database in the context of recruitment, given the accuracy of 75% in the coherence of the topics generated.

Resumo.
Diante do grande volume de informação digitalizada atualmente no contexto organizacional, surgem desafios relacionados à gestão destes documentos, especialmente na classificação de currículos no processo de recrutamento. O objetivo deste estudo é avaliar a utilização da modelagem de tópicos em uma base não classificada de currículos. Para isto, foi construída uma aplicação web, utilizando o algoritmo lattent Dirichlet allocation (LDA) em um conjunto de currículos. Para validação da coerência dos tópicos gerados, foi utilizado a técnica de word intrusion com um especialista da área. Como resultado, apresentou-se favorável a utilização de modelagem de tópicos dentro de um banco de dados no contexto de recrutamento, dado a acurácia de 75% na coerência dos tópicos gerados.

1. Introdução

Existe um grande número de documentos digitais gerados na atualidade. Por isso, administrar uma coleção de documentos é considerado de extrema importância em diversas esferas como ciência, cultura e indústria [Witten 2004].

No contexto organizacional, Chiavenato [2008] cita a dificuldade de as empresas gerirem os documentos relativos ao processo de recrutamento e seleção para encontrar os candidatos certos para as funções em aberto. Em específico, Xavier [2006] ressalta que a problemática na etapa de recrutamento é organizar e tratar a grande quantidade de currículos recebidos.

A busca de informações em grandes conjuntos de documentos é custosa e, organizá-los manualmente, é impraticável. Classificar os documentos não estruturados em temas subjacentes pode auxiliar na busca de informações [Witten 2004].

Nesse contexto, a modelagem de tópicos apresenta-se como uma ferramenta favorável a ser utilizada [Witten 2004], pois tem como objetivo criar um conjunto de termos que mostram temas ou assuntos macros de uma coleção de documentos, facilitando a busca e análise dos mesmos [Nolasco e Oliveira 2016].

Desta maneira, o presente trabalho tem como objetivo avaliar a utilização da modelagem de tópicos em uma base não classificada de currículos. Para atingi-lo, os objetivos específicos são:

1. Construir uma aplicação web para facilitar a visualização e interpretação dos dados a serem gerados.
2. Gerar tópicos a partir de uma base não classificados de currículos.
3. Avaliar a acurácia dos tópicos gerados.

A contribuição do trabalho consiste em facilitar com que recrutadores de empresas organizem suas bases de currículos e tenham mais assertividade na busca de candidatos potenciais.

O restante do trabalho apresenta na seção 2 a revisão teórica dos conceitos aplicados, na seção 3 o método proposto para alcance dos objetivos, na seção 4 os resultados gerados e, por último, na seção 5 a conclusão do estudo.

2. Revisão da Literatura

2.1. Inteligência Artificial

A inteligência artificial é definida por Kurzwiel [1990] como “a arte de criar máquinas que executam funções que exigem inteligência quando executadas por pessoas”. Para Charniak e McDermott [1985], ela é “o estudo das faculdades mentais pelo uso de modelos computacionais”.

Segundo Russel e Norving [2004], ela tem por objetivo construir entidades inteligentes que podem perceber, compreender, prever e manipular um mundo.

Dentre as disciplinas que contribuíram para a sua criação estão filosofia, matemática, economia, neurociência, psicologia, engenharia de computadores, teoria de controle e cibernética, linguística, entre outras [Russel e Norving 2004].

O campo de atuação da inteligência artificial é vasto, sendo utilizada em dentro e fora de organizações, em veículos autônomos, reconhecimento de voz, planejamento autônomo e escalonamento, jogos, combate a spam, planejamento logístico, robótica, tradução automática [Luger 2008].

2.2. Mineração de Texto

A mineração de texto, também chamada de mineração de dados textuais ou descoberta de conhecimento de bases de dados textuais [Aranha e Passos 2006], é um processo de extração de informação desconhecidas e úteis de documentos escritos em linguagem natural [Pezzini 2017].

A mineração de texto é considerada uma técnica interdisciplinar que abrange áreas como recuperação de informação, aprendizagem de máquina, inteligência artificial, estatística, linguística computacional e mineração de dados [Hotho, Nurnberger e Paass 2005].

A mineração de texto se difere da mineração de dados no que diz respeito à clareza da apresentação da informação. Enquanto na mineração de texto os dados estão claros e explicitamente declarados, na mineração de dados o conteúdo é apresentado de forma implícita, previamente desconhecida [Witten 2004].

Segundo Morais e Ambrósio [2007], sua principal aplicação é na busca de informações específicas em documentos de diversos tipos, como e-mails, páginas da web e textos eletrônicos. Além disso, contribui para melhor compreensão e realização de análises qualitativas e quantitativas em grandes volumes de texto.

A técnica é aplicada a contextos organizacionais, como demonstra o estudo de Moura [2004], que a utilizou para otimizar a seleção, classificação e qualificação de documentos da Embrapa com objetivo de dar suporte a tomada de decisões da Agência.

2.3. Modelagem de Tópicos

A modelagem de tópicos é um conjunto de algoritmos, desenvolvida com o objetivo de compreender, classificar, sintetizar e explorar arquivos digitais de forma automática [Blei, 2012].

Esse conjunto de algoritmos é um método estatístico que analisa os termos em um conjunto de documentos e gera tópicos automaticamente, sem necessitar de nenhuma classificação prévia destes documentos [Blei, 2012].

De acordo com Faleiros e Lopes [2016], um tópico é o conjunto de termos que ocorrem com frequência em documentos semanticamente correlacionados. Tais tópicos podem ser utilizados para ordenar um conjunto de termos em documentos ou um conjunto de documentos.

Como os tópicos são gerados automaticamente, eles podem não ter coerência para um avaliador humano [Steven et al. 2012]. Por isso, o presente artigo utiliza técnicas que visam dar maior assertividade aos tópicos gerados, que serão apresentadas na seção “Metodologia Científica”.

2.4. *Latent Dirichlet Allocation* (LDA)

Existem alguns algoritmos para realizar modelagem de tópico [Lee e Seung 2001, Lancichinetti et al. 2015, Zuo, Zhao e Xu 2014] dentre eles o LDA - *Latent Dirichlet Allocation* [Blei, Ng e Jordam 2003], que foi escolhido para ser aplicado na pesquisa em questão.

O *Latent Dirichlet Allocation* (LDA) é um modelo probabilístico generativo, isto é, um modelo que gera dados a partir de variáveis latentes [Faleiros e Lopes 2016]. Em um modelo probabilístico generativo, que cria dados aleatoriamente a partir de variáveis latentes, não é necessária a realização do treinamento prévio da base de dados em comparação a outros modelos de tópicos.

O LDA utiliza um modelo *baysiano* completo baseado nas distribuições de Dirichlet [Faleiros e Lopes 2016], no qual divide-se uma coleção de documentos em tópicos, apresentando cada documento como uma sub coleção dos mesmos [Chaney e Blei 2012].

O seu objetivo é identificar tópicos latentes em um conjunto de documentos. Cada um dos tópicos é definido por uma distribuição de palavras presentes nos documentos. Sendo assim, um documento pode ser classificado como um grupo de tópicos [Blei, Ng e Jordan 2003].

Para identificar os tópicos latentes, são construídas 3 matrizes relacionando documento-palavra, tópico-palavra e documento-tópico, conforme apresentado na Figura 1 a seguir.

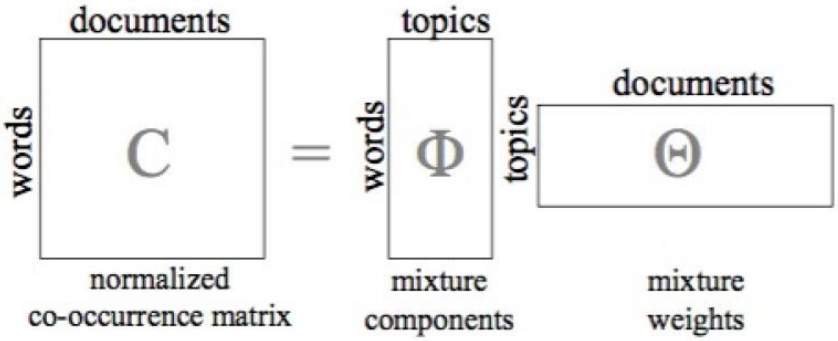

Figura 1. Ilustração da fatoração desejada para a matriz de documentos.
Fonte: [Steyvers e Griffiths 2007]

2.5. Recrutamento de Pessoal

O recrutamento de pessoal nas organizações, desempenhado usualmente pela área de recursos humanos, é um conjunto de ações realizadas para captar candidatos capacitados para determinadas posições em aberto na empresa [Schermerhorn, Hunt e Osborn 2000].

Chiavenato [2008] aborda que o processo do recrutamento é iniciado com o recebimento de currículos de potenciais candidatos, pela empresa. No currículo são informados dados pessoais, escolaridade, experiência profissional, conhecimentos gerais. As organizações os recebem majoritariamente via internet e, a partir disso, eles são triados de acordo com as necessidades.

Segundo o mesmo autor, a dificuldade presente nas organizações é de localizar os candidatos certos para o preenchimento de vagas, aqueles com as características necessárias. A tecnologia da informação (TI) tem auxiliado neste processo, porém sua

velocidade de desenvolvimento não tem sido suficiente. Desta maneira, a busca por talentos aumenta e a dificuldade de encontrá-los também.

Xavier (2006) vai além e aponta que as organizações estão sendo desacreditadas pelos candidatos em potenciais, pois o enorme volume de currículos recebidos pela internet impossibilita o correto tratamento dos mesmos. Assim, as organizações não utilizam os documentos enviados e não dão nenhum retorno aos candidatos, o que gera gerando desperdício e perda de tempo.

Sendo assim, este estudo visa segmentar uma base de currículos não classificada por meio da aplicação do algoritmo LDA, visando facilitar aos recrutadores a pesquisa por candidatos potenciais.

3. Metodologia Científica

Etapa 1: Construção da base de dados

A fonte da base de dados de currículos utilizada na pesquisa é o Banco Nacional de Empregos (BNE), um website que disponibiliza acesso irrestrito e gratuito a diversos currículos a nível nacional.

Para extração da base de dados, foi desenvolvido um *web scraping*, um conjunto de técnicas usadas para conseguir de forma automática informações de um website [Vargiu e Mirko 2012]. O *web scraping* foi desenvolvido com o *framework scrapy*, que é "Uma estrutura de código aberto e colaborativa para extrair os dados que você precisa dos sites. De maneira rápida, simples e extensível" [Scrapy, 2018].

Em 20 de abril de 2018 estavam cadastrados no website 18.484.508 currículos. Para efeitos da pesquisa, foram extraídos 500 currículos para aplicação do algoritmo LDA.

Etapa 2: Aplicação do algoritmo

Na segunda etapa, foi desenvolvida uma aplicação web com o intuito de aplicar o algoritmo LDA e facilitar a visualização dos tópicos pelos recrutadores. Esta aplicação foi construída com o *framework flask* e o banco de dados *sqlite*.

O algoritmo LDA foi utilizado na aplicação para gerar 20 tópicos com os dados dos 500 currículos. Para melhorar a classificação das palavras dos currículos, foram utilizadas técnicas de lematização (conversão de palavras para seus radicais comuns), tokenização (separação de palavras e caracteres desnecessários), remoção e pontuação de *stop words* (palavras consideradas irrelevantes).

Etapa 3: Teste de validação da solução

Para avaliar a assertividade dos tópicos será utilizada a técnica de *word intrusion*, utilizada também pelos seguintes autores em seus trabalhos: Chang et al. [2009], Lau, Newman e Baldwin [2014] e Mimno et al. [2011].

A técnica do *word intrusion* consiste em apresentar a um avaliador de forma aleatória uma lista com seis palavras dentre as quais cinco delas tem relação com um tópico gerado e uma destas palavras não tem. Por exemplo, no conjunto *{maçã, banana, laranja, lápis, manga, uva}* a maioria das pessoas identifica de forma rápida que a palavra *lápis* não tem relação com o conjunto. Já em um conjunto de palavras como *{ferrari, chocolate, viagem, brasília, roupa, médico}* não é intuitivo selecionar a palavra intrusa, fazendo a escolha seja feita de forma arbitrária [Chang et al. 2009].

Para este experimento, foram executadas as seguintes etapas:

1) Escolha de um tópico aleatório.
2) Seleção de cinco palavras com maior relação ao tópico anterior.
3) Definição da palavra intrusa, que tem baixa relação com o tópico selecionado no item 1 e alta relação com outro tópico.

Deste modo, as chances de a palavra intrusa derivar do mesmo grupo semântico e de ela ser descartada de forma imediata são baixas.

O teste de validação foi aplicado com um especialista na área de Recursos Humanos, que escolheu uma palavra intrusa de cada tópico gerado. O resultado consiste em o especialista ter acertado a palavra intrusa no tópico. Após a aplicação, foram analisados os resultados, que são apresentados a seguir.

4. Resultados

A tabela abaixo apresenta a matriz que relaciona os tópicos gerados, as cinco palavras com maior relação com os tópicos, a palavra intrusa e resultado da validação do especialista.

Tabela 1: Matriz de tópicos gerados por palavras e resultados da validação

Tópico	Palavra 1	Palavra 2	Palavra 3	Palavra 4	Palavra 5	Palavra Intrusa	Resultado
Tópico 1	farmaceutico	medicamento	producao	farmaceutica	lavagem	pedagogicas	Acertou
Tópico 2	informatica	sistema	equipamento	elaboracao	testar	lavagem	Acertou
Tópico 3	condominio	copeiro	porteiro	copar	pastar	saude	Acertou
Tópico 4	pessoal	admissao	departamento	folhar	ferir	lavagem	Acertou
Tópico 5	coordenador	pedagogicas	acompanhamento	corpo	educacao	projetista	Acertou
Tópico 6	administrativo	auxiliar	atendimento	cliente	controlo	obrar	Acertou
Tópico 7	projetista	desenvolvimento	dar	mover	desenvolver	prescricao	Acertou
Tópico 8	contramestre	sanitarias	construção	eletricas	hidraulicas	producao	Acertou
Tópico 9	despacho	fiscalizar	pessoa	dner	movimentacao	ensinar	Errou
Tópico 10	institucional	relacoes	estrar	rodoviaria	pronasci	testar	Errou
Tópico 11	paciente	exame	saude	prescricao	medicar	eletricas	Acertou
Tópico 12	obrar	construcao	seguranca	engenheiro	trabalhar	farmaceutica	Acertou

Tópico 13	educacao	professorar	aula	ensinar	professor	projetista	Acertou
Tópico 14	manutencao	juridico	tecnico	advogar	preventivo	professor	Acertou
Tópico 15	saude	tecnico	enfermagem	paciente	enfermeiro	farmaceutico	Errou
Tópico 16	obrar	execucao	sanitario	pavimentacao	sistema	construcao	Errou
Tópico 17	producao	comunicacao	educacao	marketing	publicar	conservacao	Acertou
Tópico 18	limpeza	gerar	servicos	limpar	conservacao	aula	Acertou
Tópico 19	grafica	exercicios	prescricao	saude	orientacao	desenvolver	Errou
Tópico 20	tubulacoes	instalacoes	hidraulico	bombeiro	fazer	educacao	Acertou

Como apresentado anteriormente o *word intrusion* mensura se os tópicos gerados pelo LDA são compatíveis com os conceitos humanos, ou seja, o quão fácil as palavras intrusas são detectadas pelos sujeitos. Tendo em vista o teste executado, observou-se uma acurácia de 75%, indicando que a maioria dos tópicos gerados adequados ao contexto.

Dado os resultados encontrados, supõem-se que a modelagem de tópico se apresenta como uma possível solução para auxiliar recrutadores de empresas a organizarem suas bases de currículos.

5. Conclusão

De forma geral, o aumento da informação digitalizada dificulta a constante classificação do conteúdo, prejudicando a seleção de currículos em grandes bases de dados Xavier [2006].

O desenvolvimento do presente estudo possibilitou uma análise da utilização da modelagem de tópicos em um conjunto não classificado de currículos. Espera-se que isso facilite a busca de currículos em um grande banco de dados tendo em vista que isto é um problema para as organizações.

Para análise foi desenvolvida uma aplicação web utilizando o *framework flask* e o banco de dados *sqlite*, na qual permitiu-se a aplicação do algoritmo LDA em uma base de dados composta por 500 currículos gerando um total de 20 tópicos, desta forma facilitou-se a visualização e interpretação dos dados gerados.

Pela validação da coerência dos tópicos com um especialista na área de recrutamento de seleção, por meio da técnica de *word intrusion*, verificou-se uma acurácia de 75%. Dessa forma, supõe-se que a modelagem de tópicos pode ser utilizada no contexto em questão para aumento da eficiência na escolha de currículos uma vez que possibilita ao recrutador executar suas atividades de forma mais rápida e precisa, diminuindo também as chances de um currículo não ser avaliado.

Dada a importância do assunto, trabalhos futuros poderão aprimorar a seleção se currículos, ranqueando os currículos relacionados a um mesmo tópico, permitindo assim com que o recrutador tenha maior assertividade. Também é possível automatizar a validação da coerência dos tópicos por meio de algoritmos de inteligência artificial.

6. Bibliografia

A. Lancichinetti, et al. (2015) High-reproducibility and high-accuracy method for automated topic classification. Physical Review X, 5(1):011007.

Aranha, C. e Passos, E. (2006). A tecnologia de mineração de textos. Revista Eletrônica de Sistemas de Informação n. 2.

Banco Nacional de Empregos (BNE). <http://www.bne.com.br/>. Acesso em: 20/04/2018.

Blei, D. (2012). Probabilistic topic models. Communications of the ACM, 55(4), 77-84.

Chaney, A. e Blei. D. (2012). "Visualizing Topic Models." ICWSM.

Chang, J. et al. (2009) Reading tea leaves: How humans interpret topic models. Advances in neural information processing systems.

Charniak, E. e McDermott, D. (1985) *Introduction to Artificial Intelligence*, Reading, MA: Addison-Wesley.

Chiavenato, I. (2008). Gestão de pessoas. Elsevier Brasil.

Blei, D. Ng, A. e Jordan, M. (2003) Latent dirichlet allocation. the Journal of machine Learning research, 3:993–1022.

Faleiros, T. e Lopes, A. (2016) Modelos probabilísticos de tópicos: desvendando o latent Dirichlet allocation.

Flask <http://flask.pocoo.org/> Acesso em: 21/04/2018

Hotho A., Nurnberger, A. e Paass, G. (2005) A brief survey of text mining. LDV Forum - GLDV Journal for Computational Linguistics and Language Technology, v. 20, n. 1. páginas 19-62.

Kurzweil Ray (1990) The age of intelligent machines. MIT Press, Cambridge, MA.

Lau, J. Newman, D. e Baldwin, T. (2014) Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics.

Lee, D. e Seung. H. (2001). Algorithms for non-negative matrix factorization. Advances in neural information processing systems.

Luger, G. (2008). Artificial Intelligence: Structures and Strategies for Complex Problem Solving. Editora Addison Wesley. 6 edição.

Mimno, D. et al. (2011) Optimizing semantic coherence in topic models. Proceedings of the conference on empirical methods in natural language processing. Association for Computational Linguistics.

Morais, E. e Ambrósio, A. (2007). Mineração de textos. Relatório Técnico–Instituto de Informática (UFG).

Moura, M. (2004). Proposta de utilização de mineração de textos para seleção, classificação e qualificação de documentos. Embrapa Informática Agropecuária-Documentos (INFOTECA-E).

Pezzini, A. (2017). Mineração de textos: conceitos, processo e aplicações. Revista Eletrônica do Alto Vale do Itajaí 5.8. páginas 58-61.

Russel, S. e Norving, P. (2004). Inteligência artificial. Editora Campus. 2 edição.

Schermerhorn, J., Hunt, J., & Osborn, R. (2000). Organizational behavior. John Wiley & Sons Inc., New York.

Scrapy. <https://scrapy.org/>. Acesso em: 18/04/2018.

STEVENS, K. et al. (2012) Exploring topic coherence over many models and many topics. In: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 952–961, Association for Computational Linguistics.

STEYVERS, M., GRIFFITHS, T. (2007) Probabilistic topic models. Handbook of latent semantic analysis, v. 427, n. 7, pp. 424–440,

Vargiu, E. e Mirko, U. (2012) Exploiting web scraping in a collaborative filtering-based approach to web advertising. Artificial Intelligence Research 2.1. 44.

Witten, H. (2004). Text Mining. 198.

Xavier, R. (2006). Gestão de pessoas na prática; Os desafios e soluções. Editora gente.

Y. Zuo, J. Zhao, and K. Xu. (2014) Word network topic model: a simple but general solution for short and imbalanced texts. Knowledge and Information Systems, páginas 1–20,