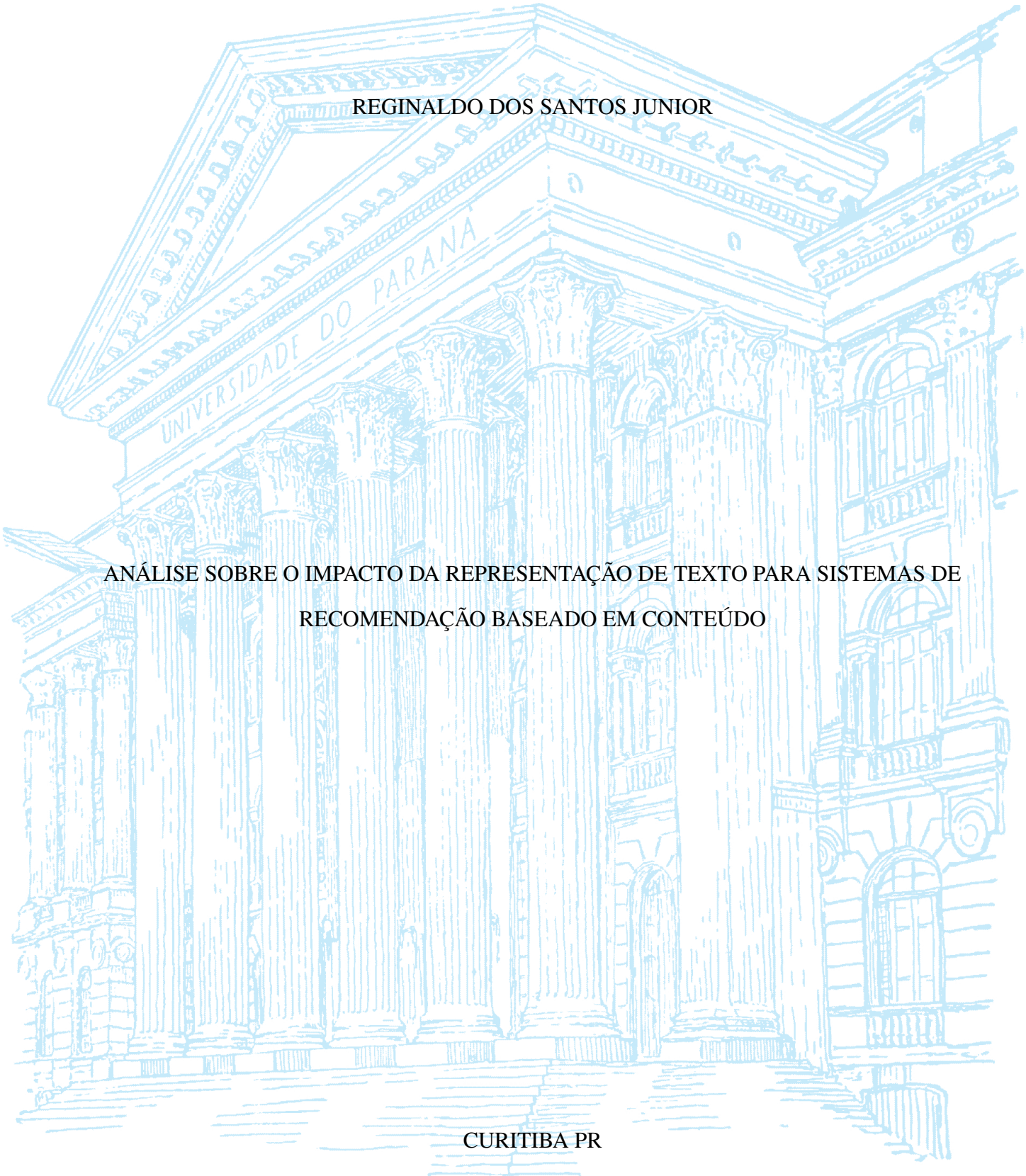
UNIVERSIDADE FEDERAL DO PARANÁ

REGINALDO DOS SANTOS JUNIOR



ANÁLISE SOBRE O IMPACTO DA REPRESENTAÇÃO DE TEXTO PARA SISTEMAS DE RECOMENDAÇÃO BASEADO EM CONTEÚDO

CURITIBA PR

2019

REGINALDO DOS SANTOS JUNIOR

ANÁLISE SOBRE O IMPACTO DA REPRESENTAÇÃO DE TEXTO PARA SISTEMAS DE RECOMENDAÇÃO BASEADO EM CONTEÚDO

Trabalho apresentado como requisito parcial à conclusão do Curso de Bacharelado em Ciência da Computação, Setor de Ciências Exatas, da Universidade Federal do Paraná.

Área de concentração: *Ciência da Computação*.

Orientador: Eduardo J. Spinosa.

CURITIBA PR

2019

"Tudo o que você sempre quis está do outro lado da linha do medo."
George Addair

AGRADECIMENTOS

Agradeço especialmente, a minha namorada Bianca Berwig Silva que sempre acreditou em mim mesmo quando eu não mais acreditava. A minha família, que me apoiou incondicionalmente e sempre me incentivou a estudar. Ao professor orientador Eduardo Spinosa, por toda paciência, atenção e orientação, sempre me incentivando a persistir. A todos os professores do programa de ciência da computação que me auxiliaram durante a trajetória dentro da universidade. Aos meus amigos que me aconselharam, motivaram e escutaram durante todos esses anos, em especial ao Rafael Gomes de Castro e Alexandre Peres Arias. Agradeço a todos que contribuíram para o desenvolvimento deste trabalho direta ou indiretamente.

RESUMO

O sistema de recomendação auxilia os usuários filtrando o conteúdo existente e construindo sugestões baseadas nos seus principais interesses. Diversos Sistemas de recomendações utilizam-se da informação textual como entrada para seus algoritmos. Este trabalho propõe uma análise quantitativa nas recomendações geradas pelos modelos dos algoritmos do paragraph vector, distributed memory e distributed bag of words, utilizando filtragem por conteúdo com a temática de filmes.

Palavras-chave: 1. Sistema de Recomendação 2. Processamento de Linguagem Natural 3. Filtragem de conteúdo

ABSTRACT

The recommendation system helps users by filtering existing content and building suggestions based on their main interests. Several recommendation systems use textual information as input for their algorithms. This work aims a quantitative analysis in the recommendations generated by the paragraph vector, distributed memory and distributed bag of words algorithm models, using content filtering with the theme of movies.

Keywords: 1. Recommender System 2. Natural Language Processing 3. content-based recommender systems

LISTA DE FIGURAS

2.1 Modelo PV-DM, Fonte: Le e Mikolov (2014) 13
2.2 Modelo PV-DBOW, Fonte: Le e Mikolov (2014) 13

LISTA DE TABELAS

2.1 Resumo dos conceitos abordados em Sistema de Recomendação - Fonte: Adaptado de Aggarwal (2016) 15

4.1 Valor do MAP para os algoritmos PV-DB e PV-DBOW para o processamento de texto 18

4.2 PD-DM - Resultados dos experimentos. 19

4.3 PD-DBOW - Resultados dos experimentos 19

LISTA DE ACRÔNIMOS

DINF	Departamento de Informática
PPGINF	Programa de Pós-Graduação em Informática
UFPR	Universidade Federal do Paraná
NLP	Natural Language Processing
PV	Paragraph Vector
PV-DBOW	Paragraph Vector - Distributed Bag of Words
PV-DM	Paragraph Vector - Distributed Memory
RS	Recommender System
SC	Similaridade do Cosseno

LISTA DE SÍMBOLOS

$\sum$	Somatório de números
$\prod$	Produtório de números

SUMÁRIO

1 INTRODUÇÃO 11
2 FUNDAMENTOS TEÓRICOS 12
2.1 PROCESSAMENTO DE LINGUAGEM NATURAL 12
2.2 PARAGRAPH VECTOR 12
2.3 SISTEMA DE RECOMENDAÇÃO 12
2.3.1 Sistema de Recomendação com Base em Conteúdo 14
2.3.2 Sistema de Recomendação Colaborativo 14
2.3.3 Sistema de Recomendação Híbrido 15
3 ABORDAGEM 16
3.1 NORMATIZAÇÃO DE TEXTO 16
3.2 RECOMENDAÇÃO 16
3.3 MÉTRICA 16
3.4 IMPLEMENTAÇÃO 17
3.5 BASE DE DADOS 17
4 EXPERIMENTOS 18
4.1 AVALIAÇÃO DA NORMATIZAÇÃO DE TEXTO 18
4.2 AVALIAÇÃO DOS MODELOS DO PARAGRAPH VECTOR 18
4.3 AVALIAÇÃO SOBRE O GÊNERO 19
5 CONCLUSÃO 20
REFERÊNCIAS 21
APÊNDICE A – LISTA DE *STOP WORDS* 23

1 INTRODUÇÃO

À medida que a quantidade de informações disponíveis para os usuários aumenta, são necessários métodos e ferramentas para auxiliarem os usuários a encontrar informações relacionadas aos seus interesses (Pazzani, 1999, p. 14). Neste sentido, Sarwar et al. (2002) corrobora afirmando que atualmente os sistemas de recomendações estão presentes em inúmeros sites, sendo uma ferramenta importante na Web.

Os sistemas de recomendações são uma ferramenta que podem contribuir para a solução do problema acima descrito (Pazzani, 1999). Em um sistema de recomendação, os usuários fornecem classificações como entrada, e o sistema agrega e direciona a outros usuários as novas sugestões (Resnick e Varian, 1997, p. 56).

Dentre os métodos de filtragem de dados do sistema de recomendação, destaca-se o baseado em conteúdo. De acordo com Pazzani (1999), o sistema de recomendações baseado em conteúdo analisa a descrição dos itens que foram classificados pelo usuário e a descrição dos itens a serem recomendados. Existe uma variedade de algoritmos que analisam o conteúdo dos textos para encontrar regularidades no conteúdo, que servem como base para fazer recomendações (Pazzani, 1999). Destacam-se os algoritmos Paragraph Vector - Distributed Memory (PV-DM) e Paragraph Vector - Distributed Bag of Words (PV-DBOW). Desse modo, o presente trabalho propõe analisar o impacto que diferentes formas de representação causam nos sistemas de recomendação baseados em conteúdo para filmes.

A monografia está estruturada em capítulos, que reúnem diversos temas complementares e o objetivo do trabalho. No segundo capítulo são abordados os principais conceitos relacionados à pesquisa, como processamento de linguagem natural, sistema de recomendações e paragraph vector. No terceiro capítulo é apresentada a abordagem propostas pela pesquisa, a base de dados, o processamento de texto e as métricas. Na sequência, no quarto capítulo, são apresentados os resultados encontrados pelos experimentos. No último capítulo são descritas as principais considerações e conclusões da presente pesquisa.

2 FUNDAMENTOS TEÓRICOS

Neste trabalho serão abordados os principais conceitos relacionados ao objetivo da pesquisa com o intuito de embasar os experimentos e o trabalho proposto. Primeiramente será discutido o conceito de processamento de linguagem natural. Na sequência, serão analisados os conceitos de paragraph vector e sistema de recomendações.

2.1 PROCESSAMENTO DE LINGUAGEM NATURAL

O conceito de Processamento de Linguagem Natural surgiu na década de 1950 com a inserção da inteligência artificial na linguística (Nadkarni et al., 2011). Compreende-se como Processamento de Linguagem Natural (NLP) as diversas técnicas e métodos utilizados para analisar textos em um ou mais níveis de análise linguística (Liddy, 2001). Este campo de estudo da ciência da computação tem como objetivo adotar técnicas para aprender, compreender e produzir conteúdo de linguagem humana (Hirschberg e Manning, 2015).

Hirschberg e Manning (2015) apontam algumas das principais finalidades do processamento de linguagem natural como auxiliar na comunicação humana-humana, tradução automática e contribuir na comunicação humano-máquina. Liddy (2001) afirma que o processamento de linguagem possui dois objetivos diferentes, sendo um o processamento de linguagem e o outro a geração de linguagem.

Outro ponto destacado por Liddy (2001) são os níveis de linguagem, sendo os principais os seguintes: fonológico, morfológico, lexical, sintático, semântico, discursivo e pragmático. Neste trabalho, se torna relevante comentar brevemente sobre a semântica. Segundo Liddy (2001), o nível de linguagem da semântica tem como principal objetivo verificar os possíveis significados de uma frase.

2.2 PARAGRAPH VECTOR

O *paragraph vector* (PV) proposto por Le e Mikolov (2014), é um algoritmo não supervisionado para representar textos de tamanhos diferentes em um vetor numérico, de tamanho fixo, através de dois modelos: o *Paragraph Vector - Distributed Memory* (PV-DM), e o modelo *Paragraph Vector - Distributed Bag of Words* (PV-DBOW).

No modelo PV-DM os vetores das palavras W são concatenados com o vetor do documento D, para prever a próxima palavra, mostrado na figura 2.1. Portanto, ao treinar os vetores W, o vetor D também é treinado. Ao final do processo o vetor D é o vetor de características do documento.

Para o modelo PD-DBOW é feito o processo inverso, isto é, os vetores W é utilizado como treino para prever palavras que o seu documento contenha, através de uma rede neural, figura 2.2. O vetor de características será gerado do peso dos neurônios da rede neural.

2.3 SISTEMA DE RECOMENDAÇÃO

O sistema de recomendação começou a se tornar um tópico amplamente pesquisado a partir da década de 1990, conjuntamente com o surgimento da Web e do E-commerce (Aggarwal, 2016). Compreende-se como sistema de recomendação, o instrumento utilizado em softwares

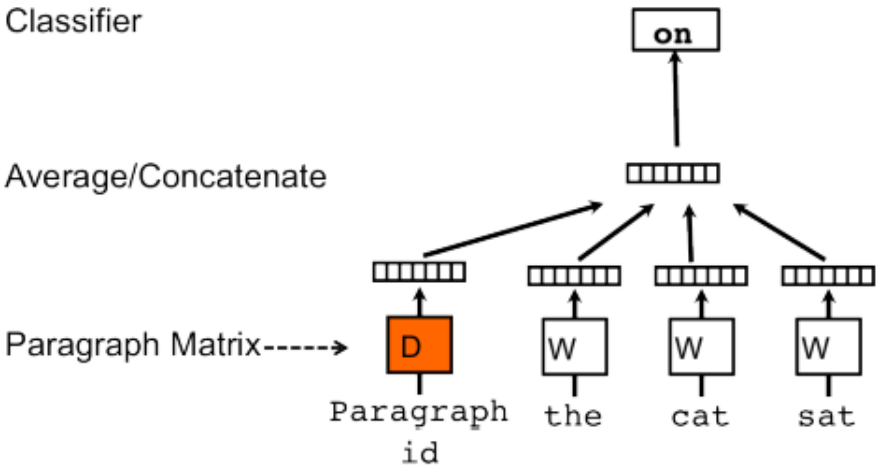


Figura 2.1: Modelo PV-DM, Fonte: Le e Mikolov (2014)

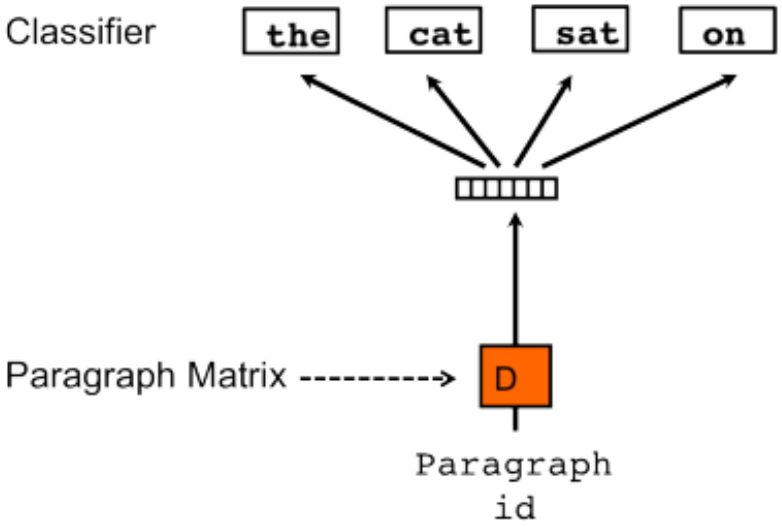


Figura 2.2: Modelo PV-DBOW, Fonte: Le e Mikolov (2014)

que tem como objetivo fornecer sugestões ao usuário de forma eficaz e eficiente (Ricci et al., 2011).

Aggarwal (2016) complementa ao afirmar que o sistema de recomendação é muito diverso, permitindo com que os dados de preferência dos usuários sejam analisados para fazer as recomendações. As sugestões feitas pelo sistema auxiliam os usuários no processo de tomada de decisão, já que esta ferramenta filtra os dados e possibilita aos indivíduos lidar com o excesso de informação (Ricci et al., 2011).

Sarwar et al. (2002) ressalta que os sistemas de recomendação são implantados em diversos sites atualmente, atendendo inumeros usuários. Os autores afirmam que é considerada uma ferramenta crucial na Web e no E-commerce. Neste sentido, o sistema de recomendação é visto como umas das ferramentas mais importantes no e-commerce e umas das técnicas mais adotadas nos últimos tempos em diversos sites de empresas como Amazon.com, YouTube, Netflix, Yahoo, Tripadvisor, Last.fm e IMDb (Ricci et al., 2011).

Ricci et al. (2011) comentam que essa tecnologia é frequentemente adotada pelos prestadores de serviços para aumentar o número de itens vendidos, vender uma maior variedade

de itens, aumentar a satisfação e a fidelidade dos usuários e compreender suas necessidades. Sarwar et al. (2002) corroboram esta noção ao salientarem que o sistema de recomendação agrega valor para um negócio a partir da base de dados de seus clientes. De acordo com os autores essa ferramenta ajuda os clientes a encontrarem os produtos que querem comprar, ao mesmo tempo que ajuda o negócio gerando mais vendas.

É importante destacar a necessidade de abordar e desenvolver o sistema de recomendação a partir de um olhar interdisciplinar, no qual são envolvidas diversas áreas como inteligência artificial, interação humana com computadores, tecnologia da informação, mineração de dados, estatística, interfaces adaptativas de usuário, sistemas de apoio à decisão, marketing ou comportamento do consumidor (Ricci et al. (2011)).

Ricci et al. (2011) afirmam que o sistema de recomendações levanta diversos dados para construir suas sugestões aos usuários. Os autores salientam que a fonte de dados e o conhecimento disponível podem ser analisados por meio de diversas técnicas de recomendação. Neste mesmo sentido, Aggarwal (2016) afirma que os principais sistemas de recomendações possuem os seguintes métodos de filtragem dos dados: colaborativa, baseada em conteúdo ou híbrida. Por fim, Sarwar et al. (2002) destaca que é preciso desenvolver novas tecnologias que possam melhorar a escalabilidade dos sistemas de recomendação. A seguir serão apresentados e descritos os principais sistemas de recomendação.

2.3.1 Sistema de Recomendação com Base em Conteúdo

O sistema de recomendação baseado no método de filtragem por conteúdo tem como objetivo principal recomendar itens semelhantes aos que o usuário gostou no passado (Ricci et al., 2011). Os autores afirmam que as similaridades dos itens analisados são calculadas a partir das características associadas a esses itens, ou seja, o sistema analisa um conjunto de descrições de itens previamente classificados por um usuário. Dessa forma, o sistema constrói um perfil de interesses novos do usuário com base nas características dos objetos classificados por esse usuário no passado (Ricci et al., 2011).

Aggarwal (2016) afirma que nesse método os atributos descritivos dos itens são usados para fazer recomendações, ou seja, as classificações e o perfil do usuário são cruzados com as informações de conteúdo disponíveis nos descritivos dos itens. Neste sentido, Ricci et al. (2011) comentam que o processo de recomendação compara os atributos do perfil do usuário com os atributos de um determinado item.

2.3.2 Sistema de Recomendação Colaborativo

A filtragem colaborativa é considerada a técnica mais popular e amplamente adotada em diversos sistemas de recomendação (Ricci et al., 2011, p. 11). Aggarwal (2016) descreve que as classificações fornecidas pelos usuários são utilizadas pela filtragem colaborativa para encontrar a similaridade entre si. Ricci et al. (2011) corrobora esta ideia ao afirmar que essa ferramenta recomenda ao usuário itens que outros usuários com gostos parecidos tenham classificado de forma positiva no passado.

De acordo com os autores (2011), a semelhança entre os usuários é calculada com base no histórico de suas classificação. Neste método de filtragem é fundamental tanto os dados dos itens quanto os dados de mais de um usuário, sendo que as principais abordagens são feitas a partir dos seguintes métodos: a abordagem de vizinhos mais próximos e os modelos de fator latente (Ricci et al., 2011, p. 146).

2.3.3 Sistema de Recomendação Híbrido

Os sistemas de recomendação híbridos são combinações de dois ou mais sistemas de recomendações, com o objetivo de melhorar o desempenho da recomendação (Burke, 2007). Ricci et al. (2011) descreve a recomendação híbrida como a combinação de fontes de dados. Dessa forma, Aggarwal (2016) afirma que os sistemas híbridos de recomendação estão diretamente ligados à análise de conjuntos, tornando o sistema mais preciso, isto é, gerar recomendações com maior relevância para o usuário.

Por fim, a partir da análise dos três principais métodos de filtragem de dados, apresenta-se a seguir um quadro comparativo no qual são ilustradas as principais características de cada um.

Abordagem	**Conceito**	**Entrada**
Colaborativa	Fornece recomendações baseadas numa abordagem colaborativa. As avaliações consideram ações e classificações dos meus pares e minhas próprias	Classificação do usuário + classificação de uma determinada comunidade
Baseada em Conteúdo	As recomendações são baseadas no conteúdo (atributos). As sugestões são baseadas no que eu classifiquei em minhas avaliações e ações passadas.	Classificação do usuário + os atributos do item (conteúdo)
Híbrida	As recomendações são baseadas na combinação de mais de um método de filtragem	A entrada dependerá dos métodos adotados para a análise dos dados dos usuários

Tabela 2.1: Resumo dos conceitos abordados em Sistema de Recomendação - Fonte: Adaptado de Aggarwal (2016)

A partir dessa discussão, pode-se ter uma visão geral sobre a temática de sistema de recomendação. A seguir serão apresentados os métodos de abordagem utilizados no presente trabalho como também os resultados obtidos por meio dos experimentos.

3 ABORDAGEM

Para analisar o impacto do processamento de texto dentro do sistema de recomendações, utilizou-se a temática de filmes. Foi abordado a filtragem por conteúdo, isto é, as recomendações seriam feitas com base apenas nos itens com os quais o usuário interagiu. Avaliamos as abordagens das recomendações com base da métrica aqui descrita.

3.1 NORMATIZAÇÃO DE TEXTO

Neste tópico serão descritos os critérios utilizados para o processamento dos dados textuais. É importante destacar que essa etapa transforma a linguagem humana textual em um formato de texto compreensível à máquina, ou seja, é realizada uma normatização dos textos (Davydova, 2018).

A seguir, serão apresentadas as etapas realizadas:

1. Conversão de todas as letras em minúscula.
2. Remoção de números.
3. Remoção de pontuação e acentos.
4. Remoção de espaços em branco.
5. Remoção de palavras irrelevantes (stop words), Lista no Apêndice A.
6. Lematização - analisar o campo lexical e mapear as palavras com o mesmo radical Davydova (2018) e Koenig (2019).

3.2 RECOMENDAÇÃO

Para a recomendação, utilizou-se dos vetores de características, extraídos do PV-DM e PV-DBOW (Seção 3.2). Através da média dos vetores de treinamento positivos, calculou-se a similaridade do cosseno (SC) entre vetores que representavam cada filme.

A SC é obtida através do cálculo do cosseno entre os ângulos dos vetores que desejamos comparar, utilizando-se a fórmula do produto interno entre dois vetores (Equação 3.1). O resultado é um valor entre -1 e 1. Quanto mais perto de 1, maior é a similaridade.

$$cos(a, b) = \frac{< a, b >}{||a|| \times ||b||} \tag{3.1}$$

As recomendações foram feitas com base nos 1000 filmes mais similares em ordem crescente da SC.

3.3 MÉTRICA

Adotou-se o *mean average precision* (MAP) para avaliação deste trabalho. MAP é um método de classificação que leva em conta a o ranking da recomendação (Aggarwal, 2016). Essa métrica não penaliza pelo aumento do número de recomendações e tem valor diretamente

proporcional ao ranking. Dessa forma, podemos calcular o ganho das recomendações de acordo com os parâmetros de interesses.

Primeiramente, precisamos calcular a $P(k)$, da equação 3.2, onde k é o número de recomendações e T são as recomendações relevantes, logo $|k \cap T|$ é o total de recomendações corretas.

$$P(k) = k \cap T/k \tag{3.2}$$

Em seguida, calcula-se a *average precision* (AP), através da Equação 3.3, onde rel(k) é 0 quando o k-ésimo item recomendado não seja relevante ou 1 caso contrário, e m é o número de recomendações corretas.

$$AP@N = \frac{1}{m}\sum_{k=1}^{N}(P(k) \text{ if} k^{th} \text{item was relevant}) = \frac{1}{m}\sum_{k=1}^{N} P(k) * rel(k) \tag{3.3}$$

Por fim, calculamos o MAP (Equação 3.4), que é a média do AP feita para cada usuário(U) (Hopfgartner, 2010).

$$MAP@N = \frac{1}{U}\sum_{u=1}^{|} U|(AP@N)u \tag{3.4}$$

3.4 IMPLEMENTAÇÃO

Este trabalho foi implementado na linguagem de programação Python, utilizando-se da biblioteca Gensim (Řehůřek, 1990), onde eles possuem uma versão implementada dos algoritmos PV-DBOW e PV-DM. Outra biblioteca utilizada foi a NLTK (Project, 1990), para a normatização de texto.

3.5 BASE DE DADOS

A base de dados adotada para o desenvolvimento do presente trabalho foi a MovieLens. Harper e Konstan (2015) afirmam que o MovieLens foi difundido a partir do final da década de 90, permitindo que os usuários tivessem acesso aos dados que descreviam a preferência das pessoas por determinados filmes. Harper e Konstan (2015) comentam que a preferência era expressa por dados do usuário, o filme, a classificação e a data e o horário, sendo que os indivíduos classificam os filmes entre 0-5 estrelas.

De acordo com o site MovieLens (GroupLens, 2019), a base de dados possui mais de 20 milhões de classificações, 465.000 aplicações de tags, 27.000 filmes e 138.000 usuários. Para o desenvolvimento do experimento, foi adotada a base de dados "*latest small dataset*"com 100.000 classificações, 3.600 aplicações de tags, 9.000 filmes e 600 usuários (GroupLens, 2019).

Definiu-se como positiva as avaliações que receberam notas acima de 3.0, ou seja, os filmes que o usuário interagiu e gostou. Foram escolhidos os usuários que possuiam 50 ou mais avaliações após esta filtragem. Separam-se os as avaliações em treinamento e teste em proporção de 80% e 20% respectivamente.

Na sequência, foi utilizado o site Imdb (1990) por meio da biblioteca Python ImdbPie para extrair os seguintes dados: sinopse, gênero e crítica (O'Dwyer, 1990).

4 EXPERIMENTOS

Primeiramente analisaram-se os impactos do processamento de texto (Seçao 3.1) para os algoritmos PV-DBOW e PV-DM. Aplicaram-se as etapas da normatização 1, 2, 3 e 4, apresentadas na seção 3.1, em todos os testes, visto que são necessários para os modelos do PV. Após isso foram feitos testes da etapa 5, remoção das stop words, e da etapa 6, lematização.

Em seguida, comparam-se os dois algoritmos variando a janela como os valores 2, 4, 6, 8 e 10; e variando o tamanho do vetor para os valores 25, 50, 100 e 200.

Por fim, utilizando-se dos melhores resultados dos testes anteriores, aplicou-se um peso sobre os gêneros, isto é, foram replicados os textos de gênero dos filmes. Os resultados foram avaliados comparando a métrica do mean average precision, descrita na Seção 3.3.

A seguir, serão apresentados os resultados de cada experimento.

4.1 AVALIAÇÃO DA NORMATIZAÇÃO DE TEXTO

Aplicaram-se diferentes etapas da normatização de texto. Os resultados são apresentados na tabela 4.1. Na coluna "Base", foram aplicadas as etapas da normatização 1,2, 3 e 4. Já na coluna "Lematização", acrescentou-se a etapa 6. E, por fim, a coluna "*Stop Words*", acrescentou-se a etapa 5 em relação a coluna a base. Os valores são as métricas do MAP (Seção 3.3) e os maiores valores estão destacado em negrito.

Algoritmo	**Base**	**Lematizado**	**Stop Words**
PV-DM	0,00185	**0,00188**	0,00180
PV-DBOW	**0,00177**	0,00176	0,00156

Tabela 4.1: Valor do MAP para os algoritmos PV-DB e PV-DBOW para o processamento de texto

Pode-se observar um aumento de 1,39% no resultado do algoritmo PV-DM aplicando a lematização em relação a coluna "Base". Este resultado positivo se deve por gerar vetores de características parecidas entre si, uma vez que palavras como amei, amou e ame, tornam se uma única palavra, amor.

Por outro lado, podemos ver uma piora removendo as *stop words* em relação aos outros experimentos. Tal fato ocorre por remover palavras que agregam significado ao texto. Ao remover determinadas palavras do algoritmo, o texto perde ordem e relações, afetando seu significado (Liddy, 2001).

4.2 AVALIAÇÃO DOS MODELOS DO PARAGRAPH VECTOR

Variaram-se os paramêtros tamanho da janela (J) e o tamanho do vetor de características (V) que são parâmetros do PV, com o intuito de encontrar as melhores representações.

A Tabela 4.2 apresenta os resultados do modelo PV-DM. O melhor resultado está na janela 8 e tamanho do vetor de características 25.

	2	4	6	8	10
25	0,00188	0,00179	0,00183	0,00192	0,00188
50	0,00106	0,00104	0,00099	0,00098	0,00096
100	0,00061	0,00058	0,00060	0,00060	0,00061
200	0,00046	0,00047	0,00052	0,00050	0,00046

Tabela 4.2: PD-DM - Resultados dos experimentos

A Tabela 4.3, contém os resultados do modelo PV-DBOW. O melhor resultado se encontra na janela 2 e tamanho do vetor de características 25.

	2	4	6	8	10
25	0,00172	0,00145	0,00124	0,00116	0,00098
50	0,00096	0,00087	0,00084	0,00070	0,00070
100	0,0006	0,00052	0,00053	0,00050	0,00049
200	0,0005	0,00046	0,00043	0,00042	0,00042

Tabela 4.3: PD-DBOW - Resultados dos experimentos

Comparando o resultado das duas tabelas podemos observar que o algoritmo PV-DM se sobressai ao PV-DBOW para os mesmos parâmetros. O melhor resultado do PV-DM tem uma desempenho de 11,8% maior em relação ao melhor resultado do PV-DBOW. Este resultado já era esperado, uma vez que o PV-DM preserva a ordem das palavras (Le e Mikolov, 2014), gerando assim uma representação mais específica do texto.

4.3 AVALIAÇÃO SOBRE O GÊNERO

A partir dos testes anteriores, definimos para este experimento o modelo PV-DM com parâmetros do tamanho da janela igual a 8 e vetor com tamanho 25. O resultado do MAP foi de 0.00532, um aumento de 173% em relação ao melhor desempenho do experimento anterior. Uma das prováveis causas, são filmes como “Very Potter Musical” que é uma paródia dos filmes do Harry Potter, e tem características semelhantes dentro das sinopse, porém gêneros completamente opostos. Ou seja, os vetores de características ficam mais distantes entre si, segundo a similaridade do cosseno(Seção 3.2), melhorando o ranking das recomendações positivas. Outro possível motivo, é a baixa informação textual extraída da base de dados IMDB.

5 CONCLUSÃO

Este trabalho analisou o impacto da representação textual para o sistema de recomendações baseado em conteúdo. Foi abordada a temática de filmes para o experimento, extraídos das bases de dados MovieLens (GroupLens, 2019) e Imdb (1990), os filmes e as informações (sinopse, gênero e crítica) respectivamente. Aplicamos os dois modelos do paragraph vector, PV-DM e PV-DBOW e utilizou-se a métrica *mean average precision* para avaliar os resultados.

Analisamos as diferentes formas de processamento de texto, bem como o impacto nos algoritmos de PV-DM e PV-DBOW. Alteramos alguns parâmetros destes algoritmos, como a janela e o tamanho do vetor, em busca do melhor resultado, utilizando a métrica MAP. Na sequência adicionou-se peso nos gêneros para o modelo que teve melhor desempenho.

A partir dos resultados, foi constatado que ao removermos as *stop words* houve redução no valor do MAP, enquanto que aplicando a lematização obteve-se melhores resultados. O algoritmo PV-DM se sobressaiu em todos os testes em relação ao PV-DBOW para o tema de filmes. O resultado mais expressivo ocorreu ao utilizar peso nos gêneros, onde foi observado uma melhora em relação aos resultados anteriores deste trabalho.

Durante o desenvolvimento da pesquisa foram verificadas algumas limitações, uma delas foi a falta de informação textual extraída da fonte do Imdb (1990). Outra dificuldade encontrada foi o baixo número de ratings, que gerou uma amostra de usuários pequena para o experimento.

Para trabalhos futuros, uma sugestão seria a inclusão dos campos de artistas, escritores e diretores, uma vez que eles desempenham um grande papel dentro desse ramo. Caberia também uma análise sobre outras abordagens de classificação.

REFERÊNCIAS

Aggarwal, C. C. (2016). *Recommender Systems: The Textbook*. Springer International Publishing.

Burke, R. (2007). *Hybrid Web Recommender Systems*, página 377–408. Springer Berlin / Heidelberg.

Davydova, O. (2018). Text preprocessing in python: Steps, tools, and examples. https://medium.com/@datamonsters/text-preprocessing-in-python-steps-tools-and-examples-bf025f872908. Acessado em 01/12/2019.

GroupLens (2019). Movielens. https://grouplens.org/datasets/movielens. Acessado em 01/12/2019.

Harper, F. M. e Konstan, J. A. (2015). The movielens datasets: History and context. *ACMTrans. Interact. Intell. Syst.*, V, N, Article XXXX (2015):20 pages.

Hirschberg, J. e Manning, C. D. (2015). Advances in natural language processing. *American Association for the Advancement of Science*, 349(6245):261–266.

Hopfgartner, F. (2010). *Personalised Video Retrieval:Application of Implicit Feedback and Semantic User Profiles*. Tese de doutorado, University of Glasgow.

Imdb (1990). Imdb - movies, tv and celebrities. https://www.imdb.com/. Acessado em 01/12/2019.

Koenig, R. (2019). Nlp for beginners: Cleaning & preprocessing text data. https://towardsdatascience.com/nlp-for-beginners-cleaning-preprocessing-text-data-ae8e306bef0f. Acessado em 01/12/2019.

Le, Q. e Mikolov, T. (2014). Distributed representations of sentences and documents. Em *31st International Conference on Machine Learning*, Pequim, China.

Liddy, E. D. (2001). *Natural language processing, 2nd edn*. Encyclopedia of Library and Information Science, Marcel Decker.

Nadkarni, P. M., Ohno-Machado, L. e Chapman, W. W. (2011). Natural language processing: an introduction. *Journal of the American Medical Informatics Association : JAMIA*, 18 5:544–51.

O'Dwyer (1990). Imdbpie. https://www.imdb.com/. Acessado em 01/12/2019.

Pazzani, M. (1999). A framework for collaborative, content-based and demographic filtering. *Artificial Intelligence Review*, 13(6):393–408.

Project, N. (1990). Natural language toolkit. https://www.nltk.org/. Acessado em 01/12/2019.

Řehůřek, R. (1990). Gensim. https://radimrehurek.com/gensim/index.html. Acessado em 01/12/2019.

Resnick, P. e Varian, H. (1997). Recommender systems. *ACM*, 40(3):56–58.

Ricci, F., Rokach, L., Shapira, B. e Kantor, P. B. (2011). *Recommender Systems Handbook*. Springer Science+Business Media.

Sarwar, B. M., Karypis, G., Konstan, J. e Riedl, J. (2002). Recommender systems for large-scale e-commerce: Scalable neighborhood formation using clustering. Em *5^{th}th International Conference on Computer and Information Technology (ICCIT)*, Dhaka - Bangladesh.

APÊNDICE A – LISTA DE *STOP WORDS*

whenever, full, ourselves, sixty, whose, found, third, always, seem, him, this, per, we, moreover, couldnt, therefore, anywhere, which, hers, without, four, amoungst, across, ten, thru, over, amongst, describe, am, side, out, through, every, hence, has, amount, yet, what, were, yourself, he, besides, meanwhile, was, thereupon, please, would, upon, ever, six, nowhere, too, become, within, here, made, because, three, somewhere, much, used, whoever, due, anyhow, your, onto, thick, nor, bottom, her, whereas, thereby, don, anything, between, about, where, one, their, otherwise, no, around, namely, sometimes, off, mill, co, also, part, km, go, before, get, who, some, still, by, since, thence, whither, during, among, empty, more, at, take, if, less, anyway, eg, are, became, these, give, least, been, an, move, us, under, else, with, see, its, however, con, almost, well, beforehand, sometime, interest, they, afterwards, nine, hasnt, everyone, enough, those, and, rather, put, make, top, than, system, either, doing, etc, cant, until, of, have, another, that, the, none, therein, thereafter, even, latterly, she, against, as, whom, you, next, others, two, although, myself, anyone, will, hereupon, whereupon, other, is, several, himself, fire, fill, had, might, could, everywhere, i, someone, a, toward, last, being, ours, does, both, whereby, cannot, re, twelve, fifteen, thin, together, many, doesn, again, latter, whole, yours, ltd, all, such, do, not, any, my, seemed, mostly, former, so, further, herein, ie, find, using, hereafter, throughout, indeed, somehow, on, whatever, except, already, in, though, own, when, but, back, beside, elsewhere, serious, front, our, once, herself, whereafter, detail, nothing, seeming, various, formerly, from, name, bill, wherever, just, must, it, wherein, did, themselves, most, hundred, via, done, hereby, whether, keep, be, never, whence, while, five, perhaps, his, eight, first, becoming, same, then, forty, mine, everything, how, inc, very, didn, each, up, for, now, computer, above, itself, behind, few, to, becomes, towards, de, often, neither, seems, nobody, them, twenty, eleven, something, really, along, unless, down, noone, only, kg, sincere, should, fifty, beyond, un, quite, there, yourselves, may, or, show, thus, me, into, can, below, after, why, nevertheless, say, call, regarding, cry, alone.