Grau de pertencimento como insumo para classificação automática de textos: uma abordagem sintática

André Fabiano Dyck
Doutorando em Ciência da Informação pela Universidade Federal de Santa Catarina (UFSC) - Florianópolis, SC - Brasil. Mestre em Ciências da Computação pela Universidade Federal de Santa Catarina (UFSC) - Brasil. Analista de Tecnologia da Informação da Universidade Federal de Santa Catarina (UFSC) - Brasil.
http://lattes.cnpq.br/7745380984531130
E-mail: andre.dyck@ufsc.br

Rogério de Aquino Silva
Mestrando em Ciência da Informação pela Universidade Federal de Santa Catarina (UFSC) - Florianópolis, SC - Brasil. Especialização em Business Intelligence pela Instituto Brasileiro de Tecnologia Avançada (IBTA) - Brasil. Cientista de dados do Instituto de Previdência do Estado de Santa Catarina (IPREV) - Florianópolis, SC - Brasil.
http://lattes.cnpq.br/2735959956037192
E-mail: rogerio.aquino@posgrad.ufsc.br

Moisés Lima Dutra
Doutor em Ciências da Computação pela Université Claude Bernarde Lyon 1 (LYON I) - França, com período co-tutela em Universidade Nova de Lisboa (UNL) – Portugal. Professor da Universidade Federal de Santa Catarina (UFSC) - Florianópolis, SC - Brasil.
http://lattes.cnpq.br/1973469817655034
E-mail: moises.dutra@ufsc.br

Gustavo Medeiros de Araújo
Doutor em Engenharia de Automação e Sistemas pela Universidade Federal de Santa Catarina (UFSC) - Florianópolis, SC – Brasil, com período sanduíche em Otto-von-Guericke-Universität Magdeburg – Alemanha. Professor da Universidade Federal de Santa Catarina (UFSC) - Florianópolis, SC - Brasil.
http://lattes.cnpq.br/2609254559240670
E-mail: gustavo.araujo@ufsc.br

Submetido em: 13/09/2020. Aprovado em: 25/11/2020. Publicado em: 28/07/2021.

RESUMO

Agrupar documentos em categorias é uma das soluções adotadas para agilizar o processo de recuperação de informação, cada vez mais relevante devido à grande oferta de informação existente nos dias atuais. A localização manual de documentos de determinada temática, disponíveis em repositórios digitais, passa pela leitura de título, resumo e palavras-chave, além de posterior avaliação mais detalhada com o intuito de identificar se a publicação pertence ao eixo temático desejado. Considerando o número de publicações existentes num repositório digital, a localização manual de todos os textos desejados de uma determinada temática pode ser trabalhosa e demorada. Esta pesquisa propõe uma técnica para classificação automática de textos que se baseia em questões sintáticas, ou seja, empreende uma comparação de n-gramas, que são combinações de *n-uplas* de palavras identificadas ao longo do texto. Realizou-se uma pesquisa aplicada, de cunho exploratório, que aplicou um tipo de aprendizagem supervisionada, baseada fundamentalmente no modelo de representação dos documentos chamado saco de palavras (*bag-of-words* - BoW). Seu objetivo-macro foi o de classificar textos de maneira geral, de acordo com categorias pré-definidas, por meio da geração e comparação de graus de pertencimento entre os textos, como um dos critérios-chave. Os resultados destas comparações, a partir da utilização de n-grama = 3, demonstram que, na utilização de classificações por n-gramas, quanto maior o número de gramas, e com a retirada das *stop words*, obtém-se um grau de pertencimento reduzido, demonstrando um rigor maior para identificar a combinação (*match*) durante a classificação. Para termos maior confiança nos resultados, é necessário um *corpus* de treinamento maior, para ampliar o número de palavras que caracterizem as categorias pré-definidas, a serem utilizadas na classificação dos textos.

Palavras-chave: Grau de pertencimento. Classificação textual. *Bag-of-Words*. N-Gramas. Ciência da Informação.

Degree of belonging as an input for automatic text classification: a syntactic approach

Abstract

Grouping documents into categories is one of the solutions adopted to streamline the information retrieval process, which is increasingly relevant due to the large amount of information available today. The manual localization of documents of a specific theme, available in digital repositories, involves reading the title, abstract and keywords, in addition to further detailed evaluation in order to identify whether the publication belongs to the desired thematic axis. Considering the number of publications in a digital repository, manually locating all the desired texts on a given topic can be laborious and time-consuming. This research proposes an architecture for automatic classification of texts that is based on syntactic questions, that is, it undertakes a comparison of n-grams, which are combinations of n-pairs of words that are identified throughout the text. An exploratory applied research was carried out, which applied a type of supervised learning, fundamentally based on the document representation model called bag-of-words (BoW). The paper's macro objective was to classify texts in general, according to pre-defined categories, by generating and comparing degrees of belonging between texts, as one of the key criteria. The results of these comparisons, using n-gram = 3, demonstrate that in the use of classifications by n-grams, the greater the number of grams, and with the removal of the stop words, we obtain a reduced degree of belonging, demonstrating greater rigor in identifying the match during the classification. In order to have greater confidence in the results, a larger training corpus is necessary to expand the number of words that characterize the pre-defined categories, to be used in the classification of the texts.

***Keywords**: Degree of belonging. Text Classification. Bag-of-Words. N-Grams. Information Science.*

Grado de pertenencia como entrada para la clasificación automática de texto: un enfoque sintáctico

RESUMEN

La agrupación de documentos en categorías es una de las soluciones adoptadas para agilizar el proceso de recuperación de información, que es cada vez más relevante debido a la gran cantidad de información disponible en la actualidad. La localización manual de documentos de un tema específico, disponibles en repositorios digitales, implica la lectura del título, resumen y palabras clave, además de una evaluación más detallada con el fin de identificar si la publicación pertenece al eje temático deseado. Teniendo en cuenta la cantidad de publicaciones en un repositorio digital, ubicar manualmente todos los textos deseados sobre un tema determinado puede resultar laborioso y llevar mucho tiempo. Esta investigación propone una arquitectura de clasificación automática de textos que se basa en preguntas sintácticas, es decir, realiza una comparación de n-gramos, que son combinaciones de n-pares de palabras que se identifican a lo largo del texto. Se realizó una investigación aplicada de carácter exploratorio, que aplicó un tipo de aprendizaje supervisado, basado fundamentalmente en el modelo de representación de documentos denominado bolsa de palabras (bag-of-words - BoW). Su macro objetivo era clasificar los textos en general, según categorías predefinidas, generando y comparando grados de pertenencia entre textos, como uno de los criterios clave. Los resultados de estas comparaciones, utilizando n-gramo = 3, demuestran que en el uso de clasificaciones por n-gramos, a mayor número de gramos, y con la eliminación de las palabras vacías, obtenemos un grado de pertenencia reducido, demostrando mayor rigor en la identificación del partido durante la clasificación. Para tener una mayor confianza en los resultados, es necesario un corpus de formación más amplio para ampliar el número de palabras que caracterizan las categorías predefinidas, para ser utilizadas en la clasificación de los textos.

***Palabras clave**: Grado de pertenencia. Classificación textual. Bag-of-Words. N-Gramos. Ciencias de la información.*

INTRODUÇÃO

Agrupar documentos em categorias é uma das soluções adotadas para agilizar o processo de recuperação de informação, cada vez mais relevante devido à enorme oferta de informação dos dias atuais. Estas categorias, ou rótulos, podem ser geradas por meio de intervenção humana, geralmente associando-se semântica, que facilitaria a recuperação, ou usando apenas algoritmos computadorizados que utilizam outras características dos textos para agrupá-los, num processo que é um tipo de classificação.

A classificação é uma capacidade inerente do ser humano, que utiliza categorias como ferramentas para entender o mundo, e este processo envolve uma série de etapas. Segundo Piaget, no construtivismo, o sujeito aprende com base na assimilação, na integração e na reorganização de estruturas que lhe permitem interpretar o mundo e interagir com ele. Ainda longe de mapear e simular este processo complexo, a classificação de texto apenas atua na organização de informação por meio de atribuição de rótulos. Em um sentido computacional, classificar é atribuir rótulos aos dados, que, no caso da classificação textual, são as palavras de um documento. A categorização, por outro lado, trata de agrupar documentos semelhantes, não rotulados, com base em alguma medida de similaridade (INGERSOLL; MORTON, 2013). Neste trabalho, concordamos com a distinção feita por Ingersoll e Morton (2013), de que a classificação textual (*text classification*) e a categorização textual (*text clustering*) são visões diferentes sobre os dados. Enquanto a primeira distingue a forma pela qual um dado pertence a uma categoria e não a outra, de modo absoluto, a segunda considera a semelhança entre os dados dentro de uma mesma categoria, atribuindo níveis de especialização.

A oferta de repositórios de documentos, nos quais podemos fazer pesquisas livres para encontrar os mais variados temas, está aumentando. A localização manual de documentos de determinada temática, disponíveis em repositórios digitais, passa pela leitura do título, do resumo e das palavras-chave e posterior avaliação mais minuciosa para identificar se esta publicação é do eixo temático[1] desejado. Considerando o número de publicações existentes num repositório digital, a localização manual de todos os textos desejados de determinada temática pode ser trabalhosa e demorada. O cenário de aplicação desta pesquisa parte do pressuposto de que os eixos temáticos de pesquisa dos Programas de Pós-Graduação (PPG) em Ciência da Informação (CI) possuem um alinhamento com os Grupos de Trabalho (GT) da Associação Nacional de Pesquisa e Pós-Graduação em Ciência da Informação (ANCIB). Assim, a realização desta pesquisa abre questões como: Quais são as temáticas mais trabalhadas pelos PPG em CI? Qual é o alinhamento das teses e dissertações da área da CI com os temas dos GTs da Ancib? Como localizar teses e dissertações alinhadas com um eixo temático específico relacionado a um GT da Ancib? Este artigo é uma versão revista e ampliada de um trabalho apresentado no III Workshop de Informação, Dados e Tecnologia (WIDaT 2019), que ocorreu em Brasília em novembro de 2019 (DYCK; DUTRA; VIERA, 2019). A estrutura do artigo segue com: (i) os procedimentos metodológicos utilizados; (ii) a análise e discussão dos resultados; e (iii) as considerações finais.

METODOLOGIA

A realização deste trabalho toma por base o modelo de representação dos documentos chamado saco de palavras (*bag-of-words* ou BoW). O BoW é um modelo de representação simplificado usado no processamento de linguagem natural[2], uma subárea da Ciência da Computação, para representar os documentos como um conjunto de palavras, sem considerar sua semântica original (HARRIS, 1954; GOLDBERG, 2017).

Ao trabalhar as coleções de textos como BoW, utilizando apenas as palavras e suas combinações presentes no texto, sem considerar questões semânticas, chamamos de n-gramas essas palavras ou suas combinações. Por exemplo, a palavra "grau", é um exemplo da representação de n-grama=1; as palavras "grau de", são exemplos da representação de n-grama=2; as palavras "grau de pertencimento", são exemplos da representação de n-grama=3; as palavras "grau de pertencimento como", são exemplos da representação de n-grama=4 e; as palavras "grau de pertencimento como insumo", são exemplos da representação de n-grama=5 (MOURA *et. al.*, 2010; JURAFSKY; MARTIN, 2018). As coleções de textos trabalhadas neste artigo como cenário de aplicação se referem: (i) aos resumos e aos textos completos de teses e dissertações do Programa de Pós-Graduação em Ciência da Informação, da Universidade Federal de Santa Catarina (PGCIN/UFSC[3]), extraídos do Repositório Institucional da UFSC – RI/UFSC[4]; e (ii) às ementas dos GTs da Ancib, extraídas do site do "Fórum de Coordenadores de Grupo de Trabalho da Ancib"[5].

No presente estudo, realizamos uma pesquisa aplicada, de cunho exploratório, cuja coleta e tabulação de dados se deram entre os dias 26/08/2019 e 09/09/2019, para os resumos, e entre os dias 12/12/2019 e 21/12/2019, para os textos completos, e que toma como cenário de aplicação o conjunto de 223 documentos das teses e dissertações em CI que estavam disponíveis no RI/UFSC naquele período. Para obtermos os resumos, ao acessar o RI/UFSC, selecionamos a comunidade "Teses e Dissertações" e, então, a coleção "Programa de Pós-Graduação em Ciência da Informação". Em seguida, exportamos os metadados[6] dos 223 documentos para um arquivo CSV que, então, foram lidos pela biblioteca Python Pandas.

[1] Eixo temático no contexto deste trabalho significa um suporte ou guia para limitar os conteúdos de um assunto principal.

[2] Linguagem natural no contexto deste trabalho são as línguas faladas pelos humanos.

[3] http://pgcin.paginas.ufsc.br/

[4] https://repositorio.ufsc.br/

[5] http://gtancib.fci.unb.br/

[6] Metadados, no contexto deste trabalho, são dados sobre as dissertações e teses, como por exemplo: título, resumo, tipo de publicação, palavras-chave etc.

Com o arquivo CSV em memória, extraímos a coluna de resumos. Os textos completos das teses e dissertações dos 223 documentos foram obtidos com a equipe que administra a ferramenta de RI da UFSC, num arquivo compactado em formato ZIP[7].

A redução dos textos às palavras que os constituem forma os unigramas. Uma possibilidade de se preservar um mínimo do significado do texto, usando ainda BoW, é utilizar também bigramas e trigramas (GOLDBERG, 2017). Assim, mantém-se a proximidade de duas e três palavras do texto original. Esta pesquisa identificou e analisou unigramas, bigramas e trigramas, e, também, n-grama = 4 e n-grama = 5, tanto para os resumos quanto para os textos completos de teses e dissertações e para as ementas de GT coletadas, para assim comparar seus resultados.

Este trabalho utilizou técnicas de Aprendizagem de Máquina, uma área de inteligência artificial que está preocupada em desenvolver algoritmos que aprendem padrões presentes em uma massa de dados (chamada de massa de dados de aprendizagem). Estes padrões aprendidos podem ser usados para predizer informações sobre dados novos, por isso a importância da massa de dados de aprendizagem ser diversa o suficiente para ampliar as chances de predições (BAEZA-YATES; RIBEIRO-NETO, 2013).

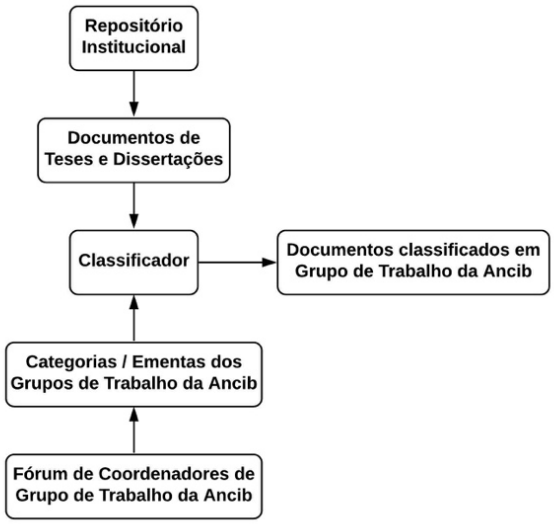
O uso deste tipo de algoritmos é extensivo em diagnóstico médico, detecção de fraudes a cartões de crédito, análise de mercado de ações e recuperação de informação. Na recuperação de informação, a classificação de textos é chave para o sucesso (BAEZA-YATES; RIBEIRO-NETO, 2013). Para automatizar a classificação de texto, podemos fazer uso de várias técnicas e conceitos. Há principalmente três tipos de técnicas de aprendizagem: (i) Aprendizagem de máquina supervisionada, quando há intervenção humana na fase de treinamento; (ii) Aprendizagem de máquina não supervisionada, quando não há intervenção humana no treinamento, como, por exemplo, a técnica chamada de clusterização (*clustering*); e (iii) Aprendizagem semi-supervisionada, na qual o conjunto inicial de dados é composto apenas por uma pequena entrada rotulada e grande parte dos dados da entrada não está rotulada, i.e., a categoria associada a eles é desconhecida. Neste caso, o objetivo é similar à classificação supervisionada, que é gerar uma relação binária, mapeando a entrada para saída (BAEZA-YATES; RIBEIRO-NETO, 2013).

Nesta pesquisa, trabalhamos com classificação textual (*text classification*), que é um tipo de aprendizagem supervisionada, pois há intervenção humana na definição prévia das categorias. Em nosso cenário de aplicação, utilizamos as categorias pré-definidas dos GTs da Ancib. Nosso objetivo de classificar documentos de textos (resumos e textos completos) das teses e dissertações, de acordo com categorias pré-definidas, pode descrever a tarefa como uma função: $DxC\ \{T,F\}$, onde $D=\{d1,d2,...,d223\}$ é o conjunto que representa o *corpus* de documentos, no nosso caso 223 documentos de teses e dissertações, e $C=\{c1,c2,...,c11\}$ é o conjunto pré-definido de categorias que são os 11 GTs da Ancib.

O valor *T* atribuído a <*dj,ci*> indica uma decisão de classificar *dj* como *ci*, e *F* indica que *dj* não é classificado como *ci* (BAEZA-YATES; RIBEIRO-NETO, 2013). A figura 1 esquematiza os módulos que constituem a técnica proposta para classificação automática de teses e dissertações.

[7] Zip (ou ZIP) é um formato de arquivo usado para compactação de dados armazenados no computador. O objetivo da compactação é reduzir o tamanho de um arquivo ou agrupar vários arquivos em um só.

Figura 1 − Proposta para a classificação automática de teses e dissertações da CI

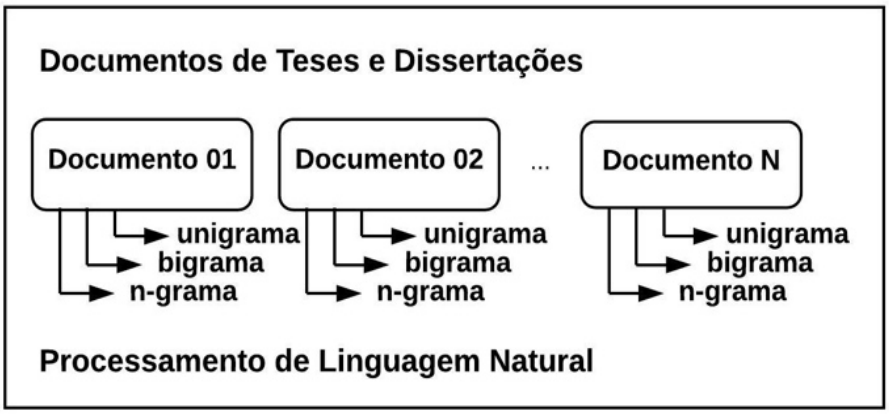


Fonte: Dyck, Dutra e Viera (2019).

1. **Repositório Institucional:** Ambiente de acesso livre e irrestrito à literatura científica e acadêmica da instituição.
2. **Documentos de Teses e Dissertações:** Os documentos (resumos e textos completos) extraídos do Repositório Institucional submetidos ao processamento de linguagem natural.
3. **Classificador:** Onde ocorre o treinamento, o cálculo do grau de pertencimento e a obtenção da função de classificação.
4. **Categorias / Ementas dos Grupos de Trabalho da Ancib:** As ementas dos GTs da Ancib extraídos do site do "Fórum de Coordenadores de Grupo de Trabalho da Ancib", submetidos a processamento de linguagem natural.
5. **Fórum de Coordenadores de Grupo de Trabalho da Ancib:** Site com a apresentação da Ancib e a descrição e ementas de cada um dos seus GTs.
6. **Documentos classificados em Grupo de Trabalho da Ancib:** Resultados da classificação automática com os documentos classificados em uma categoria (um GT da Ancib).

A partir da interação destes seis módulos, chega-se à classificação automática dos documentos das teses e dissertações com relação aos GTs da Ancib. O módulo 1 representa o RI, ambiente de acesso livre e irrestrito à literatura científica e acadêmica da instituição, que é a fonte dos documentos das publicações de teses e dissertações.

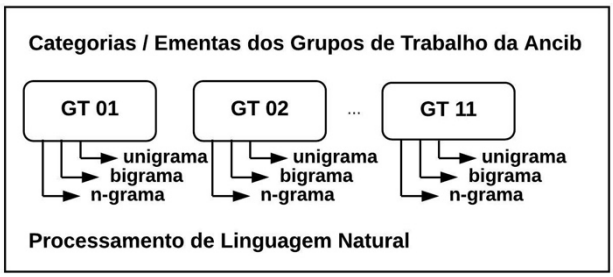
Figura 2 − Processamento de linguagem natural sobre os documentos de teses e dissertações



Fonte: Dyck, Dutra e Viera (2019).

O módulo 2, detalhado na figura 2, apresenta o processamento de linguagem natural, utilizando o modelo BoW, para a criação e limpeza de dados dos n-gramas, para cada um dos documentos, resumos e textos completos, de teses e dissertações extraídos do RI. O módulo 3 representa a classificação automática, no qual ocorre o treinamento, o cálculo do grau de pertencimento e a obtenção da função de classificação.

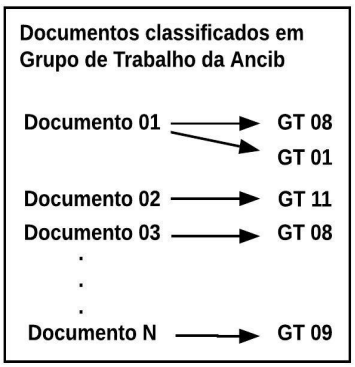
Figura 3 − Processamento de linguagem natural sobre as Categorias / Ementas dos GTs da Ancib



Fonte: Dyck, Dutra e Viera (2019).

O módulo 4, detalhando na figura 3, apresenta o processamento de linguagem natural, semelhante ao módulo 2, utilizando o modelo BoW, para a criação e limpeza de dados dos n-gramas, para cada uma das categorias / ementas dos 11 GTs da Ancib. O módulo 5 representa o "Fórum de Coordenadores de Grupo de Trabalho da Ancib", do qual foram extraídas as categorias / ementas de cada um dos GTs. O módulo 6, detalhado na figura 4, apresenta o resultado da classificação automática, em que cada um dos documentos de teses e dissertações foi classificado com uma probabilidade de pertencimento a determinado GT da Ancib.

Figura 4 − Documentos classificados em GTs da Ancib



Fonte: Dyck, Dutra e Viera (2019).

Nosso cenário de aplicação foi montado sobre as onze ementas dos GTs da ANCIB e os 223 documentos das teses e dissertações em CI, do PGCIN, obtidos no RI/UFSC. Uma vez de posse dos 223 documentos, os seus metadados nos mostraram que, de fato, tínhamos 185 resumos. E, de posse do arquivo ZIP com todos os textos completos, tínhamos 190 textos completos, aptos a serem trabalhados. Os documentos restantes não possuíam resumo e/ou texto completo hospedado no RI/UFSC. Nosso processo, desenvolvido para identificar automaticamente o grau de pertencimento dos documentos, conta com as seguintes etapas:

1. Criamos uma lista de *stop words* (palavras sem valor semântico), a partir das *stop words* relacionadas dentro da lista do NLTK[8], que é uma biblioteca do Python;

2. Geramos os dicionários com as ementas dos GTs. Para cada ementa, pegamos todo o seu texto e o colocamos em uma variável de dicionário. Pegamos, então, esse texto da ementa e o colocamos dentro de variáveis do tipo STRING[9]. Isto é, pegamos todo o texto original da ementa sem qualquer alteração. É esse texto que está na variável. Assim, temos uma variável para cada GT;

3. Aplicamos uma função, que criamos, que é aplicada sobre todas essas variáveis, para remover toda acentuação;

4. Em seguida, passamos para o processo de normalização em cada um dos documentos (ementas, resumos e textos completos). Retiramos a acentuação, caracteres que não são de A-Z, com ou sem acento. Transformamos todos os caracteres acentuados em não acentuados. Isso também é feito para os caracteres latinos, como a letra "Ç", que é transformada na letra "C".

[8] https://www.nltk.org/

[9] O termo STRING serve para identificar uma sequência de zero ou mais caracteres. Na prática, as STRINGS são usadas para representar textos.

Retiramos, então, todas as pontuações-padrão ('.', ',', ';', '!', '?', etc.), e transformamos todas as letras em maiúsculas. Esse mesmo processo de normalização também é aplicado à lista de *stop words*, criada inicialmente, a partir da lista do NLTK. Até aqui, temos nossas variáveis do tipo STRINGS normalizadas;

5. Então, as variáveis STRING são divididas (*splitted*) e são geradas listas, que são os vetores de palavras. A partir deste momento, o que é feito tem relação com o processo de mineração de dados[10]. As variáveis são colocadas em dicionários, que são uma estrutura de dados que possui os campos 'chave' e 'valor'. O campo chave armazena o número do GT e o campo valor armazena um vetor de palavras, para cada GT, criando-se um *set* com eles. Criar um *set* consiste em retirar as palavras repetidas. O *set* garante que as palavras restantes são únicas. É preciso garantir que estas palavras sejam únicas, que não se repitam, para não se gerar métricas equivocadas. Elas serão utilizadas na comparação com os documentos de resumos e textos completos de teses e dissertações. Essa lista de vetores de palavras está contida dentro do dicionário dos GTs;

6. A seguir, carregamos os resumos, numa primeira rodada, e os textos completos, numa segunda rodada. Neste momento, aplicamos uma função, chamada "VerificaTaxa", apresentada na figura 5, para identificar o grau de pertencimento entre um documento (resumo ou texto completo) um GT da Ancib. Essa verificação se repete para todos os textos, que, neste cenário, são as ementas dos GTs, os resumos e os textos completos dos documentos de teses e dissertações. Isso é necessário para podermos compará-los. A função verifica, para todas as palavras de uma ementa de GT, sua existência, ou não, na lista de palavras dos documentos (resumo ou texto completo). A seguinte verificação é feita:

a) Quantas dessas palavras existem dentro desse documento? Quantas vezes aparece essa palavra dentro do documento? Nenhuma, duas, três, etc., e incrementa-se um contador para o GT em avaliação. É utilizado um contador para cada GT. É contabilizada a quantidade de palavras do GT, encontradas dentro de um documento (resumo ou texto completo);

b) Depois de verificar todo o documento e chegar em um número final de ocorrências de palavras do vetor do GT, aplicamos a fórmula abaixo, na qual, pega-se esse número de ocorrências e divide-se pelo número total de palavras (tamanho total, *length*) do documento (resumo ou texto completo) e multiplica-se por 100, para verificar a porcentagem, ou seja, o grau de pertencimento do GT dentro daquele documento. Caso o resultado seja ZERO, automaticamente, sabemos que não há pertencimento do documento ao GT com o qual foi feita a comparação. O resultado é diferente de ZERO quando ao menos uma mesma palavra foi encontrada tanto no GT como no documento. E, nesse caso, registramos numa variável chamada "tupla" os valores: número do GT e grau de pertencimento, conforme figura 5.

$$\frac{Número\ de\ ocorrências}{Número\ total\ de\ palavras} \times 100 = Grau\ de\ pertencimento$$

Figura 5 – Código Python da função "VerificaTaxa"

```
def VerificaTaxa(texto):
    texto = remover_acentos(texto)
    texto = texto.split()
    texto = RemoveStopWord(texto)

    resultado = list()
    for gt in grupos:
        lista_frequencia = list()
        quantidade = 0
        for palavra in grupos[gt]:
            quantidade = quantidade + texto.count(palavra)

        if quantidade>0:
            freq_gt = (quantidade / len(texto))*100
            tupla = (gt, freq_gt)
            resultado.append(tupla)

    return resultado
```

Fonte: Autores (2020).

[10] Mineração de dados (também conhecida pelo termo inglês data mining) é o processo automatizado ou semiautomatizado para extrair conhecimentos e padrões, a partir de grandes quantidades de dados (OLAFSSON; LI; WU, 2008).

O processo de comparação entre os n-Gramas das ementas dos GTs com os resumos, num primeiro momento, e com os textos completos, dos documentos de teses e dissertações, num segundo momento, diversifica a quantidade de palavras a serem comparadas, de acordo com a variação do número **n** do **n**-grama utilizado, da seguinte maneira:

- Para o n = 1, os unigramas, os vetores são constituídos pelas próprias palavras, sem a necessidade de se aplicar a função para a formação dos n-gramas maiores que 1, conforme a seguir;
- Para o n = 2, os bigramas, a função inicia pegando a primeira e segunda palavras do documento. Assim, o primeiro bigrama é composto pela primeira e segunda palavras do documento. Em seguida, a função pega a segunda palavra do documento, junto com a palavra seguinte, que, neste caso, é a terceira. Agora, este novo bigrama é composto pela segunda e terceira palavras do documento. E, assim, sucessivamente, vai para a próxima palavra e pega também a palavra seguinte, e repete este processo até o final do documento;
- Para o n = 3, os trigramas, a função inicia pegando a primeira, a segunda e a terceira palavras do documento. Assim, o primeiro trigrama é composto pela primeira, segunda e terceira palavras do documento. Em seguida, a função pega a segunda palavra do documento, junto com as duas palavras seguintes, que são, agora, a terceira e a quarta palavras do documento. E, assim, sucessivamente, vai para a próxima palavra e pega também as duas palavras seguintes e repete este processo até o final do documento;
- Para o n = 4, a função inicia pegando a primeira, segunda, terceira quarta palavras do documento. Assim, o primeiro n-grama = 4 é composto pela primeira, segunda, terceira e quarta palavras do documento. Em seguida, a função pega a segunda palavra do documento, junto com as três palavras seguintes, que são, agora, a terceira, a quarta e a quinta palavras do documento. E, assim, sucessivamente, vai para a próxima palavra e pega também as três palavras seguintes, e repete este processo até o final do documento;
- E, para o n=5, a função inicia pegando a primeira, a segunda, a terceira, a quarta e a quinta palavras do documento. Assim, o primeiro n-grama = 5 é composto pela primeira, segunda, terceira, quarta e quinta palavras do documento. Em seguida, a função pega a segunda palavra do documento, junto com as quatro palavras seguintes, que são, agora, a terceira, a quarta, a quinta e a sexta palavras do documento. E, assim, sucessivamente, vai para a próxima palavra e pega também as quatro palavras seguintes, e repete este processo até o final do documento.

Isso é feito tanto para os vetores com as palavras das ementas dos GTs, como para os vetores com as palavras dos documentos, sejam os resumos ou os textos completos. É feito para todos os documentos: ementas, resumos e textos completos.

ANÁLISE E DISCUSSÃO DOS RESULTADOS

O resultado da aplicação do processo, no cenário de aplicação aqui apresentado é sintetizado nas tabelas 1 e 2. A primeira com os resultados da aplicação do processo sobre os resumos e a outra com os resultados da aplicação do processo sobre os textos completos. Considerando que o número total de resumos avaliados foi de 185, observamos que os resultados da aplicação do processo, utilizando unigramas, n-grama = 1, sem *stop words*, que aparece na primeira linha de resultados da tabela 1, mostra um grau de pertencimento bem próximo do total de 185 resumos, em todos os GTs. No GTs 2, 6 e 7, o grau de pertencimento atingiu a totalidade dos 185 resumos. Na aplicação do processo, utilizando unigramas, vetores de palavras com n-grama = 1, ou seja, com uma única palavra a ser comparada, utilizamos apenas vetores sem *stop words*. Os resultados da aplicação do processo utilizando os bigramas, n-grama = 2, foram idênticos tanto para os vetores com e sem *stop words*. Esses resultados, da aplicação de unigramas e bigramas na comparação dos termos, obtendo valores extremos, próximos da totalidade de documentos avaliados, neste cenário de aplicação, indicam não ser apropriado seu uso.

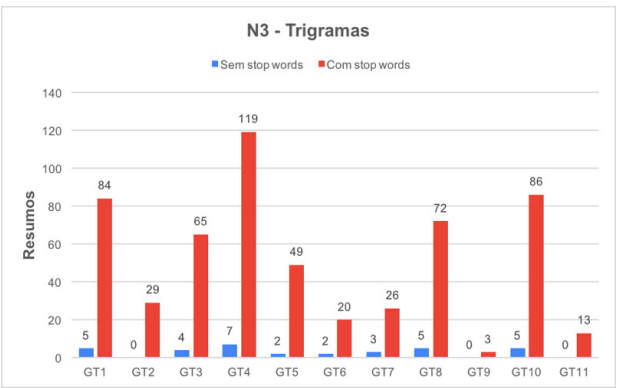
Tabela 1 – Resultados para os Resumos

Resultados para os Resumos												
N-Gramas		GT1	GT2	GT3	GT4	GT5	GT6	GT7	GT8	GT9	GT10	GT11
N-Grama = 1	Sem stop Words	184	185	182	181	177	185	185	178	179	180	178
N-Grama = 2	Sem stop words	170	167	140	180	150	165	151	158	148	162	154
	Com stop words	170	167	140	180	150	165	151	158	148	162	154
N-Grama = 3	Sem stop words	5	0	4	7	2	2	3	5	0	5	0
	Com stop words	84	29	65	119	49	20	26	72	3	86	13
N-Grama = 4	Sem stop words	0	0	0	1	0	0	0	1	0	0	0
	Com stop words	35	7	28	41	6	1	5	17	1	37	1
N-Grama = 5	Sem stop words	0	0	0	0	0	0	0	0	0	0	0
	Com stop words	3	3	1	12	3	0	0	8	0	3	0

Fonte: Autores (2020).

A aplicação do vetor com trigramas, n-grama = 3, mostra resultados com uma significativa oscilação, não ficando em zero, quando sem o uso de *stop words*, exceto para os GTs 2, 9 e 11. Há, também, uma oscilação longe do extremo de todos os resumos, 185, com resultados variando entre grau de pertencimento igual a 3 para o GT9 e grau de pertencimento igual a 119 para o GT4, conforme figura 6. Esses resultados apresentam graus bem mais reduzidos quando se utilizam vetores sem *stop words*.

Figura 6 – Resumos em Trigramas: Número de documentos pertencentes a cada GT



Fonte: Elaborado pelos autores (2020).

A aplicação do vetor com n-grama = 4, sem o uso de *stop words,* obteve resultados de grau de pertencimento igual a zero, nas comparações com todos os GTs, exceto para os GTs 4 e 8, nos quais o resultado foi igual a um. Esses resultados também indicam não ser apropriado o uso de n-gramas sem *stop words* a partir de n = 4, uma vez que no cenário de aplicação aqui utilizado não identificam pertencimento a, praticamente, nenhum dos GTs avaliados. Isso se justifica pelo fato de que, quando se trata de n-gramas, as *stop words* adquirem a importância que não possuem na mineração textual com termos simples, devido ao fato de que, combinadas com outras palavras da *n-upla*, elas, neste caso, possuem valor semântico. A aplicação do vetor com n-grama = 4 com o uso de *stop words* obteve resultados que pareceram promissores, num primeiro momento, porém, quando comparados com a totalidade de resultados, com os graus de pertencimento em relação a todos os GTs, apresenta uma grande oscilação nos resultados, variando desde um grau de pertencimento igual a um, com o GTs 6, 9 e 11, até uma taxa de quarenta e um, com o GT4.

Já os resultados para os resumos, com n=5, tanto sem como com *stop words*, apresentam combinações (*matches*) consideravelmente reduzidas.

A tabela 2 mostra os resultados da aplicação do processo com os textos completos, no total de 190 documentos, de teses e dissertações do PGCIN/UFSC. Os resultados da aplicação do processo, utilizando unigramas sem *stop words*, e bigramas tanto com, quanto sem *stop words*, que aparecem nas três primeiras linhas da tabela 2, à semelhança dos resultados obtidos com os resumos, obtendo resultados da totalidade de documentos avaliados, neste cenário de aplicação, indicam não ser apropriado seu uso. Isto ocorre porque nos textos completos a probabilidade de ocorrência de termos simples (n = 1) ou em duplas (n = 2), ainda que desconectados da questão semântica, é muito maior, ou seja, estas duas situações acabam por não servir de balizamento para a determinação do grau de pertencimento.

Tabela 2 − Resultados para os Textos completos

Resultados para os Textos completos												
N-Gramas		GT1	GT2	GT3	GT4	GT5	GT6	GT7	GT8	GT9	GT10	GT11
N-Grama = 1	Sem stop Words	190	190	190	190	190	190	190	190	190	190	190
N-Grama = 2	Sem stop words	190	190	190	190	190	190	190	190	190	190	190
	Com stop words	190	190	190	190	190	190	190	190	190	190	190
N-Grama = 3	Sem stop words	48	47	48	47	46	42	44	47	34	47	42
	Com stop words	189	184	188	189	186	166	182	190	132	190	168
N-Grama = 4	Sem stop words	9	4	10	10	0	3	3	5	0	8	0
	Com stop words	182	78	165	182	83	48	127	167	16	177	15
N-Grama = 5	Sem stop words	3	0	0	4	0	0	1	1	0	1	0
	Com stop words	79	13	4	151	41	1	34	113	1	74	1

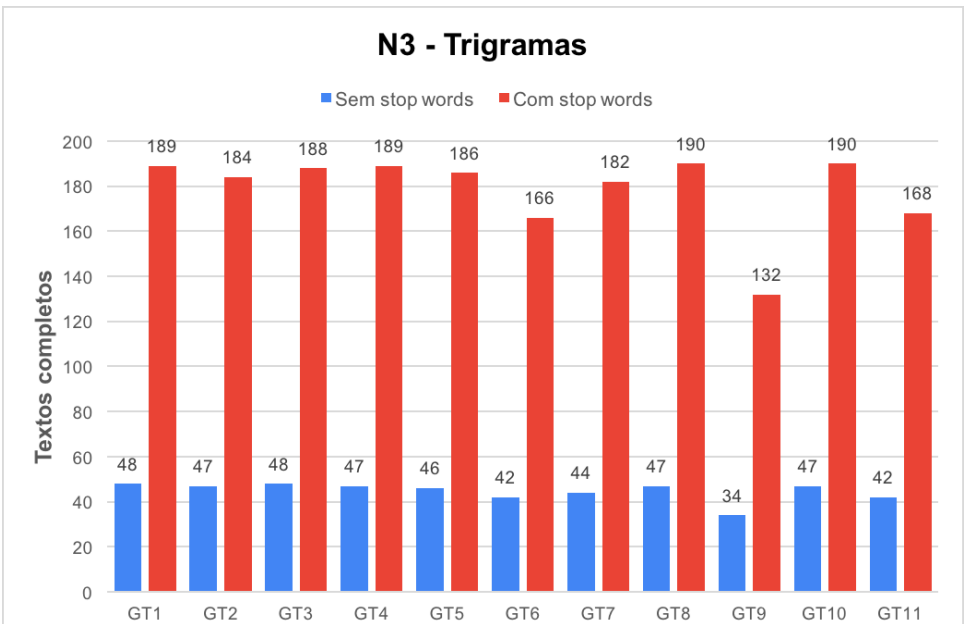
Fonte: Elaborado pelos autores (2020).

Com os resultados da aplicação do processo no vetor com trigramas, n-grama = 3, apresentado na figura 7, entramos numa possível discussão do "equilíbrio perfeito". Percebemos, com estes resultados, que existe um equilíbrio nos valores da aplicação sem *stop words*. Aqui, fazendo uma média aritmética dos resultados de cada um dos GT, sem *stop words*, e dividindo pelo total de documentos avaliados, 190, e multiplicando por 100, chega-se ao resultado de que 23% dos documentos fazem parte de algum GT. Isto é, 23% dos textos completos estão inseridos em cada GT. Este resultado mostra um equilíbrio na distribuição das teses e dissertações do PGCIN, nos 11 GTs da Ancib. Quase 1/4 dos documentos estão representados em cada GT.

Nos resultados da aplicação de trigramas com *stop words*, os valores obtidos para os GTs 8 e 10 mostram o resultado de 190, que é todo universo de documentos analisados. Esses valores aparecem na linha 5 dos resultados apresentados na tabela 2. Para isso ter acontecido, é indicação de que trigramas com palavras sem valor semântico, como, por exemplo, "E ISSO E", "POR QUE ISSO", tenham sido utilizados na comparação. Uma vez que, em nosso processo, basta o contador ser igual a um, ou seja, existir apenas uma combinação (um *match*), para considerar que existe um pertencimento daquele documento ao GT. Ou seja, quando se mantém as *stop words*, temos, potencialmente, esse tipo de resultado. Quando retiramos as *stop words*, em nosso processo de comparações, significando a inexistência de palavras com baixo, ou nenhum, peso semântico, automaticamente, percebemos resultados com números menores nos graus de pertencimento, sugerindo maior confiabilidade.

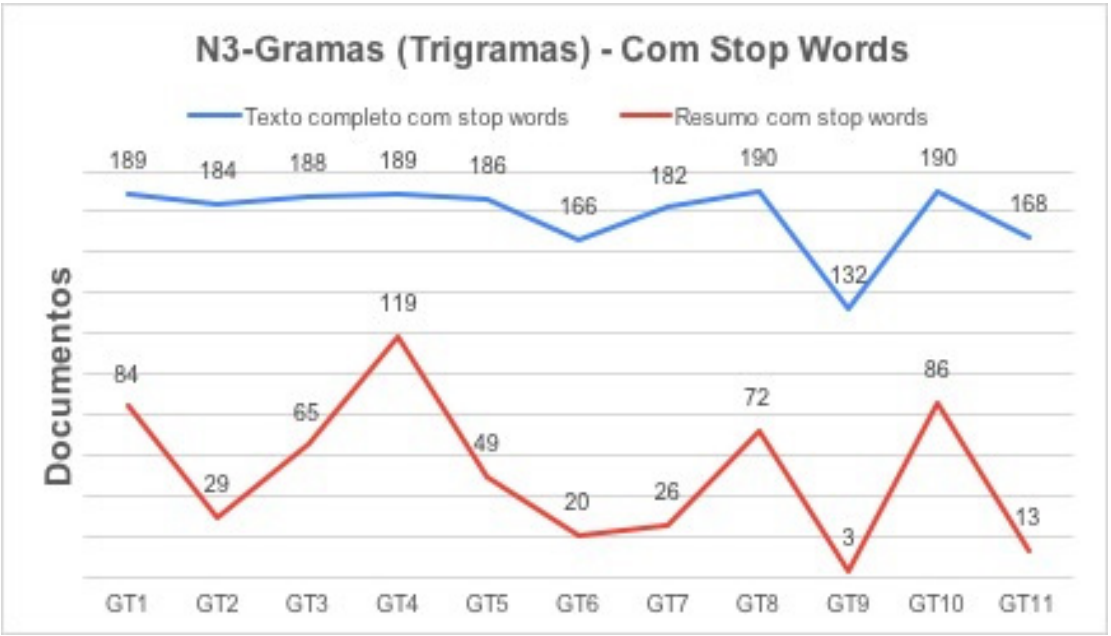
Os resultados, apresentados na tabela 2, mostram que a confiabilidade pode ser auferida com os trigramas sem *stop words*. Diante dos resultados obtidos nesta pesquisa, é a partir do n = 3, dos trigramas, que se obtém os melhores graus de pertencimento, com as melhores chances de acerto. A figura 8 mostra os resultados obtidos na aplicação de trigramas com *stop words*, tanto para resumos quanto para textos completos. A figura 9 mostra os resultados obtidos na aplicação, também de trigramas sem *stop words* tanto para os resumos e textos completos.

Figura 7 – Textos completos em Trigramas: número de documentos pertencentes a cada GT



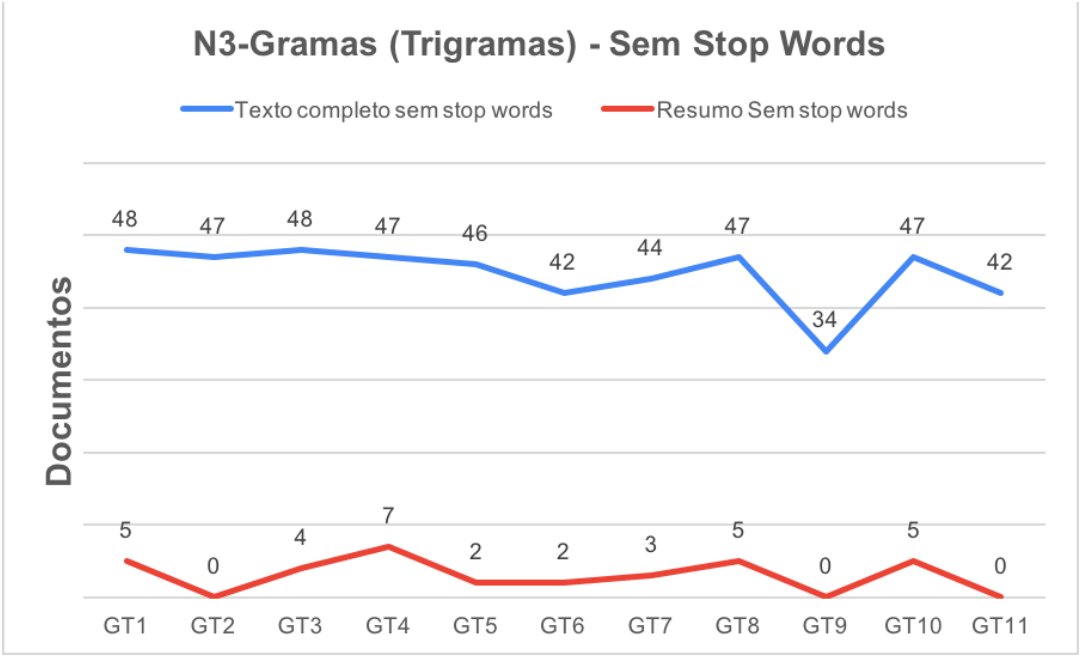
Fonte: Elaborado pelos autores (2020).

Figura 8 – Resultados dos Trigramas com stop words para Resumos e Textos completos



Fonte: Elaborado pelos autores (2020).

Figura 9 – Resultados dos Trigramas sem stop words para Resumos e Textos completos



Fonte: Elaborado pelos autores (2020).

CONCLUSÕES

Esta pesquisa utilizou um tipo de aprendizagem supervisionada para classificar textos de maneira geral, de acordo com categorias pré-definidas. Utilizamos como cenário de aplicação o universo de documentos de teses e dissertações do PGCIN/UFSC, de acordo com as ementas dos GTs da ANCIB. Utilizando o modelo BoW, propusemos uma arquitetura de classificação automática, identificando o grau de pertencimento de cada um dos documentos. Os resultados das comparações com os textos completos, a partir da utilização de n-grama = 3, demonstraram que a utilização de classificação por n-gramas implica um cenário em que quanto maior o número de gramas, sem *stop words*, menor será grau de pertencimento obtido. Ou seja, será muito mais difícil se identificar uma combinação entre documentos (*match*). É preciso, no entanto, que seja levada em consideração a semântica do cenário de aplicação trabalhado na decisão de se incluir ou não as *stop words* na formação dos n-gramas e serem buscados. Dependendo do contexto, n-gramas conhecidos e que representam expressões corriqueiras poderão se basear fortemente na utilização de elementos textuais, não obstante, considerados *stop words*. Será preciso achar um equilíbrio nas decisões de projeto de maneira a se maximizar o grau de pertencimento obtido, com a extração da maior semântica possível do corpus *textual*.

A proposta desta pesquisa foi de identificar o grau de pertencimento entre textos de maneira geral, utilizando como cenário de aplicação os documentos de teses e dissertações do PGCIN em relação aos GTs da ANCIB. Como premissa de pesquisa, é possível que um documento pertença, em maior ou menor grau, a vários GTs, pois, mesmo que a linha de pesquisa na qual a tese ou dissertação foi escrita, focada num assunto particular, que tenha relação com a ementa de um determinado GT, considerando que a área da Ciência da Informação possui inúmeras interseções com diferentes áreas do conhecimento, sendo multidisciplinar, é possível referenciar, num único trabalho, assuntos que são fortemente trabalhados em diferentes GTs. Isso também pode acontecer devido ao que é discutido na introdução dos trabalhos, muitas vezes fazendo apresentações históricas da área. Isso sugere, portanto, a necessidade de um trabalho em que seja definida uma linha de corte, para se definir o grau de pertencimento a um GT. A definição de uma linha de corte apropriada demandará um trabalho criterioso.

Outra questão a ser ponderada, foi a utilização das ementas dos GTs, encontradas no site da ANCIB, como único documento de fonte textual para caracterizar cada GT. Para termos maior confiança nos resultados, é necessário um *corpus* maior, para ampliar o número de palavras que caracterizem cada GT. Que então, por sua vez, serão utilizadas para a classificação das teses e dissertações.

Futuros trabalhos incluem a definição de um limite de corte, no número de palavras coincidentes, para então considerar que existe um pertencimento a determinado GT; comparar os resultados da classificação dos resumos e seus respectivos textos completos, de um mesmo documento, tese ou dissertação no cenário aqui utilizado, e comparar os resultados do grau de pertencimento, para então avaliar a representatividade dos resumos em relação aos seus textos completos; o desenvolvimento de um aplicativo, de livre acesso pela internet, que permita classificação de quaisquer *corpora* de documentos de acordo com categorias pré-definidas; testar o modelo com um método não-supervisionado, a Clusterização, usando volumes maiores de dados. E, também, posteriormente, substituir o método BoW por uma técnica que preserve a semântica, como, por exemplo, os vetores de palavras (WordEmbeddings).

REFERÊNCIAS

BAEZA-YATES, R.; RIBEIRO-NETO, B. *Recuperação de Informação: conceitos e tecnologia das máquinas de busca.* 2. Ed. Porto Alegre: Bookman, 2013.

DYCK, A. F.; DUTRA, M. L.; VIERA, A. F. G. Classificação automática de teses e dissertações da área da Ciência da Informação sob a ótica dos grupos de trabalho da Ancib. In: WORKSHOP DE INFORMAÇÃO, DADOS E TECNOLOGIA, 2019, Brasília. *Anais* [...] Widat 2019. Brasília: Editora da UnB, 2019. p. 48-53.

FAN, W. *et. Al*. Tapping the power of text mining. *ACM,* [*s. l.*], v. 49, n. 9, p. 76-82, 2006. DOI: https://doi.org/10.1145/1151030.1151032

GOLDBERG, Y. *Neural network methods in natural language processing*: synthesis lectures on human language technologies. [*S. l.*]: Morgan & Claypool Publishers. 2017. 310 p.

HARRIS, Z.S. Distributional Structure. *WORD*, [*s. l.*], v. 10, n. 2-3, p. 146-162, 1954. Publicado online em 04 dez. 2015. ISSN: 0043-7956. Disponível em: <https://www.tandfonline.com/doi/pdf/10.1080/00437956.1954.11659520> Acesso em: 11 fev. 2020.

INGERSOLL, G.S.; MORTON, T.S.; FARRIS, A.L. 2013. *Taming text:* how to find, organize and manipulate it. Shelter Island, NY (USA): Manning Publications Co., 2012. 298 p. ISBN: 9781933988382

JURAFSKY, D.; MARTIN, J.H. S*peech and language processing*: an introduction to natural language processing, computational linguistics, and speech recognition. 3. ed. Stanford University. 2020. Disponível em:<https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf>. Acesso em: 04 mar. 2020.

KUMAR, P. An introduction to n-grams: what are they and why do we need them?. *XRDS Crossroads The ACM Magazine for Students.* 2017. Disponível em: <https://blog.xrds.acm.org/2017/10/introduction-n-grams-need/>. Acesso em: 11 fev. 2020.

MEIRELES, M. R. G.; CENDóN, B. V. Categorização e classificação de documentos a partir de suas citações: uma proposta baseada em redes neurais artificiais. *DataGramaZero*, v. 12, n. 5, 2011. Disponível em: <http://hdl.handle.net/20.500.11959/brapci/7466>. Acesso em: 11 mar. 2021.

MOURA, M.F.; *et al. Um modelo para seleção de n-gramas significativos e não redundantes em tarefas de mineração de textos.* Campinas: Embrapa Informática Agropecuária, 2010. (Boletim de pesquisa e desenvolvimento, n. 23). ISSN 1677- 9274.

OLAFSSON, S.; LI, X.; WU, S. Operations research and data mining. *European Journal of Operational Research*, [*s. l.*], v. 187, n. 3, p. 1429-1448. 2008. ISSN 0377-2217. DOI: https://doi.org/10.1016/j.ejor.2006.09.023.

SILVA, E.L. *Metodologia da pesquisa e elaboração de dissertação.* Estera Muszkat Menezes. 4. ed. rev. atual. Florianópolis: UFSC, 2005. 138p.