**UNIVERSIDADE FEDERAL DA FRONTEIRA SUL**
**CAMPUS CHAPECÓ**
**CURSO DE CIÊNCIA DA COMPUTAÇÃO**

**MICHEL CHAGAS DA COSTA**

# AVALIAÇÃO DE ABORDAGENS PROBABILÍSTICAS DE EXTRAÇÃO DE TÓPICOS EM DOCUMENTOS CURTOS

**CHAPECÓ**
**2018**

**MICHEL CHAGAS DA COSTA**

# AVALIAÇÃO DE ABORDAGENS PROBABILÍSTICAS DE EXTRAÇÃO DE TÓPICOS EM DOCUMENTOS CURTOS

Trabalho de conclusão de curso de graduação apresentado como requisito para obtenção do grau de Bacharel em Ciência da Computação da Universidade Federal da Fronteira Sul.

Orientador: Prof. Dr. Denio Duarte

CHAPECÓ

2018

Costa, Michel Chagas da

Avaliação de abordagens probabilísticas de extração de tópicos em documentos curtos / por Michel Chagas da Costa. – 2018.
35 f.: il.; 30 cm.

Orientador: Denio Duarte
Monografia (Graduação) - Universidade Federal da Fronteira Sul, Ciência da Computação, Curso de Ciência da Computação, RS, 2018.

1. Modelagem de tópicos. 2. Textos curtos. 3. LDA. 4. Aprendizado de máquina. 5. Avaliação. I. Duarte, Denio. II. Título.

© 2018
Todos os direitos autorais reservados a Michel Chagas da Costa. A reprodução de partes ou do todo deste trabalho só poderá ser feita mediante a citação da fonte.
E-mail: costa.michel10@hotmail.com

**MICHEL CHAGAS DA COSTA**

# AVALIAÇÃO DE ABORDAGENS PROBABILÍSTICAS DE EXTRAÇÃO DE TÓPICOS EM DOCUMENTOS CURTOS

Trabalho de conclusão de curso de graduação apresentado como requisito para obtenção do grau de Bacharel em Ciência da Computação da Universidade Federal da Fronteira Sul.

Orientador: Prof. Dr. Denio Duarte

Este trabalho de conclusão de curso foi defendido e aprovado pela banca em: ____/____/____

BANCA EXAMINADORA:

Dr. Denio Duarte - UFFS

Dr. Guilherme Dal Bianco - UFFS

Me. Andressa Sebben - UFFS

em representação de:

# RESUMO

Devido ao amplo uso das redes sociais, textos pequenos se popularizaram na Web. A possibilidade de interação entre usuários, como comentários, fez com que um grande número de textos curtos surgissem dia após dia. Extrair tópicos de uma grande quantidade de textos curtos tornou-se uma tarefa crítica e desafiadora em tarefas de análise de conteúdo [2]. Foram propostos novos meios de inferir tópicos de conjuntos de dados de textos curtos em vez do uso ferramentas já conhecidas na modelagem de tópicos. Este trabalho avaliará o uso de algumas destas abordagens probabilísticas na extração de tópicos em documentos curtos.

Palavras-chave: Modelagem de tópicos. Textos curtos. LDA. Aprendizado de máquina. Avaliação.

# LISTA DE FIGURAS


Figura 2.1 – Distribuição de tópicos por palavras [1] ..... 13
Figura 2.2 – Modelo generativo de inferência de tópicos [7] ..... 13
Figura 3.1 – Os nós representam as palavras de uma coleção e os pesos das arestas representam o número de ligações entre duas palavras [9] ..... 20
Figura 5.1 – Trecho da coleção *Tag My News* antes do pré-processamento ..... 23
Figura 5.2 – Trecho da coleção *News-Short* depois do pré-processamento ..... 23
Figura 6.1 – Exemplo de tópicos gerados após a execução dos algoritmos (para 120 tópicos) ..... 26
Figura 6.2 – Pontuações sobre a coleção *News-Head* utilizando a métrica $C_V$ ..... 29
Figura 6.3 – Pontuações sobre a coleção *News-Head* utilizando a métrica $C_A$ ..... 30
Figura 6.4 – Pontuações sobre a coleção *News-Head* utilizando a métrica $C_{UMass}$ ..... 30
Figura 6.5 – Pontuações sobre a coleção *News-Short* utilizando a métrica $C_V$ ..... 31
Figura 6.6 – Pontuações sobre a coleção *News-Short* utilizando a métrica $C_A$ ..... 31
Figura 6.7 – Pontuações sobre a coleção *News-Short* utilizando a métrica $C_{UMass}$ ..... 31
Figura 6.8 – Pontuações sobre a coleção *Ohsumed* utilizando a métrica $C_V$ ..... 32
Figura 6.9 – Pontuações sobre a coleção *Ohsumed* utilizando a métrica $C_A$ ..... 32
Figura 6.10 – Pontuações sobre a coleção *Ohsumed* utilizando a métrica $C_{UMass}$ ..... 32

# LISTA DE TABELAS

Tabela 5.1 – Informações sobre conjuntos de dados antes e após pré-processamento ..... 24
Tabela 6.1 – BTM ..... 27
Tabela 6.2 – PTM ..... 27
Tabela 6.3 – SATM ..... 27
Tabela 6.4 – WNTM ..... 28
Tabela 6.5 – $C_V$ ..... 28
Tabela 6.6 – $C_A$ ..... 28
Tabela 6.7 – $C_{UMass}$ ..... 29

# LISTA DE ABREVIATURAS E SIGLAS

| | |
|---|---|
| LDA | *Latent Dirichlet Allocation* |
| BTM | *Biterm Topic Model* |
| PTM | *Pseudo-document based Topic Model* |
| SATM | *Self-Aggregation based Topic Model* |
| WNTM | *Word Network Topic Model* |

# SUMÁRIO

**1 INTRODUÇÃO** ..... 10
**2 MODELAGEM DE TÓPICOS** ..... 12
**2.1 Textos curtos** ..... 15
**3 ABORDAGENS PARA MODELAGEM DE TÓPICOS EM TEXTOS CURTOS** ... 17
**3.1 BTM** ..... 17
**3.2 SATM** ..... 18
**3.3 PTM** ..... 18
**3.4 WNTM** ..... 19
**4 TRABALHOS RELACIONADOS** ..... 21
**5 PROJETO DO EXPERIMENTO** ..... 22
**6 RESULTADOS** ..... 25
**7 CONCLUSÃO** ..... 33
**REFERÊNCIAS** ..... 35

# 1 INTRODUÇÃO

Textos curtos dominam a Web, tanto no contexto de sites tradicionais - títulos de páginas, anúncios, legendas de imagens, mensagens em fóruns, títulos de notícias, etc - quanto mídias sociais, que tiveram um grande crescimento, como *tweets* e mensagens de status [2]. Há um número muito grande de textos curtos, o qual está em rápido e constante crescimento. Um exemplo disso é o *Twitter*, que, segundo Zuo, em 2016, com 250 milhões de usuários ativos gerava aproximadamente meio bilhão de *tweets* por dia [8]. 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 probabilísticos para modelagem de tópicos têm sido usados 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. O *Latent Dirichlet Allocation* (LDA), abordagem que possui a forma mais simples de modelagem de tópicos e serve como base para outras abordagens, também assume a premissa acima citada.

Extrair tópicos de uma coleção de documentos permite gerenciar esta coleção, como por exemplo, através da categorização de temas presentes na mesma. Normalmente, há um mecanismo de busca onde o usuário digita um ou mais termos e encontra um documento, o qual pode levar a outros documentos relacionados através de links. Usando a modelagem de tópicos, seria possível organizar a coleção de documentos de tal modo que o usuário pudesse partir de um tópico específico para um tópico mais generalizado, ou ao contrário, como se pudesse aplicar "zoom" na coleção de documentos [1]. Com o surgimento de um grande número de textos curtos, grande parte devido ao massivo uso de mídias sociais, descobrir tópicos tornou-se importante para uma gama aplicações de análise de conteúdo, como classificação, perfil de interesse de usuário, auxílio no rastreamento de tópicos populares [2] [3] e sugestões de pesquisa.

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 [2] [8] [9]. Dados esparsos constituem o principal

problema na modelagem de tópicos em documentos curtos [4]. 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 irá avaliar o uso de abordagens probabilísticas para a extração de tópicos em documentos curtos. Será utilizado quatro abordagens: BTM, PTM, SATM e WNTM. Este trabalho avaliará os resultados da execução destas quatro abordagens sobre três conjuntos de dados através de três métricas de coerência apresentadas por Röder et al [5]: $C_V$, $C_{UMass}$ e $C_A$. Os conjuntos de dados utilizados possuíam tamanho médio 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.

No próximo capítulo será apresentada uma visão geral sobre modelagem de tópicos, partindo de conceitos e premissas básicas e apresentando o LDA, um modelo de tópicos que serve como base para outros. No terceiro capítulo serão apresentadas as abordagens que focam em textos curtos que serão usadas neste trabalho. No quarto capítulos serão apresentados trabalhos relacionados a este. Finalmente, serão apresentados os experimentos, os resultados e a conclusão.

# 2 MODELAGEM DE TÓPICOS

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.

Dado um conjunto de textos não-organizados, os algoritmos de modelagem de tópicos têm como objetivo descobrir os principais tópicos, que podem ser vistos como assuntos, relacionados ao conjunto. 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 [1].

Os tópicos surgem a partir da análise do documento original, e, para fins de aprendizado de máquina, não necessitam de um rótulo. Desta forma, o aprendizado de máquina é não-supervisionado. Em suma, a modelagem de tópicos provê uma solução para gerenciar grandes quantidades de texto.

A forma mais simples de modelar tópicos é através da Alocação Latente de Dirichlet - *Latent Dirichlet Allocation* (LDA) [1]. O LDA serve como base para vários outros modelos de tópicos, inclusive os modelos de tópicos descritos no próximo capítulo.

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.

Imagine que se tivesse um texto em mãos e se decidisse marcar as palavras deste texto relacionadas a cada assunto com uma cor, como na Figura 2.1. Esta intuição é a que está por trás da implementação do modelo estatístico LDA [1].

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.

Um modelo de tópicos é um modelo generativo: através de procedimento estatístico,

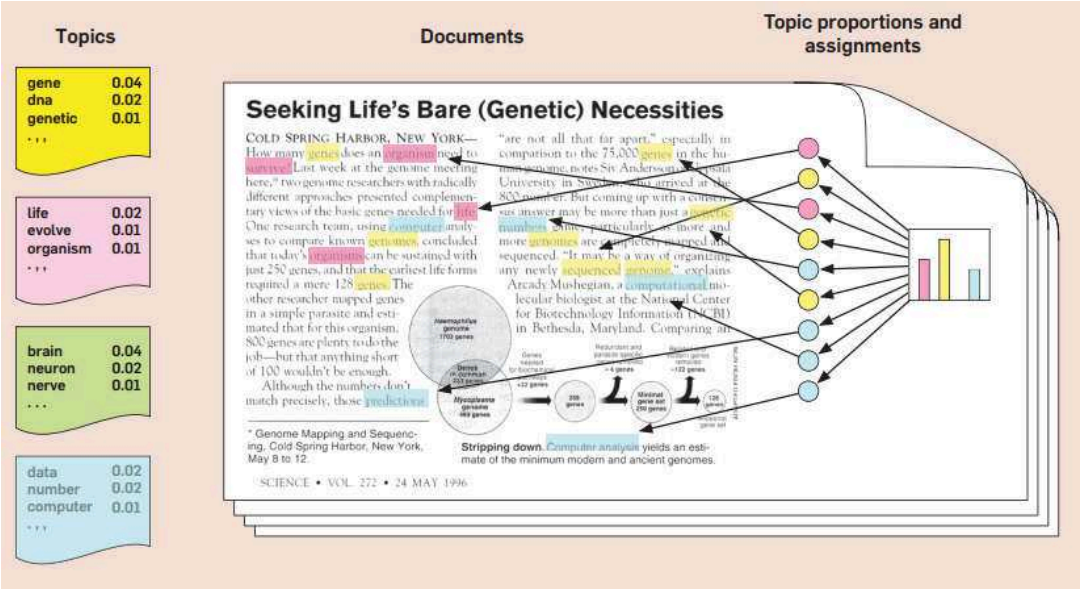


Figura 2.1: Distribuição de tópicos por palavras [1]

documentos podem ser gerados. Parte-se do pressuposto que vários tópicos geram um documento [1]. O processo inverso à geração de documentos, que é o objetivo da modelagem de tópicos, dada uma coleção de documentos, é descobrir quais tópicos foram responsáveis por gerar aqueles documentos.

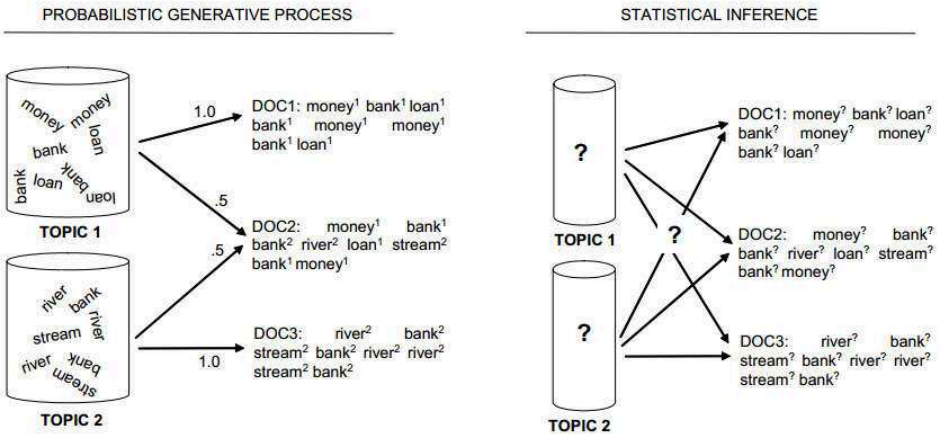


Figura 2.2: Modelo generativo de inferência de tópicos [7]

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 [1], 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.

A estrutura de tópicos, em uma coleção de documentos, é a distribuição de tópicos por documento e palavras por tópico. Os documentos são conhecidos, mas a estrutura de tópicos é algo oculto. Descobrir esta estrutura de tópicos que, dada a coleção de documentos, está oculta, é o problema computacional central da modelagem de tópicos. Isso pode ser pensado como um processo de geração reversa de tópicos.

Juntamente com outros modelos de tópicos, o LDA faz parte de um grande campo de modelagem probabilística. Na modelagem probabilística generativa, os dados são tratados levando em conta as "variáveis ocultas". Esse processo generativo define uma distribuição de probabilidade conjunta sobre as variáveis conhecidas e as variáveis ocultas. Usa-se esta distribuição conjunta para computar a distribuição condicional (também conhecida como distribuição posterior) das variáveis ocultas, dada as variáveis observadas (conhecidas).

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).

Algoritmos de modelagem de tópico geralmente se enquadram em duas categorias: algoritmos baseados em exemplos e "algoritmos variacionais" [1]. Ambos tentam achar uma estrutura de tópicos a partir de uma coleção de documentos. Algoritmos baseados em exemplos tentam coletar exemplos do posterior e fazer uma aproximação usando uma distribuição estatística. Comumente, o algoritmo usado para isso é o *Gibbs sampling* [1], no qual são construídas cadeias de Markov (sequência de variáveis randômicas, onde cada variável depende da variável anterior). Essas cadeias são aproximações da estrutura de tópicos (variáveis ocultas). Normalmente, é encontrada apenas uma cadeia com probabilidade máxima.

Algoritmos variacionais são uma alternativa determinística aos algoritmos baseados em exemplos. Em vez de coletar exemplos, eles colocam um conjunto parametrizada de distribuições sobre as variáveis ocultas. Então, se procura descobrir qual membro deste conjunto é o mais próximo do posterior, usando a divergência de *Kullback-Leiber* [1], métrica que calcula a distância entre duas distribuições probabilísticas. O LDA pode ser implementado com algoritmos baseados em exemplos ou algoritmos variacionais [1]. As quatro abordagens que serão usadas neste trabalho usam algoritmos baseados em exemplos [2, 4, 9, 8].

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. 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.

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 [2]. 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 [4]. É necessário, para textos curtos, usar outras abordagens, que se provam mais coerentes neste contexto.

As abordagens de extração de tópicos para coleções de textos curtos usados neste trabalho serão apresentados no próximo capítulo.

# 3 ABORDAGENS PARA MODELAGEM DE TÓPICOS EM TEXTOS CURTOS

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 serão apresentadas neste capítulo.

## 3.1 BTM

O *Biterm Topic Model* (BTM [2]) 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 [2], 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.

Sendo termo-par um par não ordenado de palavras em um contexto pequeno, como um texto curto, um documento com três palavras distintas, por exemplo, geraria três termos-pares:

(p1, p2, p3) => {(p1,p2), (p1,p3), (p2,p3)}.

Após extrair os termos-pares de cada documento, o *corpus* passa a ser um conjunto de termos-pares. A ideia chave por trás de disso é que se duas palavras são encontradas frequentemente juntas, então elas provavelmente pertencem ao mesmo tópico.

## 3.2 SATM

O modelo de tópicos baseado em auto-agregação – *Self-Aggregation based Topic Model* (SATM [4]) – 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.

Em particular, o SATM assume que textos curtos são a consequência do desmembramento de textos longos, que poderiam ser gerados usando um modelo de tópicos convencional (pseudo-documento). Encontrar a correspondência entre o texto curto e o pseudo-documento longo que poderia ser gerado através de modelos convencionais é parte crítica para modelar tópicos com sucesso no SATM.

Diferentemente de outras abordagens recentes, que assumem que cada texto curto poderia corresponder a mais de um tópico, o SATM assume que cada texto curto está relacionado a apenas um tópico.

## 3.3 PTM

Movido pelo potencial dos métodos de agregação, como o SATM, para lidar com os dados esparsos, um modelo de tópicos baseado em pseudo-documento – *Pseudo-document based Topic Model* (PTM [8]) – para textos curtos foi proposto por Zuo et al [8]. 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.

O PTM assume que um grandioso volume de textos curtos é gerado por uma quantidade muito menor de pseudo-documentos. Ao transformar textos curtos em pseudo-documentos, pode ser que a coleção fique muito pequena, tendendo a ter ambiguidade. Para tratar isso, Zuo

et al [8] propõem, no mesmo artigo, o SPTM, que aplica a distribuição de Spike e Slab na distribuição de tópicos por documento. O SPTM demonstrou ter melhor desempenho que o PTM apenas quando o número de pseudo-documentos é relativamente pequeno. Neste trabalho optou-se pelo uso do PTM.

## 3.4 WNTM

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 – *Word Network Topic Model* (WNTM [9]) – 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.

Usa-se o algoritmo *Gibbs sampling* do LDA para fazer a inferência da estrutura oculta (latente) e aprender a distribuição de tópicos por palavras em vez de tópicos por documento. Isso faz do WNTM menos sensível ao tamanho dos documentos e a heterogeneidade da distribuição de tópicos. Além disso, a rede de palavras pode ser construída para qualquer tipo de texto, o que faz desta abordagem simples e genérica em aplicações no mundo real.

Em uma rede de co-ocorrências de palavra (que pode ser denotada por rede de palavras), os nós representam uma palavra do *corpus* e as arestas representam a ligação ocorrida entre duas palavras, pelo menos uma vez, em um mesmo contexto. Aqui, o contexto é um documento ou uma janela de tamanho fixo.

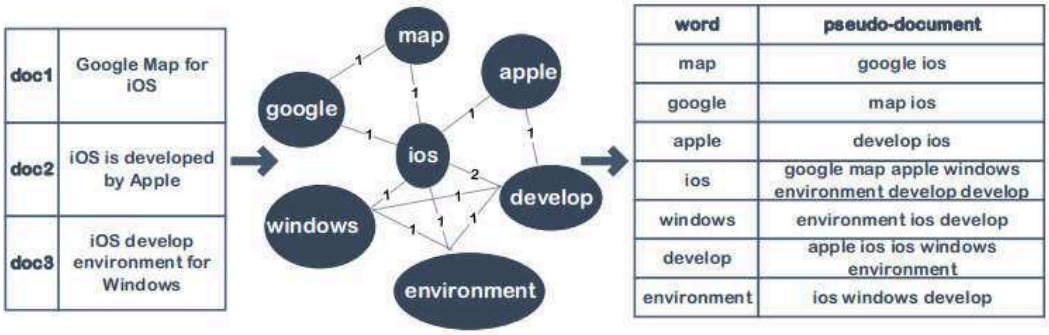


Figura 3.1: Os nós representam as palavras de uma coleção e os pesos das arestas representam o número de ligações entre duas palavras [9]

Para converter a coleção de documentos em uma rede de palavras, o primeiro passo é filtrar as *stopwords* e palavras de baixa frequência. Então, a janela de tamanho fixo é aplicada fazendo um escaneamento em cada documento. Se palavras distintas aparecem na mesma janela, uma aresta liga os nós na rede. O peso das arestas indica a quantidade de vezes que duas palavras aparecem na mesma janela (vide figura 3.1).

Antes de usar o *Gibbs sampling*, a rede de palavras é transformada em um conjunto de pseudo-documentos, onde cada palavra da rede pode ser tratada como um novo pseudo-documento, que possui em seu conteúdo as palavras (nós) ligados à palavra que originou aquele pseudo-documento. As palavras podem aparecer múltiplas vezes, conforme o seu peso. Por exemplo, se a palavra "develop" está ligada com a palavra "iOS" na rede, e o peso da aresta é dois, quando for gerar o pseudo-documento da palavra "develop", a palavra "iOS" aparecerá duas vezes, aumentando a probabilidade destas palavras estarem associadas ao mesmo tópico.

As quatro abordagens apresentadas neste capítulo terão seu uso avaliado neste trabalho, visando a extração de tópicos em conjuntos de dados de documentos curtos. Será apresentado qual abordagem se comportou melhor em cada cenário, considerando o número de tópicos e a métrica utilizada para indicar a coerência.

No próximo capítulo serão apresentados trabalhos relacionados a este.

# 4 TRABALHOS RELACIONADOS

Neste capítulo serão apresentados alguns trabalhos relacionados a este. 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 [2], 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 [8]: SATM, LDA, *Mixture of Unigrams* e *Dual Sparse Topic Model*. Para avaliação foi utilizado 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 [4], 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 artigo em que o WTNM é proposto [9], Zuo et al comparam esta abordagem com o BTM e o LDA. Com base na validação cruzada, os autores concluem que o WNTM se mostrou melhor que o BTM e o LDA, sendo o número de tópicos 100.

No próximo capítulo serão apresentados como foram projetados os experimentos realizados neste trabalho.

# 5 PROJETO DO EXPERIMENTO

Nesta seção serão apresentadas como foram projetadas as avaliações das quatro abordagens probabilísticas de modelagem de tópicos em coleções de textos curtos. Serão descritos os conjuntos de dados, as configurações máquina usada para execução, as métricas e os parâmetros para cada algoritmo. Os resultados dos experimentos serão apresentados no próximo capítulo.

Foram selecionados dois conjuntos de dados: *Ohsumed* e *Tag My News*. 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.

Foi realizado uma limpeza no conjunto de dados *Tag My News*, a fim de extrair apenas o título e as descrições curtas. Originalmente, o arquivo contava com outras informações como data, hora, categoria e outros. Também foi retirado do conjunto de dados pontuações, *stopwords* e palavras de baixa frequência. Deste processo, originou-se duas coleções: uma coleção contendo apenas os títulos das notícias, o qual será referido como *News-Head*; e outra que armazenou apenas as descrições curtas das notícias, o qual será referido como *News-Short*. A Figura 5.1 apresenta um exemplo da coleção *Tag My News* antes do pré processamento e a Figura 5.2 mostra um exemplo após o pré-processamento, que originou a coleção *News-Short*.

O tamanho do conjunto de dados original *Tag My News* era de 11,2 MB e possuía 260832 linhas. As coleções *News-Head* e *News-Short*, que foram geradas a partir do conjunto de dados *Tag My News* 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 5.1 traz informações sobre os conjuntos de dados usados neste trabalho, inclusive

```
court agrees to expedite n.f.l.'s appeal
the decision means a ruling could be made nearly tw
http://feeds1.nytimes.com/~r/nyt/rss/sports/~3/nbjc
0
04 May 2011 07:39:03
nyt
sport

investing: can you profit in agricultural commoditi
bad weather is one factor behind soaring food price
http://rssfeeds.usatoday.com/~r/usatodaycommoney-tc
1
20 May 2011 15:13:57
ut
business
```

Figura 5.1: Trecho da coleção *Tag My News* antes do pré-processamento

```
decision means ruling could made nearly two months regular season
bad weather one factor behind soaring food prices make hay farm st
though jack warners threatened soccer tsunami remains stuck doldru
joshua jacksons show goes bang plus amazing race nears finish line
fbi asking publics help deciphering two encrypted notes found man
new documentary hollywood producer music promoter jerry weintraub
disneyland paris theme park closed thunder mountain train ride mon
hewlettpackard ceo leo apotheker detailed nextgeneration strategy
follow companys expense policy held liable corporate card balance
security court sentenced prominent shiite cleric eight others year
arizonas highest court put hold execution scheduled wednesday apar
benn ferriero scored playoff debut deflected goal gave san jose sh
hollywood made plenty disaster movies earthquake tsunami struck ja
floridas little man erving walker came big again scoring points hi
authorities thursday released surveillance images said showed atta
```

Figura 5.2: Trecho da coleção *News-Short* depois do pré-processamento

propriedades que foram modificadas após o pré-processamento.

A máquina usada foi 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.

As métricas escolhidas para avaliação foram três métricas de coerência: $C_V$, $C_{UMass}$, $C_A$ [6]. Foi utilizada a ferramenta *Palmetto* para calcular os resultados com base nas métricas.

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.

Tabela 5.1: Informações sobre conjuntos de dados antes e após pré-processamento

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

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.

A métrica $C_{UMass}$ é baseada 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. A repetição de três execuções para cada cenário foi realizado a fim de mitigar a influência do fator aleatório.

Para todas as abordagens foram realizadas 100 iterações. Usou-se os hiper-parâmetros indicados pelo artigo referente a cada abordagem para execução da mesma:

- BTM: *alpha* $50/K$, sendo K o número de tópicos; *beta* 0.01.
- PTM: *alpha* 0.1; *alpha2* 0.15; *beta* 0.01.
- SATM: *beta* 0.1; *treshold* 0.001;
- BTM: *alpha* $50/K$, sendo K o número de tópicos; *beta* 0.01.

Por fim, para cada saída das execuções foi utilizado a ferramenta *Palmetto* para calcular o resultado considerando cada uma das métricas. Esta ferramenta usa como auxilio uma base de dados da *Wikipedia*. O resultado final representa a média das três execuções considerando o tripé "algoritmo-conjunto de dados-número de tópicos".

No próximo capítulo serão mostrados os resultados do experimento projetado neste capítulo.

# 6 RESULTADOS

Neste capítulo serão apresentados os resultados obtidos após a execução dos algoritmos referentes as abordagens BTM, PTM, SATM e WNTM e o cálculo utilizando as métricas de coerência $C_V$, $C_{UMass}$ e $C_A$ através da ferramenta *Palmetto*. Inicialmente, serão mostrados todos os resultados referentes a cada abordagem.

Na Figura 6.1 há um exemplo de tópicos gerados em cada abordagem, para uma execução com 120 tópicos, considerando 10 *topwords* (palavras que melhor descrevem um tópico). A escolha dos tópicos foi realizada de forma aleatória: escolheu-se uma palavra e a partir dessa palava foi buscado tópicos com essa mesma palavra em todas as abordagens. Para a coleção *News-Head* a palavra escolhida foi "nadal". Para a coleção *News-Short*, "president"foi a palavra escolhida. Para a coleção *Ohsumed* a palavra escolhida foi "treatment". As palavras foram escolhidas devido às palavras do seu tópico representarem provavelmente um mesmo tópico, com base em um conhecimento prévio. Por exemplo, as palavras da coleção *News-Head* na Figura 6.1 provavelmente indicam o tópico "Tênis". As palavras da coleção *News-Short* indicam tópicos sobre o "presidente Obama"ou sobre a "corrida presidencial". O SATM foi a única abordagem que não gerou tópicos explícitos com as palavras "president Obama". As palavras do conjunto de dados *Ohsumed* indicam um provável "estudo de tratamento de uma doença".

Nas Tabelas 6.1, 6.2, 6.3 e 6.4, K é o número de tópicos e cada métrica mostra a média das três execuções no mesmo cenário (Abordagem-Coleção-Número de tópicos) e o desvio padrão.

A Tabela 6.1 apresenta os resultados das execuções do BTM. Considerando a métrica $C_V$, o BTM teve maior pontuação para: *News-Head* a maior pontuação foi com 30 tópicos; as execuções sobre o conjunto de dados *News-Short* obtiveram maior pontuação com 30 tópicos; e para *Ohsumed*, as execuções com 120 tópicos foram as que obtiveram maior coerência. A métrica $C_A$ mostrou resultados diferentes da métrica $C_V$: as execuções para os conjuntos de dados *News-Head* e *Ohsumed* obtiveram maior coerência com 30 tópicos; as execuções do BTM para a coleção *News-Short* para 60 tópicos obtiveram melhor pontuação que 30 ou 120 tópicos. Considerando a métrica $C_{UMass}$, a melhor pontuação, levando em conta coleção e número de tópicos foi: para a coleção *News-Head* 60 tópicos; e para as coleções *News-Short* e *Ohsumed* 30 tópicos.

Os resultados para as execuções do PTM são apresentados na Tabela 6.2. Com base

na métrica $C_V$, os resultados obtidos apontam o seguinte: para *News-Head* e *News-Short*, as execuções com 120 tópicos foram mais coerentes; e para a coleção *Ohsumed* as execuções com 30 tópicos foram melhores que 60 e 120 tópicos. Considerando a métrica $C_A$, as execuções com 30 tópicos tiveram pontuações mais altas que as execuções com 60 e 120 tópicos, independente do conjunto de dados. E para a métrica $C_{UMass}$ as execuções com 30 tópicos demonstraram ser mais coerentes nas coleções *News-Head* e *Ohsumed*, enquanto as execuções com 60 tópicos foram melhores no conjunto de dados *News-Short*.

Na Tabela 6.3 é mostrado os resultados das execuções para a abordagem SATM. Para a métrica $C_V$, os seguintes resultados foram obtidos: para os conjuntos de dados *News-Head* e *Ohsumed* as execuções com 120 tópicos foram mais coerentes que as execuções com 30 e 60 tópicos; e para a coleção *News-Short* as execuções com 60 tópicos apresentaram melhor pontuação. Com base na métrica $C_A$, as execuções com melhor pontuação foram as seguintes: 120 tópicos para as coleções *News-Head* e *News-Short*; e 30 tópicos para *Ohsumed*. E para $C_{UMass}$ as execuções com 30 tópicos obtiveram pontuação melhor independente do conjunto de dados.

Os resultados das execuções do WNTM são mostrados na Tabela 6.4. Considerando a métrica de coerência $C_V$, observou-se o seguinte: para os conjuntos de dados *News-Head* e *Ohsumed* as execuções com 120 tópicos obtiveram melhores pontuações que as execuções para 30 e 60 tópicos; e para *News-Short* as melhores execuções foram com 60 tópicos. Com base em $C_A$, obteve-se como pontuações mais altas: para as coleções *News-Head* e *Ohsumed* 60 tópicos; e para *News-Short* 30 tópicos. E para a métrica de coerência $C_{UMass}$, todas as execuções com 30 tópicos, em todas as coleções usadas neste trabalho, obtiveram pontuação melhor que as execuções com 60 e 120 tópicos.

| 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 6.1: Exemplo de tópicos gerados após a execução dos algoritmos (para 120 tópicos)

Tabela 6.1: BTM

| **Coleção** | **K** | $C_V$ | $C_A$ | $C_{UMass}$ |
|---|---|---|---|---|
| News-Head | 30 | **0.3965** $\pm$ 0.0009 | **0.1705** $\pm$ 0.0163 | -3.3416 $\pm$ 0.2437 |
| News-Head | 60 | 0.3958 $\pm$ 0.0025 | 0.1643 $\pm$ 0.0096 | **-3.2748** $\pm$ 0.1342 |
| News-Head | 120 | 0.3955 $\pm$ 0.0044 | 0.1598 $\pm$ 0.0025 | -3.4143 $\pm$ 0.1814 |
| News-Short | 30 | **0.4405** $\pm$ 0.0053 | **0.1882** $\pm$ 0.0140 | **-3.6217** $\pm$ 0.2223 |
| News-Short | 60 | 0.4317 $\pm$ 0.0034 | 0.1831 $\pm$ 0.0085 | -3.7769 $\pm$ 0.0780 |
| News-Short | 120 | 0.4288 $\pm$ 0.0014 | 0.1791 $\pm$ 0.0051 | -3.6311 $\pm$ 0.0213 |
| Ohsumed | 30 | 0.4076 $\pm$ 0.0067 | **0.1716** $\pm$ 0.0076 | **-3.7935** $\pm$ 0.4409 |
| Ohsumed | 60 | 0.4072 $\pm$ 0.0042 | 0.1672 $\pm$ 0.0041 | -3.8294 $\pm$ 0.1782 |
| Ohsumed | 120 | **0.4249** $\pm$ 0.0032 | 0.1559 $\pm$ 0.0017 | -3.9789 $\pm$ 0.0581 |

Tabela 6.2: PTM

| **Coleção** | **K** | $C_V$ | $C_A$ | $C_{UMass}$ |
|---|---|---|---|---|
| News-Head | 30 | 0.3772 $\pm$ 0.0119 | **0.1455** $\pm$ 0.0071 | **-3.1731** $\pm$ 0.1219 |
| News-Head | 60 | 0.3860 $\pm$ 0.0048 | 0.1429 $\pm$ 0.0047 | -3.4928 $\pm$ 0.0241 |
| News-Head | 120 | **0.3891** $\pm$ 0.0037 | 0.1404 $\pm$ 0.0091 | -3.8028 $\pm$ 0.1613 |
| News-Short | 30 | 0.4019 $\pm$ 0.0047 | **0.1564** $\pm$ 0.0103 | -3.4563 $\pm$ 0.0413 |
| News-Short | 60 | 0.4005 $\pm$ 0.0039 | 0.1521 $\pm$ 0.0060 | **-3.2831** $\pm$ 0.0347 |
| News-Short | 120 | **0.4030** $\pm$ 0.0017 | 0.1449 $\pm$ 0.0027 | -3.6838 $\pm$ 0.0658 |
| Ohsumed | 30 | **0.4298** $\pm$ 0.0079 | **0.2247** $\pm$ 0.0070 | **-3.2095** $\pm$ 0.2055 |
| Ohsumed | 60 | 0.4096 $\pm$ 0.0013 | 0.2050 $\pm$ 0.0071 | -3.3351 $\pm$ 0.1544 |
| Ohsumed | 120 | 0.4058 $\pm$ 0.0047 | 0.1834 $\pm$ 0.0054 | -3.6581 $\pm$ 0.1717 |

Tabela 6.3: SATM

| **Coleção** | **K** | $C_V$ | $C_A$ | $C_{UMass}$ |
|---|---|---|---|---|
| News-Head | 30 | 0.3733 $\pm$ 0.0083 | 0.1447 $\pm$ 0.0081 | **-3.0576** $\pm$ 0.1389 |
| News-Head | 60 | 0.3712 $\pm$ 0.0079 | 0.1450 $\pm$ 0.0032 | -3.1802 $\pm$ 0.2030 |
| News-Head | 120 | **0.3842** $\pm$ 0.0005 | **0.1490** $\pm$ 0.0020 | -3.6563 $\pm$ 0.1544 |
| News-Short | 30 | 0.3909 $\pm$ 0.0044 | 0.1475 $\pm$ 0.0135 | **-3.1756** $\pm$ 0.2095 |
| News-Short | 60 | **0.3973** $\pm$ 0.0065 | 0.1438 $\pm$ 0.0031 | -3.4906 $\pm$ 0.1058 |
| News-Short | 120 | 0.3922 $\pm$ 0.0050 | **0.1504** $\pm$ 0.0059 | -3.4707 $\pm$ 0.1693 |
| Ohsumed | 30 | 0.3425 $\pm$ 0.0026 | **0.1736** $\pm$ 0.0116 | **-1.5555** $\pm$ 0.0181 |
| Ohsumed | 60 | 0.3521 $\pm$ 0.0070 | 0.1627 $\pm$ 0.0068 | -1.7188 $\pm$ 0.1154 |
| Ohsumed | 120 | **0.3663** $\pm$ 0.0022 | 0.1534 $\pm$ 0.0036 | -2.0969 $\pm$ 0.0553 |

Tabela 6.4: WNTM

| **Coleção** | **K** | $C_V$ | $C_A$ | $C_{UMass}$ |
|---|---|---|---|---|
| News-Head | 30 | 0.3816 ± 0.0046 | 0.1569 ± 0.0062 | **-2.8735** ± 0.0862 |
| News-Head | 60 | 0.3891 ± 0.0044 | **0.1680** ± 0.0085 | -2.9835 ± 0.1014 |
| News-Head | 120 | **0.3896** ± 0.0036 | 0.1502 ± 0.0040 | -3.4213 ± 0.0742 |
| News-Short | 30 | 0.4149 ± 0.0067 | **0.1874** ± 0.0015 | **-2.9649** ± 0.1104 |
| News-Short | 60 | **0.4158** ± 0.0039 | 0.1793 ± 0.0102 | -3.1748 ± 0.2178 |
| News-Short | 120 | 0.4118 ± 0.0043 | 0.1729 ± 0.0005 | -3.4334 ± 0.0575 |
| Ohsumed | 30 | 0.3904 ± 0.0020 | 0.1324 ± 0.0085 | **-3.9141** ± 0.0929 |
| Ohsumed | 60 | 0.3993 ± 0.0026 | **0.1411** ± 0.0047 | -3.9944 ± 0.0584 |
| Ohsumed | 120 | **0.4120** ± 0.0021 | 0.1348 ± 0.0022 | -3.9718 ± 0.1070 |

Nas Tabelas 6.5, 6.6 e 6.7 é avaliado o desempenho de cada abordagem por coleção. Cada tabela representa uma métrica de coerência. O valor exposto na tabela é a média de todas as execuções de cada abordagem para 30, 60 e 120 tópicos.

Tabela 6.5: $C_V$

| **Média entre execuções (30, 60 e 120 tópicos) – Métrica** $C_V$ | | | |
|---|---|---|---|
| Abordagem/ Coleção | News-Head | News-Short | Ohsumed |
| BTM | **0.3959** | **0.4337** | 0.4132 |
| PTM | 0.3841 | 0.4018 | **0.4151** |
| SATM | 0.3762 | 0.3935 | 0.3536 |
| WNTM | 0.3868 | 0.4142 | 0.4006 |

Tabela 6.6: $C_A$

| **Média entre execuções (30, 60 e 120 tópicos) – Métrica** $C_A$ | | | |
|---|---|---|---|
| Abordagem/ Coleção | News-Head | News-Short | Ohsumed |
| BTM | **0.1649** | **0.1835** | 0.1649 |
| PTM | 0.1429 | 0.1511 | **0.2044** |
| SATM | 0.1462 | 0.1472 | 0.1632 |
| WNTM | 0.1583 | 0.1799 | 0.1361 |

Nos gráficos apresentados nas Figuras 6.2, 6.3, 6.4, 6.5, 6.6, 6.7, 6.8, 6.9 e 6.10 é demonstrado as pontuações médias das execuções das abordagens sobre cada um dos conjuntos de dados, considerando para 30, 60 e 120 tópicos bem como uma média geral de todas as execuções. Nos gráficos das Figuras acima citadas cada cor representa uma abordagem, conforme a legenda dos gráficos. No eixo vertical é exibido a pontuação enquanto no eixo horizontal é mostrado o número de tópicos das execuções e também a média geral (rótulo "Média"). Cada gráfico demonstra uma métrica distinta em relação ao conjunto de dados.

Tabela 6.7: $C_{UMass}$

| **Média entre execuções (30, 60 e 120 tópicos) – Métrica $C_{UMass}$** | | | |
|---|---|---|---|
| Abordagem/ Coleção | News-Head | News-Short | Ohsumed |
| BTM | -3.3436 | -3.6766 | -3.8673 |
| PTM | -3.4896 | -3.4744 | -3.4009 |
| SATM | -3.2981 | -3.3790 | **-1.7904** |
| WNTM | **-3.0928** | **-3.1910** | -3.9601 |

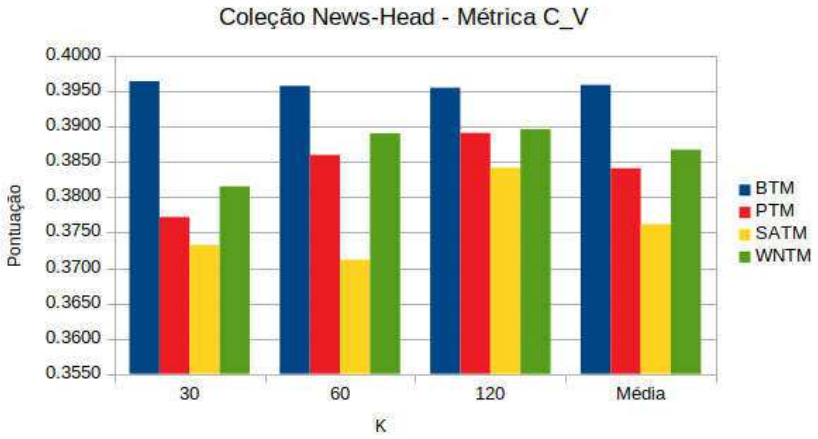


Figura 6.2: Pontuações sobre a coleção *News-Head* utilizando a métrica $C_V$

As Figuras 6.2, 6.3 e 6.4 apresentam a média das pontuações das execuções das abordagens sobre a coleção *News-Head*. É apresentado conforme o número de tópicos e também a média de todas as execuções de determinada abordagem sobre aquela coleção.

As pontuações da média das execuções das abordagens sobre o conjunto de dados *News-Short* são mostrados em gráfico na Figuras 6.5, 6.6 e 6.7. Além de exibir graficamente as pontuações médias das execuções pelo número de tópicos, é também mostrado a média geral de todas as execuções de uma abordagem sobre a coleção *News-Short*.

Nas Figuras 6.8, 6.9 e 6.10 são apresentadas graficamente as pontuações médias das abordagens considerando a coleção *Ohsumed*. Além de mostrar a média considerando o número de tópicos, é exibido uma média geral.

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.

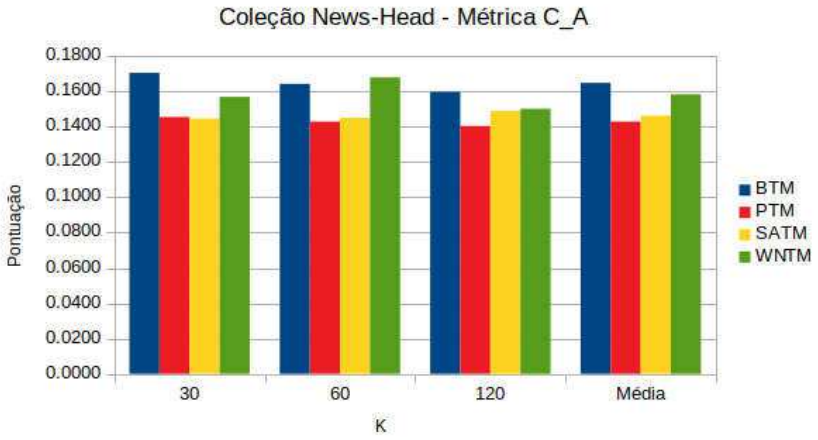


Figura 6.3: Pontuações sobre a coleção *News-Head* utilizando a métrica $C_A$

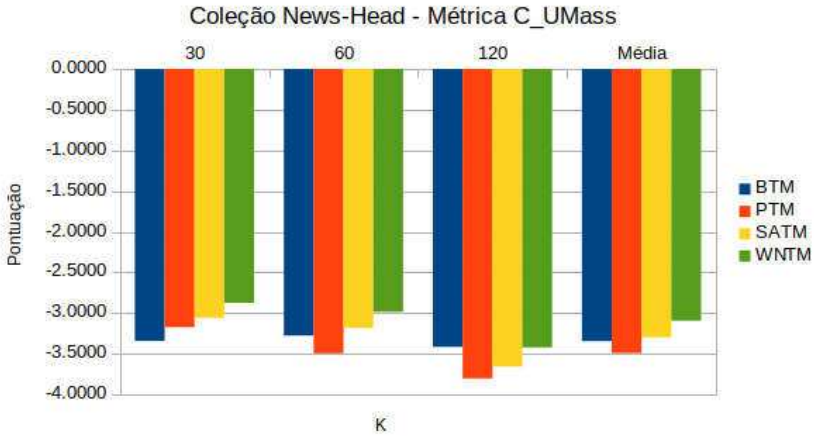


Figura 6.4: Pontuações sobre a coleção *News-Head* utilizando a métrica $C_{UMass}$

Outro ponto notado foi que, em geral, as abordagens usadas neste trabalho obtiveram 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.

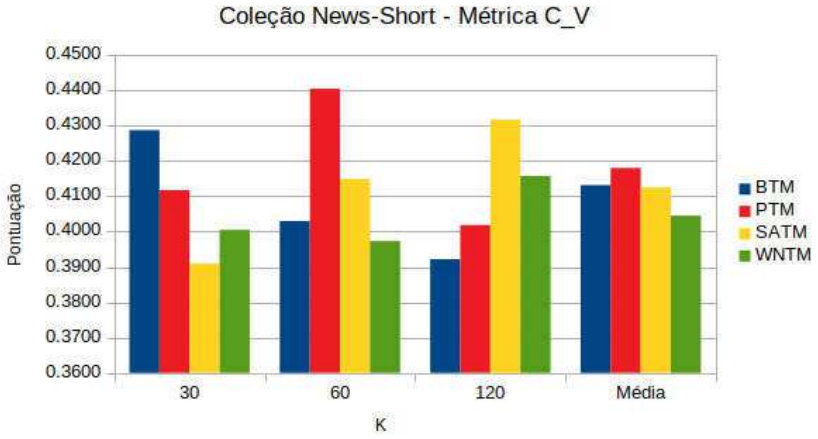


Figura 6.5: Pontuações sobre a coleção *News-Short* utilizando a métrica $C_V$

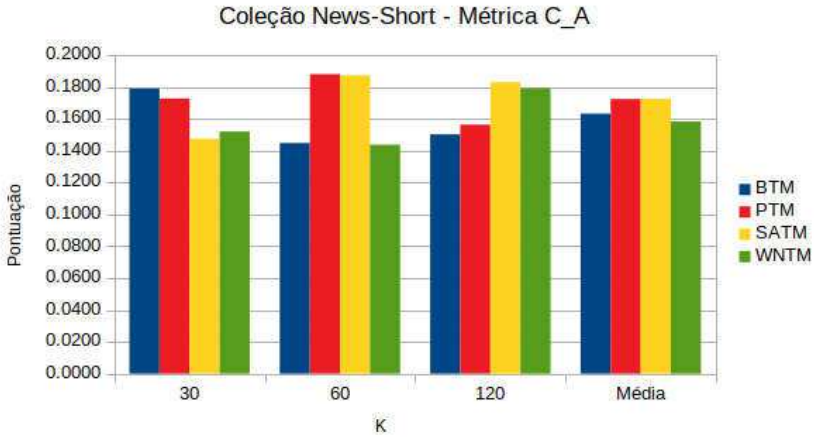


Figura 6.6: Pontuações sobre a coleção *News-Short* utilizando a métrica $C_A$

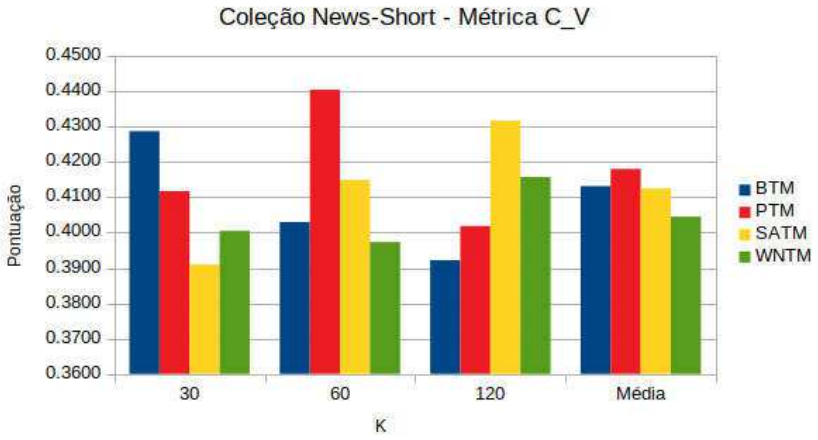


Figura 6.7: Pontuações sobre a coleção *News-Short* utilizando a métrica $C_{UMass}$

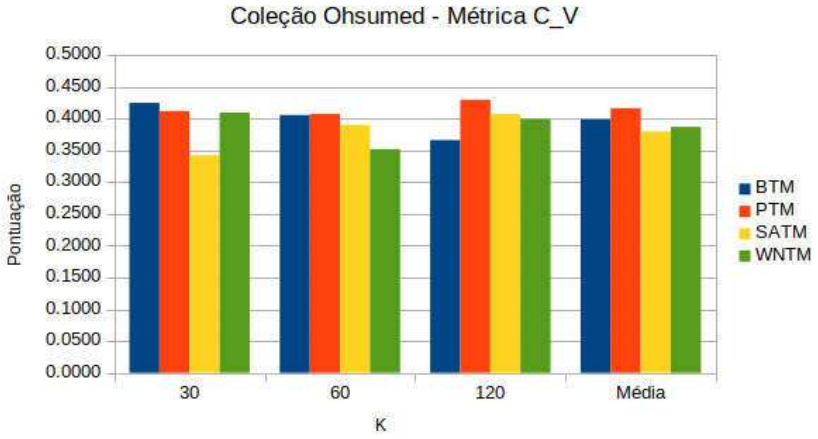


Figura 6.8: Pontuações sobre a coleção *Ohsumed* utilizando a métrica $C_V$

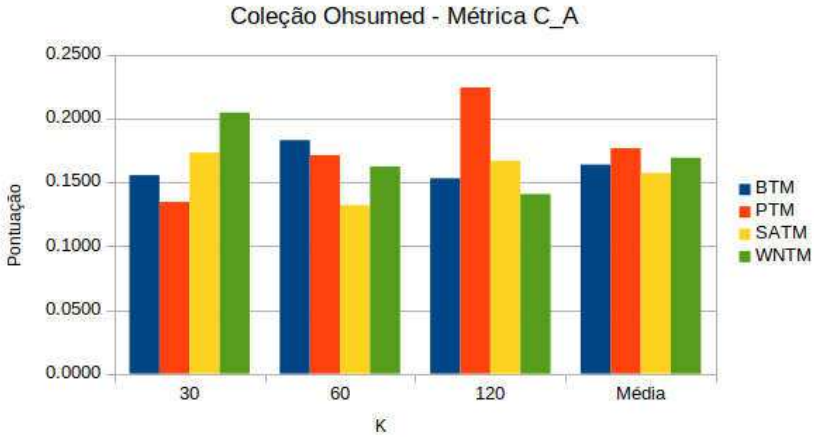


Figura 6.9: Pontuações sobre a coleção *Ohsumed* utilizando a métrica $C_A$

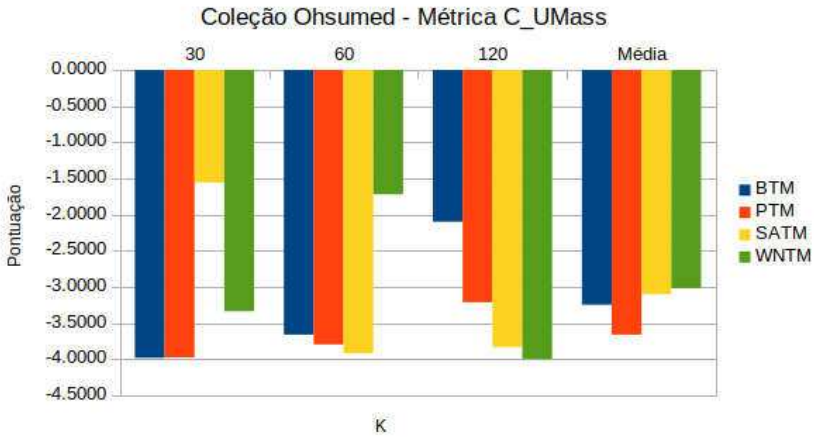


Figura 6.10: Pontuações sobre a coleção *Ohsumed* utilizando a métrica $C_{UMass}$

# 7 CONCLUSÃO

A modelagem de tópicos possibilitou a extração de tópicos em grandes coleções de documentos através de abordagens como o LDA, auxiliando em tarefas como classificação de conteúdo. O LDA é uma abordagem tradicional que serve como base para outras abordagens de extração de tópicos em coleções de documentos, inclusive às abordagens probabilísticas que visam extrair tópicos de documentos curtos que foram usadas neste trabalho. A co-ocorrência de uma palavra dentro de uma coleção de documentos é o que permite ao LDA alocar tópicos através de probabilidade usando a distribuição de Dirichlet. Essa abordagem possui algumas premissas que norteiam sua implementação. Uma das premissas, por exemplo, considera que a coleção inteira de documentos é uma "sacola de palavras"(termo original, *bag-of-words*) e que a ordem das palavras não importa. Mas as abordagens que têm o LDA como base podem relaxar algumas premissas para adaptar o funcionamento ao seu fim.

Devido à dominância de textos curtos na Web, extrair tópicos de documentos curtos tornou-se uma tarefa cada vez mais importante. Tradicionalmente, já existia um número grande textos curtos na Web, como títulos de páginas, legendas de fotos, manchetes de notícias, etc. 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.

Cada uma das quatro abordagens se diferencia em relação às outras em aspectos cruciais. Houve, então, a necessidade de saber qual abordagem se comportaria melhor em cada cenário apresentado.

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).

Os resultados aqui apresentados foram capazes de indicar qual abordagem se comportou melhor em um dado cenário. Para avaliação foram usadas três métricas de coerência a fim de estabelecer uma pontuação para cada execução de uma abordagem sobre uma coleção e um diferente número de tópicos: $C_V$, $C_A$ e $C_{UMass}$. As abordagens foram executadas três vezem em cada cenário (número de tópicos e coleção de documentos) para diminuir a influência do fator aleatório e o resultado para cada cenário foi a média destas três execuções.

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).

Este trabalho pode servir de auxílio para um usuário que busque uma solução para extrair tópicos de coleções de documentos curtos e também para futuros trabalhos que tenham seu foco na modelagem de tópicos. Um trabalho que se proponha a projetar uma abordagem voltada a extrair tópicos de textos muito pequenos, por exemplo, poderia apoiar-se neste trabalho para escolher uma abordagem que poderia servir como base para implementação. Também outros trabalhos que se proponham a avaliar outras abordagens probabilísticas para extração de tópicos em coleções de textos curtos poderiam usar este trabalho como referência.

# REFERÊNCIAS

[1] D. M. Blei. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84, 2012.

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

[3] C. X. Lin, B. Zhao, Q. Mei, and J. Han. Pet: a statistical model for popular events tracking in social communities. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 929–938. ACM, 2010.

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

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

[6] M. Röder, A. Both, and A. Hinneburg. 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, 2015.

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

[8] Y. Zuo, J. Wu, H. Zhang, H. Lin, F. Wang, K. Xu, and H. Xiong. Topic modeling of short texts: A pseudo-document view. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 2105–2114. ACM, 2016.

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