UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO DE EDUCAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO

Luana da Silva Sousa

Avaliação do efeito de normalização de corpus na coerência de tópicos extraídos usando Latent Dirichlet Allocation

Florianópolis
2022

Luana da Silva Sousa

Avaliação do efeito de normalização de corpus na coerência de tópicos extraídos usando Latent Dirichlet Allocation

Dissertação submetida ao Programa de Pós-Graduação em Ciência da Informação da Universidade Federal de Santa Catarina para a obtenção do título de Mestre em Ciência da Informação.
Orientador: Gustavo Medeiros de Araújo, Dr.

Florianópolis
2022

Ficha de identificação da obra elaborada pelo autor,
através do Programa de Geração Automática da Biblioteca Universitária da UFSC.

Sousa, Luana da Silva
Avaliação do efeito de normalização de corpus na coerência de tópicos extraídos usando Latent Dirichlet Allocation / Luana da Silva Sousa ; orientador, Gustavo Medeiros de Araújo, 2022.
77 p.

Dissertação (mestrado) - Universidade Federal de Santa Catarina, Centro de Ciências da Educação, Programa de Pós Graduação em Ciência da Informação, Florianópolis, 2022.

Inclui referências.

1. Ciência da Informação. 2. Processamento de Linguagem Natural. 3. Topic Modeling. 4. Normalização de Corpus. 5. Recuperação de Informação. I. Araújo, Gustavo Medeiros de. II. Universidade Federal de Santa Catarina. Programa de Pós Graduação em Ciência da Informação. III. Título.

Luana da Silva Sousa

Avaliação do efeito de normalização de corpus na coerência de tópicos extraídos usando Latent Dirichlet Allocation

O presente trabalho em nível de mestrado foi avaliado e aprovado por banca examinadora composta pelos seguintes membros:

Prof. Márcio Matias, Dr.
Universidade Federal de Santa Catarina

Prof. Ricardo Alexandre Moraes, Dr.
Universidade Federal de Santa Catarina

Certificamos que esta é a **versão original e final** do trabalho de conclusão que foi julgado adequado para obtenção do título de Mestre em Ciência da Informação.

Documento assinado digitalmente
Edgar Bisset Alvarez
Data: 03/05/2022 14:25:24-0300
CPF: 236.516.228-27
Verifique as assinaturas em https://v.ufsc.br

Coordenação do Programa de
Pós-Graduação

Documento assinado digitalmente
Gustavo Medeiros de Araujo
Data: 03/05/2022 14:46:24-0300
CPF: 828.754.991-49
Verifique as assinaturas em https://v.ufsc.br

Gustavo Medeiros de Araújo, Dr.
Orientador

Florianópolis, 2022.

Dedico este trabalho ao Deus criador, a quem tudo pertence.

AGRADECIMENTOS

Gostaria de agradecer em primeiro lugar a Deus, que me capacitou e possibilitou finalizar este trabalho. Ao meu marido, Vinícius, que contribuiu com as ideias e discussões apresentadas no texto, além de me apoiar e suportar em todos os momentos, sempre trazendo alegria e paz em relação às decisões. Aos meus pais, Gilberto e Nelci, por sempre me incentivarem a estudar e ir um passo a mais. Ao meu orientador, Gustavo, pela orientação durante esse tempo. Também à UFSC e à CAPES, por terem possibilitado que eu passasse um ano estudando em tempo integral.

“Tudo quanto fizerdes, fazei-o de todo o coração, como para o Senhor e não para homens.”
Colossenses 3.23, ARA

RESUMO

Uma das formas de tornar possível o acesso e recuperação da grande quantidade de informação sendo produzida nos últimos anos é com ferramentas para melhor entender o conteúdo de documentos de texto. O algoritmo de *Topic Modeling* é amplamente usado para esse tipo de problema, pois tem a capacidade de sumarizar e categorizar documentos de texto. Devido ao fato de ser um método estatístico e baseado em dados, ele pode produzir tópicos que nem sempre são interpretáveis (coerentes). Este trabalho é baseado na hipótese de que, dado que o LDA leva em consideração o número de ocorrências de palavras, é possível afetar a qualidade dos tópicos por meio de uma normalização semântica do texto, na qual os conceitos poderiam ser representados pela mesma palavra. Podemos encontrar uma descrição formal de conceitos usando uma base de conhecimento (da Web Semântica) ou conhecimento de domínio sobre um determinado tema, extraindo diversas formas de expressar um determinado conceito, a fim de normalizar o corpus. Foi usada a métrica de coerência dos tópicos para quantificar a influência da normalização semântica, dado que essa métrica representa a interpretabilidade semântica dos termos usados para descrever um tópico. Foram testadas duas hipóteses: (i) bases de conhecimento da web semântica para normalizar dois corpora de teste genéricos de forma automática, e (ii) conhecimento de domínio para efetuar a normalização em um corpus específico, a fim de aceitar ou rejeitar a hipótese de que a normalização afeta ou não a coerência dos tópicos extraída. Uma amostragem aleatória com um número variável de documentos (dependendo do corpus) foi selecionada para aplicar o teste estatístico de Mann-Whitney com a métrica C_V. Os resultados mostraram que a normalização semântica de corpus afeta, de forma positiva com significância estatística, a coerência dos tópicos extraídos via algoritmo LDA de *Topic Modeling* em um corpus de domínio específico, caso haja um percentual minimamente considerável de texto normalizado. É possível concluir também que as bases de conhecimento da Web Semântica ainda são incipientes para este tipo de aplicação.

Palavras-chave: Normalização de Corpus. LDA. Coerência de Tópicos. Processamento de Linguagem Natural.

ABSTRACT

One of the ways to make it possible to access and retrieve the large amount of information being produced in recent years is with tools to better understand the content of text documents. The *Topic Modeling* algorithm is widely used for this type of problem as it has the ability to summarize and categorize text documents. Due to the fact that it is a statistical and data-based method, it can produce topics that are not always interpretable (coherent). This work is based on the hypothesis that, given that the LDA takes into account the number of occurrences of words, it is possible to affect the quality of topics through a semantic normalization of the text, in which the concepts could be represented by the same word. We can find a formal description of concepts using a knowledge base (from Semantic Web) or domain knowledge on a given topic, extracting different ways of expressing a given concept in order to normalize the corpus. The topic coherence metric was used to quantify the influence of semantic normalization, since this metric represents the semantic interpretability of the terms used to describe a topic. Two hypotheses were tested: (i) semantic web knowledge bases to automatically normalize two generic test corpora, and (ii) domain knowledge to perform normalization on a specific corpus, in order to accept or reject the hypothesis that normalization affects or not the coherence of the extracted topics. A random sample with a variable number of documents (depending on the corpus) was selected to apply the Mann-Whitney statistical test with the metric C_V. The results showed that semantic corpus normalization positively affects the coherence of topics extracted via the LDA algorithm of *Topic Modeling* in a domain-specific corpus, if there is a minimally considerable percentage of normalized text. It is also possible to conclude that the Semantic Web knowledge bases are still incipient for this type of application.

Keywords: Corpus Normalisation. LDA. Topic Coherence. Natural Language Processing.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Desde períodos históricos, em que livros e registros materiais do conhecimento eram escassos, perdidos ou intencionalmente destruídos, havia o desafio da aquisição da informação. Ao longo do tempo, principalmente depois da criação da imprensa, esse desafio foi sendo cada vez mais superado, até existirem grandes coleções e acervos de conhecimento. Nas últimas décadas, devido à grande quantidade de informação sendo produzida e armazenada, a recuperação e uso dessa informação têm se tornado o verdadeiro desafio (ARAÚJO, 2009).

Durante o período pós-guerra, na segunda metade do século XX, a competição entre EUA e URSS acelerou o processo de evolução científica e tecnológica. Com a necessidade de aumentar a produtividade e a velocidade de produção científica e tecnológica, a informação foi percebida como um recurso importante (ARAÚJO, 2009). A Ciência da Informação (CI) nasceu aproximadamente nessa época, a fim de encontrar soluções para lidar com grande quantidade de informação. O artigo de Bush apresentou a ideia de uma máquina que serviria para auxiliar a memória e guardar conhecimento (BUSH *et al.*, 1945). Nesse artigo, o autor propôs o principal problema da incipiente CI: tornar acessível às pessoas o crescente acervo de conhecimento armazenado.

Dentro do campo da CI houve inúmeras tentativas de definir diretrizes para estudar essa crescente quantidade de informações, que se consolidaram nas seguintes correntes teóricas: (i) estudos de natureza matemática, como a Recuperação de Informação (RI) e a bibliometria; (ii) a teoria sistêmica, com modelos biológicos; (iii) a teoria crítica, baseada na desigualdade de acesso à informação; (iv) as teorias de representação, cujos estudos visam encontrar uma representação útil para recuperar a informação; (v) estudos de comunicação científica, com foco no fluxo e gestão da informação e conhecimento; e (vi) de estudo de usuários, buscando entender a informação do ponto de vista do usuário (ARAÚJO, 2009). As correntes teóricas de natureza matemática (i) e de representação (iv) fundamentam esse estudo, pois o aprimoramento da representação de informação, por meio da normalização dos textos (corrente iv) e da extração de informação de corpus, usando um método matemático (corrente i), tem o fim de promover uma melhor recuperação de informação.

Uma das formas de tornar possível o acesso e recuperação dessa grande quantidade de informação é extraindo informações relevantes que possam categorizá-la, e algoritmos de Modelagem de Tópicos (em inglês *Topic Modeling* - TM) são extensivamente usados para lidar com esse tipo de problema (ALLAHYARI, 2016). *Topic Modeling* é uma forma de encontrar uma estrutura semântica latente dentro de uma coleção de documentos de texto. Modelos probabilísticos como o *Latent Dirichlet Allocation* (LDA) têm se tornado o modelo mais comumente empregado para implementar

a modelagem tópicos (BLEI, D. M.; NG; JORDAN, 2003; O'CALLAGHAN *et al.*, 2015).

Apesar de ser muito usado, o modelo LDA tem sido criticado por favorecer palavras muito gerais e que aparecem frequentemente na construção de um tópico (O'CALLAGHAN *et al.*, 2015). Devido ao fato de ser um método puramente estatístico e baseado na co-ocorrência de palavras, é possível que produza tópicos que nem sempre sejam interpretáveis (ALLAHYARI, 2016). Esse é um fato que nos leva à pergunta: é possível melhorar a qualidade dos tópicos extraídos via TM? Duas possíveis soluções são desbravadas neste trabalho, unindo a recuperação de informação com a web semântica e conhecimentos de domínio.

Um desafio comum aos algoritmos de Processamento de Linguagem Natural (em inglês, *Natural Language Processing* - NLP) é a complexidade e ambiguidade que a linguagem humana apresenta. Nossa língua tende naturalmente à ambiguidade, pois facilita a comunicação para nós, humanos, que possuímos um sofisticado mecanismo de desambiguação que permite que compreendamos o significado de uma mesma palavra em diversos contextos diferentes, pois conseguimos entender justamente isso: o contexto (PIANTADOSI; TILY; GIBSON, 2012). É difícil sistematizar e criar lógica computacional que permita facilmente a desambiguação de palavras em diferentes contextos, embora haja um grande avanço nessa área nos últimos anos (POPOV, 2018; SCARLINI; PASINI; NAVIGLI, 2020; KADDOURA; D. AHMED, 2022).

Avanços na área de Web Semântica nesse sentido trouxeram um conceito (que pode ser considerado uma ferramenta) chamada Ontologia. Uma ontologia é a descrição explícita e precisa de conceitos e relações entre esses conceitos que existem em um certo domínio de conhecimento (GRUBER, Thomas R., 1993). Assim, uma ontologia pode gerar um ambiente com informações documentadas, confiáveis e de fácil manutenção e reutilização (DIAS; SANTOS, 2003). Logo, ao conectar informações não estruturadas, em documentos de texto, com dados semânticos estruturados, disponíveis a partir das ontologias, é possível auxiliar diversas tarefas de processamento de linguagem natural, como a extração de informação (SUGANYA; PORKODI, 2018), Recuperação de Informação (WAITELONIS, 2018; VALLET; FERNÁNDEZ; CASTELLS, 2005), classificação de texto (DE MELO; SIERSDORFER, 2007), extração de características dos textos (GARLA; BRANDT, 2012), etc. pois sistematiza uma definição e a relação entre conceitos de forma que um computador possa entender e processar.

Quando exemplos (instâncias) são atribuídos aos conceitos descritos em uma ontologia, se forma uma base de conhecimento (em inglês, *Knowledge Base* - KB) (NOY; MCGUINNESS, 2001). A representação de um conceito por uma ou mais palavras pode ser encontrada em uma base de conhecimento. Um exemplo de base de conhecimento é a DBpedia, construída com um esforço comunitário de extrair informação estruturada da Wikipedia e tornar essa informação disponível na Web (AUER *et al.*, 2007). Diferentes formas de representar esse conceito (diferentes palavras) podem ser

extraídos dessa KB. Por exemplo, o conceito "país Estados Unidos da América"pode ser representado de diversas formas: EUA, USA, *United States*, America, *United States of America*, etc.

Assim, a fim de melhorar os tópicos extraídos via TM, é proposta a ideia de normalizar o corpus de entrada do modelo. A *normalização* se refere a padronizar todas as representações textuais de um determinado conceito, de modo a reduzir a ambiguidade das palavras. Dessa forma, dado que conseguiríamos obter diferentes formas de representação de um conceito por meio da KB ou por meio de um conhecimento de domínio, seria possível substituir (normalizar) todas as aparições de um conceito por apenas uma palavra (ou conjunto de palavras). Então todas as vezes que um conceito fosse mencionado, haveria uma única palavra (será chamada aqui de "forma de superfície") para descrevê-lo. Em termos mais gerais, estamos reduzindo a complexidade gerada pela ambiguidade da linguagem de forma que o algoritmo computacional possa entender que se trata do mesmo conceito.

1.1 HIPÓTESE

A intuição do TM é que pares de termos descritores (formas de superfície) que co-ocorrem frequentemente ou estão próximos em um espaço semântico são mais propensos a contribuir com maiores níveis de interpretabilidade para um tópico específico (O'CALLAGHAN *et al.*, 2015). Este trabalho é baseado na hipótese de que, como o LDA leva em consideração o número de co-ocorrência de palavras, seria possível afetar a qualidade dos tópicos por meio de uma normalização semântica do texto, no qual cada conceito poderia ser representado pela mesma palavra (forma de superfície). Se o mesmo conceito é representado por duas formas de superfície diferentes em textos diferentes, o algoritmo terá mais dificuldade de encontrar tópicos coerentes e interpretáveis.

1.2 PROBLEMA DE PESQUISA

Assim, a questão de pesquisa que essa dissertação se propõe a responder é: a normalização semântica de documentos de texto afeta a interpretabilidade de tópicos extraídos via algoritmo LDA de *Topic Modeling*?

1.3 OBJETIVOS

O objetivo geral desta dissertação é avaliar quantitativamente o efeito de uma normalização semântica de um corpus na interpretabilidade dos tópicos extraídos desse corpus via *Topic Modeling*, usando o algoritmo LDA. A interpretabilidade é aqui definida pela métrica de coerência dos tópicos. Para isso, os objetivos específicos são:

1. Definir e desenvolver uma arquitetura de normalização semântica de documentos;
2. Identificar e definir quais bases de dados serão usadas para a avaliação;
3. Analisar o resultado da normalização de corpus na coerência dos tópicos;
4. Comparar os resultados dos tópicos extraídos dos corpora normalizados com não normalizados.

O trabalho está organizado da seguinte forma: este primeiro capítulo introduz e explicita a questão de pesquisa e os objetivos do trabalho; o segundo capítulo apresenta revisão sistemática da literatura e a fundamentação teórica; o terceiro capítulo apresenta os métodos de normalização propostos e sua respectiva arquitetura de implementação; o quarto capítulo apresenta as bases de dados e os resultados desta pesquisa; e, ao final, o último capítulo apresenta as conclusões e os próximos passos desta pesquisa.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 TRABALHOS RELACIONADOS

A fim de levantar os trabalhos relacionados no contexto de normalização de corpus e *Topic Modeling*, foi realizada uma revisão sistemática da literatura. A revisão tem por objetivo responder a pergunta:

- Quais são as formas de tratamento de ambiguidade de conceitos em corpus que existem para melhorar os resultados de algoritmos de NLP baseados em co-ocorrência, como o *topic modeling?*

Foram selecionadas três bases de dados para a busca: (i) IEEE, por ser a principal base de periódicos de tecnologia e engenharia do mundo; (ii) *Scopus*, por ser uma base abrangente e (iii) ACM, por ser uma base da computação. Primeiramente, uma estratégia de busca foi definida. Após a definição, a operação de busca foi realizada por meio do mecanismo de busca das bases de dados.

Para realizar a busca, uma *string* de busca foi formada com as seguintes palavras-chave:

1. information retrieval
2. “natural language processing” OR NLP
3. ambiguity
4. “topic modeling” OR “Latent Dirichlet Allocation”
5. "corpus normalisation"OR "corpus normalization"OR normalization

Foram selecionados apenas artigos em periódicos e em anais de eventos. O idioma escolhido foi o inglês. Os critérios de inclusão e exclusão são descritos a seguir:

- Inclusão:
  - **CI1:** Deve propor um método de tratamento de ambiguidade de corpus;
  - **CI2:** Deve testar o tratamento em algum algoritmo de NLP;
  - **CI3:** Deve ter experimentação e resultados;
- Exclusão:
  - **CE1:** Não tenham propostas de um método de tratamento de ambiguidade de corpus;
  - **CE2:** Não testa em um algoritmo de NLP;
  - **CE3:** Não tem experimentação e resultados;
  - **CE4:** Livros e capítulos de livros;
  - **CE5:** Propõe somente uma métrica.

Após a definição dos critérios e das palavras-chave, as *strings* de busca foram montadas para cada base e os resultados foram recuperados. As consultas foram

realizadas no dia 23 de abril de 2022, com o espaço temporal de 2016 a 2021. Foram recuperados um total de 328 artigos, sendo 17 na *Scopus*, 98 na ACM e 213 na IEEE. Após remoção de artigos duplicados nas bases, restaram 276 artigos.

Os artigos foram selecionados por meio da leitura do título e resumo, analisando a quais critérios de inclusão e exclusão pertenciam. Os artigos a serem usados para comparação são os que possuem ao menos um critério de inclusão e nenhum critério de exclusão. Uma tabela explicitando os artigos e os critérios de inclusão e exclusão se encontra no Apêndice 3. Ao final, 04 artigos tiveram ao menos um critério de inclusão e nenhum critério de exclusão (YADAV; SARKAR, 2018; NANNI; PONZETTO; DIETZ, 2018; KEKEÇ; MAATEN; TAX, 2018; AZIMI; VEISI; AMOUIE, 2019).

Yadav e Sarkar (2018) abordaram a tarefa de análise de sentimentos, de forma a construir um léxico que os ajude a categorizar a polaridade das palavras de acordo com os sentimentos, tratando da ambiguidade de polaridade das palavras dependente do contexto em que elas estão inseridas. Por exemplo, "*growth*"(crescimento) é uma palavra neutra, mas no contexto financeiro, a mesma palavra tem um sentido positivo. Então, os autores criaram um corpus de notícias relacionadas a um evento financeiro para formar um léxico a ser usado junto com um léxico de propósito geral (SentiWordNet), a fim de classificar sentimentos de *tweets* de um evento relevante a esse domínio. Os autores concluíram que o uso do léxico de domínio específico melhorou a classificação de sentimentos do corpus testado.

Nanni, Ponzetto e Dietz (2018) propõem um método para melhorar a ligação de entidades nomeadas, a fim de trazer mais contexto para as entidades que são encontradas nos textos. As entidades são ligadas a sentenças, parágrafos e seções de páginas da Wikipedia. Os autores testam a proposta em (i) sugestão de entidades relacionadas a uma consulta, expandir consultas, (ii) previsão de entidades relevantes a um evento e (iii) previsão de aspectos de eventos para *tweets*.

Kekeç, Maaten e Tax (2018) desenvolveram um novo modelo de *word embeddings*, que trata uma propriedade natural da língua: a polissemia das palavras (múltiplos significados para uma mesma forma de superfície). Uma representação que leva em consideração a polissemia das palavras ajuda a desambiguar o significado das palavras por meio do desacoplamento dos significados em diferentes mapas. Os autores testaram o método em tarefas de similaridade de palavras, e mostraram que consegue distinguir diferentes sentidos, além de ter uma performance melhor que os modelos *embedding* que usam apenas a Wikipedia. A abordagem detecta variações semânticas e sintáticas inerentes à linguagem natural.

Azimi, Veisi e Amouie (2019) propõem um método para construir um corpus que auxilie tarefas de desambiguação de acrônimos. No processamento de linguagem natural, a desambiguação de acrônimos pode ser vista como um caso particular de desambiguação de sentido de palavras. Os autores usaram o contexto de textos téc-

nicos para testar o método, tratando dois tipos de padrões de acrônimos em texto, e chegaram a uma acurácia de 86% na detecção correta de expansões de acrônimos.

Conforme visto, existem abordagens diversas para o tratamento de ambiguidade das palavras na área de processamento de linguagem natural. Entretanto, até o conhecimento da autora, a abordagem proposta neste trabalho não encontra nenhum precedente na literatura recente, e pode ser considerada uma novidade. Ainda, as tarefas de desambiguação de sentido geralmente tratam de desambiguação de uma única palavra com vários sentidos. A proposta deste trabalho, no entanto, foca principalmente em tratar casos onde palavras diferentes representam o mesmo conceito, podendo causar degradação da qualidade dos resultados de algoritmos que se baseiam em co-ocorrência de palavras.

2.2 *TOPIC MODELING* E INTERPRETABILIDADE DOS TÓPICOS

A utilização de modelos probabilísticos na área de RI tem se tornado cada vez mais popular para o aprendizado de tópicos latentes (categorias) em coleções de documentos (BLEI, D.; LAFFERTY, 2009). O principal objetivo do modelo é encontrar descrições curtas de documentos dentro de grandes coleções. Essas descrições têm o propósito de facilitar tarefas como a classificação, sumarização, cálculo de similaridade e relevância dos documentos e, por consequência, a recuperação de informação. É um método usado em vários campos, como na análise de redes sociais, imagens e processamento de textos (BLEI, D.; CARIN; DUNSON, 2010), área da saúde (SONG; JUNG; CHUNG, 2017), e até na política (GREENE; CROSS, 2017). O modelo mais difundido, e usado neste trabalho, para fazer *Topic Modeling*, é o modelo probabilístico generativo *Latent Dirichlet Allocation* (LDA). Esse modelo é amplamente aplicável para coleções genéricas de dados discretos, como documentos de texto, nos quais os dados discretos são as palavras (BLEI, D. M.; NG; JORDAN, 2003).

Um tópico é definido como uma distribuição de probabilidade sobre palavras, e define-se documentos como misturas de tópicos. Logo, um modelo de tópicos pode ser considerado um modelo generativo para documentos (STEYVERS; GRIFFITHS, 2007). No LDA, assume-se que há k tópicos latentes subjacentes a partir dos quais os documentos são gerados. Cada tópico é representado como uma distribuição multinomial sobre as $|V|$ palavras do vocabulário. Um documento é gerado por amostragem de uma mistura desses tópicos e, em seguida, a amostragem de palavras dessa mistura (BLEI, D. M.; NG; JORDAN, 2003).

Assume-se que um documento com N palavras $\mathbf{d} =< k_1, ..., k_N >$ é gerado pelo seguinte processo:

1. A mistura de tópicos θ é amostrado de uma distribuição de Dirichlet($\alpha_1, ..., \alpha_k$);
2. Então, para cada uma das N palavras, um tópico $z_n \in \{1, ..., K\}$ é amostrado

de uma distribuição Mult(θ), onde $p(z_n = 1|\theta) = \theta_i$;

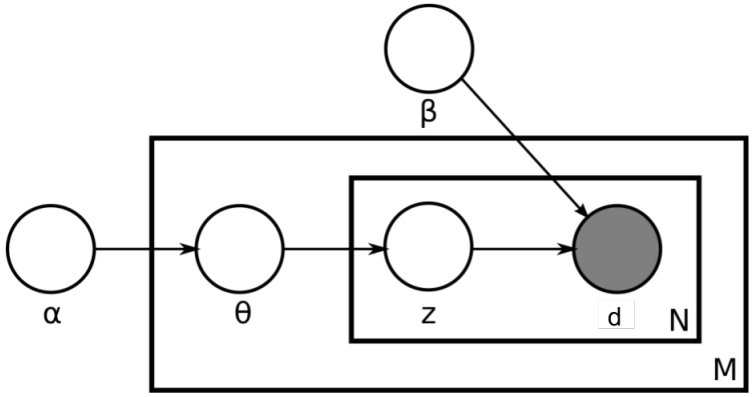
3. Finalmente, cada palavra k_n é amostrada, condicionada ao z_n-ésimo tópico, a partir da distribuição multinomial $p(k|z_n)$.

É possível pensar em θ_i como o grau que um tópico se refere a um documento. Então, a probabilidade de um documento é dada pela equação (1):

$$p(\mathbf{d}) = \int_\theta \left(\prod_{n=1}^{N} \sum_{z_n=1}^{K} p(k_n|z_n;\beta)p(z_n|\theta) \right) p(\theta;\alpha)d\theta, \quad (1)$$

onde $p(\theta;\alpha)$ é uma distribuição Dirichlet, $p(z_n|\theta)$ é uma distribuição multinomial parametrizada por θ, e $p(k_n|z_n;\beta)$ é uma distribuição multinomial sobre as palavras. Esse modelo é parametrizado pelos parâmetros $\alpha = \langle \alpha_1, ..., \alpha_k \rangle$ e β, que é uma matriz com dimensões K x $|V|$. A representação gráfica do LDA é mostrado na Figura 1.

Figura 1 – Representação gráfica do algoritmo *Latent Dirichlet Allocation*.



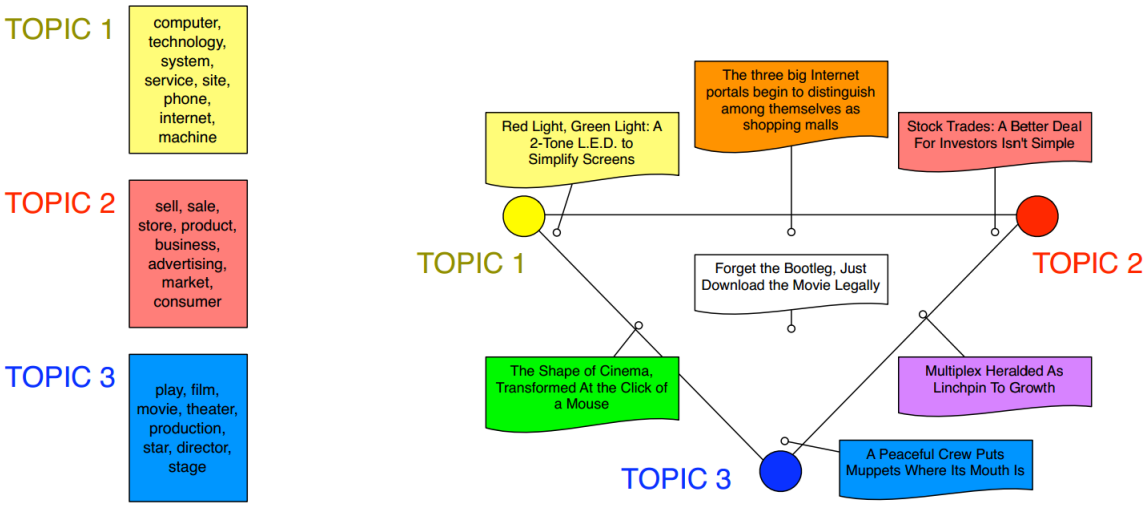
Fonte: https://upload.wikimedia.org/wikipedia/commons/d/d3/Latent_Dirichlet_allocation.svg

No modelo gráfico, cada nó representa uma variável aleatória e tem um papel no processo generativo. O nó cinza representa a única variável observada e os nós brancos as variáveis latentes (não observadas). Neste caso, o conjunto de palavras do corpus é o único dado que é possível observar. As variáveis latentes determinam a mistura aleatória de tópicos na coleção de documentos θ e a distribuição de palavras nos documentos z. O objetivo do LDA é usar as palavras observadas **d** para inferir a estrutura do tópico latente.

Como uma forma de tornar os conceitos mais claros, a Figura 2 mostra a distribuição de palavras sobre os tópicos e a distribuição de tópicos sobre os documentos. Um tópico é representado por uma distribuição de palavras e um documento é representado por uma distribuição de tópicos.

A medida de coerência de tópicos é usada para se referir à interpretabilidade semântica dos termos usados para descrever um tópico particular (O'CALLAGHAN

Figura 2 – Exemplo de tópicos extraídos usando *Latent Dirichlet Allocation*.



Distribuição palavras-tópicos na esquerda e distribuição documento-tópicos na direita. Os três tópicos representam os primeitos três tópicos de um conjunto de 50 tópicos de um modelo LDA treinado em artigos do New York Times (CHANG *et al.*, 2009).

et al., 2015). A qualidade dos tópicos (qualidade significa interpretabilidade e significância) é baseada na hipótese de que palavras com significados similares tendem a co-ocorrer dentro de um contexto similar. Cada distribuição de tópicos contém todas as palavras, mas atribui diferentes probabilidades a cada uma delas. As palavras com as maiores probabilidades dentro de um tópico são aquelas que tendem a co-ocorrer mais frequentemente. Então, as primeiras 10 a 15 palavras com maior probabilidade são geralmente usadas para interpretar e semanticamente nomear os tópicos (SYED; SPRUIT, 2017).

Há outras métricas que podem ser utilizadas, como a perplexidade ou a verossimilhança. Entretanto, tais métricas são usadas somente para avaliar a capacidade preditiva do modelo estatístico e não se referem aos objetivos explanatórios do TM (CHANG *et al.*, 2009). A tarefa de quantificar a coerência de um conjunto de tópicos tem sido estudada para remediar o desafio de que o TM não garante a interpretabilidade da sua saída (RÖDER; BOTH; HINNEBURG, 2015).

Além disso, muitas medidas de coerência têm sido propostas recentemente, baseada em abordagens que incluem a frequência de co-ocorrência de termos dentro de um corpus de referência (RÖDER; BOTH; HINNEBURG, 2015; NEWMAN *et al.*, 2010; LAU; NEWMAN; BALDWIN, 2014). O estudo de (RÖDER; BOTH; HINNEBURG, 2015) sistematicamente e empiricamente explorou a multitude de medidas de coerência de tópicos e suas correlações com dados de ranqueamento de tópicos feito por humanos. (RÖDER; BOTH; HINNEBURG, 2015) propuseram um framework que representa a

métrica de coerência como uma composição de partes que podem ser combinadas. Essas partes são: segmentação, estimativa de probabilidade, medida de confirmação e agregação.

Esse espaço de configurações é representado pelo produto de quatro conjuntos: $C = S \times M \times P \times \Sigma$. A primeira parte, S se refere ao conjunto de tipos de segmentação que são usados para dividir um conjunto de palavras em conjuntos menores. A segunda parte, M, é o conjunto de medidas de confirmação que pontua a concordância entre cada par de palavras. Essa medida de concordância é a NPMI (*Normalized Pointwise Mutual Information*). As medidas de confirmação usam a probabilidade das palavras, que pode ser calculada de diferentes formas. O conjunto de formas pelas quais a probabilidade pode ser calculada é chamada de P. Por último, o conjunto de métodos usados para agregar valores escalares computados pelos métodos de confirmação compõe a quarta parte e é chamada de Σ.

Primeiro, um conjunto de palavras w é segmentado em um conjunto de pares de subconjuntos de palavras s. Em seguida, as probabilidades p são computadas baseadas em um corpus de referência. Ambos o conjunto s e as probabilidades p são consumidas pela métrica de confirmação para calcular as concordâncias. Enfim, os valores são agregados em um único valor c.

Na métrica C_V, cada uma das partes é definida a seguir: A segmentação S para esse caso compara cada palavra com o conjunto total de palavras W usando vetores de contexto, S^{one}_{set}, simplificada aqui por apenas S. Sejam W' e W^* subconjuntos de palavras (de um tópico, por exemplo),

$$S = (W', W^*) | W' = w_i; w_i \in W; W^* = W \tag{2}$$

A estimativa das probabilidades usa janelas deslizantes (*sliding windows*), P_{sw}, e tem o tamanho da janela definido como 110, representado por $P_{sw(110)}$. A janela se move através dos documentos uma palavra por vez. Cada passo define um novo documento virtual copiando o conteúdo da janela. Essa probabilidade captura um grau de proximidade entre as palavras.

A medida de confirmação usa um ou mais pares $S_i = (W', W^*)$ de palavras, junto com as probabilidades correspondentes para computar quanto um subconjunto W^* semanticamente suporta W'. Por exemplo, X e Y são duas marcas de carro que semanticamente suportam uma a outra, ou seja, faz sentido aparecerem juntas. Porém, não necessariamente isso é verdade dentro do corpus em análise. Para isso, existem duas formas de fazer a avaliação: direta e indireta. A medida direta avalia a confirmação de um único par S_i de palavras ou subconjuntos de palavras. Já a medida indireta calcula a similaridade de palavras em W' e W^* em função da confirmação de todas as palavras. A medida direta usada na métrica C_V é a NPMI, dada pelas seguintes equações:

$$m_{lr}(S_i) = log\frac{P(W', W^*) + \varepsilon}{P(W') * P(W^*)} \tag{3}$$

$$m_{nlr}(S_i) = \frac{m_{lr(S_i)}}{-log(P(W', W^*) + \varepsilon)} \tag{4}$$

A ideia da medida indireta pode ser formalizada pela representação de conjuntos de palavras W' e W^* como vetores com dimensão do tamanho total do conjunto de palavras W. No caso de W' e W^* conterem mais de uma palavra cada, o vetor de elementos é a soma das confirmações diretas de cada palavra. Dado que η é um parâmetro usado para dar mais ênfase em palavras com confirmações maiores (ALETRAS; STEVENSON, 2013), o vetor de elementos é definido como:

$$\vec{v}_{m,\eta}(W') = \sum_{w_i \in W'} m(w_i, w_j)^{\eta}_{j=1,...,|W|} \tag{5}$$

Logo, dados os vetores $\vec{u} = \vec{v}(W')$ e $\vec{w} = \vec{v}(W^*)$ para os conjuntos de palavras de um par $S_i = (W', W^*)$, a confirmação indireta é computada como uma similaridade de vetores. A similaridade de cosseno, usada na métrica C_v, é definida como segue:

$$s_{cos}(\vec{u}, \vec{w}) = \frac{\sum_{i=1}^{|W|} u_i * w_i}{||\vec{u}||_2 ||\vec{w}||_2} \tag{6}$$

Então, dada a medida de similaridade s, a medida de confirmação direta m e um valor para η, a medida de confirmação indireta é:

$$\hat{m}_{s(m,\eta)}(W', W^*) = s_{cos}(\vec{v}_{m,\eta}(W'), \vec{v}_{m,\eta}(W^*)) = m_{cos(nlr,1)} \tag{7}$$

Todas as confirmações de todos os pares de subconjuntos S_i são agregados em um único *score* de coerência, e a média aritmética σ_a é calculada.

Enfim, a métrica C_v pode ser descrita como:

$$C_v = (S, P_{sw(110)}, m_{cos(nlr,1)}, \sigma_a), \tag{8}$$

2.3 WEB SEMÂNTICA

A ideia da Web Semântica foi descrita em 2001 por Tim Berners-Lee et al. como "uma nova forma de conteúdo web que é significativa para computadores"(BERNERS-LEE; HENDLER; LASSILA *et al.*, 2001). Eles introduziram a ideia como uma extensão da web atual, na qual as informações recebem um significado bem definido, permitindo que computadores e pessoas trabalhem em cooperação (BERNERS-LEE; HENDLER; LASSILA *et al.*, 2001).

A Web Semântica é baseada no *Resource Description Framework* (RDF), uma linguagem formal para descrever uma informação estruturada (HITZLER; KROTZSCH;

RUDOLPH, 2009). Um documento RDF descreve uma especificação formal de um domínio arbitrário. A especificação é modelada por um grafo rotulado e direcionado, no qual cada aresta representa uma ligação entre dois recursos, representados pelos nós (MANOLA; MILLER; MCBRIDE *et al.*, 2004). A ligação é expressa como triplas RDF (*sujeito*, *relação*, *objeto*). A identificação dos recursos e relações RDF é feita por *Uniform Resource Identifiers* (URI).

Para acessar e consultar um grafo RDF um protocolo foi desenvolvido: o *Protocol and RDF Query Language* (SPARQL) (PRUD'HOMMEAUX; SEABORNE, 2008). Os resultados das consultas SPARQL podem ser novos grafos RDF ou apenas um conjunto de recursos.

Os relacionamentos e propriedades que os recursos RDF podem ter podem ser especificados por uma linguagem de descrição de vocabulário, chamado RDF *Schema* (RDFS) (BRICKLEY; GUHA; MCBRIDE, 2014). O RDFS permite criar vocabulários personalizados para organizar o conhecimento. Já que as URIs permitem identificar recursos RDF globalmente, é interessante combinar vocabulários compartilhados por diferentes criadores através de diferentes domínios. Quando compartilhado, um vocabulário RDF pode ser denotado como uma *ontologia*. Uma ontologia é uma especificação formal e explícita de uma conceitualização compartilhada, e define os termos usados para descrever e representar uma área de conhecimento (GRUBER, Thomas R, 1995). De forma mais didática, uma ontologia define um vocabulário comum para pesquisadores que precisam compartilhar informações em um domínio. Isso inclui definições interpretáveis por máquina de conceitos básicos nesse domínio e as relações entre elas (NOY; MCGUINNESS, 2001).

Uma ontologia descreve formal e explicitamente conceitos em um determinado domínio (medicina, indústria de automóveis, biologia, etc.), assim como define propriedades de cada conceito, descrevendo várias características e atributos do conceito, além de restrições a essas propriedades. Por exemplo, uma ontologia de vinho e comida e as combinações apropriadas de vinho com refeições (NOY; MCGUINNESS, 2001). Essa ontologia pode ser usada como uma base para algumas aplicações como ferramentas de gerenciamento de restaurantes: uma aplicação poderia criar sugestões de vinho para o menu do dia ou responder consultas dos garçons ou clientes. Outra aplicação poderia analisar uma lista de inventário de uma adega e sugerir quais categorias de vinho deveriam ser expandidas e quais vinhos particulares comprar para os próximos menus.

O conceito de ontologia nos leva ao projeto *Linked Open Data* (LOD). Esse projeto tem como objetivo (i) identificar conjuntos de dados na web que estão disponíveis sob licenças abertas, (ii) republicar esses dados em RDF e (iii) interligá-los com outros (BIZER *et al.*, 2008). O termo *Linked Data* se refere ao conjunto de princípios para publicar e ligar os dados estruturados na Web. Uma das ontologias disponíveis na

Web é a YAGO [1] (*Yet Another Great Ontology*). YAGO é uma base de conhecimento, derivada da Wikipedia, WordNet, WikiData, GeoNames, entre outras fontes de dados (REBELE *et al.*, 2016). Atualmente, a YAGO tem mais de 17 milhões de entidades (como pessoas, organizações, cidades, etc.) e contém mais de 150 milhões de fatos sobre essas entidades.

2.4 LIGAÇÃO DE ENTIDADES NOMEADAS

A Ligação de Entidades Nomeadas (em inglês, *Named Entity Linking* - NEL) é a tarefa de identificar entidades em um texto e ligá-los a entidades em uma base de conhecimento. Uma breve introdução à terminologia e aos conceitos se faz necessária, baseada em (WAITELONIS, 2018). O termo *entidade* se refere a alguma coisa que é cognitivamente representável. Uma *menção à entidade* se refere à parte do texto onde a referência àquela entidade é feita. A *forma de superfície* é a representação sintática específica de um lexema (unidade básica de significado). Uma *entidade na base de conhecimento* se refere à representação da entidade, geralmente identificada por uma URI.

Seja K uma base de conhecimento formal, $d \in D$ um documento em um corpus D, $W \subseteq d$ as palavras do documento d, $M \subseteq 2^W$ o conjunto de menções à entidade, e $m = (s,l,d,c) \in M$ denota uma menção à entidade em um documento d, com posição de início s, tamanho l e *score* de confiança $c \in [0,1]$. Então, o problema da Ligação de Entidades Nomeadas pode ser descrito segundo a seguinte definição:

1. Uma função de extração $f_{ex} : W \rightarrow M$ para extrair as menções às entidades M de um conjunto de documentos D;
2. Uma função de mapeamento $f_{map} : M \rightarrow 2^W \cup NIL$ para compilar a lista $C \in 2^K$ de potenciais entidades candidatas da base de conhecimento para cada lexema;
3. Uma função de *score* $f_{score} : C \rightarrow \mathbb{R}$ para calcular um *score* que indica o grau de certeza que a URI candidata está de ser selecionada como a correta;
4. Uma função de seleção $f_{sel} : C \rightarrow K$ para selecionar o candidato certo de acordo com os *scores* calculados.

No melhor dos casos, a lista de candidatos C contém apenas uma possibilidade. "*NIL*"é incluído caso nenhum candidato seja encontrado. O tamanho da lista de candidatos C pode ser considerada como um indicador do grau de ambiguidade. Então, a tarefa de desambiguação é definida por meio das funções de mapeamento, *scoring* e seleção juntas. A implementação dessas funções observa o contexto inteiro. Assim como na teoria da comunicação e linguística, o contexto é essencial ao se interpretar

[1] https://yago-knowledge.org/

pedaços de informação, na Ligação de Entidades Nomeadas também é (WAITELONIS, 2018).

O principal desafio da ligação de entidades é a ambiguidade natural da língua (ZHANG *et al.*, 2018), porque uma entidade nomeada pode ter múltiplas formas de superfície. Por exemplo: macaxeira, mandioca ou aipim. Além disso, diferentes entidades nomeadas podem compartilhar o mesmo nome, por exemplo a palavra "manga"pode ser uma fruta ou uma parte de uma camisa.

Bases de conhecimento (KBs) como DBpedia (AUER *et al.*, 2007) ou Yago (REBELE *et al.*, 2016), usadas para essa tarefa de ligar entidades, contém bilhões de fatos sobre o mundo. Essas bases de conhecimento têm aplicações nas mais diversas áreas, como recuperação de informação, tradução, manutenção de dados, *question answering*, entre outras (GALÁRRAGA *et al.*, 2017). A utilidade dessas aplicações depende da qualidade das bases de conhecimento usadas (LAJUS; SUCHANEK, 2018). Entretanto, a qualidade dos dados nas KBs não é sempre perfeita. Os problemas incluem dados falsos, dados faltantes ou até inconsistências no esquema. Mesmo que exista métricas que indiquem a proporção do KB que está correto, é difícil encontrar indicações da proporção do mundo real que elas cobrem (GALÁRRAGA *et al.*, 2017; ZHANG *et al.*, 2018). Por exemplo, entre 69% e 99% das instâncias em KBs populares têm ao menos uma propriedade que outras entidades na mesma classe não têm (GALÁRRAGA *et al.*, 2017). Logo, as KBs ainda são bastante incompletas.

Por exemplo, na pesquisa de *Question Answering* há dois paradigmas envolvidos para resolver o problema: (i) responder as questões em uma base de conhecimento ou (ii) responder as questões usando texto (DAS *et al.*, 2017). Apesar de as bases de conhecimento serem uma ótima solução para responder questões composicionais, a performance dessa solução geralmente é bastante afetada pela incompletude das bases de conhecimento (DAS *et al.*, 2017; DONG *et al.*, 2014).

Há algumas iniciativas para encontrar formas de completar as bases de conhecimento. Por exemplo: (i) determinar a obrigatoriedade de alguns atributos, a fim de indicar onde estão faltando dados (LAJUS; SUCHANEK, 2018); ou (ii) predizer a completude de uma base de conhecimento automaticamente (GALÁRRAGA *et al.*, 2017), mas ainda é um campo a ser desenvolvido.

2.5 PADRÃO *PUBLISHER/SUBSCRIBER*

Um sistema *Publisher/Subscriber* (pub/sub, ou produtor/consumidor) é um sistema de mensagens no qual há um ou mais módulos que publicam mensagens em um canal de comunicação comum e um ou mais módulos podem se inscrever para receber informações postadas. Chama-se o módulo que envia mensagens de publicador (em inglês *publisher*, ou produtor) e o módulo que recebe a mensagem de assinante (em inglês *subscriber*, ou consumidor).

O desacoplamento entre o produtor e o consumidor é a funcionalidade mais fundamental de um sistema pub/sub (DOBBELAERE; ESMAILI, 2017). Outra funcionalidade importante nesses sistemas é a lógica de roteamento (ou modelo de assinatura), que decide se e onde uma mensagem vinda de um produtor vai chegar até o consumidor. Essas diferentes formas fazem um balanço entre flexibilidade e performance. Os dois principais tipos de lógica de roteamento são: baseada em tópicos (*topic-based*) ou baseada em conteúdo (*content-based*) (DOBBELAERE; ESMAILI, 2017). Quando é baseada em tópico, o consumidor se inscreve em um tópico e passa a receber todas as mensagens relacionadas àquele tópico. Já quando é baseada em conteúdo, o consumidor recebe apenas mensagens que correspondem a algum padrão ou regra pré-definido (TIAN *et al.*, 2004).

Uma das principais ferramentas que implementam esse padrão é o RabbitMQ. O RabbitMQ é basicamente um gerenciador de filas de mensagens. É um software onde as filas são definidas e onde as aplicações (publicadoras e consumidoras) se conectam a fim de trocar mensagens. Uma mensagem pode incluir qualquer tipo de informação. Poderia, por exemplo, ter informações sobre um processo ou tarefa que deve começar em outra aplicação (que poderia estar em outro servidor) ou simplesmente uma mensagem de texto. O gerenciador da fila armazena a mensagem até que a aplicação consumidora se conecte e consuma a mensagem da fila para então processá-la (WHAT IS RABBITMQ?. . . , 2019). A Figura 3 ilustra a relação do produtor com o consumidor.

Figura 3 – Arquitetura básica do RabbitMQ.

Arquitetura básica de uma fila de mensagens: há uma aplicação cliente, chamada de produtor que cria mensagens e envia para um intermediário (em inglês, *broker*); outra aplicação, chamada de consumidor, se conecta ao intermediário e se inscreve para receber as mensagens a serem processadas. Fonte: (WHAT IS RABBITMQ?. . . , 2019)

2.6 TESTE DE MANN-WHITNEY

O teste de Mann-Whitney (KRUSKAL, 1957) foi desenvolvido primeiramente por F. Wilcoxon em 1945, para comparar tendências centrais de duas amostras independentes de tamanhos iguais. Em 1947, H.B. Mann e D.R. Whitney generalizaram a técnica para amostras de tamanhos diferentes. O teste de Mann-Whitney (Wilcoxon rank-sum test) é indicado para comparação de dois grupos não pareados para verificar

se pertencem ou não à mesma população e cujos requisitos para aplicação do teste t de *Student* não foram cumpridos. (TESTE DE MANN-WHITNEY. . . , 2012).

Quando se dispõe de uma amostra pequena e a variável numérica não apresenta sabidamente uma variação normal (ou não é possível verificadar satisfatoriamente, pode-se utilizar o teste não paramétrico de Mann-Whitney. Nesse teste, verifica-se se há evidências para acreditar que valores de um grupo A são superiores aos valores de um grupo B.

Calcula-se uma certa estatística de teste e obtém-se o p-valor a partir da distribuição amostral dessa estatística sob a hipótese nula. A diferença é que ao invés de construir essa estatística com dados originais, eles são previamente convertidos em postos (ordenações). A vantagem é que, com isso, as suposições de normalidade e homogeneidade das variâncias não são necessárias, permitindo mais generalidade aos resultados.

3 MÉTODO DE NORMALIZAÇÃO DE CORPUS

Há dois passos principais que compõem a avaliação do impacto da normalização de documentos de texto na coerência de tópicos: (i) a normalização dos textos (simples ou semântica) e (ii) extração e avaliação de tópicos. No primeiro passo, um conjunto de textos é normalizado, ou seja, várias formas de superfície são substituídas por apenas uma, que representa o conceito subjacente àquelas formas de superfície. É possível normalizar manualmente, quando existe um conjunto conhecido de formas de superfície a serem substituídas, devido a um conhecimento de domínio, e também automaticamente, usando bases de conhecimento semântico (KBs). Enfim, no segundo passo, a coerência de tópicos extraídos usando LDA é avaliada. Este trabalho não tem a intenção de explorar todas as formas/métodos de executar a normalização, e propõe duas possíveis formas a fim de testar a hipótese principal.

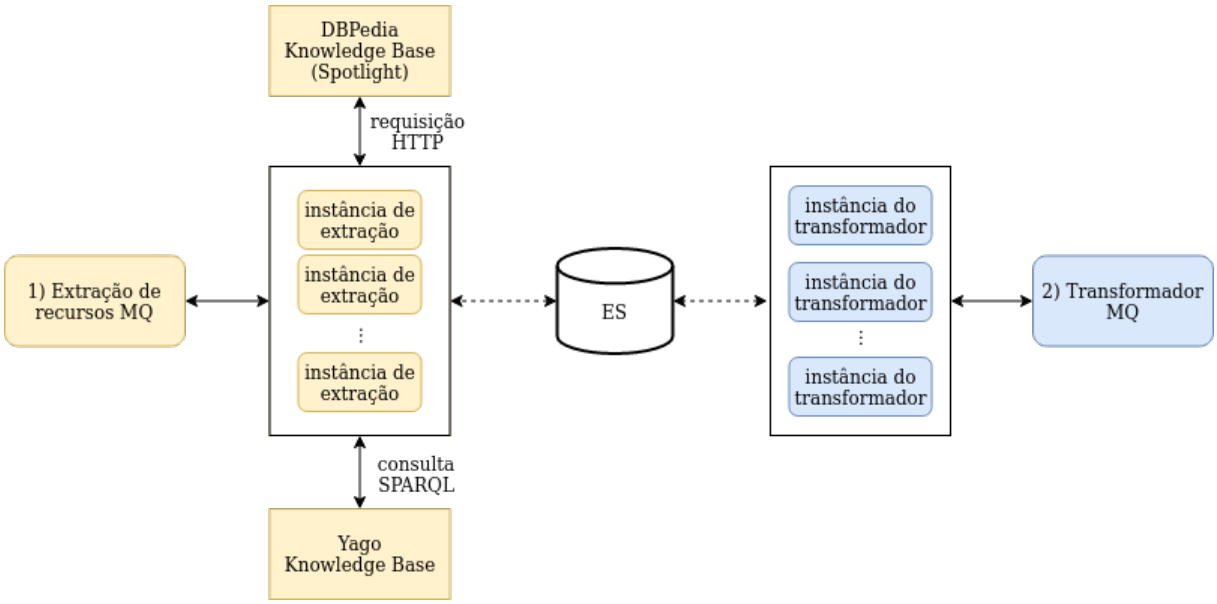
Este capítulo tem um foco maior em mostrar as possíveis formas de normalizar documentos de texto e o teste que valida o efeito dessa normalização nos tópicos extraídos via LDA. Não é a intenção otimizar o próprio algoritmo de LDA, e sim otimizar o corpus que serve de entrada para o algoritmo. Dessa forma, as duas primeiras seções explicam as formas de normalização propostas: automática, usando web semântica, e manual, usando conhecimento de domínio; a terceira seção explica brevemente o procedimento usado para execução do LDA e a quarta seção explica em detalhes o procedimento para o teste estatístico realizado.

3.1 NORMALIZAÇÃO AUTOMÁTICA USANDO BASES DE CONHECIMENTO

A normalização automática tem um custo maior na definição e implementação da arquitetura, mas a ideia é, como o nome diz, automatizar a tarefa, a fim de possivelmente entrar no fluxo de execução do LDA. A Figura 4 mostra a arquitetura do método de normalização automática proposta, e cada bloco é explicado a seguir.

A normalização automática proposta é composta de três passos: (i) identificação e ligação dos recursos/conceitos contidos no texto e (ii) criação da estrutura de substituição dos conceitos por uma mesma forma de superfície e (iii) transformação (substituição das formas de superfície). No primeiro passo (a extração dos conceitos) encontram-se todos os conceitos nos textos e os ligam (associam) a recursos da DBpedia. Os conceitos são anotados (associados a um recurso na KB) pelo processo de NEL, usando uma ferramenta de anotação, como a *DBpedia Spotlight* (que usa a KB DBpedia). Uma vez que os possíveis recursos mencionados no texto e suas respectivas URIs são obtidos, é possível encontrar informações e metadados relacionados a esses recursos. Um dos metadados necessários para a normalização é um conjunto de nomes alternativos para um determinado conceito. Esses nomes alternativos, aqui chamados de formas de superfície, que são palavras diferentes usadas para descrever

Figura 4 – Arquitetura de normalização automática. Os blocos amarelos (à esquerda do banco de dados ES) se referem ao passo de extração de recursos, enquanto os blocos azuis (direita do banco de dados ES) se referem ao passo de transformação.



um mesmo recurso ou conceito.

Duas bases de conhecimento (DBpedia e Yago) podem ser usadas para encontrar potenciais formas de superfície. A base de conhecimento YAGO é usada nesse passo como uma base complementar, pois nessa KB os campos "nome alternativo"(*alternateName*) e "identificador"*label* são mais comumente encontrados. Então, primeiramente o recurso foi buscado na DBpedia e, a partir da ligação que existe entre a DBpedia e a Yago (via *Linked Data*) foi possível buscar também nessa segunda base, mesmo ela não possuindo uma ferramenta pronta de ligação de entidades como a DBpedia possui o *Spotlight*.

O segundo passo é criar uma estrutura de dados para substituição, onde todos as possíveis formas de superfície de um recurso são substituídos por apenas uma. Essa estrutura de dados pode ser exemplificada pelo recurso "água"(*water*) no código 3.1. A URI desse recurso na KB YAGO é `https://yago-knowledge.org/resource/Water`.

Listing 3.1 – Exemplo de estrutura de dados de substituição para o recurso "water".

```
{
    'dihydridooxygen': 'water',
    'pure water': 'water',
    'dihydrogen monoxid': 'water',
    'dihydrogen oxide': 'water',
```

```
    ...
}
```

Finalmente, o último passo é realizar a substituição de todas as possíveis formas de superfície de um recurso por apenas uma. Assim, a variação de palavras usadas para descrever um mesmo conceito diminui, e a hipótese é de que essa substituição (que culmina no que chamamos de *normalização*) ajudará o algoritmo de extração de tópicos a encontrar tópicos mais coerentes.

Como a extração de recursos é obtida por meio de solicitações HTTP e consultas SPARQL para cada documento separadamente, é possível e conveniente modularizar e paralelizar o processo. Os documentos são salvos em um banco de dados, cada um com um identificador exclusivo associado. A estrutura dos documentos no banco é mostrada no código 3.2. A ferramenta Elasticsearch [1] (ELASTICSEARCH..., 2019) foi usada como banco de dados. O identificador dos documentos é usado para rastrear quais documentos já foram processados. A ferramenta RabbitMQ [2] pode ser usada para coordenar a extração, criando uma fila de documentos a serem processados. O módulo de extração de recursos pode ser executado em várias instâncias (lado esquerdo da Fig. 4), a fim de acelerar a extração. Cada instância consome da mesma fila de documentos, representada na Fig. 4 como *Extração de recursos MQ* para saber qual texto deve ser processado em seguida.

A extração funciona desta forma para cada texto: (i) extrai todos os recursos encontrados no texto usando *DBpedia Spotlight*; (ii) seleciona apenas as URIs e formas de superfície; (iii) para cada recurso encontrado, faz uma consulta SPARQL na base de conhecimento YAGO procurando nomes alternativos e identificadores registrados para aquele recurso; por último, agrega todas as formas de superfície possíveis para um recurso e os salva no banco de dados.

Listing 3.2 – Exemplo da estrutura de dados do banco de dados de recursos.

```
{
    ...,
    '_id': <doc_id>,
    '_source': {
        'content': <raw text>,
        'resources': [
                'http://dbpedia.org/resource/<resource>',
                'http://dbpedia.org/resource/<resource>',
                ...
        ]
        'labels': [
```

[1] https://www.elastic.co/

[2] https://www.rabbitmq.com/

```
            {
                <surface_form>: [
                    <label>,
                    <label>
                ]
            }, {...}
        ]
    }

}
```

A etapa do transformador, que é realizada quando toda a extração de recursos termina, coleta todos os recursos e identificadores e os organiza em uma grande lista de mapeamento. Essa lista mapeia todos os identificadores possíveis de um conceito para um identificador principal, que vai substituir todas as menções possíveis desse conceito. Uma vez que a lista de mapeamento é construída, uma tarefa de substituição via expressão regular é realizada para fazer todas as substituições, incluindo palavras com hífen. Uma fila é usada para gerenciar todos os textos que estão sendo processados, semelhante à fase de extração de recursos.

O processo de normalização pode ser resumido nos seguintes passos:

1. Indexar os documentos a serem normalizados em um índice e somente os IDs desses documentos em outro índice;
2. Criar uma fila de documentos e uma fila recursos;
3. Extrair os recursos dos textos, controlando os documentos que já foram processados usando a fila de documentos;
4. Criar um dicionário de substituição de recursos;
5. Transformação dos textos usando o dicionário de recursos.

3.2 NORMALIZAÇÃO MANUAL USANDO CONHECIMENTO DE DOMÍNIO

A normalização do corpus pode ser realizada usando possíveis formas de superfície para substituir diferentes representações de um conceito em documentos de texto. Como foi explorado na seção 2.4, abordagens usando bases de conhecimento podem ser prejudicadas pela falta de completude da base de conhecimento. Devido à web semântica ser uma forma recente de gerenciar conhecimento, e de as bases de conhecimento disponíveis ainda serem incipientes e incompletas, é razoável trazer outra abordagem, na qual as formas de superfície para substituição são manualmente selecionadas e agrupadas, usando um conhecimento de domínio, a fim de testar com mais precisão a hipótese deste trabalho.

Na normalização manual, dado um conhecimento de domínio pré existente sobre o contexto de um determinado corpus, define-se um conjunto de conceitos e as respectivas formas de superfície relacionadas a esse conceito, para substituição. A normalização é manual pois as formas de superfície para substituição são definidas manualmente. Porém, o resto do processo, como a substituição das palavras no texto, pode ser automática. Ainda, caso haja alguma base de informações de onde se possa extrair essas informações automaticamente, esse processo se tornaria automático também, somente sem usar a base de conhecimento. A mesma estrutura de substituição pode ser criada, mas não precisa se restringir ao formato JSON apresentado na lista 3.1.

3.3 TOPIC MODELING

Já na extração dos tópicos, a primeira etapa é o pré processamento, conforme as seguintes etapas: (i) remover caracteres inválidos e pontuação; (ii) transformar as palavras em minúsculas; (iii) transformar o texto em um vetor de palavras (*tokenizar*); (iv) remover palavras irrelevantes, chamadas de *stopwords* (palavras muito comuns que não agregam significado); e (v) lematizar (remover inflexões de palavras, retornando-as à sua forma de raiz, por exemplo, "dito"para "dizer"), mantendo apenas as palavras que são substantivos, advérbios e adjetivos. Além das *stopwords*, uma lista de palavras muito frequentes também pode ser removida, caso existam especificidades no corpus sendo usado. Após o pré-processamento, o vocabulário de palavras está pronto servir de entrada para o algoritmo LDA.

Então, para cada conjunto de dados pré-processado, estimam-se as probabilidades de cada palavra em cada tópico e cada tópico em cada documento. Como resultado, temos a distribuição de palavras em cada tópico e a distribuição de tópicos para cada documento. Com isso é possível calcular a coerência, usando os tópicos extraídos do corpus.

- Para cada corpus pré-processado, estimam-se as probabilidades de cada palavra em cada tópico e de cada tópico em cada documento;
- Como resultado, temos a distribuição de palavras por tópico e a distribuição de tópicos por documento;
- Com isso é possível calcular a coerência, usando os tópicos extraídos do corpus.

.

3.4 TESTE ESTATÍSTICO

Para o teste estatístico foi utilizado o teste não paramétrico de Mann-Whitney. Os resultados da coerência dos tópicos foram obtidos utilizando a estratégia descrita a seguir:

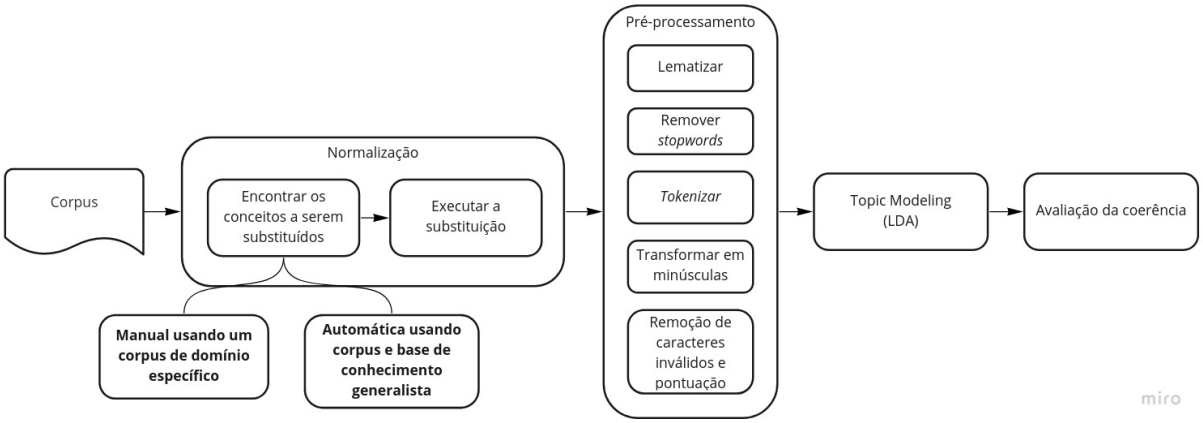
1. Após a transformação, o corpus foi amostrado em subconjuntos aleatórios de textos. Os textos foram amostrados sem remover os textos já escolhidos, ou seja, com reposição, devido ao tamanho do conjunto de documentos disposto para teste;
2. Para cada corpus de teste criado, o *score* de coerência é calculado, usando sempre a mesma configuração de hiper-parâmetros do LDA, específico para cada corpus, mas sem mudar nos testes do mesmo corpus.

Os hiper-parâmetros foram fixados porque o objetivo deste trabalho é analisar a os resultados (coerência dos tópicos) do LDA para diferentes entradas (corpus original e corpus normalizado).

3.5 EXPERIMENTOS CONDUZIDOS

A Figura 5 apresenta o fluxo dos experimentos conduzidos, explicitando que a normalização dos corpora é realizada antes do pré-processamento no processo de geração dos tópicos.

Figura 5 – Fluxo dos experimentos de normalização automática e manual.



4 RESULTADOS

O objetivo deste trabalho é verificar se a normalização semântica de documentos de texto afeta a coerência de tópicos extraídos via algoritmo LDA de Topic Modeling, conforme exposto na seção 1.2. Neste capítulo serão mostrados os resultados para as duas propostas de normalização: automática, usando web semântica e manual, usando conhecimento de domínio.

Foram executados alguns testes a fim de validar ou rejeitar as hipóteses, que são resumidos abaixo e explicado com mais detalhes nas seções a seguir.

Hipótese 1: Corpus Genérico
A normalização dos corpora 20 News-group e Reuters afetará a métrica de coerência Cv dos tópicos, extraídos com o algoritmo LDA, em relação ao corpus não normalizado.

Teste

1. Cria-se para cada documentos do corpus uma versão normalizada a partir da DBPedia e Yago;
2. Para cada corpus:
   a. Seleciona-se aleatoriamente 30 vezes uma amostra de 1500 documentos;
   b. Para cada uma das 30 amostras estima-se o conjunto de tópicos;
   c. Para cada estimado calcula-se a métrica Cv;
3. Repete-se o item 2 para 40, 60, 80 e 100 amostras;
4. Aplica-se para cada número de amostras (40, …, 100) o teste Mann-Whitney com as métricas Cv de forma a rejeitar ou aceitar a hipótese.

Hipótese 2: Corpus Domínio Específico
A normalização de 4 conceitos importantes no Antigo Testamento afetará a métrica de coerência Cv dos tópicos, extraídos com o algoritmo LDA, em relação ao corpus não normalizado.

Teste 1

1. Cria-se para cada documento do corpus uma versão normalizando os 4 conceitos;
2. Seleciona-se aleatoriamente 30 vezes uma amostra de 500 documentos;
3. Para cada uma das 30 amostras estima-se o conjunto de tópicos;
4. Para cada conjunto de tópicos estimado calcula-se a métrica Cv;
5. Aplica-se o teste Mann-Whitney com as métricas Cv de forma a rejeitar ou aceitar a hipótese.

Teste 2

1. Repete-se os itens de 2 a 5 do teste 1, selecionando apenas documentos que tiveram conceitos normalizados no item 1.

4.1 NORMALIZAÇÃO AUTOMÁTICA USANDO BASES DE CONHECIMENTO

Dois corpora amplamente utilizados, conhecidos de tarefas de NLP foram usados: 20-Newsgroups [1] e Reuters [2] a fim de testar a hipótese de que a normalização do texto afeta a coerência dos tópicos extraídos via LDA. O corpus 20-Newsgroups tem mais de 18.000 posts agrupados em 20 tópicos já definidos. Está dividido em treinamento e teste, embora para este trabalho foram usados ambos como um conjunto de dados único. Como já existiam 20 tópicos mapeados, foram utilizados também 20 para o hiperparâmetro de número de tópicos. O corpus Reuters contém mais de 10.000 documentos de notícias, totalizando 1.3 milhão de palavras. Os documentos foram classificados originalmente em 90 tópicos e agrupados em dois conjuntos, denominados "treinamento"e "teste". No entanto, para este trabalho, ambos os conjuntos de

[1] https://scikit-learn.org/0.19/datasets/twenty_newsgroups.html
[2] https://www.nltk.org/book/ch02.html

treinamento e teste foram usados como um só corpus para extrair os tópicos e realizar os testes. Além disso, como o corpus original foi anotado em 90 tópicos, então também foi 90 o hiperparâmetro de número de tópicos.

Na Fig. 6 podemos ver a distribuição das palavras por documento em cada corpus utilizado. Há uma grande variedade de comprimento de texto no corpus 20-Newsgroup, assim como a média do comprimento de cada documento é maior antes do pré-processamento. Após o pré-processamento, as palavras úteis restantes eram semelhantes entre os dois corpus.

Na Tabela 1 existem exemplos de ambos os corpus. O corpus 20-Newsgroup tem uma forma de e-mails, textos curtos e o corpus Reuters tem uma forma de notícia. Pode-se observar na Fig. 6 que, após o pré-processamento, o 20-Newsgroup perdeu mais palavras do que o Reuters, pois grande parte dos caracteres não eram letras ou dígitos, que são removidos na etapa de pré-processamento. É possível que isso influencie negativamente o algoritmo entender o contexto de determinadas entidades.

Após algumas análises dos corpora utilizados, os recursos e possíveis formas de superfície associadas foram extraídos do texto e salvos no banco de dados. Com todas as formas de superfície encontradas salvas, uma lista de mapeamento foi construída e usada para transformar os textos, na qual cada possível variação de um conceito foi substituído por uma única forma de superfície.

A biblioteca *Gensim* foi usada para implementar o LDA, pois ela implementa o framework proposto por (RÖDER; BOTH; HINNEBURG, 2015) para calcular métricas de coerência [3]. Os hiper-parâmetros *chunksize* e iterações foram definidos como 100, 10 passes e o alfa foi definido como simétrico.

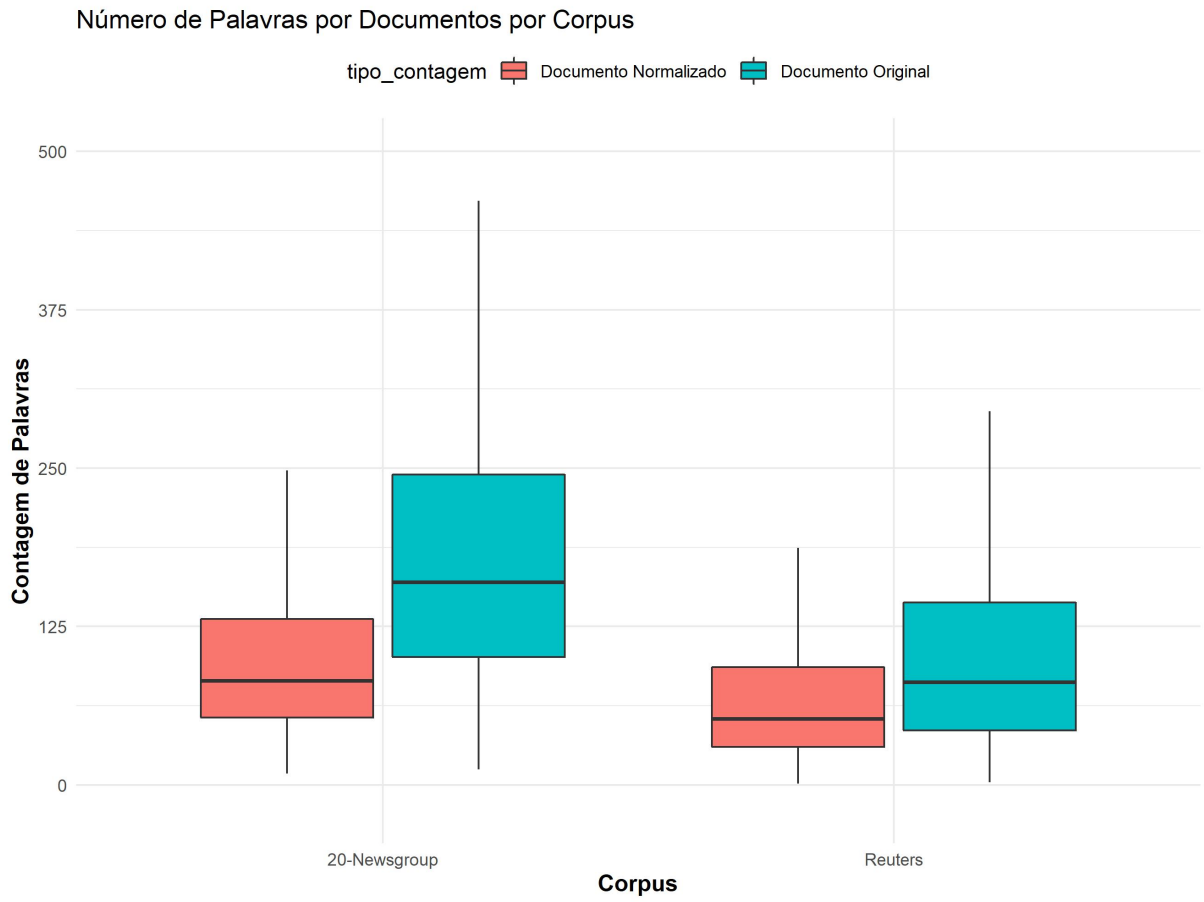
Os resultados para os experimentos são mostrados na Tabela 2. A coerência C_V dos tópicos para o corpus 20-Newsgroup diminuiu com a normalização do corpus, já para o corpus da Reuters a coerência do tópico aumentou com significância estatística para corpora de teste com 80 e 100 amostras de dados. Cada amostra representa um conjunto de 1500 textos.

Sobre o valor da coerência: é um valor bom ou estamos tentando otimizar algo que já está ruim? O artigo referência da métrica (RÖDER; BOTH; HINNEBURG) expõe alguns testes em alguns corpora, incluindo o 20-newsgroup. O valor médio da métrica C_V nos corpora testados nesse trabalho é 0.628, mas o autor não expõe o método de pré processamento dos dados ou mesmo os hiper-parâmetros usados no LDA, sendo impossível reproduzir os resultados.

Pode-se perceber, pelos resultados na Tabela 2, que há um efeito positivo na coerência do tópico no corpus da Reuters, enquanto no 20-Newsgroup parece ter diminuído. É possível elencar algumas hipóteses para essas diferenças: (i) o tamanho dos documentos é importante, porque em textos pequenos é mais difícil obter recursos

[3] https://radimrehurek.com/gensim/models/coherencemodel.html

Figura 6 – Número de palavras por documento por corpus. O boxplot vermelho mostra o contador de todas as palavras tokenizadas. O azul mostra os documentos pré-processados, onde *stopwords* e palavras muito frequentes foram retiradas. Este pré-processamento é o mesmo que os documentos são expostos antes do algoritmo de modelagem de tópicos.



devido ao fato de que há pouco contexto para o algoritmo desambiguar recursos; (ii) a natureza do texto, já que o 20-Newsgroup tem um uma forma de escrita de e-mail e a Reuters é composta por notícias; ou (iii) a abrangência da base de conhecimento é limitada e não trouxe formas de superfície suficientes.

Sobre a primeira hipótese do tamanho dos documentos, é possível dizer que quando um documento é muito pequeno, o algoritmo de *entity linking* pode não ter certeza suficiente de que uma palavra corresponda a um recurso, por isso não o captura. Embora o 20-Newsgroup tenha um tamanho maior de documentos, ambos os corpora são pequenos, com uma média de menos de 200 palavras por documento. Ainda após o pré-processamento, os documentos ficam com menos de 125 palavras. Além disso, pela Fig. 6 podemos ver que o número de palavras válidas diminui muito mais no 20-Newsgroup do que no Reuters corpus. Assim, podemos inferir que, embora o número total de palavras seja maior no 20-Newsgroup, a quantidade de palavras

Quadro 1 – Exemplo de documento de cada corpus.

20-Newsgroup	"From: sxs@extol.Convergent.Com (S. Sridhar) Subject: Re: tvtwm icon manager Organization: Unisys Open Systems Group, San Jose Lines: 22 In article <13960@risky.Convergent.COM>, sxs@extol.Convergent.Com (S. Sridhar) writes: |> Keywords: tvtwm icon manager |> |> Need help on resource bindings for tvtwm. Here's what I'd like to |> see the icon manager do. |> |> Say I iconify a window and this shows up on the icon list. Now when I |> pan into another section of the virtual desktop and try to deiconify |> the window that I iconed (sp ?) earlier, I'd like this window to |> deiconify in the current region. |> |> Any resources that I can use to do this ? Or more important, can I |> do this ? Rather find it painful to remember where I iconified a |> window, go back there and deiconify. Or simply, it is a pain to |> pan around to get to a deiconified window. |> |> Thanks, |> |> ssridhar@convergent.com |> |> Just opened up the distribution. "
Reuters	"AMERICAN CENTURY <ACT> RESTATES EARNING-SAmerican Century Corp said ithas restated its earnings for the fiscal year ended June 30,1986 to provide an additional five mln dlrs to its loan lossallowance, causing a restated year-end net loss of 14,937,000dlrs, instead of 9,937,000 dlrs.The company said the change came after talks with theSecurities and Exchange Commission on the company's judgementin considering the five mln dlrs collectible.In the note to its 1986 financial statement, AmericanCentury said it considered the five mln dlrs collectible,making its loan loss provision less than required.The company said in spite of the SEC decision, it stillfeels its allowance for possible loan losses at June 30, 1986was adequate and that it has considered all relevantinformation to determine the collectibility of the five mln dlrreceivable.But, it said continued disagreement with the SEC staffwould not be in its best interest."

Fonte: Elaborado pela autora.

válidas para o algoritmo de extração de recursos é muito próxima ao corpus da Reuters. Ademais, como o contexto é importante, e o contexto é o conjunto de palavras em torno de um conceito, é muito difícil para o algoritmo de NEL vincular um recurso útil se apenas algumas palavras forem válidas.

Na segunda hipótese sobre a natureza do texto, podemos verificar pela Tabela 1 que os textos têm naturezas muito diferentes. Um texto no formato e-mail é muito mais sujeito a ter símbolos e iniciais ou acrônimos, como visto no primeiro texto do

Tabela 1 – Médias das coerências, usando a métrica C_V de coerência, com as top-10 palavras por tópico e *p-valor* correspondentes. Cada amostra corresponde a 1500 textos.

Nº amostras	Corpus	Normalizado	Original	p-valor
30	20news	0.5129	0.5209	0.1809
30	reuters	0.4195	0.4182	0.728265
40	20news	0.5166	0.5225	0.246248
40	reuters	0.4199	0.4190	0.613431
60	20news	0.5148	0.5214	0.080955
60	reuters	0.4213	0.4172	0.087556
80	20news	0.5116	0.5248	0.00013
80	reuters	**0.4224**	**0.4176**	**0.033356**
100	20news	0.5111	0.5246	2.4e-05
100	reuters	**0.4224**	**0.4184**	**0.04604**

20-Newsgroup da Tabela 1. Na Reuters, pode-se verificar que o texto está mais fluente e sem muitos símbolos.

Em relação à terceira hipótese, durante a execução dos testes notou-se que muitos recursos do Yago vinculados na página da DBpedia não estavam disponíveis. De um total de 342019 entidades localizadas nos documentos, via ligação de entidades (DBPedia Spotlight), apenas 99878 entidades ligadas às base da DBPedia foram encontradas na base Yago. Ou seja, menos de 30% das entidades tiveram possíveis nomes alternativos para serem substituídos. Não são 30% das palavras ao todo, mas sim das entidades que puderam ser encontradas na DBPedia. A quantidade (e qualidade) dos nomes alternativos encontradas pela Yago são parte fundamental do processo, pois possibilita que os mesmos conceitos em outras partes do texto possam ser substituídas.

4.2 NORMALIZAÇÃO MANUAL USANDO CONHECIMENTO DE DOMÍNIO

Para realizar os testes dessa proposta foi utilizado o Antigo Testamento da Bíblia cristã. O conhecimento de domínio está no entendimento da língua hebraica, língua original dos textos do antigo testamento, e os termos originais para a tradução dos conceitos que foram usados para normalizar o texto. Esse corpus foi escolhido devido ao conhecimento de domínio prévio da autora.

O corpus foi o Antigo Testamento da bíblia cristã na versão King James[4] (King James Version - KJV), em inglês. Para ter um entendimento completo da Bíblia Sagrada, teólogos e historiadores recorrem aos manuscritos originais dos textos. No caso do Antigo Testamento, os escritos são em hebraico, e no Novo Testamento, em grego. Em seu livro Inteligência Humilhada, Jonas Madureira expõe no Capítulo 4 sobre a Antropologia Bíblica, na qual o autor considera quatro palavras, na língua hebraica, que

[4] https://www.kaggle.com/oswinrh/bible

melhor definem o que é o ser humano. Na perspectiva bíblica, essas quatro palavras, a saber, *néfesh* (alma), *lebab* (coração), *basar* (carne) e *ruah* (espírito), estão presentes e são recorrentes no Antigo Testamento, sendo fundamentais para a compreensão do homem, biblicamente. Nas palavras de (MADUREIRA):

> Segundo as Escrituras, o homem é uma *alma* que não se satisfaz com as coisas deste mundo. Além disso, esse homem insatisfeito é governado por um centro existencial chamado *coração*, um centro que reúne todas as grandes decisões de uma pessoa. As Escrituras também dizem que o ser humano é finito, marcado pela *carne*, que, de tempos em tempos, anuncia ao homem que sua vida é breve. Por fim, as Escrituras também revelam que a vida (ou o *espírito*) do homem - bem como a vida de todos os seres vivos - sempre dependeu da vontade de Deus. Portanto, o homem é "alma", "coração", "carne"e "espírito".

Na língua hebraica existiam menos palavras do que existe hoje no inglês, por exemplo. Era comum que uma mesma palavra fosse usada em diferentes contextos, tendo a mesma raiz de significado, mas com conotações diferentes. Partindo do fato de que os textos são traduzidos de uma língua mais antiga (com menos palavras, mas não menos rica), é possível usar algumas palavras dessa língua para substituir diversas palavras que usamos hoje. Nesse contexto, entender que uma mesma palavra era utilizada com outras conotações nos ajuda a entender de forma mais profunda o real significado dos textos bíblicos. Logo, o conhecimento de domínio utilizado é o conhecimento da Bíblia e sua tradução da língua original para o inglês.

A fim de explicitar, vemos a palavra *néfesh*, que é traduzida em sua maior parte como "alma". Hoje em dia podemos pensar em alma como sendo o espírito que vive dentro do homem, segundo o dualismo grego que permeia nosso entendimento. Porém, a Bíblia traz uma visão diferente desse conceito, como se pode notar em Gênesis 2.7 ("E o senhor Deus formou o homem do pó da terra e soprou-lhe nas narinas o fôlego de vida; e o homem tornou-se alma vivente."). Então, no entendimento de ser humano que a Bíblia traz, o homem *é* uma alma, não *tem* uma alma (PALAU, 2007).

Além disso, a palavra *néfesh* também é traduzida como "garganta". Podemos ver outras passagens nas quais essa palavra é traduzida nessa conotação, como Isaías 5.14 ("Por isso o Sheol aumentou o *apetite* [néfesh] e abriu totalmente a boca; [...]") e Habacuque 2.5 ("O homem arrogante não permanece. Seu *desejo* [néfesh] impetuoso é como o Sheol; [...]"). Assim, a imagem de "garganta faminta"comunica mais adequadamente o que o homem é, segundo as Escrituras, do que "substância incorpórea", que é o conceito que alma elucida nos dias de hoje (MADUREIRA, 2017). Enfim, tanto "apetite", quanto "desejo", quanto "alma"podem ser substituídos no texto por *néfesh*, e isso é a base do dicionário de substituição para normalizar os textos.

O site utilizado para encontrar as referências dos versículos onde deveria acontecer a substituição foi o Blue Letter Bible [5]. Devido aos termos de uso da plataforma impedirem o uso de *web scrapers*, a seleção dos versículos foi realizada manualmente. Por exemplo a palavra *néfesh*[6] pode substituir "alma"em Gênesis 2.7, "apetite"em Eclesiastes 6.7, "desejo"em Jeremias 22.27, etc.

No pré-processamento do LDA foram removidas outras palavras que não entraram nas *stopwords* ou na lematização, mas era muito comuns no texto. São elas: '*thee*', '*also*', '*thing*', '*thus*', '*therefore*', '*thereof*', '*even*'.

Quanto ao número de tópicos, como não havia um pré conhecimento definido sobre quantos tópicos existiam nos textos, um *grid search* foi aplicado, otimizando a coerência C_V, a fim de encontrar o número de tópicos adequado para este corpus. Os valores escolhidos como hiper-parâmetros foram 8 tópicos, *alpha* de 0.6 e *beta* de 0.9, e usando as mesmas 100 iterações, *chunksize* de 100 e 10 passes do experimento anterior.

Após definir os hiper-parâmetros do LDA, foi executado o pré-processamento dos dados e a estimação das distribuições de probabilidades das palavras nos tópicos e dos tópicos nos documentos. Tendo a distribuição das palavras em cada tópico, foram usadas as 10 primeiras palavras (as primeiras palavras tem a maior probabilidade de estarem naquele tópico) para calcular a coerência.

Foram realizados dois experimentos: (i) AT 1, que usou todo o texto do antigo testamento e (ii) AT 2, que usou apenas os textos que tiveram alguma substituição. A abordagem usando todo o antigo testamento não resultou em coerências com uma diferença estatística significativa. A hipótese é que, pelo motivo de a construção do dicionário de substituição ser manual e demorada, e terem sido substituídos apenas 4 conceitos, não houve substituição o suficiente para representar alguma diferença real. Esses 4 conceitos correspondem a 1425 palavras substituídas de um total de 789634 palavras, totalizando 0.18% . Quanto somente os versículos que sofreram uma substituição são selecionados, são 39766 palavras, ficando então 3.58% do texto. É mais fácil de mensurar esse percentual nesse corpus, com a substituição, pois as substituições foram apenas palavra por palavra, sendo as palavras conhecidas e a substituição direta. No caso da substituição automática, é possível que algumas formas de superfície contenham mais de uma palavra e também sejam substituídas por uma ou mais palavras.

Podemos ver no experimento AT-1 que a coerência aumenta 0,006872 (1,88%), mas essa diferença não é estatisticamente significativa. Já no experimento AT-2, usando apenas os versículos com substituição, a coerência aumenta 0,029156 (10,73%), e a diferença é estatisticamente significativa.

[5] https://www.blueletterbible.org
[6] https://www.blueletterbible.org/lexicon/h5315/kjv/wlc/0-1/

Tabela 2 – Média da coerência das top-10 palavras em cada tópico, usando a métrica C_V, e *p-valor* correspondentes para o conjunto de dados inteiro (todo o antigo testamento, AT-1) e para o conjunto de dados contendo somente os versículos que sofreram alteração (AT-2).

Experimento	% subst.	Original	Normalizado	p-valor
AT-1	0.18%	0.3657	0.3725	0.610008
AT-2	3.58%	**0.2715**	**0.3007**	**0.000077**

Então, respondendo a pergunta de pesquisa exposta no Capítulo 1: a normalização semântica de documentos de texto afeta a interpretabilidade de tópicos extraídos via algoritmo LDA de *Topic Modeling*? Esse resultado mostra que a normalização semântica afeta sim, e de forma positiva, a coerência de tópicos extraídos via LDA, se a substituição for percentualmente significativa.

5 CONCLUSÃO

Este trabalho teve como objetivo explorar a possibilidade de normalização semântica de corpus a fim de avaliar se essa normalização pode afetar a coerência de tópicos extraídos via algoritmo LDA de *Topic Modeling*. Foram propostas duas abordagens para testar essa possibilidade: uma normalização semântica automática, usando bases de conhecimento, e uma normalização semântica manual, usando conhecimento de domínio.

A normalização semântica automática de corpus foi testada usando dois corpora conhecidos: 20-newsgroup e Reuters. O corpus foi submetido à extração de entidades nomeadas, no qual cada conceito encontrado no texto foi ligado a um recurso na DBpedia; essas entidades (recursos) foram ligadas a outros recursos na base de conhecimento Yago, extraindo outras possíveis formas de superfície para o mesmo conceito; essas novas formas de superfície foram organizadas em um dicionário de substituição e substituídas no corpus original, formando o corpus normalizado. Após a normalização, os corpora original e normalizado formaram um novo corpus via amostragem. Foram amostrados 30, 40, 60, 80 e 100 amostras de 1500 textos, com reposição, a fim de testar a reprodutibilidade do cálculo da normalização. Após, o pré-processamento para o LDA foi realizado e as distribuições de probabilidade das palavras nos tópicos e dos tópicos nos documentos foram estimadas. Com os tópicos estimados, a coerência foi calculada para cada conjunto de textos amostrados e o teste estatístico Mann-Whitney foi aplicado. Os resultados mostraram que, para 80 e 100 amostras, a coerência dos tópicos extraídos do corpus normalizado aumentou com significância estatística para um dos corpora sendo testado.

A normalização semântica manual de corpus foi testada usando o corpus do Antigo Testamento da Bíblia Cristã, na versão King James. Foram usados 4 conceitos importantes do Antigo Testamento para realizar a substituição. A relação dos conceitos com as formas de superfície foi extraída de forma manual. Com essa relação, a substituição foi realizada, gerando o corpus normalizado. Depois da normalização, ambos os corpora original e normalizado formaram um novo corpus via amostragem. Como o tamanho desse corpus era menor, foram amostradas somente 30 amostras de 500 textos. Dois conjuntos de dados de experimento foram construídos: um com o corpus todo e outro apenas com os textos que tiveram alguma substituição. O pré-processamento foi o mesmo usado no teste anterior, com exceção de algumas palavras específicas que foram removidas no processo. Após o pré-processamento, as distribuições de probabilidade foram calculadas e os tópicos estimados. Com os tópicos, a coerência pode ser calculada para cada conjunto de textos amostrados e o teste estatístico foi aplicado. Para o conjunto de dados correspondente ao corpus todo, o resultado mostra que houve um aumento na coerência, mas não é estatisticamente significativa. Porém, no

conjunto de dados correspondente ao corpus contendo apenas os textos que sofreram alteração, houve um aumento significante na coerência.

Portanto, este trabalho mostra que a normalização semântica de corpus afeta, de forma positiva, a coerência dos tópicos extraídos via algoritmo LDA de *Topic Modeling* caso haja um percentual minimamente considerável de texto normalizado. Podemos concluir também que as bases de conhecimento ainda são incipientes para este tipo de aplicação.

REFERÊNCIAS

ALETRAS, Nikolaos; STEVENSON, Mark. Evaluating topic coherence using distributional semantics. *In*: PROCEEDINGS of the 10th International Conference on Computational Semantics (IWCS 2013)–Long Papers. [*S.l.*: *s.n.*], 2013. P. 13–22.

ALLAHYARI, Mehdi. **Semantic Web Topic Models: Integrating Ontological Knowledge and Probabilistic Topic Models**. 2016. Tese (Doutorado) – University of Georgia.

ARAÚJO, Carlos Alberto Ávila. Correntes teóricas da ciência da informação. **Ciência da informação**, SciELO Brasil, v. 38, n. 3, 2009.

AUER, Sören; BIZER, Christian; KOBILAROV, Georgi; LEHMANN, Jens; CYGANIAK, Richard; IVES, Zachary. Dbpedia: A nucleus for a web of open data. *In*: THE semantic web. [*S.l.*]: Springer, 2007. P. 722–735.

AZIMI, Sasan; VEISI, Hadi; AMOUIE, Reyhaneh. A method for automatic detection of acronyms in texts and building a dataset for acronym disambiguation. *In*: IEEE. 2019 5th Iranian Conference on Signal Processing and Intelligent Systems (ICSPIS). [*S.l.*: *s.n.*], 2019. P. 1–4.

BERNERS-LEE, Tim; HENDLER, James; LASSILA, Ora *et al.* The semantic web. **Scientific american**, New York, NY, USA: v. 284, n. 5, p. 28–37, 2001.

BIZER, Christian; HEATH, Tom; IDEHEN, Kingsley; BERNERS-LEE, Tim. Linked data on the web (LDOW2008). *In*: PROCEEDINGS of the 17th international conference on World Wide Web. [*S.l.*: *s.n.*], 2008. P. 1265–1266.

BLEI, David; CARIN, Lawrence; DUNSON, David. Probabilistic Topic Models: A focus on graphical model design and applications to document and image analysis. **IEEE signal processing magazine**, NIH Public Access, v. 27, n. 6, p. 55, 2010.

BLEI, David; LAFFERTY, John. Text mining: Theory and applications, chapter topic models. **Taylor and Francis**, 2009.

BLEI, David M; NG, Andrew Y; JORDAN, Michael I. Latent dirichlet allocation. **Journal of machine Learning research**, v. 3, Jan, p. 993–1022, 2003.

BRICKLEY, Dan; GUHA, Ramanathan V; MCBRIDE, Brian. RDF Schema 1.1. **W3C recommendation**, W3C, v. 25, p. 2004–2014, 2014.

BUSH, Vannevar *et al.* As we may think. **The atlantic monthly**, v. 176, n. 1, p. 101–108, 1945.

CHANG, Jonathan; GERRISH, Sean; WANG, Chong; BOYD-GRABER, Jordan L; BLEI, David M. Reading tea leaves: How humans interpret topic models. *In*: ADVANCES in neural information processing systems. [*S.l.*: *s.n.*], 2009. P. 288–296.

DAS, Rajarshi; ZAHEER, Manzil; REDDY, Siva; MCCALLUM, Andrew. Question answering on knowledge bases and text using universal schema and memory networks. **arXiv preprint arXiv:1704.08384**, 2017.

DE MELO, Gerard; SIERSDORFER, Stefan. Multilingual text classification using ontologies. *In*: SPRINGER. EUROPEAN Conference on Information Retrieval. [*S.l.*: *s.n.*], 2007. P. 541–548.

DIAS, Tatiane Domingos; SANTOS, Neide. Web semântica: conceitos básicos e tecnologias associadas. **Cadernos do IME-Série Informática**, v. 14, p. 80–92, 2003.

DOBBELAERE, Philippe; ESMAILI, Kyumars Sheykh. Kafka versus RabbitMQ: A comparative study of two industry reference publish/subscribe implementations: Industry Paper. *In*: PROCEEDINGS of the 11th ACM international conference on distributed and event-based systems. [*S.l.*: *s.n.*], 2017. P. 227–238.

DONG, Xin; GABRILOVICH, Evgeniy; HEITZ, Geremy; HORN, Wilko; LAO, Ni; MURPHY, Kevin; STROHMANN, Thomas; SUN, Shaohua; ZHANG, Wei. Knowledge vault: A web-scale approach to probabilistic knowledge fusion. *In*: PROCEEDINGS of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. [*S.l.*: *s.n.*], 2014. P. 601–610.

ELASTICSEARCH. [*S.l.*: *s.n.*], 2019. Disponível em: `https://www.elastic.co/pt/`. Acesso em: 29 jul. 2019.

GALÁRRAGA, Luis; RAZNIEWSKI, Simon; AMARILLI, Antoine; SUCHANEK, Fabian M. Predicting completeness in knowledge bases. *In*: PROCEEDINGS of the tenth acm international conference on web search and data mining. [*S.l.*: *s.n.*], 2017. P. 375–383.

GARLA, Vijay N; BRANDT, Cynthia. Ontology-guided feature engineering for clinical text classification. **Journal of biomedical informatics**, Elsevier, v. 45, n. 5, p. 992–998, 2012.

GREENE, Derek; CROSS, James P. Exploring the political agenda of the European parliament using a dynamic topic modeling approach. **Political Analysis**, Cambridge University Press, v. 25, n. 1, p. 77–94, 2017.

GRUBER, Thomas R. Toward principles for the design of ontologies used for knowledge sharing? **International journal of human-computer studies**, Elsevier, v. 43, n. 5-6, p. 907–928, 1995.

GRUBER, Thomas R. A translation approach to portable ontology specifications. *In*.

HITZLER, Pascal; KROTZSCH, Markus; RUDOLPH, Sebastian. **Foundations of semantic web technologies**. [*S.l.*]: Chapman e Hall/CRC, 2009.

KADDOURA, Sanaa; D. AHMED, Rowanda. A comprehensive review on Arabic word sense disambiguation for natural language processing applications. **Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery**, Wiley Online Library, e1447, 2022.

KEKEÇ, Taygun; MAATEN, Laurens van der; TAX, David MJ. PAWE: Polysemy aware word embeddings. *In*: PROCEEDINGS of the 2nd International Conference on Information System and Data Mining. [*S.l.*: *s.n.*], 2018. P. 7–13.

KRUSKAL, William H. Historical notes on the Wilcoxon unpaired two-sample test. **Journal of the American Statistical Association**, Taylor & Francis Group, v. 52, n. 279, p. 356–360, 1957.

LAJUS, Jonathan; SUCHANEK, Fabian M. Are all people married? Determining obligatory attributes in knowledge bases. *In*: PROCEEDINGS of the 2018 World Wide Web Conference. [*S.l.*: *s.n.*], 2018. P. 1115–1124.

LAU, Jey Han; NEWMAN, David; BALDWIN, Timothy. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. *In*: PROCEEDINGS of the 14th Conference of the European Chapter of the Association for Computational Linguistics. [*S.l.*: *s.n.*], 2014. P. 530–539.

MADUREIRA, Jonas. **Inteligência humilhada**. [*S.l.*]: Sociedade Religiosa Edições Vida Nova, 2017.

MANOLA, Frank; MILLER, Eric; MCBRIDE, Brian *et al.* RDF primer. **W3C recommendation**, v. 10, n. 1-107, p. 6, 2004.

NANNI, Federico; PONZETTO, Simone Paolo; DIETZ, Laura. Entity-aspect linking: providing fine-grained semantics of entities in context. *In*: PROCEEDINGS of the 18th ACM/IEEE on Joint Conference on Digital Libraries. [*S.l.*: *s.n.*], 2018. P. 49–58.

NEWMAN, David; LAU, Jey Han; GRIESER, Karl; BALDWIN, Timothy. Automatic evaluation of topic coherence. *In*: HUMAN language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics. [*S.l.*: *s.n.*], 2010. P. 100–108.

NOY, Natalya Fridman; MCGUINNESS, Deborah L. Ontology Development 101 : A Guide to Creating Your First Ontology. *In*.

O'CALLAGHAN, Derek; GREENE, Derek; CARTHY, Joe; CUNNINGHAM, Pádraig. An analysis of the coherence of descriptors in topic modeling. **Expert Systems with Applications**, Elsevier, v. 42, n. 13, p. 5645–5657, 2015.

PALAU, José Roberto Fortes. **A Força Salvıfica da Mortificação**. 2007. Tese (Doutorado) – PUC-Rio.

PIANTADOSI, Steven T; TILY, Harry; GIBSON, Edward. The communicative function of ambiguity in language. **Cognition**, Elsevier, v. 122, n. 3, p. 280–291, 2012.

POPOV, Alexander. Neural network models for word sense disambiguation: an overview. **Cybernetics and information technologies**, v. 18, n. 1, p. 139–151, 2018.

PRUD'HOMMEAUX, Eric; SEABORNE, Andy. SPARQL query language for RDF. W3C recommendation, W3C. **Retrieved November**, v. 16, p. 2009, 2008.

REBELE, Thomas; SUCHANEK, Fabian M.; HOFFART, Johannes; BIEGA, Joanna; KUZEY, Erdal; WEIKUM, Gerhard. YAGO: A Multilingual Knowledge Base from Wikipedia, Wordnet, and Geonames. *In*: THE Semantic Web - ISWC 2016 - 15th International Semantic Web Conference, Kobe, Japan, October 17-21, 2016, Proceedings, Part II. [*S.l.*: *s.n.*], 2016. P. 177–185.

RÖDER, Michael; BOTH, Andreas; HINNEBURG, Alexander. Exploring the space of topic coherence measures. *In*: PROCEEDINGS of the eighth ACM international conference on Web search and data mining. [*S.l.*: *s.n.*], 2015. P. 399–408.

SCARLINI, Bianca; PASINI, Tommaso; NAVIGLI, Roberto. With more contexts comes better performance: Contextualized sense embeddings for all-round word sense disambiguation. *In*: PROCEEDINGS of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). [*S.l.*: *s.n.*], 2020. P. 3528–3539.

SONG, Chang-Woo; JUNG, Hoill; CHUNG, Kyungyong. Development of a medical big-data mining process using topic modeling. **Cluster Computing**, Springer, p. 1–10, 2017.

STEYVERS, Mark; GRIFFITHS, Tom. Probabilistic topic models. **Handbook of latent semantic analysis**, v. 427, n. 7, p. 424–440, 2007.

SUGANYA, G; PORKODI, R. Ontology Based Information Extraction-A Review. *In*: IEEE. 2018 International Conference on Current Trends towards Converging Technologies (ICCTCT). [*S.l.*: *s.n.*], 2018. P. 1–7.

SYED, Shaheen; SPRUIT, Marco. Full-text or abstract? Examining topic coherence scores using latent dirichlet allocation. *In*: IEEE. 2017 IEEE International conference on data science and advanced analytics (DSAA). [*S.l.*: *s.n.*], 2017. P. 165–174.

TESTE DE MANN-WHITNEY. [*S.l.*: *s.n.*], 2012. Disponível em: `http://www.inf.ufsc.br/~vera.carmo/Testes_de_Hipoteses/Testes_nao_parametricos_Mann-Whitney.pdf`. Acesso em: 21 jan. 2022.

TIAN, Feng; REINWALD, Berthold; PIRAHESH, Hamid; MAYR, Tobias; MYLLYMAKI, Jussi. Implementing a scalable XML publish/subscribe system using relational database systems. *In*: PROCEEDINGS of the 2004 ACM SIGMOD international conference on Management of data. [*S.l.*: *s.n.*], 2004. P. 479–490.

VALLET, David; FERNÁNDEZ, Miriam; CASTELLS, Pablo. An ontology-based information retrieval model. *In*: SPRINGER. EUROPEAN Semantic Web Conference. [*S.l.*: *s.n.*], 2005. P. 455–470.

WAITELONIS, Jörg. **Linked Data Supported Information Retrieval**. 2018. Tese (Doutorado) – Karlsruher Institut für Technologie.

WHAT IS RABBITMQ? [*S.l.*: *s.n.*], 2019. Disponível em: `https://www.cloudamqp.com/blog/part1-rabbitmq-for-beginners-what-is-rabbitmq.html`. Acesso em: 13 out. 2021.

YADAV, Subham; SARKAR, Madhulina. Enhancing sentiment analysis using domain-specific lexicon: A case study on GST. *In*: IEEE. 2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI). [*S.l.*: *s.n.*], 2018. P. 1109–1114.

ZHANG, Shaohua; LOU, Jiong; ZHOU, Xiaojie; JIA, Weijia. Entity Linking Facing Incomplete Knowledge Base. *In*: SPRINGER. INTERNATIONAL Conference on Web Information Systems Engineering. [*S.l.*: *s.n.*], 2018. P. 325–334.

APÊNDICE A – RESULTADOS DA REVISÃO SISTEMÁTICA DA LITERATURA

Este Apêndice contém a tabela com os resultados da seleção dos artigos, com os critérios de inclusão e exclusão de cada um dos artigos.

Tabela 3 – Relação de Trabalhos da RSL.

Author	Publication Year	Inclusion	Exclusion
Vallina, Pelayo; Le Pochat, Victor; Feal, Álvaro; Paraschiv, Marius; Gamba, Julien; Burke, Tim; Hohlfeld, Oliver; Tapiador, Juan; Vallina-Rodriguez, Narseo	2020	ci3	ce1,ce2
Reza Sadeghi, Hamid Reza; Shiry Shiry Ghidary, Saeed; Bakhtiari Bakhtiari Bastaki, Benhur	2020	ci3	ce1,ce2
Bekkali, Mohammed; Lachkar, Abdelmonaime	2019	ci3	ce1,ce2
Boudaer, Glenn; Loeckx, Johan	2016	ci3	ce1,ce2
Guo, Jiafeng; Fan, Yixing; Ai, Qingyao; Croft, W. Bruce	2016	ci3	ce1,ce2
Yin, Peifeng; Liu, Zhe; Xu, Anbang; Nakamura, Taiga	2017	ci3	ce1,ce2
Rouhizadeh, Masoud; Magge, Arjun; Klein, Ari; Sarker, Abeed; Gonzalez, Graciela	2018	ci3	ce1,ce2
Mousavi, Maryam; Steiner, Elena; R.Corman, Steven; Ruston, Scott; Weber, Dylan; Davulcu, Hasan	2021	ci3	ce1,ce2
Shi, Bei; Lam, Wai; Jameel, Shoaib; Schockaert, Steven; Lai, Kwun Ping	2017	ci3	ce1,ce2
Yazdavar, Amir Hossein; Al-Olimat, Hussein S.; Ebrahimi, Monireh; Bajaj, Goonmeet; Banerjee, Tanvi; Thirunarayan, Krishnaprasad; Pathak, Jyotishman; Sheth, Amit	2017	ci3	ce1,ce2
Liu, Hongyu; He, Ruifang; Wang, Haocheng; Wang, Bo	2020	ci3	ce1,ce2
Marukatat, Rangsipan	2020	ci3	ce1,ce2
Färber, Michael; Sampath, Ashwath	2020	ci3	ce1,ce2
Yan, Rui; Song, Yiping; Zhou, Xiangyang; Wu, Hua	2016	ci3	ce1,ce2
Mattis, Toni; Rein, Patrick; Hirschfeld, Robert	2019	ci3	ce1,ce2
Nanni, Federico; Ponzetto, Simone Paolo; Dietz, Laura	2018	ci1,ci2,ci3	
Patwari, Ayush; Goldwasser, Dan; Bagchi, Saurabh	2017	ci3	ce1,ce2
Yan, Rui; Song, Yiping; Wu, Hua	2016	ci3	ce1,ce2

Poghosyan, Gevorg; Ifrim, Georgiana	2019	ci3	ce1,ce2
Kekeç, Taygun; van der Maaten, Laurens; Tax, D. M. J.	2018	ci1,ci2,ci3	
Ahmadvand, Ali; Sahijwani, Harshita; Choi, Jason Ingyu; Agichtein, Eugene	2019	ci3	ce1,ce2
Van Gysel, Christophe; de Rijke, Maarten; Kanoulas, Evangelos	2017	ci3	ce1,ce2
Jayarajah, Kasthuri; Radhakrishnan, Meera; Zakaria, Camellia	2016	ci3	ce1,ce2
Rexha, Andi; Dragoni, Mauro; Kern, Roman	2020	ci3	ce1,ce2
Yu, Hai-Tao; Jatowt, Adam; Blanco, Roi; Joho, Hideo; Jose, Joemon; Chen, Long; Yuan, Fajie	2017	ci3	ce1,ce2
Liu, Gang; Wang, Kai; Liu, Wangyang; Cao, Yang; Li, Guang	2019	ci3	ce1,ce2
Yan, Rui; Zhao, Dongyan	2018	ci3	ce1,ce2
Borah, Angana; Barman, Manash Pratim; Awekar, Amit	2021	ci3	ce1,ce2
Botev, Viktor; Marinov, Kaloyan; Schäfer, Florian	2017	ci3	ce1,ce5
Alhaj, Yousif A.; Wickramaarachchi, Wiraj Udara; Hussain, Aamir; Al-Qaness, Mohammed A. A.; Abdelaal, Hammam M.	2018	ci3	ce1,ce2
Liu, Yozen; Shi, Xiaolin; Pierce, Lucas; Ren, Xiang	2019	ci3	ce1,ce2
Vedula, Nikhita; Parthasarathy, Srinivasan; Shalin, Valerie L.	2017	ci3	ce1,ce2
Zhang, Yanzhao; Zhang, Richong; Kim, Jaein; Liu, Xudong; Mao, Yongyi	2021	ci3	ce1,ce2
Zhang, Chao; Tao, Fangbo; Chen, Xiusi; Shen, Jiaming; Jiang, Meng; Sadler, Brian; Vanni, Michelle; Han, Jiawei	2018	ci3	ce1,ce2
Ai, Qingyao; Zhang, Yongfeng; Bi, Keping; Chen, Xu; Croft, W. Bruce	2017	ci3	ce1,ce2
Gunes, Omer	2016	ci3	ce1,ce2
Rong, Xin; Chen, Zhe; Mei, Qiaozhu; Adar, Eytan	2016	ci3	ce1,ce2
Siddiqui, Tarique; Ren, Xiang; Parameswaran, Aditya; Han, Jiawei	2016	ci3	ce1,ce2
Zhou, Wubai; Xue, Wei; Baral, Ramesh; Wang, Qing; Zeng, Chunqiu; Li, Tao; Xu, Jian; Liu, Zheng; Shwartz, Larisa; Ya. Grabarnik, Genady	2017	ci3	ce1,ce2
Bekkali, Mohammed; Lachkar, Abdelmonaime	2018	ci1,ci2,ci3	ce5

Jiang, Zhengbao; Wen, Ji-Rong; Dou, Zhicheng; Zhao, Wayne Xin; Nie, Jian-Yun; Yue, Ming	2017	ci3	ce1,ce2
Ponza, Marco; Ferragina, Paolo; Chakrabarti, Soumen	2017	ci3	ce1,ce2
Panchendrarajan, Rrubaa; Ahamed, Nazick; Sivakumar, Prakhash; Murugaiah, Brunthavan; Ranathunga, Surangika; Pemasiri, Akila	2017		ce1,ce2,ce3
Jiang, He; Zhang, Jingxuan; Ren, Zhilei; Zhang, Tao	2017	ci3	ce1,ce2
Di Tommaso, Giorgia	2018	ci3	ce1,ce2
Mulder, Mats; Inel, Oana; Oosterman, Jasper; Tintarev, Nava	2021	ci3	ce1,ce2
Guo, Jin; Rahimi, Mona; Cleland-Huang, Jane; Rasin, Alexander; Hayes, Jane Huffman; Vierhauser, Michael	2016	ci3	ce1,ce2
Qin, Xubo; Dou, Zhicheng; Wen, Ji-Rong	2020	ci3	ce1,ce2
Li, Pengfei; Lu, Hua; Zheng, Gang; Zheng, Qian; Yang, Long; Pan, Gang	2019	ci3	ce1,ce2
Chen, Lihan; Liang, Jiaqing; Xie, Chenhao; Xiao, Yanghua	2018	ci3	ce1,ce2
Chang, Che-Chia; Chiu, Shu-I; Hsu, Kuo-Wei	2017	ci3	ce1,ce2
Krishnan, Adit; Sankar, Aravind; Zhi, Shi; Han, Jiawei	2017	ci3	ce1,ce2
Chauhan, Uttam; Shah, Apurva	2021		ce1,ce2,ce3
Yuan, Bo; Gao, Xinbo; Niu, Zhenxing; Tian, Qi	2019	ci3	ce1,ce2
Ma, Tinghuai; Al-Sabri, Raeed; Zhang, Lejun; Marah, Bockarie; Al-Nabhan, Najla	2020	ci3	ce1,ce2
Gupta, Amulya; Zhang, Zhu	2021	ci3	ce1,ce2
Tamine, Lynda; Goeuriot, Lorraine	2021		ce1,ce2,ce3
Uprety, Sagar; Gkoumas, Dimitris; Song, Dawei	2020		ce1,ce2,ce3
Seifollahi, Sattar; Piccardi, Massimo; Jolfaei, Alireza	2021	ci3	ce1,ce2
Joshi, Aditya; Karimi, Sarvnaz; Sparks, Ross; Paris, Cécile; Macintyre, C. Raina	2019		ce1,ce2,ce3

Wilson, Shomir; Schaub, Florian; Liu, Frederick; Sathyendra, Kanthashree Mysore; Smullen, Daniel; Zimmeck, Sebastian; Ramanath, Rohan; Story, Peter; Liu, Fei; Sadeh, Norman; Smith, Noah A.	2018	ci3	ce1,ce2
Chandrasekaran, Dhivya; Mago, Vijay	2021		ce1,ce2,ce3
Tiwari, Akanksha; Weth, Christian Von Der; Kankanhalli, Mohan S.	2018	ci3	ce1,ce2
Antoniak, Maria; Walsh, Melanie; Mimno, David	2021	ci3	ce1,ce2
Meo, Rosa; Sulis, Emilio	2017	ci3	ce1,ce2
Ruder, Sebastian; Vulić, Ivan; Søgaard, Anders	2019		ce1,ce2,ce3
Zhang, Dan; Hulsebos, Madelon; Suhara, Yoshihiko; Demiralp, Çağatay; Li, Jinfeng; Tan, Wang-Chiew	2020	ci3	ce1,ce2
Morstatter, Fred; Liu, Huan	2017	ci3	ce1,ce2
Krishnamurthi, Karthik; Panuganti, Vijayapal Reddy; Bulusu, Vishnu Vardhan	2016	ci3	ce1,ce2
Al-Sallab, Ahmad; Baly, Ramy; Hajj, Hazem; Shaban, Khaled Bashir; El-Hajj, Wassim; Badaro, Gilbert	2017	ci3	ce1,ce2
Tymoshenko, Kateryna; Moschitti, Alessandro	2018	ci3	ce1,ce2
Naili, Marwa; Chaibi, Anja Habacha; Ghezala, Henda Hajjami Ben	2018	ci3	ce1,ce2
Liu, Peng; Zhang, Lemei; Gulla, Jon Atle	2021	ci3	ce1,ce2
Camacho-Collados, Jose; Pilehvar, Mohammad Taher	2018		ce1,ce2,ce3
Yang, Tianchi; Hu, Linmei; Shi, Chuan; Ji, Houye; Li, Xiaoli; Nie, Liqiang	2021	ci3	ce1,ce2
Fernández, Alejandro Moreo; Esuli, Andrea; Sebastiani, Fabrizio	2016	ci3	ce1,ce2
Can, Burcu; Manandhar, Suresh	2018	ci3	ce1,ce2
Vuong, Tung; Andolina, Salvatore; Jacucci, Giulio; Ruotsalo, Tuukka	2021	ci3	ce1,ce2
Jauhiainen, Tommi; Lui, Marco; Zampieri, Marcos; Baldwin, Timothy; Lindén, Krister	2019		ce1,ce2,ce3
Cantini, Riccardo; Marozzo, Fabrizio; Bruno, Giovanni; Trunfio, Paolo	2021	ci3	ce1,ce2
Yadollahi, Ali; Shahraki, Ameneh Gholipour; Zaiane, Osmar R.	2017		ce1,ce2,ce3

Guo, Yangyang; Cheng, Zhiyong; Nie, Liqiang; Wang, Yinglong; Ma, Jun; Kankanhalli, Mohan	2019	ci3	ce1,ce2
Obin, Nicolas; Roebel, Axel	2016	ci3	ce1,ce2
Nguyen, Quoc Viet; Duong, Chi Thang; Nguyen, Thanh Tam; Weidlich, Matthias; Aberer, Karl; Yin, Hongzhi; Zhou, Xiaofang	2017	ci3	ce1,ce2
Abukmeil, Mohanad; Ferrari, Stefano; Genovese, Angelo; Piuri, Vincenzo; Scotti, Fabio	2021		ce1,ce2,ce3
Giachanou, Anastasia; Crestani, Fabio	2016	ci3	ce1,ce2
Chatzakou, Despoina; Leontiadis, Ilias; Blackburn, Jeremy; Cristofaro, Emiliano De; Stringhini, Gianluca; Vakali, Athena; Kourtellis, Nicolas	2019	ci3	ce1,ce2
Rudra, Koustav; Ganguly, Niloy; Goyal, Pawan; Ghosh, Saptarshi	2018	ci3	ce1,ce2
Srba, Ivan; Bielikova, Maria	2016		ce1,ce2,ce3
Der Weth, Christian Von; Abdul, Ashraf; Kashyap, Abhinav R.; Kankanhalli, Mohan S.	2019	ci3	ce1,ce2
Molino, Piero; Aiello, Luca Maria; Lops, Pasquale	2016	ci3	ce1,ce2
Jiang, Yexi; Perng, Chang-Shing; Sailer, Anca; Silva-Lepe, Ignacio; Zhou, Yang; Li, Tao	2016	ci3	ce1,ce2
Chen, Qiuyuan; Xia, Xin; Hu, Han; Lo, David; Li, Shanping	2021	ci3	ce1,ce2
Jiang, He; Chen, Xin; He, Tieke; Chen, Zhenyu; Li, Xiaochen	2018	ci3	ce1,ce2
Fang, Ruogu; Pouyanfar, Samira; Yang, Yimin; Chen, Shu-Ching; Iyengar, S. S.	2016		ce1,ce2,ce3
Verma, Pradeepika; Pal, Sukomal; Om, Hari	2019	ci3	ce1,ce2
Kumar, Abhishek; Braud, Tristan; Kwon, Young D.; Hui, Pan	2020	ci3	ce1,ce2
Liu, Xuanzhe; Ai, Wei; Li, Huoran; Tang, Jian; Huang, Gang; Feng, Feng; Mei, Qiaozhu	2017	ci3	ce1,ce2
Zoya; Latif, Seemab; Shafait, Faisal; Latif, Rabia	2021	ci3	ce1,ce2
Gupta, Hritvik; Patel, Mayank	2021	ci3	ce1,ce2
Lin, Xiaoguang; Liu, Mingxuan; Zhang, Ju	2020	ci3	ce1,ce2
Patel, Jay; Makvana, Kamlesh; Shah, Parth	2019		ce1,ce2,ce3

Al-Zubi, Sawsan; Awaysheh, Feras M.; Al-Shboul, Bashar Awad	2021	ci3	ce1,ce2
Alhawarat, M.; Hegazi, M.	2018	ci3	ce1,ce2
Otter, Daniel W.; Medina, Julian R.; Kalita, Jugal K.	2021		ce1,ce2,ce3
Xu, Yue; Nguyen, Hanh; Li, Yuefeng	2020	ci3	ce1,ce2
Ramina, Mayank; Darnay, Nihar; Ludbe, Chirag; Dhruv, Ajay	2020	ci3	ce1,ce2
Buenaño-Fernandez, Diego; González, Mario; Gil, David; Luján-Mora, Sergio	2020	ci3	ce1,ce2
Galassi, Andrea; Lippi, Marco; Torroni, Paolo	2021		ce1,ce2,ce3
Chen, Yi-Hui; Lu, Eric Jui-Lin; Ou, Ting-An	2021	ci3	ce1,ce2
Xu, Bo; Lin, Hongfei; Lin, Yuan; Diao, Yufeng; Yang, Liang; Xu, Kan	2019	ci3	ce1,ce2
Wang, Yuan; Liu, Jie; Huang, Yalou; Feng, Xia	2016	ci3	ce1,ce2
Bounhas, Ibrahim; Ben Guirat, Souheila	2019	ci3	ce1,ce2
Almazrua, Amal; Almazrua, Manal; Alkhalifa, Hend	2020	ci3	ce1,ce2
Ray, Santosh K.; Shaalan, Khaled	2016		ce1,ce2,ce3
Anandika, Amrita; Mishra, Smita Prava	2019		ce1,ce2,ce3
Thieu, Thanh; Camacho, Jonathan; Ho, Pei-Shu; Porcino, Julia; Ding, Min; Nelson, Lisa; Rasch, Elizabeth; Zhou, Chunxiao; Chan, Leighton; Brandt, Diane; Newman-Griffis, Denis; Yuan, Ao; Lai, Albert M.	2017	ci3	ce1,ce2
Chen, Yuxin; Bordes, Jean-Baptiste; Filliat, David	2018	ci3	ce1,ce2
Saranya, M S; Selvi, M; Ganapathy, S.; Muthurajkumar, S; Ramesh, L. Sai; Kannan, A.	2017		ce1,ce2,ce3
Ramesh, Vignav; Kolonin, Anton	2020	ci3	ce1,ce2
Yunianto, Ide; Permanasari, Adhistya Erna; Widyawan, Widyawan	2020		ce1,ce2,ce3
Atapour-Abarghouei, Amir; Bonner, Stephen; McGough, Andrew Stephen	2021	ci3	ce1,ce2
Larabi Marie-Sainte, Souad; Alalyani, Nada; Alotaibi, Sihaam; Ghouzali, Sanaa; Abunadi, Ibrahim	2019		ce1,ce2,ce3
AlAni, Jabir Alshehabi; Fasli, Maria	2019	ci3	ce1,ce2
Zahui, Asma; Elhor, Wahiba; Cheragui, Mohamed Amine	2017	ci3	ce1,ce2
Van, Toan Pham; Thanh, Ta Minh	2017	ci3	ce1,ce2

Salah, Ramzi Esmail; Binti Zakaria, Lailatul Qadri	2018	ci3	ce1,ce2
Hashem, Rawdah Abu	2021		ce1,ce2,ce3
Bara, George Antoniu	2021		ce1,ce2,ce3
Khurpia, Naman	2021	ci3	ce1,ce2
Najafabadipour, Marjan; Zanin, Massimiliano; Rodríguez-González, Alejandro; Gonzalo-Martín, Consuelo; Nuñez García, Beatriz; Calvo, Virginia; Luis Cruz Bermudez, Juan; Provencio, Mariano; Menasalvas, Ernestina	2019	ci3	ce1,ce2
Naseem, Usman; Musial, Katarzyna; Eklund, Peter; Prasad, Mukesh	2020	ci3	ce1,ce2
Panahandeh, Mahnaz; Ghanbari, Shirin	2019	ci3	ce1,ce2
Zhang, Yizhou; Ma, Xiaojun; Song, Guojie	2018	ci3	ce1,ce2
Farrah, Soufiane; El Manssouri, Hanane; Ziyati, El Houssaine; Ouzzif, Mohammed	2018	ci3	ce1,ce2
Ruambo, Francis A.; Nicholaus, Mrindoko R.	2019		ce1,ce2,ce3
Veras Magalhães Junior, Gilvan; Albuquerque Vieira, João Paulo; Lira de Sales Santos, Roney; Nascimento Barbosa, Jardeson Leandro; de Alcântara dos Santos Neto, Pedro; Santos Moura, Raimundo	2019	ci3	ce1,ce2
Zhang, Tingting; Lee, Baozhen; Zhu, Qinghua; Han, Xi; Ye, Edwin Mouda	2020	ci3	ce1,ce2
Biltawi, Mariam M.; Tedmori, Sara; Awajan, Arafat	2021		ce1,ce2,ce3
Wang, Chengzhi; Sun, Xian; Yu, Hongfeng; Zhang, Wenkai	2019	ci3	ce1,ce2
Mridha, M. F.; Keya, Ashfia Jannat; Hamid, Md. Abdul; Monowar, Muhammad Mostafa; Rahman, Md. Saifur	2021		ce1,ce2,ce3
Almuzaini, Huda Abdulrahman; Azmi, Aqil M.	2020	ci3	ce1,ce2
Chandrasekaran, Dhivya; Mago, Vijay	2021	ci3	ce1,ce2
Appiktala, Nirupama; Huang, SansWord; Sankar, Balachandar; Tripathi, Shweta; Goldman, Eyan	2021	ci3	ce1,ce2
Azimi, Sasan; Veisi, Hadi; Amouie, Reyhaneh	2019	ci1,ci2,ci3	
Zhang, Ming; Palade, Vasile; Wang, Yan; Ji, Zhicheng	2019	ci3	ce1,ce2

Li, Xiang; Zhang, Kewen; Zhu, Quanyin; Wang, Yuanyuan; Ma, Jialin	2021	ci3	ce1,ce2
Sen, Ovishake; Fuad, Mohtasim; Islam, Md. Nazrul; Rabbi, Jakaria; Masud, Mehedi; Hasan, Md. Kamrul; Awal, Md. Abdul; Ahmed Fime, Awal; Hasan Fuad, Md. Tahmid; Sikder, Delowar; Raihan Iftee, Md. Akil	2022		ce1,ce2,ce3
Lei, Jianjun; Song, Yuxin; Peng, Bo; Ma, Zhanyu; Shao, Ling; Song, Yi-Zhe	2020	ci3	ce1,ce2
Pundge, Ajitkumar Meshram; Namrata Mahender, C.	2018	ci3	ce1,ce2
koochari, Abbas; Alavi Gharahbagh, Abdorreza; Hajihashemi, Vahid	2020	ci3	ce1,ce2
Besdouri, Fatma Zahra; Mekki, Asma; Zribi, Inès; Ellouze, Mariem	2021	ci3	ce1,ce2
Zhu, Hongyin; Tiwari, Prayag; Ghoneim, Ahmed; Hossain, M. Shamim	2022	ci3	ce1,ce2
Chang, Yung-Chun; Shih, Cheng-Wei; Hsu, Wen-Lian	2018	ci3	ce1,ce2
Ablimit, Mijit; Parhat, Sardar; Hamdulla, Askar; Zheng, Thomas Fang	2018	ci3	ce1,ce2
Zeng, Ping; Tan, Qingping; Yan, Ying; Xie, Qinzheng; Xu, Jianjun; Cao, Wei	2017	ci3	ce1,ce2
Madi, Nora; Al-Mutlaq, Nourah; Al-Khalifa, Hend S.	2019	ci3	ce1,ce2
Baviskar, Dipali; Ahirrao, Swati; Potdar, Vidyasagar; Kotecha, Ketan	2021		ce1,ce2,ce3
Mohammadi, Shahin; Kylasa, Sudhir; Kollias, Giorgos; Grama, Ananth	2016	ci3	ce1,ce2
Khaleghi, Tannaz; Murat, Alper; Arslanturk, Suzan; Davies, Eric	2020	ci3	ce1,ce2
Papadopoulos, Helene; Tzanetakis, George	2017	ci3	ce1,ce2
Sernadela, Pedro; Oliveira, José L.	2017	ci3	ce1,ce2
Ariestya, Winda Widya; Astuti, Ida; Wiryana, I Made	2018	ci3	ce1,ce2
Hu, Fan; Xia, Gui-Song; Yang, Wen; Zhang, Liangpei	2020	ci3	ce1,ce2
Frisoni, Giacomo; Moro, Gianluca; Carbonaro, Antonella	2021		ce1,ce2,ce3
Huang, Heyan; Wang, Yashen; Feng, Chong; Liu, Zhirun; Zhou, Qiang	2018	ci3	ce1,ce2
Liu, Linyan; Yuan, Tangxiao; Song, Haojie; Wang, Huifen	2019		ce1,ce2,ce3

Ali, Farman; Kwak, Daehan; Khan, Pervez; Ei-Sappagh, Shaker Hassan A.; Islam, S. M. Riazul; Park, Daeyoung; Kwak, Kyung-Sup	2017	ci3	ce1,ce2
Yafoz, Ayman; Mouhoub, Malek	2020		ce1,ce2,ce3
El-Defrawy, Mahmoud; Belal, Nahla A.; El-Sonbaty, Yasser	2017	ci3	ce1,ce2
Zhang, Dongxiang; Wang, Lei; Zhang, Luming; Dai, Bing Tian; Shen, Heng Tao	2020		ce1,ce2,ce3
Herwando, Raditya; Jiwanggi, Meganingrum Arista; Adriani, Mirna	2017	ci3	ce1,ce2
Guo, Lantian; Cai, Xiaoyan; Hao, Fei; Mu, Dejun; Fang, Changjian; Yang, Libin	2017	ci3	ce1,ce2
Laudy, Claire; Dragos, Valentina	2020		ce1,ce2,ce3
Mohasseb, Alaa; Bader-El-Den, Mohamed; Liu, Han; Cocea, Mihaela	2017	ci3	ce1,ce2
Sharma, Debraj Rudra; Bhaskaran, Sreebha	2019	ci3	ce1,ce2
Francis, Danny; Pidou, Paul; Merialdo, Bernard; Huet, Benoit	2017	ci3	ce1,ce2
Bahojb Imani, Maryam; Khan, Latifur; Thuraisingham, Bhavani	2019	ci3	ce1,ce2
Ren, Fuji; Xue, Siyuan	2020	ci3	ce1,ce2
Davenport, Mark A.; Romberg, Justin	2016		ce1,ce2,ce3
Alharbi, Najla Hamandi; Alkhateeb, Jawad Hassan	2021	ci3	ce1,ce2
Xu, Dan; Song, Jingkuan; Alameda-Pineda, Xavier; Ricci, Elisa; Sebe, Nicu	2016	ci3	ce1,ce2
Rong, Xuejian; Yi, Chucai; Tian, Yingli	2022	ci3	ce1,ce2
Siriyasatien, P.; Chadsuthi, S.; Jampachaisri, K.; Kesorn, K.	2018		ce1,ce2,ce3
Jiang, Shanshan; Hagelien, Thomas F.; Natvig, Marit; Li, Jingyue	2019	ci3	ce1,ce2
Fang, Quan; Xu, Changsheng; Sang, Jitao; Hossain, M. Shamim; Ghoneim, Ahmed	2016	ci3	ce1,ce2
Li, Jing; Sun, Aixin; Han, Jianglei; Li, Chenliang	2022		ce1
Yadav, Subham; Sarkar, Madhulina	2018	ci1,ci2,ci3	
Singh, Thoudam Doren; Divyansha, Divyansha; Singh, Apoorva Vikram; Sachan, Anubhav; Khilji, Abdullah Faiz Ur Rahman	2020	ci3	ce1,ce2
Phan, Viet Anh; Chau, Ngoc Phuong; Nguyen, Minh Le	2016	ci3	ce1,ce2

Zhu, Qile; Ma, Xiyao; Li, Xiaolin	2019		ce1,ce2,ce3
Alnajran, Noufa; Crockett, Keeley; McLean, David; Latham, Annabel	2018	ci3	ce1,ce2
Berquand, Audrey; Murdaca, Francesco; Riccardi, Annalisa; Soares, Tiago; Generé, Sam; Brauer, Norbert; Kumar, Kartik	2019		ce1,ce2,ce3
Das, Siddhartha Shankar; Serra, Edoardo; Halappanavar, Mahantesh; Pothen, Alex; Al-Shaer, Ehab	2021	ci3	ce1,ce2
Al-Azani, Sadam; El-Alfy, El-Sayed M.	2021	ci3	ce1,ce2
Franco-Riquelme, Jose N.; Bello-Garcia, Antonio; Ordieres-Meré, Joaquín	2019	ci3	ce1,ce2
Jian, Huang; Bai, Yu; Zhang, Guiping; Miu, Wanwan	2019	ci3	ce1,ce2
Jeong, Yujin; Kim, Sunhye; Yoon, Byungun	2018		ce1,ce2,ce3
Elsaid, Asmaa; Mohammed, Ammar; Ibrahim, Lamiaa Fattouh; Sakre, Mohammed M.	2022		ce1,ce2,ce3
Nassif, Mathieu; Treude, Christoph; Robillard, Martin P.	2020	ci3	ce1,ce2
Zhang, Kun; Lv, Guangyi; Wu, Le; Chen, Enhong; Liu, Qi; Wu, Han; Xie, Xing; Wu, Fangzhao	2021	ci3	ce1,ce2
Kugathasan, Archchana; Sumathipala, Sagara	2020		ce1,ce2,ce3
Chen, Xinying; Cong, Peimin; Lv, Shuo	2022	ci3	ce1,ce2
Madyatmadja, Evaristus Didik; Yahya, Bernardo Nugroho; Wijaya, Cristofer	2022	ci3	ce1,ce2
Ma, Handong; Hou, Jiawei; Zhu, Chenxu; Zhang, Weinan; Tang, Ruiming; Lai, Jincai; Zhu, Jieming; He, Xiuqiang; Yu, Yong	2021	ci3	ce1,ce2
Guo, Yulan; Wang, Hanyun; Hu, Qingyong; Liu, Hao; Liu, Li; Bennamoun, Mohammed	2021		ce1,ce2,ce3
Specia, Lucia; Barrault, Loic; Caglayan, Ozan; Duarte, Amanda; Elliott, Desmond; Gella, Spandana; Holzenberger, Nils; Lala, Chiraag; Lee, Sun Jae; Libovicky, Jindrich; Madhyastha, Pranava; Metze, Florian; Mulligan, Karl; Ostapenko, Alissa; Palaskar, Shruti; Sanabria, Ramon; Wang, Josiah; Arora, Raman	2020		ce1,ce2,ce3

Mosharraf, Maedeh; Taghiyareh, Fattaneh	2020	ci3	ce1,ce2
Tudu, Ronald; Saha, Shaibal; Pritam, Prasun Nandy; Palit, Rajesh	2018	ci3	ce1,ce2
Chen, Yiye; Xu, Ruinian; Lin, Yunzhi; Vela, Patricio A.	2021	ci3	ce1,ce2
Ellaky, Zineb; Benabbou, Faouzia; Ouahabi, Sara; Sael, Nawal	2021	ci3	ce1,ce2
Hani, Anoud Bani; Adedugbe, Oluwasegun; Al-Obeidat, Feras; Benkhelifa, Elhadj; Majdalawieh, Munir	2020	ci3	ce1,ce2
Bavota, Gabriele	2016		ce1,ce2,ce3
Ye, Deheng; Xing, Zhenchang; Foo, Chee Yong; Ang, Zi Qun; Li, Jing; Kapre, Nachiket	2016	ci3	ce1,ce2
Putra Perdana, B.B Sakti; Irawan, Budhi; Setianingsih, Casi	2019	ci3	ce1,ce2
Zakraoui, Jezia; Saleh, Moutaz; Al-Maadeed, Somaya; Alja'am, Jihad Mohamed	2021		ce1,ce2,ce3
Gharouit, Kenza; Nfaoui, El Habib	2017	ci3	ce1,ce2
Xu, Dianlei; Li, Tong; Li, Yong; Su, Xiang; Tarkoma, Sasu; Jiang, Tao; Crowcroft, Jon; Hui, Pan	2021		ce1,ce2
Shajalal, Md; Aono, Masaki; Azim, Muhammad Anwarul	2018	ci3	ce1,ce2
Aqlan, Fares; Fan, Xiaoping; Alqwbani, Abdullah; Al-Mansoub, Akram	2019	ci3	ce1,ce2
Wang, Chengyu; He, Xiaofeng; Zhou, Aoying	2021	ci3	ce1,ce2
Dirar, Aamer Idris; Salih, Insaaf Juma; Alrasheed, Mosab Ibrahim; Elamin, Haysam E.	2017	ci3	ce1,ce2
Londhe, Deepali D.; Kumari, Aruna; Emmanuel, M.	2021		ce1,ce2,ce3
Awwad, Hunaida; Alpkocak, Adil	2017	ci3	ce1,ce2
Wang, Chu; Wang, Daling; Feng, Shi; Zhang, Yifei; Liu, Hongchen	2017	ci3	ce1,ce2
Dershowitz, Nachum; Labenski, Daniel; Silberpfennig, Adi; Wolf, Lior; Tsur, Yaron	2017	ci3	ce1,ce2
Amelio, L.; Amelio, A.	2019		ce1,ce2,ce3
Diaz, Juan Sebastian Beleno; Bauzer Medeiros, Claudia	2017	ci3	ce1,ce2
Aguiar, Z. Camila; Cury, Davidson	2016		ce1,ce2,ce3

Correia, António; Jameel, Shoaib; Schneider, Daniel; Paredes, Hugo; Fonseca, Benjamim	2020	ci3	ce1,ce2
Rizun, Nina; Taranenko, Yurii	2018	ci3	ce1,ce2
Huang, Qiao; Xia, Xin; Lo, David; Murphy, Gail C.	2020	ci3	ce1,ce2
Blanco, Pedro Almagro; Caparrini, Fernando Sancho	2017	ci3	ce1,ce2
Xu, Yingkun; Qin, Lei; Huang, Qingming	2016	ci3	ce1,ce2
Bibi, Nazia; Rana, Tauseef; Qurat-ul-Ain; Naseer, Ayesha	2021		ce1,ce2,ce3
Wang, Haoyi; Sanchez, Victor; Li, Chang-Tsun	2022	ci3	ce1,ce2
Han, Jialong; Sun, Aixin; Cong, Gao; Zhao, Wayne Xin; Ji, Zongcheng; Phan, Minh C.	2018	ci3	ce1,ce2
Subramani, Sudha; Vu, Huy Quan; Wang, Hua	2017	ci3	ce1,ce2
Roh, Yuji; Heo, Geon; Whang, Steven Euijong	2021		ce1,ce2,ce3
Kidmose, Egon; Stevanovic, Matija; Brandbyge, Søren; Pedersen, Jens M.	2020	ci3	ce1,ce2
Mumtaz, Sara; Rodriguez, Carlos; Benatallah, Boualem; Al-Banna, Mortada; Zamanirad, Shayan	2020	ci3	ce1,ce2
Baltrušaitis, Tadas; Ahuja, Chaitanya; Morency, Louis-Philippe	2019		ce1,ce2,ce3
He, Jun-Yan; Liang, Shi-Hua; Wu, Xiao; Zhao, Bo; Zhang, Lei	2021	ci3	ce1,ce2
Silva, Paulo; Monteiro, Edmundo; Simões, Paulo	2021		ce1,ce2,ce3
Abdelminaam, Diaa Salama; Neggaz, Nabil; Gomaa, Ibrahim Abd Elatif; Ismail, Fatma Helmy; Elsawy, Ahmed A.	2021	ci3	ce1,ce2
Ito, Hiroyoshi; Komamizu, Takahiro; Amagasa, Toshiyuki; Kitagawa, Hiroyuki	2018	ci3	ce1,ce2
Liang, Gongbo; Greenwell, Connor; Zhang, Yu; Xing, Xin; Wang, Xiaoqin; Kavuluru, Ramakanth; Jacobs, Nathan	2022	ci3	ce1,ce2
Jiang, Wei; Lin, Junyu; Wang, Huiqiang; Zou, Shichen	2020	ci3	ce1,ce2
Zheng, Yixian; Wu, Wenchao; Chen, Yuanzhe; Qu, Huamin; Ni, Lionel M.	2016		ce1,ce2,ce3
Dai, Tao; Gao, Tianyu; Zhu, Li; Cai, Xiaoyan; Pan, Shirui	2018	ci3	ce1,ce2

Rinaldi, Antonio M.; Russo, Cristiano	2018	ci3	ce1,ce2
Al-Anzi, Fawaz S.; AbuZeina, Dia	2017	ci3	ce1,ce2
Zheng, Xin; Han, Jialong; Sun, Aixin	2018		ce1,ce2,ce3
Liu, Hong; Ren, Congyaxu	2019	ci3	ce1,ce2
Neubert, Peer; Schubert, Stefan	2021	ci3	ce1,ce2
	2017		ce1,ce2,ce3,ce4
Zhang, Chaoyun; Patras, Paul; Haddadi, Hamed	2019		ce1,ce2,ce3
Satvat, Kiavash; Gjomemo, Rigel; Venkatakrishnan, V.N.	2021	ci3	ce1,ce2
Haeb-Umbach, Reinhold; Heymann, Jahn; Drude, Lukas; Watanabe, Shinji; Delcroix, Marc; Nakatani, Tomohiro	2021	ci3	ce1,ce2
Yan, Z.; Yang, C.; Hu, L.; Zhao, J.; Jiang, L.; Gong, J.	2021	ci3	ce1,ce2
Mladenovic, M.; Ošmjanski, V.; Stankovic, S.V.	2021		ce1,ce2,ce3
Aman, M.; Abdulkadir, S.J.; Aziz, I.A.; Alhussian, H.; Ullah, I.	2021	ci3	ce1,ce2
Batanović, V.; Cvetanović, M.; Nikolić, B.	2020	ci3	ce1,ce2
Yildirim, A.; Uskudarli, S.	2020	ci3	ce1,ce2
Azad, H.K.; Deepak, A.	2019		ce1,ce2,ce3
Zhao, J.; Guan, Z.; Sun, H.	2019	ci3	ce1,ce2
Alnajran, N.; Crockett, K.; McLean, D.; Latham, A.	2019		ce1,ce2,ce3
Jauhiainen, T.; Lui, M.; Zampieri, M.; Baldwin, T.; Lindén, K.	2019		ce1,ce2,ce3
Du, K.-L.; Swamy, M.N.S.	2019		ce1,ce2,ce3,ce4
Li, Y.; Schulze, S.; Saake, G.	2018	ci3	ce1,ce2
Goyal, A.; Gupta, V.; Kumar, M.	2018		ce1,ce2,ce3
Hoogeveen, D.; Wang, L.; Baldwin, T.; Verspoor, K.M.	2018		ce1,ce2,ce3
Krallinger, M.; Rabal, O.; Lourenço, A.; Oyarzabal, J.; Valencia, A.	2017		ce1,ce2,ce3
Wang, C.; Song, Y.; Roth, D.; Zhang, M.; Han, J.	2016	ci3	ce1,ce2
Witten, I.H.; Frank, E.; Hall, M.A.; Pal, C.J.	2016		ce1,ce2,ce3,ce4
Vieira, H.S.; Da Silva, A.S.; Calado, P.; Cristo, M.; De Moura, E.S.	2016	ci3	ce1,ce2

Fonte: Elaborado pela autora.

ANEXO A – ARTIGO PUBLICADO

A.1 EVALUATING THE EFFECT OF CORPUS NORMALISATION IN TOPICS COHERENCE

Abstract: Probabilistic topic models are extensively used to better understand the content of documents. Due to the fact that topic models are totally unsupervised, statistical and data driven, they may produce topics not always meaningful. This work is based on the hypothesis that, since LDA takes into account the number of occurrences of words, we could affect the quality of topics by semantically normalising the text, where each concept would be represented by the same word. We can find a formal description of lexemes found in text using a knowledgebase and extract the several forms of mentioning a lexeme to normalize a corpus. We use topic coherence metric, as it represents the semantic interpretability of the terms used to describe a particular topic, to quantify the influence of semantic corpus normalisation in topics. The first tests on the semantic normalisation framework of texts showed prominent results, and shall be investigated in depth in future.

Evaluating the Effect of Corpus Normalisation in Topics Coherence

Luana da Silva Sousa(✉), Vinicius Melquiades de Sousa, Rogerio de Aquino Silva, and Gustavo Medeiros de Araújo

Engineering and Data Science Lab, Federal University of Santa Catarina, Florianópolis, Brazil
gustavo.araujo@ufsc.br

Abstract. Probabilistic topic models are extensively used to better understand the content of documents. Due to the fact that topic models are totally unsupervised, statistical and data driven, they may produce topics not always meaningful. This work is based on the hypothesis that, since LDA takes into account the number of occurrences of words, we could affect the quality of topics by semantically normalising the text, where each concept would be represented by the same word. We can find a formal description of lexemes found in text using a knowledgebase and extract the several forms of mentioning a lexeme to normalize a corpus. We use topic coherence metric, as it represents the semantic interpretability of the terms used to describe a particular topic, to quantify the influence of semantic corpus normalisation in topics. The first tests on the semantic normalisation framework of texts showed prominent results, and shall be investigated in depth in future.

Keywords: Corpus normalisation · LDA · Topic coherence · Ontology · Natural language processing

1 Introduction

Extracting useful information from large collections of text documents has become more challenging in recent years.

Understanding and modeling the content of documents can be very useful in many applications, such as information retrieval, natural language processing (NLP), document classification, text summarization, etc. [2].

The foundation in statistics and its capability to be extended and combined with other models make probabilistic topic model one of the most used algorithms to deal with these problems [2]. Topic modeling is a form of finding latent semantic structure within a collection of documents, and probabilistic models, such as Latent Dirichlet Allocation (LDA), have become the standard method employed [6, 20]. The intuition is that pairs of descriptor terms that co-occur frequently or are close to each other within a semantic space are likely to contribute to higher levels of coherence for a specific topic [20]. LDA model has been criticized for favoring highly frequent, general words in topic descriptors [20]. Due to the fact that topic models are totally unsupervised, statistical and data driven, they may produce topics not always meaningful [2].

E. Bisset Álvarez (Ed.): DIONE 2021, LNICST 378, pp. 197–208, 2021.
https://doi.org/10.1007/978-3-030-77417-2_15

Higher interconnectivity between information sources has the potential of increasing the utility of information. By connecting unstructured information in text documents with structured semantic data available on the internet, facts from this huge Web of Data can be used to enhance several tasks such as information extraction [25], information retrieval [27, 28], text classification [10], feature extraction [13], etc.

The goal of this work is to present the first results on a text semantic normalisation framework. Our work was based on the hypothesis that, since LDA takes into account the number of word occurrences, we could affect the quality of topics by semantically normalising the text, where each concept would be represented by the same word. If the same concept is represented by two different words in different texts, the algorithm would probably struggle more to find coherent topics. We can find a formal description of lexemes (unit of meaning, composed of one or more words) found in text using a knowledge base (KB) and extract the several forms of mentioning a lexeme to normalise our corpus. The topic coherence measure is used to address the semantic interpretability of the terms used to describe a particular topic [20], and it is the measure we used to quantify the influence of semantic corpus normalisation in topics.

1.1 Contributions

- This paper proposes a framework to semantically normalise texts, and show experimental results on topic modeling task using two widely used datasets;
- Topic coherence improvement compared to traditional LDA.

1.2 Organization of the Work

The work is organized as follows: Sect. 2 presents a background content of the methods approached in this paper. Section 3 brings related work and compares our approach to others in literature. Section 4 exposes the proposed method to semantically normalise text and how it was applied to topic modeling. Section 5 presents the results and a discussion. And finally, Sect. 6 concludes our work.

2 Methods

2.1 Topic Modeling and Topic Coherence

A topic is a probability distribution over words and documents are mixtures of topics. Hence, a topic model can be considered a generative model for documents [24]. A more formal description of the Topic Modeling problem using LDA model is described as follows.

In LDA, it is assumed that there are K underlying topics from which the documents are generated and that each topic is represented as a multinomial distribution over the V words in the vocabulary. Therefore, a document is generated by sampling a mixture of these topics and then sampling words from that mixture [6].

A document with N words $d = w_1, \ldots, w_N >$ is generated by the following process:

1. The mix of topics θ is sampled from a Dirichlet distribution $(\alpha_1, \ldots, \alpha_k)$;
2. For each of the N words, a topic $z_n \in \{1, \ldots K\}$ is sampled from a $Mult(\theta)$ distribution, where $p(z_n = 1 \vee \theta) = \theta_i$;

3. Each word w_n is sampled, conditioned to the z_n-th topic, from the multinomial distribution $p(k \vee z_n)$.

It is possible to think of θ_i as the degree that a topic refers to a document. So, the probability of a document is the following mix:

$$p(d) = \int_{\theta} \left(\prod_{n=1}^{N} \sum_{z_n=1}^{K} p(w_n|z_n; \beta) p(z_n|\theta) \right) p(\theta|; \alpha) d\theta \tag{1}$$

Where $p(\theta; \alpha)$ is Dirichlet, $p(z_n \vee \theta)$ is a multinomial distribution parametrized by θ, and $p(w_n|z_n; \beta$ is a multinomial distribution over words. This model is parametrized by the parameters $\alpha = \langle \alpha_1, \ldots, \alpha_k \rangle$ and a matrix β with dimensions $K \times |V|$. The per-word topic assignment, per-document topic distribution and topics are all latent variables and are not observed. The only observed variable is words within the documents, to infer the hidden structure (latent variables) with statistical inference [26].

As a way of making it clearer, Fig. 1 depicts the word distribution over the topics and a topic distribution over documents. As it was said before, a document is a distribution of topics and a topic is a distribution of words.

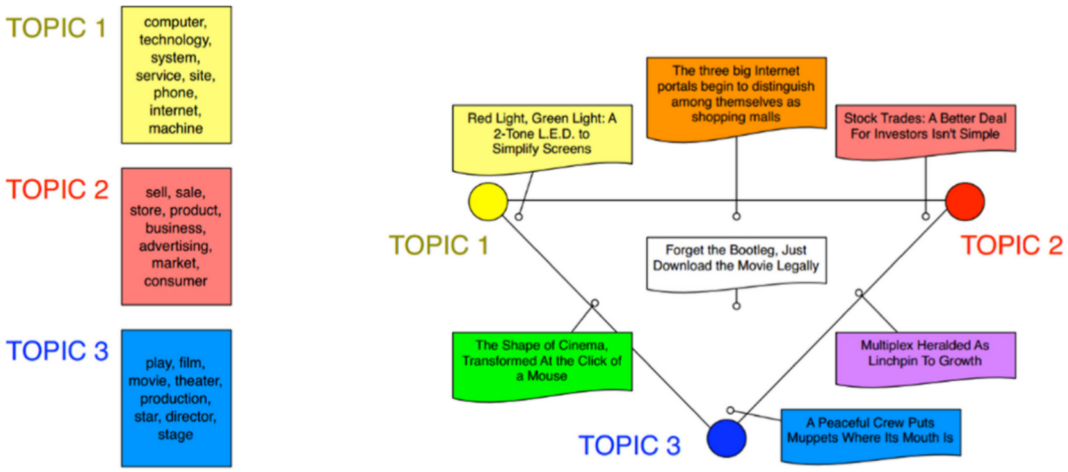


Fig. 1. Word-Topic distribution on the left and Document-Topic distribution on the right. These three topics represent the first three topics from a fifty topic LDA model trained on articles from the New York Times [8].

The topic quality (quality means interpretable and meaningful), measured as topic coherence, is based on the hypothesis that words with similar meaning tend to co-occur within a similar context. Each topic distribution contains every word but assigns a different probability to each of the words. The words with the highest probabilities within a topic are those that tend to co-occur more frequently. So, the top 10 or 15 high-probability words are usually used to interpret and semantically label the topics [26].

Researchers use several metrics of model fit, such as perplexity or held-out likelihood. However, such measures are only useful for evaluating the predictive model and do not address the explanatory goals of topic modeling [8]. The task of quantifying the coherence of a set of topics have been studied to remedy the problem that topic models give no guaranty on the interpretability of their output [23].

Many measures of coherence have been proposed recently, based on approaches that include co-occurrence frequencies of terms within a reference corpus [16, 19, 23]. A recent study [23] systematically and empirically explored the multitude of topic coherence measures and their correlation with available human topic ranking data. Their approach revealed a new coherence measure, called C_V, which achieved the highest correlation compared with all human ranking data. Hence, this study adopts the C_V coherence measure for topic coherence calculations.

2.2 Semantic Web

The idea of Semantic Web was described in 2001 by Tim Berners-Lee et al. as "A new form of web content that is meaningful to computers". In this new form of web content, introduced as an extension of the current web, information is given a well-defined meaning, where computers and people can work in cooperation [4].

The Semantic Web is based on the Resource Description Framework (RDF), a formal language for describing structured information [15]. An RDF document describes a formal specification of an arbitrary domain. This specification is modeled by a directed, labeled graph where each edge represent a link between two resources, represented by the graph nodes [17]. The link is expressed as RDF triples (*subject*, *relation*, *object*). Uniform Resource Identifiers (URI) are used to identify RDF resources and relations. To access and query RDF graphs the Protocol And RDF Query Language (SPARQL) was developed [21]. The results of SPARQL queries can be new RDF graphs or sets of resources.

The relationships and properties RDF resources may have can be specified by the vocabulary description language RDF Schema (RDFS) [7]. RDFS allows to create custom defined vocabularies to organize knowledge. Since URIs enable to identify RDF resources globally, it seems reasonable to combine vocabularies shared by different creators and across different domains. When shared, an RDF vocabulary can be denoted as an *ontology*. An ontology is an explicit, formal specification of a shared conceptualization and defines the terms used to describe and represent an area of knowledge [14].

The concept of ontology brings us to the Linking Open Data (LOD) project. It aims to identify datasets in the web that are available under open licenses, re-publish these datasets in RDF and interlink them with each other [5]. The term Linked Data refers to a set of principles to publish and interlink structured data on the web. One of the ontologies available on the web is YAGO (Yet Another Great Ontology - https://yago-knowledge.org/). YAGO is a large semantic knowledge base, derived from Wikipedia, WordNet, WikiData, GeoNames, and other data sources [22]. Currently, YAGO knows more than 17 million entities (like persons, organizations, cities, etc.) and contains more than 150 million facts about these entities. SPARQL queries are used in this work to query Yago Knowledge base in order to fetch alternative words for the same lexeme.

2.3 Named Entity Linking

Named Entity Linking can be described as the task of identifying lexemes in a text and linking them to the entity they name in a knowledge base, such as DBPedia. Before going

too deep, an introduction of terminology and concepts is established. The term *entity* refers to something which is cognitively representable. An entity *mention* refers to the part of the text where a reference to an entity is made. It is also called *lexeme*, which is the basic unit of meaning. The *surface form* is a specific syntactic representation of the lexeme (the exact character string). A *knowledge base entity* refers to a representation of the entity, usually identified by an *URI* [28].

Now, let K be a formal knowledge base, $d \in D$ a document of the corpus D, $W \subseteq d$ the words of document d, $M \subseteq 2^W$ the set of entity mentions, and $m = (s, l, d, c) \in M$ denote an entity mention in a document d with start position s, length l and confidence score $c \in [0, 1]$. The *named entity linking problem* can be described as this [28]:

Definition 1 (Name Entity linking Problem)

- An extraction function $f_{ex} : W \rightarrow M$ to extract the entity mentions M from a document set D.
- A mapping function $f_{map} : M \rightarrow 2^W \cup NIL$ to compile a list $C \in 2^K$ of potential knowledge base entity candidates for every lexeme.
- A scoring function $f_{score} : C \rightarrow R$ to calculate a score, which indicates the degree of certainty that the candidate URI is to be selected as the correct one.
- A selection function $f_{sel} : C \rightarrow K$ to select the right candidate according to the calculated scores.

The degree of ambiguity is indicated by the size of the candidate list C. Hence, the *disambiguation task* is described by putting the mapping, scoring and selection functions together. The entire *context* is observed when processing the analysis items in the implementation of these functions. Just like in communication theory and linguistics the context is essential when interpreting pieces of information, in NEL it is as well. Examining context is crucial for NEL, because some context items can be very decisive when interpreting the context information [28].

There are some options of automated entity linking, and one of them is DBpedia Spolight (https://www.dbpedia-spotlight.org/) [18]. It is and open source project developing a system for automatic annotation of DBpedia entities in natural language text. It provides an interface for phrase spotting (recognition of phrases to be annotated) and disambiguation (entity linking) as well as various output formats (XML, JSON, RDF, etc.) in a REST-based web service [9]. DBpedia Spotlight is used in this work as the tool to find resources (URIs corresponding to formal descriptions of a concept) in text. These resources have several information and metadata about the concept, as well as links to other knowledge bases.

3 Related Work

There are related works of a type of topic modeling called of knowledge-based topic modeling. The main difference with the known knowledge-based topic modeling [3, 12] is that in this work the knowledge-based content is nor on the sampling neither on the inference steps [11, 29], it is a preprocessing step, applied to the input text.

Furthermore, there have been lots of works trying to solve different NLP tasks using semantics [18]. Short text classification was dealt by [12]. They exposed the use of DBpedia ontology to better represent short texts, so that semantically similar texts with no words in common can have similar context [12]. Their approach consisted in three steps: (i) identify concepts in text using DBpedia Spotlight and annotate them as resources; (ii) select the concepts with higher similarity; and (iii) extract additional knowledge, like categories, types or topics of identified concepts. The main dissimilarity between [12] and this word is that they added the additional knowledge (additional words) to the text and did not normalise the text. Moreover, they tested their hypothesis in a classification task, and not in a topic modeling context.

[29] proposed a knowledge-based topic modeling based on multi-relational knowledge graphs. They proposed a method that models document-level word co-occurrence with knowledge encoded by entity vectors automatically learned from external knowledge graphs. In other words, they do not consider only lexemes recognized in text, but from triples in external knowledge graphs. Our work is different from [29] because they add semantic knowledge into the generative process and not in preprocessing. Yet to the best of our knowledge, this is the first work that semantically normalise documents using a semantic replacement methodology.

4 The Proposed Method

There are two big steps that compose this method: (i) semantic corpus normalisation and (ii) topic modeling. The first one is the main contribution of this work, where we semantically normalised a corpus in order to benefit from the explicit semantics of Linked Data to evaluate the effect on the coherence of topics; while the second is the method used to show how semantically normalised texts affect semantic coherence of NLP tasks such as topic modeling using small texts.

4.1 Corpus Normalisation

Figure 2 depicts the normalisation method architecture, where each box is explained in the following paragraphs.

The normalisation is composed of two steps: (i) Resource extraction and (ii) Transformer. This first step is to find all lexemes in texts and associate them with resources from DBpedia. Lexemes are annotated by the process of NEL, using the DBPedia Spotlight annotation tool. Once we have the possible resources mentioned in the text and its respective URIs, we can find additional knowledge related to this resource. We decided to search in another knowledge base in order to find potential alternative surface forms, such as other labels used to describe that resource. The Yago KB is used in this step as the authors found more options in *alternate Name* and *label* fields in this KB. The second step is to create a replacement data structure, where all possible labels of a resource would be replaced by only one. Finally, the last step is to replace them all.

Since the resource extraction is achieved by making HTTP requests and SPARQL queries for each document separately, it is modularized and convenient to parallelize. The documents are saved in a database, each one with an associated unique identifier. We

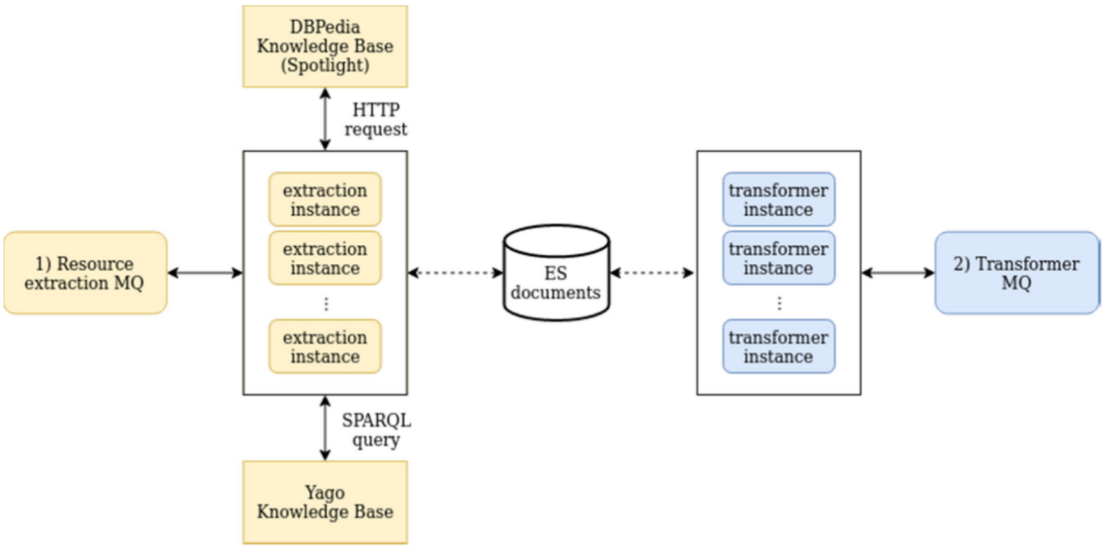


Fig. 2. Normalisation Architecture. The yellow blocks refer to the resource extraction step, as the blue block refers to the transformation step.

used Elasticsearch (https://www.elastic.co/) [1] as database. This identifier is used to keep track of which documents had already been processed. The RabbitMQ (https://www.rabbitmq.com/) tool is used to coordinate the extraction, creating a queue of documents to be processed. The resource extraction module runs in several instances (left side of Fig. 2), in order to accelerate the extraction. Each instance consumes from the queue, represented in the Fig. 2 as *Resource extraction MQ* in order to know which text it should process next.

The extraction works in this way for each text: first, it annotates all resources found in text using DBpedia Spotlight; second, for each resource, it makes a SPARQL query to Yago Knowledge base searching for alternative labels and the labels registered for that resource; lastly, it aggregates all possible labels for a resource and save them in the database.

The transformer step, which is done once all resource extraction is over, collects all resources and labels and organize them in a big mapping list. The mapping list maps all possible labels of a concept to a main label, which is going to replace all possible mentions of that concept. Once the mapping list is built, a regex substitution task is performed in order to make all substitutions. A queue is used to manage all texts that are being processed, similar to the resource extraction phase.

4.2 Topic Modeling

The first step to extract topics is the preprocessing one. The following preprocessing is done: (i) remove invalid characteries and punctuation; (ii) lowercase; (iii) tokenize (transform text into a word vector); (iv) remove stopwords (too common words that do not aggregate meaning); (v) form bigrams (composed words, e.g. “United States”) and (vi) lemmatize (remove word inflections, returning it to its root form, e.g. “said” to “say”). Besides usual stopwords, a list of too frequent words is removed too. From 20-Newsgroup: from, subject, re, edu, use, not, would, say, could, _, be, know, good,

go, get, do, done, try, many, some, nice, thank, think, see, rather, easy, easily, lot, lack, make, want, seem, run, need, even, right, line, even, also, may, take, come; and from Reuters: from, subject, re, edu, use, say, inc, -PRON-.

After preprocessing, the vocabulary of words is ready to compose the word-document matrix that serves as input to LDA algorithm.

5 Results and Discussion

We used two very known corpus of NLP tasks: 20-Newsgroups (https://scikit-learn.org/0.19/datasets/twenty_newsgroups.html) and Reuters (https://www.nltk.org/book/ch02.html). The 20-Newsgroups has more than 18.000 newsgroups posts on 20 topics. Its is divided in training and testing, although for this work we used both as an unique dataset. As the news were from 20 topics, we also used 20 for the hyper-parameter of topics. The Reuters Corpus contains more than 10.000 news documents totaling 1.3 million words. The documents have been classified into 90 topics, and grouped into two sets, called "training" and "test". However, for this work we use both training and test set to extract the topics. Also, as the original corpus was annotated in 90 topics, we used 90 for the hyper-parameter of topics.

In Fig. 3 we can see the distribution of words per document in each corpus used. There is a bigger variety of sizes in 20-Newsgroup corpus, as well as the document length mean is higher before preprocessing. After preprocessing the remaining useful words were similar between both corpora.

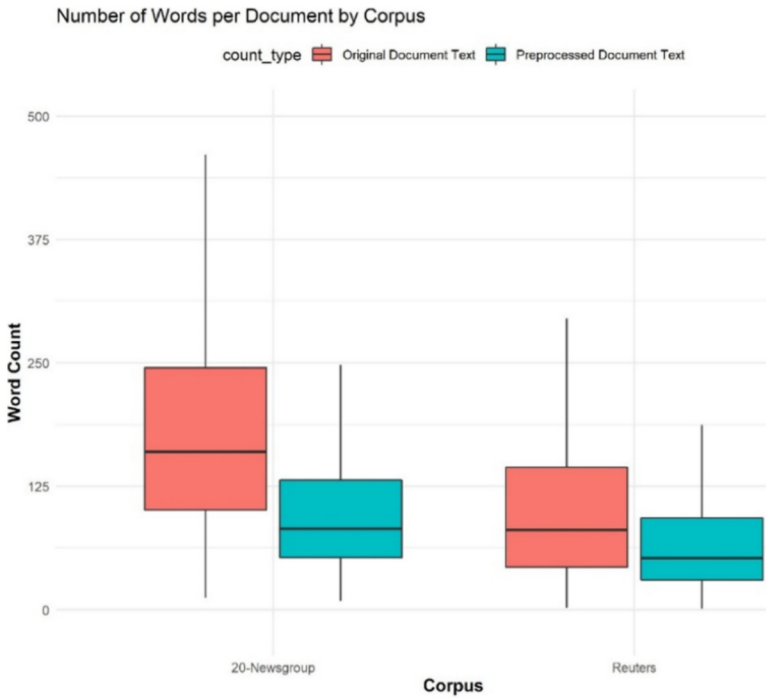


Fig. 3. Number of words per document by corpus. The red boxplot shows the counter of all words, separated only by spaces. The blue one shows the preprocessed documents, where stopwords and bigrams were built. This preprocessing is the same the documents are exposed before topic modeling algorithm.

In Table 1 there are examples of both corpora. The 20-Newsgroup has a form of e-mails, short texts and Reuters has a form of article documents. It can be seen on Fig. 3 that after preprocessing, 20-Newsgroup has lost more words than Reuters, because a big amount of characters were not letters or digits, which are removed on the preprocessing step.

Table 1. Document examples of each corpus.

20-Newsgroup	Reuters
"From: Edwin Gans Subject: Atheism Nntp-Posting-Host: 47.107.76.97 Organization: Bell-Northern Research Lines: 1"	"AMERICAN CENTURY \< ACT > RESTATES EARNINGS\n American Century Corp said it\n has restated its earnings for the fiscal year ended June 30,\n 1986 to provide an additional five mln dlrs to its loan loss\n allowance, causing a restated year-end net loss of 14,937,000\n dlrs, instead of 9,937,000 dlrs.\n The company said the change came after talks with the\n Securities and Exchange Commission on the company's judgement\n in considering the five mln dlrs collectible.\n In the note to its 1986 financial statement, American\n Century said it considered the five mln dlrs collectible,\n making its loan loss provision less than required.\n The company said in spite of the SEC decision, it still\n feels its allowance for possible loan losses at June 30, 1986\n was adequate and that it has considered all relevant\n information to determine the collectability of the five mln dlr\n receivable.\n But, it said continued disagreement with the SEC staff\n would not be in its best interest.

After a minimal analysis of the corpora used, the resources and possible labels were extracted from text and saved into the database. With all possible labels saved, the mapping list was built and used to transform the texts. The results for this experiment are shown on Table 2. The topic coherence for 20-Newsgroup corpus decreased with the corpus normalisation, as for Reuters corpus the coherence increased from 0.456 to 0.475.

Table 2. Topic coherence of the top 10 words in topic using C_V measure.

Dataset	Original corpus	Normalised corpus
Reuters	0.456	0.475
20-Newsgroup	0.672	0.667

As it can be seen by the results in Table 1, there is a positive effect on topic coherence on Reuters corpus, while on 20-Newsgroup it seems to have decreased the metric. From this results, we can leverage a number of hypothesis for these differences: (i) the size of the documents matters, because in small texts it is more difficult to get resources due to the fact that there is little context for the algorithm to disambiguate resources; (ii) the nature of text, as 20-Newsgroup has an e-mail like writing and Reuters is more article like; or (iii) the completeness of the knowledge base in specific topics. On the first hypothesis on the size of documents we can say that when a document it too small, the algorithm cannot be confident enough that a lexeme corresponds to a resource, so it does not capture it. Although 20-Newsgroup has a higher length of documents, both corpora are small, with a mean of less than 200 words per document. Also, by the Fig. 3 we can see that the number of valid words decrease much more on 20-Newsgroup than on Reuters corpus. Hence, we can infer that, although the total number of words is bigger on 20-Newsgroup, the number of valid lexemes to the algorithm to extract resources is very close to Reuters corpus. Besides that, as the context matters, and the context is the set of words around a lexeme, it is very difficult to the NEL algorithm to link a useful resource to the lexemes in text if only just a few words are valid.

This leads to the second hypothesis on the nature of text. It can be seen by Table 1 that the texts have very different natures. An e-mail like text is much more prone to have symbols and initials or acronyms, as seen in the first text of 20-Newsgroup of Table 1. On Reuters, it can be seen that the text is more fluent and without many symbols.

Related to the third hypothesis, we can explore in the future implementations a more complete log to track the resources that exist in the KB or not. The authors noticed during the execution of tests that many resources from Yago linked in the DBpedia page were not available anymore.

6 Conclusion

In this work we presented a framework to semantically normalise texts using resources from the Semantic Web. Our framework was tested in a topic modeling problem using two known corpora in order to have the first results and take insights for improvements.

The framework for normalisation is capable of improving the topic coherence of one of the corpora being tested.

So, the first tests on the semantic normalisation framework of texts showed prominent results and shall be investigated in depth in future. The authors plan to test this normalisation framework on a larger corpus from scientific articles or Wikipedia pages, in order to improve the analysis on the first and second hypothesis.

References

1. ELASTICSEARCH (2019). https://www.elastic.co/pt/
2. Allahyari, M.: Semantic Web Topic Models: Integrating Ontological Knowledge and Probabilistic Topic Models. Ph.D. thesis, University of Georgia (2016)
3. Allahyari, M., Kochut, K.: Semantic tagging using topic models exploiting wikipedia category network. In: 2016 IEEE Tenth International Conference on Semantic Computing (ICSC), pp. 63–70. IEEE (2016)

4. Berners-Lee, T., Hendler, J., Lassila, O., et al.: The semantic web. Sci. Am. **284**(5), 28–37 (2001)
5. Bizer, C., Heath, T., Idehen, K., Berners-Lee, T.: Linked data on the web(ldow2008). In: Proceedings of the 17th International Conference on World WideWeb, pp. 1265–1266 (2008)
6. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. **3**(Jan), 993–1022 (2003)
7. Brickley, D., Guha, R.V., McBride, B.: RDF schema 1.1. W3C Recomm. **25**, 2004–2014 (2014)
8. Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J.L., Blei, D.M.: Reading tea leaves: How humans interpret topic models. In: Advances in Neural Information Processing Systems, pp. 288–296 (2009)
9. Daiber, J., Jakob, M., Hokamp, C., Mendes, P.N.: Improving efficiency and accuracy in multilingual entity extraction. In: Proceedings of the 9th International Conference on Semantic Systems, pp. 121–124 (2013)
10. De Melo, G., Siersdorfer, S.: Multilingual text classification using ontologies. In: European Conference on Information Retrieval, pp. 541–548. Springer (2007)
11. Doshi-Velez, F., Wallace, B., Adams, R.: Graph-sparse IDA: a topic model with structured sparsity. arXiv:1410.4510 (2014)
12. Flisar, J., Podgorelec, V.: Document enrichment using dbpedia ontology for short text classification. In: Proceedings of the 8th International Conference on Web Intelligence, Mining and Semantics, pp. 1–9 (2018)
13. Garla, V.N., Brandt, C.: Ontology-guided feature engineering for clinical text classification. J. Biomed. Inform. **45**(5), 992–998 (2012)
14. Gruber, T.R.: Toward principles for the design of ontologies used for knowledge sharing? Int. J. Hum.-Comput. Stud. **43**(5–6), 907–928 (1995)
15. Hitzler, P., Krotzsch, M., Rudolph, S.: Foundations of Semantic Web Technologies. Chapman and Hall/CRC (2009)
16. Lau, J.H., Newman, D., Baldwin, T.: Machine reading tea leaves: automatically evaluating topic coherence and topic model quality. In: Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, pp. 530–539 (2014)
17. Manola, F., Miller, E., McBride, B., et al.: RDF primer. W3C Recomm. **10**(1–107), 6 (2004)
18. Mendes, P.N., Jakob, M., Garcia-Silva, A., Bizer, C.: Dbpedia spotlight: shedding light on the web of documents. In: Proceedings of the 7th International Conference on Semantic Systems (I-Semantics) (2011)
19. Newman, D., Lau, J.H., Grieser, K., Baldwin, T.: Automatic evaluation of topic coherence. In: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp. 100–108 (2010)
20. O'callaghan, D., Greene, D., Carthy, J., Cunningham, P.: An analysis of the coherence of descriptors in topic modeling. Exp. Syst. Appl. **42**(13),5645–5657 (2015)
21. Prud'hommeaux, E., Seaborne, A.: SPARQL query language for RDF. W3C Recommendation, W3C. Retrieved on 16 Nov 2009 (2008)
22. Rebele, T., Suchanek, F.M., Hoffart, J., Biega, J., Kuzey, E., Weikum, G.: YAGO: a multilingual knowledge base from wikipedia, wordnet, and geonames. In: The Semantic Web - ISWC 2016 - 15th International Semantic Web Conference, Kobe, Japan, 17-2 Oct 2016, Proceedings, Part II, pp. 177–185 (2016). https://doi.org/10.1007/978-3-319-46547-019
23. Röder, M., Both, A., Hinneburg, A.: Exploring the space of topic coherence measures. In: Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, pp. 399–408 (2015)
24. Steyvers, M., Griffiths, T.: Probabilistic topic models. Handb. Latent Seman. **427**(7), 424–440 (2007)

25. Suganya, G., Porkodi, R.: Ontology based information extraction-a review. In: 2018 International Conference on Current Trends towards Converging Technologies (ICCTCT), pp. 1–7. IEEE (2018)
26. Syed, S., Spruit, M.: Full-text or abstract? Examining topic coherence scores using latent dirichlet allocation. In: 2017 IEEE International conference on data science and advanced analytics (DSAA), pp. 165–174. IEEE (2017)
27. Vallet, D., Fernández, M., Castells, P.: An ontology-based information retrieval model. In: European Semantic Web Conference, pp. 455–470. Springer (2005)
28. Waitelonis, J.: Linked Data Supported Information Retrieval. Ph.D. thesis, Karlsruher Institut für Technologie (2018)
29. Yao, L., et al.: Incorporating knowledge graph embeddings into topic modeling. In: Thirty-first AAAI Conference on Artificial Intelligence (2017)