Avaliação de Abordagens Probabilísticas de Extração de Tópicos em Documentos Curtos

Michel Chagas da Costa[1], **Denio Duarte**[1]

[1]Universidade Federal da Fronteira Sul
Campus Chapecó
Chapecó – SC – Brazil

costa.michel10@gmail.com, duarte@uffs.edu.br

Abstract. *Short texts are very popular in social media. Comments and reviews are examples of common short texts found in the Web. Topics extraction from text is a challenging task for content analysis. Lately, probabilistic topic modelling has been used as a tool for topic extraction. To extract topics from short documents is more challenging since the word co-occurrence is more sparse. The aim of this work is, thus, evaluate some short documents topic modelling to identify which one is more suitable in the scenarios proposed. We conduct experiments on three short text collections, and results show that the approaches have similar performances.*

Resumo. *Devido ao amplo uso das redes sociais, textos pequenos se popularizaram na Web. Extrair tópicos de uma grande quantidade de textos curtos tornou-se uma tarefa crítica e desafiadora em tarefas de análise de conteúdo. Neste contexto, várias abordagens foram propostas para inferir tópicos a partir de conjuntos de coleções de textos curtos. Este trabalho tem como objetivo avaliar o uso de algumas destas abordagens probabilísticas na extração de tópicos em documentos curtos utilizando métricas para este fim. Os experimentos realizados em três coleções mostram que as abordagens estudadas tem resultados similares nos cenários propostos.*

1. Introdução

Textos curtos dominam a Web, tanto no contexto de sites tradicionais - títulos de páginas, anúncios, legendas de imagens, mensagens em fóruns e títulos de notícias - quanto mídias sociais, que tiveram um grande crescimento, como *tweets* e mensagens de status [Cheng et al. 2014]. Há um número muito grande de textos curtos, o qual está em rápido e constante crescimento. Um exemplo disso é o *Twitter* que com 250 milhões de usuários ativos gerava aproximadamente meio bilhão de *tweets* por dia [Zuo et al. 2016a]. E este grande volume de textos curtos contém informações que dificilmente são encontradas nas fontes tradicionais de busca e que trazem informações sofisticadas do mundo real.

Abordagens probabilísticas para modelagem de tópicos têm sido usadas de modo amplo para extrair automaticamente tópicos de uma grande coleção de documentos. Abordagens usuais assumem a premissa de que um documento é gerado a partir de múltiplos tópicos. A abordagem *Latent Dirichlet Allocation* (LDA) [Blei 2012] possui a forma mais simples de modelagem de tópicos e serve como base para outras abordagens, também

assume a premissa acima citada. Apesar de se mostrar como uma abordagem de sucesso para textos grandes, como notícias, artigos científicos e blogs, abordagens clássicas, como o LDA, se mostraram limitadas quanto a textos curtos. Em essência, elas descobrem tópicos capturando, implicitamente, a co-ocorrência de padrões de palavras em um documento. Como em textos curtos a co-ocorrência de padrões de palavras em um documento é algo esparso, as abordagens convencionais não são eficientes para extrair tópicos neste cenário [Cheng et al. 2014, Zuo et al. 2016a, Zuo et al. 2016b]. Dados esparsos constituem o principal problema na modelagem de tópicos em documentos curtos [Quan et al. 2015]. São necessárias, então, abordagens que se adaptaram para ter seu foco em textos curtos. Cada uma delas usa diferentes premissas para procurar resolver o problema dos dados esparsos.

Desta forma, este trabalho tem como objetivo avaliar uso de abordagens probabilísticas para a extração de tópicos em documentos curtos. Serão utilizadas quatro abordagens: *Biterm Topic Model* (BTM), *Pseudo-document Topic Model* (PTM), *Self-Aggregation based Topic Model* (SATM) e *Word Network Topic Model* (WNTM). Este trabalho avaliará os resultados da execução destas quatro abordagens sobre três conjuntos de dados através de três métricas de coerência das sete apresentadas por Röder *et al.* [Röder et al. 2015a]: C_V, C_{UMass} e C_A que representam as métricas com melhor e pior desempenhos e desempenho mediano, respectivamente. Os conjuntos de dados utilizados possuem tamanhos médios de 6, 15 e 84 palavras por documento. Cada algoritmo será executado no cenário de 30, 60 e 120 tópicos. Por fim, este trabalho apresentará os resultados avaliando o uso destas quatro abordagens probabilísticas para modelagem de tópicos em textos curtos. O objetivo da análise é identificar qual abordagem se comporta melhor em cada cenário proposto.

A próxima seção apresenta o referencial teórico. Em seguida, alguns trabalhos relacionados são apresentados. A Seção 4 apresenta o projeto e os resultados dos experimentos. Finalmente, a Seção 5 apresenta conclusão.

2. Referencial Teórico

Na área de aprendizado de máquina há uma subárea que visa extrair tópicos de uma coleção de textos. Estes tópicos podem ser usados para categorizar textos, auxiliando na definição de quais temas são abordados em um texto ou conjunto de textos. Por meio de métodos probabilísticos, esses algoritmos são a base desta subárea chamada de modelagem de tópicos [Blei 2012, Steyvers and Griffiths 2007]. A modelagem de tópico é uma técnica não supervisionada, assim métricas tradicionais como precisão, revocação e acurácia não se aplicam.

Dada uma coleção de textos não-organizados, os algoritmos de modelagem de tópicos têm como objetivo descobrir os principais conjuntos de palavras, que podem ser vistos como assuntos, relacionados à coleção. Esses algoritmos podem ser adaptados para os mais diversos tipos de dados. Entre outras aplicações, eles vêm sendo usados para descobrir padrões em dados genéticos, imagens e redes sociais [Blei 2012].

A forma mais simples de modelar tópicos é através da Alocação Latente de Dirichlet - *Latent Dirichlet Allocation* (LDA) [Blei 2012]. O LDA serve como base para vários outras abordagens de extração de tópicos, inclusive abordagens para textos curtos discutidas mais adiante.

Um texto geralmente apresenta múltiplos tópicos, tratando sobre assuntos diversos que têm ligação entre si. Um mesmo texto pode, por exemplo, tratar sobre futebol, cultura e medicina. É nesse pressuposto de que um mesmo texto pode tratar sobre uma variedade de assuntos que se apoia o LDA.

Outro pressuposto desta abordagem é que um documento é uma mistura de tópicos e um tópico é uma distribuição probabilística sobre as palavras. Por exemplo, considere as palavras "televisão" e "competição". A palavra "televisão" tem uma probabilidade pequena de aparecer em um tópico sobre esportes e tem uma probabilidade maior de aparecer em um tópico sobre eletrodomésticos. E a palavra "competição" tem uma probabilidade maior de aparecer em um tópico relacionado a esportes.

Cada documento exibe tópicos em diferentes proporções e cada palavra presente no documento está associada a um destes tópicos exibidos no documento. A alocação de tópicos por documento, de forma estatística, é feita usando a distribuição de Dirichlet [Blei 2012], o que explica o nome LDA. Cada documento, em uma coleção de documentos, compartilha os mesmos tópicos. O que muda é a proporção com que cada tópico aparece no documento.

No caso do LDA, as variáveis observadas são as palavras dos documentos e as variáveis ocultas são a estrutura de tópicos. Portanto, o problema computacional de inferir a estrutura de tópicos é o problema de computar a distribuição condicional das variáveis ocultas, dada as variáveis observadas. O cálculo da distribuição condicional é computacionalmente intratável. Geralmente, os algoritmos de modelagem de tópicos são adaptações para se aproximar da distribuição condicional (posterior).

O LDA possui algumas premissas que norteiam sua implementação. Como dito anteriormente, o LDA serve como base para outras abordagens que tem como objetivo a extração de tópicos em um conjunto de dados. Conforme o objetivo, essas outras abordagens podem relaxar algumas destas premissas, a fim de adaptar o LDA para o que seja mais interessante no contexto daquele outro modelo de tópico.

Uma das premissas é a sacola de palavras - *bag-of-words*. As palavras são vistas de modo independente, soltas [Steyvers and Griffiths 2007]. Segundo esta premissa, a ordem das palavras não importa. Isto pode ser um problema com palavras que causam ambiguidade, onde a mesma palavra tem mais de um sentido semântico (polissemia). Como exemplo, a palavra "vela", que pode ao mesmo tempo significar um barco à vela; a vela feita de cera, para iluminar; ou ainda uma conjugação do verbo velar, que significa estar vigilante. Por isso, algumas outras abordagens procuram adaptar esta premissa.

Outra premissa é que a ordem dos documentos de uma coleção também não importa. Essa premissa pode ser relaxada em outros modelos de tópicos, em que a ordem dos documentos importa, como por exemplo, ao verificar a mudança de um tópico durante uma linha de tempo. Uma abordagem que contemplaria isso é o modelo dinâmico de tópicos, que respeita a ordem dos documentos [Blei 2012].

Assume-se também que o número de tópicos é conhecido e não muda: esta é a terceira premissa do LDA. Ou seja, ao organizar uma coleção de documentos, o número de tópicos já é definido e permanece fixo. Como alternativa, o modelo Bayesiano não-parametrizado de tópicos determina o número de tópicos durante o aprendizado, quando há a inferência do posterior.

2.1. Textos curtos

Os modelos de tópicos convencionais, como LDA, conseguem modelar tópicos de forma satisfatória em uma coleção de textos longos. Porém, na Web, os textos curtos prevalecem [Cheng et al. 2014]. Títulos de páginas, anúncios, legenda de imagens, títulos de notícias, *tweets*, mensagens em redes sociais são apenas alguns exemplos da variedade de textos curtos encontrados na Web. Devido à grande quantidade de textos curtos, tornou-se importante modelar tópicos de textos curtos para várias aplicações de análise de conteúdo, como, por exemplo, descobrir o perfil de interesse do usuário.

Quanto à modelagem de tópicos em textos curtos, as abordagens como o LDA apresentam uma limitação. Nelas, o número de ocorrências de uma palavra em um documento ou uma coleção é fundamental para inferir os tópicos. Entretanto, textos curtos, devido ao seu tamanho, são muito mais esparsos em termos de ocorrência de palavras. Esse problema dos dados esparsos é o principal desafio na modelagem de tópicos em textos curtos [Quan et al. 2015].

As coleções de textos curtos demandaram algumas adaptações na modelagem de tópicos, devido aos dados esparsos. A combinação do LDA com outras técnicas resultou em novas ferramentas para modelagem de tópicos em conjuntos de dados de textos curtos. Por trás de cada cada abordagem há uma intuição básica que busca resolver o problema dos dados esparsos. Algumas destas são: agrupar pares de palavras em vez de palavras soltas (BTM); criar redes de palavras valorizando as ligações entre elas (WNTM); agregar vários textos curtos com tópicos possivelmente similares (SATM); criar textos longos a partir de textos curtos considerando que este texto longo seja híbrido (PTM). Estas abordagens são brevemente apresentadas a seguir.

BTM [Cheng et al. 2014]: o BTM extrai tópicos de textos curtos modelando a geração de termos-pares na coleção de documentos. Termo-par é um par de palavras não ordenadas em um texto curto. É uma forma de explicitar a co-ocorrência de palavras relacionadas em documentos. O BTM assume que duas palavras em um termo-par compartilham o mesmo tópico tendo em vista a coleção de documentos. Segundo Cheng et al [Cheng et al. 2014], se forem agregados todos os padrões de co-ocorrências de uma palavra no *corpus* (conjunto de exemplos), suas frequências são mais estáveis e revelam mais claramente a correlação entre as palavras.

Comparado aos modelos de tópicos convencionais, o BTM apresenta duas vantagens: (i) modelar explicitamente os padrões de co-ocorrências de uma palavra, e (ii) o BTM usa os padrões de co-ocorrência de termos-pares na coleção para descobrir tópicos, visando acabar com o problema de dados esparsos. Por exemplo, um documento com três palavras $(w_1,\ w_2,\ w_3)$ se tornaria $(w_1w_2, w_1w_3,\ w_2w_3)$

SATM [Quan et al. 2015]: o modelo de tópicos baseado em auto-agregação é motivado pela agregação de textos curtos em mídias sociais, como, por exemplo, as *hashtags*, e busca prover uma solução generalizada para extrair tópicos em textos curtos de vários tipos. A ideia da agregação é que as palavras mais usadas podem criar um cluster de textos curtos com tópicos similares, levando a uma solução para o problema dos dados esparsos.

Esta abordagem assume que cada trecho de um texto é parte de um outro texto longo que não está explícito na coleção. Durante a inferência de tópicos, há uma

integração orgânica entre a modelagem de tópicos e a auto-agregação de textos.

PTM [Zuo et al. 2016a]: movido pelo potencial dos métodos de agregação, como o SATM, para lidar com os dados esparsos, um modelo de tópicos baseado em pseudo-documento para textos curtos foi proposto por Zuo et al [Zuo et al. 2016a]. Nesta abordagem, um pseudo-documento é essencialmente um tópico híbrido que combina tópicos específicos de vários textos curtos.

A chave desta abordagem, para lidar com os dados esparsos, é a introdução de pseudo-documentos através da agregação implícita de textos curtos. Desta forma, a modelagem de tópicos de uma coleção grande e esparsa é transformada em uma coleção menor, visando melhorar a eficácia e a eficiência.

WNTM [Zuo et al. 2016b]: diferentemente de abordagens como o LDA, que modela tópicos com base na co-ocorrência de palavras dentro de um documento, o que o torna extremamente sensível ao tamanho de documentos e ao número de documentos relacionados a cada tópico, o modelo de tópico de rede de palavras baseia-se na co-ocorrência de palavras dentro de uma rede de palavras.

O WNTM foi proposto para lidar com o problema dos dados esparsos e com o desbalanceamento de documentos por tópico. A principal ideia desta abordagem vem das seguintes observações: 1) quando os textos são curtos, o espaço de palavra por documento é muito esparso, enquanto o espaço de palavra por palavras é mais denso. Então desde que a qualidade dos tópicos possa ser garantida, a escolha de uma rede de palavras em vez de uma coleção de documentos é mais razoável, 2) a distribuição de tópicos por palavras, em vez de tópicos por documento, pode revelar tópicos raros que não seriam revelados em uma abordagem que usa tópicos por documento, já que o número de palavras relacionadas a tópicos raros geralmente excede o número de documentos relacionados a estes tópicos, 3) já que a distribuição de tópicos por documentos não é aprendida de forma acurada em textos curtos ou desbalanceados, deve-se distribuir os tópicos por palavras em vez de tópicos por documentos, e 4) diferentemente de outras soluções, o WNTM visa garantir a escalabilidade em diferentes cenários.

As quatro abordagens apresentadas nesta seção terão seu uso avaliado neste trabalho, visando a extração de tópicos em conjuntos de dados de documentos curtos.

3. Trabalhos Relacionados

Os trabalhos relacionados apresentados aqui são os artigos onde as abordagens que serão utilizadas neste trabalho foram propostas. No artigo em que o BTM foi proposto por Cheng et al [Cheng et al. 2014], o LDA foi comparado com o BTM, utilizando duas coleções de documentos curtos e a métrica *PMI-Score*. Foram realizados teste com 20, 40, 60, 80 e 100 tópicos. Em todos os cenários o BTM se mostrou mais coerente do que o LDA.

O PTM e o SPTM foram comparados com outras quatro abordagens, segundo Zuo [Zuo et al. 2016a]: SATM, LDA, *Mixture of Unigrams* e *Dual Sparse Topic Model*. Para avaliação, foi utilizada a validação cruzada e 100 tópicos para todas as abordagens em todas as coleções. Em duas das quatro coleções testadas, o PTM teve melhor pontuação do que as outras abordagens. Em uma das coleções o SPTM obteve maior pontuação e SATM se mostrou melhor em um dos quatro conjunto de dados.

Quan et al [Quan et al. 2015], ao apresentar o SATM, comparam a nova abordagem proposta com o BTM e com o LDA. Foram utilizadas duas coleções e executadas estas abordagens para 50, 100, 150, 200, 250 e 300 tópicos, além de utilizar duas novas métricas apresentadas no artigo para avaliação. Os autores do artigo que apresenta o SATM, concluem que esta abordagem se mostrou masi eficiente que o BTM e o LDA naquele cenário proposto.

No trabalho da proposta da WTNM [Zuo et al. 2016b], a mesma é comparada com as abordagens BTM e LDA. Com base na validação cruzada, os autores concluem que o WNTM se mostrou melhor que o BTM e o LDA, utilizando 100 tópicos.

4. Experimentos

Foram selecionados dois conjuntos de dados: *Ohsumed*[1] e *Tag My News*[2]. O conjunto de dados *Ohsumed* consiste em títulos, autores e resumos de artigos da área de medicina, com base em 270 periódicos durante 5 anos (1987-1991). O conjunto de dados Tag My News são notícias de sites de língua inglesa obtidas através de *feeds RSS* de jornais populares.

Para cada um dos conjuntos de dados foi necessário um pré-processamento. Para o conjunto de dados *Ohsumed* foi realizado uma limpeza, de modo a deixar apenas os resumos de artigos, remover pontuações, *stopwords* e palavras de baixa frequência. Enquanto a coleção *Ohsumed* original era de 151,1 MB, após o processo de limpeza para que ficasse no arquivo apenas os resumos dos artigos, o arquivo ficou com 40,7 MB. O arquivo original contava com 155807 linhas e a média de 484 palavras por linha. O arquivo obtido após a redução e limpeza atingiu 56984 linhas e tamanho médio de 84 palavras por linha. A partir daqui, o termo *Ohsumed* será usado para se referir à coleção obtida após o pré-processamento.

O conjunto *Tag My News* também foi pré-processado. Inicialmente, o tamanho do conjunto de dados era de 11,2 MB e possuía 260832 linhas. As coleções *News-Head* (apenas as manchetes) e *News-Short* (resumo da notícia) foram geradas a partir do conjunto de dados *Tag My News* e obtiveram tamanhos de 1,4 MB e 3,7 MB, respectivamente. Ambos arquivos gerados após o processo de limpeza tiveram 32604 linhas. O conjunto de dados *News-Head* ficou, em média, com 6 palavras por linha, enquanto o *News-Short* obteve 15 palavras por linha. A Tabela 1 traz informações sobre os conjuntos de dados usados neste trabalho, inclusive propriedades que foram modificadas após o pré-processamento.

4.1. Resultados

Os experimentos foram executados em um notebook ASUSTek K45A com um processador Intel(R) Core(TM) i5-3210M CPU @ 2.50GHz e memória RAM de 8GB. O sistema operacional utilizado foi o Linux Ubuntu 18.04.1 LTS, instalado em uma partição de 34GB. O *Java Runtime Environment*, necessário para execução dos algoritmos, rodava com a versão 10.0.2. Os algoritmos foram disponibilizados pelos autores das abordagens.

As métricas escolhidas para avaliação foram três métricas de coerência: C_V, C_{UMass}, C_A [Röder et al. 2015a], assim os resultados são avaliados em termos de coerência segundo as métricas citadas. Foi utilizada a ferramenta *Palmetto*

[1] www.mat.unical.it/OlexSuite/Datasets/SampleDataSets-about.htm

[2] http://acube.di.unipi.it/tmn-dataset

Atributo/ coleção	News-Head	News-Short	Ohsumed
Tamanho do arquivo (antes)	11,2 MB	11,2 MB	151,1 MB
Tamanho do arquivo (depois)	1,4 MB	3,7 MB	40,7 MB
Número de documentos	32604	32604	56984
Número de palavras (antes)	1340835	1340835	75405134
Número de palavras (depois)	197040	501655	4780938
Número de palavras únicas (antes)	159033	159033	155807
Número de palavras únicas (depois)	23710	37231	6982
Numero médio de palavras por documentos	6	15	84

Tabela 1. Características das coleções utilizadas neste trabalho.

(aksw.org/Projects/Palmetto.html) para calcular os resultados com base nas métricas. Essa ferrramenta utiliza uma base de dados com documentos extraídos da *Wikipedia* para avaliar os tópicos. A métrica C_V é baseada numa janela deslizante, um conjunto segmentado de *topwords*, uma confirmação indireta que usa informação mútua de pontos normalizados e similaridade do cosseno. Quanto maior o valor desta métrica, maior a coerência dos tópicos. Já a métrica C_A é baseada numa janela de contexto, uma comparação de pares de *topwords*, uma confirmação indireta que usa informação mútua de pontos normalizados e similaridade do cosseno. Quanto maior o valor desta métrica, maior a coerência dos tópicos. Por fim, a métrica C_{UMass} se baseia na contagem de co-ocorrências e uma probabilidade condicional logarítmica como medida de confirmação. Quanto maior o valor desta métrica, maior a coerência dos tópicos.

Para cada conjunto de dados foram feitas nove execuções das abordagens: três para 30 tópicos; três para 60 tópicos; e três para 120 tópicos ($i.e.$, K igual a 30, 60 e 120). A repetição de três execuções para cada cenário foi realizado a fim de mitigar a influência do fator aleatório. Os tópicos gerados foram avaliados segundo as métricas e o resultado final representa a média das três execuções considerando o tripé "algoritmo-conjunto de dados-número de tópicos".

Para todas as abordagens foram realizadas 100 iterações. Usou-se os hiperparâmetros indicados pelo artigo referente a cada abordagem para execução da mesma: (i) $\alpha = 50/K$ para BTM, WNTM, e $\alpha = 0.1/K$ para PTM (K corresponde ao número de tópicos), (ii) $\beta = 0.01$ para BTM, PTM e WNTM, (iii) $\alpha_2 = 0.15$ para PTM, e (iv) SATM com um *treshold* de 0.001.

A Figura 1 apresenta as top-10 palavras de um tópico escolhido a partir de $K = 120$ para todas as abordagens e coleções. Observando as palavras dos tópicos gerados, *tenis* seria o tópico para a coleção *News-Head*, *política americana* para a coleção *News-short* e *tratamento de patologia* para a coleção *Ohsumed*.

A Figura 2 apresenta o desempenho das abordagens na coleção *News-Short* para o número de tópicos definidos para a métricas C_V e C_A, respectivamente. Também apresenta a média das métricas utilizando os resultados nos três diferentes K.

A Figura 3 apresenta, como na figura anterior, o desempenho das abordagens na coleção *News-Head* para o número de tópicos definidos para as métricas C_V e C_A, respectivamente, além da média geral do desempenho.

Coleção/ abordagem	BTM	PTM	SATM	WNTM
News-Head	nadal open djokovic federer final french win wozniacki round lead	nadal djokovic federer win murray beats reach rome monte advance	murray officials final kentucky madrid nadal barcelona federer advance derby	nadal djokovic federer final indian win last murray advance wells
News-Short	president obama barack us said united states secretary would obamas	president obama federal us barack friday tuesday court deal program	ollanta vote showed race percent candidate poll june election presidential	president obama barack republican obamas governor presidential white run campaign
Ohsumed	treatment therapy treated two survival three study time years weeks	patients skin tissue one treatment two three study factors patient	cells one treatment two study may less disease group patients	treatment therapy three treated two time total study symptoms survival

Figura 1. Exemplo de tópicos gerados após a execução dos algoritmos (K=120).

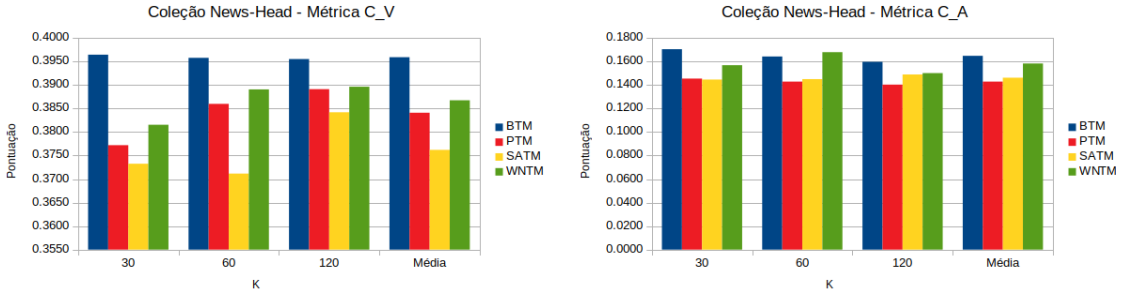

Figura 2. Resultados da C_V e C_A nos tópicos da coleção News-Head

.

Os resultados das métricas dos tópicos gerados pelas abordagens na coleção *Ohsumed* são apresentados na Figura 4. Finalmente, a Tabela 2 apresenta o desempenho das abordagens baseado na métrica C_{umass}.

Desempenho das abordagens utilizando a Métrica C_{UMass}									
Abordagem	**News-Head**			**News-Short**			**Ohsumed**		
	30	60	120	30	60	120	30	60	120
BTM	-3.34	-3.28	-3.41	-3.62	-3.78	-3.63	-3.79	-3.83	-3.98
PTM	-3.17	-3.49	-3.80	-3.46	-3.28	-3.68	-3.21	-3.34	-3.66
SATM	-3.06	-3.18	-3.66	-3.18	-3.49	-3.47	-1.56	-1.72	-2.10
WNTM	-2.87	-2.98	-3.42	-2.96	-3.17	-3.43	-3.91	-3.99	-3.97

Tabela 2. Desempenhos com a métrica C_{UMass}

As abordagens obtiveram um desempenho similar na maioria dos cenários e número de tópicos Foram distintas as abordagem que ficaram melhor ranqueadas em cada cenário. Considerando o quadro geral, o BTM e o PTM obtiveram mais coerência quando as métricas usadas foram C_V e C_A. Por outro lado, o SATM e o WNTM se mostraram mais coerentes quando a métrica usada foi a C_{UMass}. Para o quadro geral, considerando métricas, conjuntos de dados e número de tópicos, as execuções do BTM obtiveram os resultados mais coerentes em mais cenários do que as outras abordagens. A coerência

Outro ponto notado foi que, em geral, as abordagens usadas neste trabalho obtive-

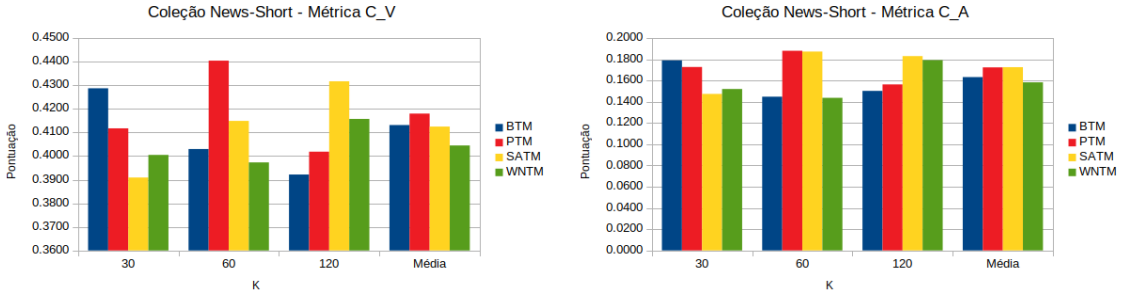

Figura 3. Resultados da C_V e C_A nos tópicos da coleção News-Short

.

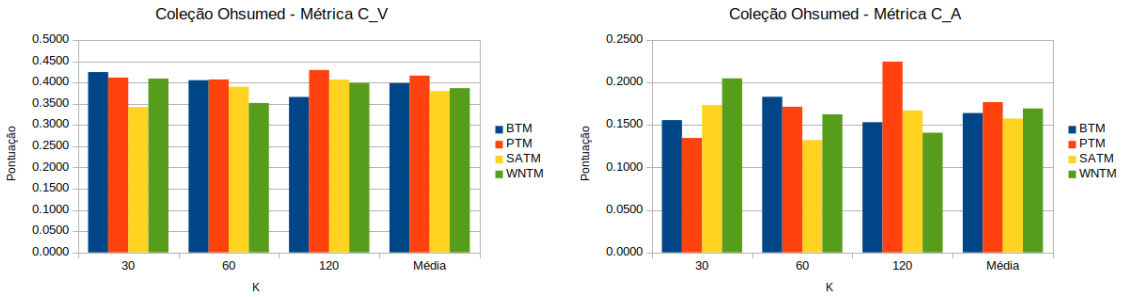

Figura 4. Resultados da C_V e C_A nos tópicos da coleção Ohsumed

.

ram melhor resposta com 30 ou 60 tópicos para as duas coleções com menor número de palavras por documento, *News-Head* e *News-Short*, com média de 6 e 15 palavras por documento, respectivamente, e obtiveram melhor resposta para o maior número de tópicos (120) com a coleção de maior número médio de palavras por documento, *Ohsumed*, com número médio de 84 palavras por documento.

5. Conclusão

Devido à dominância de textos curtos na Web, extrair tópicos de documentos curtos tornou-se uma tarefa cada vez mais importante e desafiadora. Com a ascensão das redes sociais na Web o número de textos aumentou consideravelmente. Em 2016, meio bilhão de *tweets* eram gerados por dia. Entretanto, devido a falta de co-ocorrência de palavras em coleções de documentos curtos, foi necessário o surgimento de novas abordagens, a fim de superar o problema dos dados esparsos em conjuntos de dados de textos curtos. Essas abordagens usaram diferentes premissas para atingir a finalidade de extrair tópicos de documentos curtos de modo coerente.

O BTM se apoiou na ideia de que se duas palavras que aparecem em um mesmo contexto fossem agrupadas (”termo-par”) e houvesse co-ocorrência de termos-pares na coleção, isso indicaria maior probabilidade destas duas palavras pertencerem ao mesmo tópico. O PTM baseou-se na premissa de que documentos de textos curtos pertencem a um pseudo-documento grande, mas que esse pseudo-documento grande era composto de vários tópicos distintos. Assim como o PTM, o SATM se respaldou no conceito de que pequenos textos formam um pseudo-documento grande, mas com uma distinção fundamental com relação ao PTM: para o SATM, cada pseudo-documento era composto de

pequenos documentos que integravam um único tópico. O WNTM usou como base a concepção de que era possível formar redes de palavras, ligando as palavras que aparecem próximas, como se fosse um grafo, e então gerando um pseudo-documento para assim diminuir o problema dos dados esparsos e desbalanceados.

Este trabalho avaliou o uso destas quatro abordagens que surgiram visando resolver o problema dos dados esparsos e possibilitar a extração de tópicos em conjuntos de dados de textos curtos de um modo coerente. Para isso, cenários diferentes foram apresentados, variando no número de tópicos (30, 60 e 120) e no número médio de palavras por documento (6, 15, 84).

Considerando o quadro geral, o BTM foi a abordagem que mais superou as outras na maior quantidade de casos. Apoiado pelas métricas C_V e C_A o BTM foi a que teve maior pontuação média nas execuções sobre os dois conjuntos de dados com menor número médio de palavras por documento (6 e 15) e o PTM foi a abordagem melhor ranqueada sobre a coleção com maior número médio de palavras por documento (84). A métrica C_{UMass} apontou como melhor abordagem o WNTM, se tratando dos dois conjuntos de dados com menor número médio de palavras por documento (6 e 15), e o SATM, referindo-se à coleção de documentos com maior número médio de palavras por documento (84).

Pode-se citar algumas direções para trabalhos futuros: (i) utilizar todas as métricas propostas em [Röder et al. 2015b] e avaliar a correlação entre os resultados das mesmas e (ii) utilizar uma abordagem não paramétrica para extração dos tópicos (K é calculado automaticamente).

Referências

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4).

Cheng, X., Yan, X., Lan, Y., and Guo, J. (2014). Btm: Topic modeling over short texts. *IEEE Transactions on Knowledge and Data Engineering*, 26(12):2928–2941.

Quan, X., Kit, C., Ge, Y., and Pan, S. J. (2015). Short and sparse text topic modeling via self-aggregation. In *IJCAI*, pages 2270–2276.

Röder, M., Both, A., and Hinneburg, A. (2015a). Exploring the space of topic coherence measures. In *Proceedings of the eight International Conference on Web Search and Data Mining, Shanghai, February 2-6*.

Röder, M., Both, A., and Hinneburg, A. (2015b). Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining*, pages 399–408. ACM.

Steyvers, M. and Griffiths, T. (2007). Probabilistic topic models. *Handbook of latent semantic analysis*, 427(7):424–440.

Zuo, Y., Wu, J., Zhang, H., Lin, H., Wang, F., Xu, K., and Xiong, H. (2016a). Topic modeling of short texts: A pseudo-document view. In *Proceedings of the 22nd ACM SIGKDD*, pages 2105–2114. ACM.

Zuo, Y., Zhao, J., and Xu, K. (2016b). Word network topic model: a simple but general solution for short and imbalanced texts. *Knowledge and Information Systems*, 48(2):379–398.