ANALISE E PERFILAMENTO EM DADOS DA AMANZON

Antonio Carlos Farias Ferreira
Equipe *sin(args)^0*

TEMA

Descobrimento de temas latentes no consumo de produtos da Amazon para inferência de perfis de consumo dos usuários

OBJETIVO

Explorar relações semânticas envolvidas no conteúdo dos produtos da Amazon e os seus respectivos consumos através da construção de perfis dos consumidores.

QUESTÕES DE PESQUISA E HIPÓTESES

QUE TÓPICOS EMERGEM DA DESCRIÇÃO DOS PRODUTOS?

1 Aspectos técnicos da mídia emergem como tópicos

2 Gêneros de conteúdo emergem como tópicos

TÓPICOS SÃO CAPAZES DE DESCREVER RELAÇÕES COM A COMPRA DE OUTROS PRODUTOS?

1 Os tópicos extraídos podem ser usados para predizer consumo de outros produtos

2 Os tópicos extraídos pode ser usados para predizer o consumo de outros tópicos

TÓPICOS PODEM SE RELACIONAR COM AS CATEGORIAS DO PRODUTO E REFLETIR ASPECTOS DE SUA AVALIAÇÃO PELOS CONSUMIDORES?

1 Usuários pertencentes a um determinado grupo de perfis atribuem notas menores na avaliação de categorias pouco consumidas pelo seu grupo

TRABALHOS RELACIONADOS

Alekseev e Nikolenko (2017)

- Propõem aprimorar o perfil de usuários em redes sociais via análise de texto e classificação ativa de itens utilizando tópicos. Utilizam Word2Vec para modelagem de tópicos e regressão logística para gerar vetores de perfil usados em sistemas de recomendação.

Guan, Li e Guan (2017):

- Apresentam o Enhanced SVD (ESVD), um modelo de fatoração matricial com aprendizagem ativa, que lida com a baixa densidade de interações usuário-item em sistemas de recomendação.

Shah, Shah e Banerjee (2017):

- Aplicam fatoração matricial em recomendações de cursos universitários, incorporando similaridade entre usuários para melhorar a previsão de preferências.

DADOS

Foi utilizado o Amazon Reviews Dataset (Ni, Li & McAuley, 2019) - categoria Movies & TV, que cotem 3,4M de avaliações, 203766 produtos e 156713 usuário.

PRODUTOS

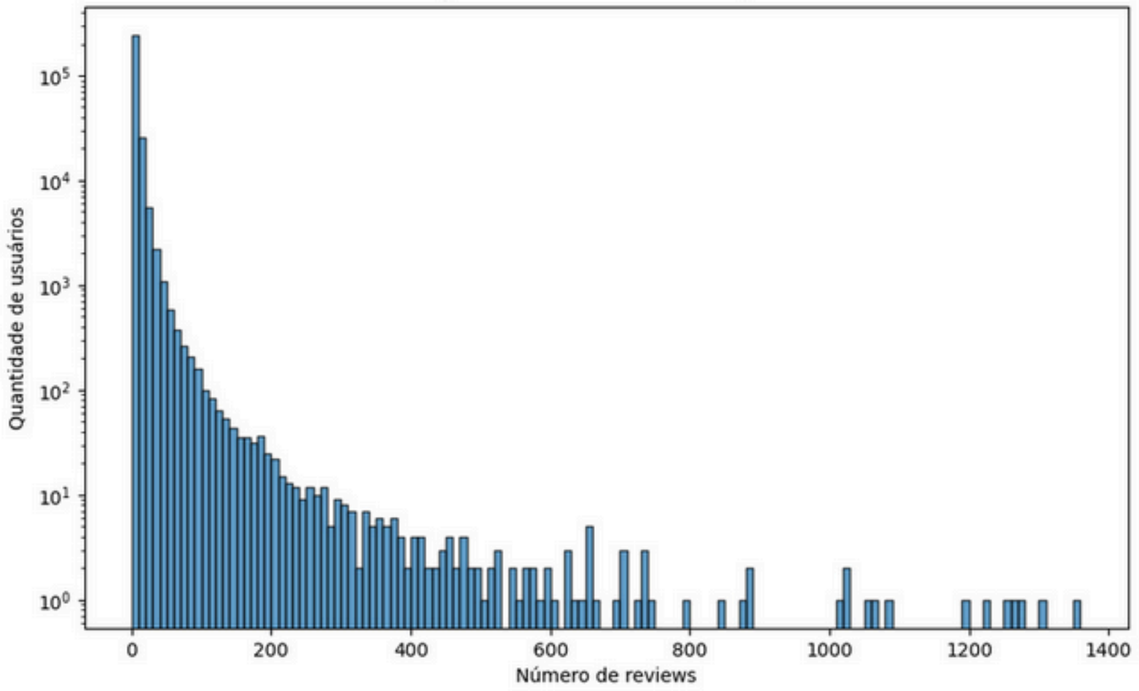
USUÁRIOS

AVALIAÇÕES

CORTE E LIMPEZA

Após remoção de produtos sem descrição, foi selecionado um conjunto de usuário com um número denso de avaliações evitando outliers. Com este primeiro corte, foram mantidos 2278 usuários com um número de avaliações entre 50 e 400 e 24725 produtos.

Gráfico 1. Distribuição de número de avaliações por usuário

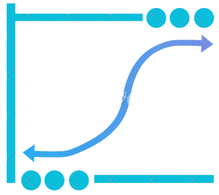


MODELOS

Vetores das probabilidades de cada documento pertencer a cada tópico foram utilizados como embeddings de produtos para produzir embeddings de perfil dos usuários. Os perfis produzidos foram usados para clusterizar os usuários e os tópicos encontrados também foram usados em um modelos de fatoração matricial para recomendar outros tópicos.

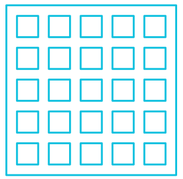
BERTOPIC

Para modelagem de tópicos



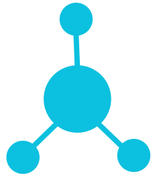
REGREÇÃO LOGISTICA

Para a construção de perfis baseados em tópicos



FATORAÇÃO MATRICIAL

Para a recomendação de tópicos



CLUSTERING

Para o agrupamento de usuários

MODELAGEM DE TÓPICOS

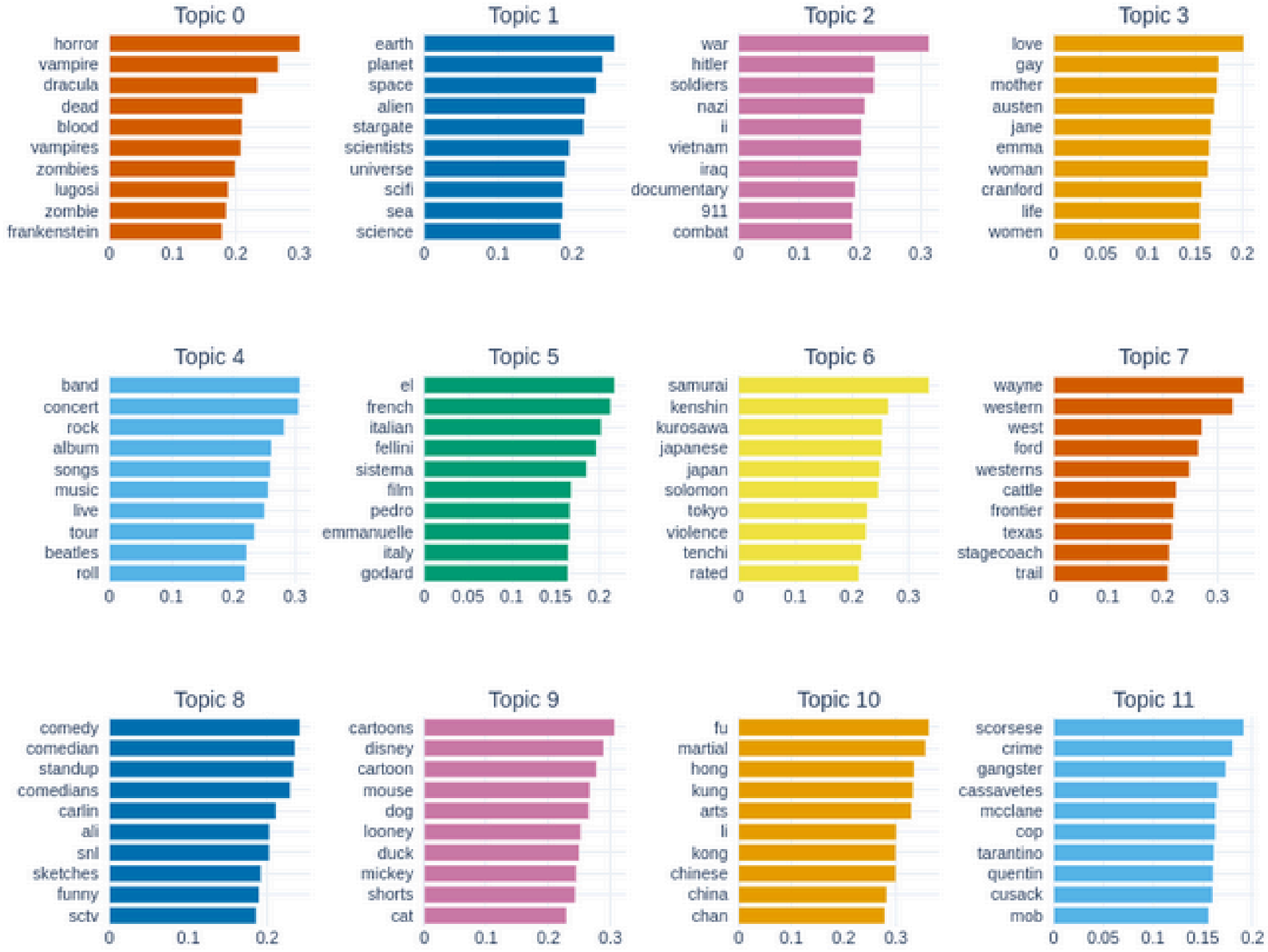
- 20 palavras por tópico
- Tamanho mínimo dos tópicos igual a 50
- Não eliminação de stopwords e uso de ClassTFIDFTransformer para diminuir o peso de palavras frequentes

54 TÓPICOS ENCONTRADOS

12553 OUTLIERS

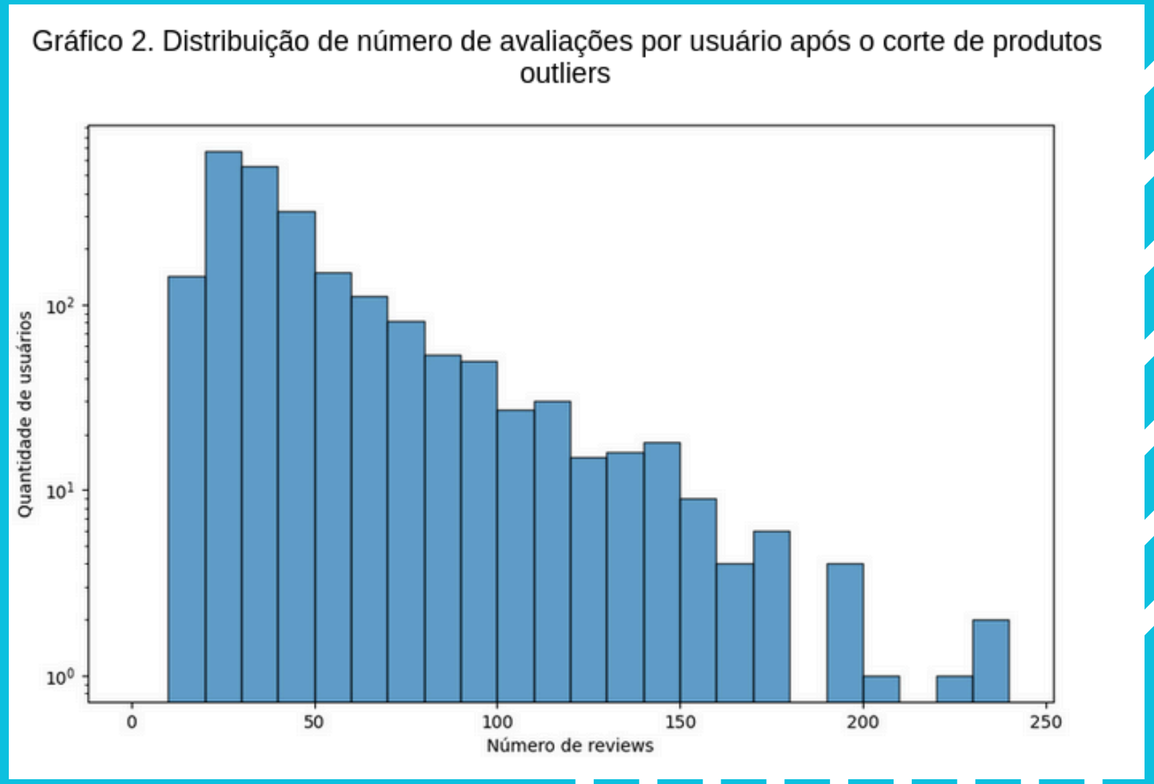
TÓPICOS MUITO COERENTES E INTERPRETAVEIS

Imagem 1. “Topic Word Scores”



Com o resultado da modelagem de tópicos e o cortes de outliers apontados por ela, foi realizado um novo corte no conjunto de usuários - mantendo aqueles com número de avaliações entre 20 e 180.

Gráfico 2. Distribuição de número de avaliações por usuário após o corte de produtos outliers



PERFILAMENTO

A ideia é que ao aprender a classificar produtos como consumidos ou não consumidos pelo usuário, a regressão logística aprenda pesos para cada tópico consumido.

Esse vetor de pesos está no mesmo espaço vetorial de produtos e é um perfil de usuário que foi usado para recomendar outros produtos através de um rankeamento de similaridade cosseno.

PARA CADA USUÁRIO

Entrada: embeddings dos produtos e rótulo de consumo

Dados: 70% dos produtos consumidos e 5x de produtos aleatórios não consumidos

AVALIAÇÃO DO PERFILAMENTO

Fórmula 1

$$\mathrm{DCG}_p = \sum_{i=1}^{p} \frac{\mathrm{liked}_i}{\log_2(i+1)},$$

Fórmula 2

$$nDCG = \frac{DCG}{IDCG}$$

Tabela 1. nDCG@K para os K elementos do topo do ranking.

Modelo	nDCG@1	nDCG@5	nDCG@10
Baseado em regressão logística	0.8472	0.8167	0.7877
Baseado em média	0.5663	0.5703	0.5660

RECOMENDAÇÃO DE TÍPICOS

Foram utilizados três modelos de fatoração matricial para reconstruir uma matriz de contagem de consumo usuários-tópicos, em que foram ocultados aleatoriamente uma 70% dos valores para cada usuário. Dessa forma, é reproduzido um cenário em que o usuário consumiu produtos de alguns tópicos e pretende-se predizer - ou recomendar - o consumo dos tópicos com os quais o usuário não interagiu.

- **ALTERNATING LEAST SQUARES (ALS)**
- **SINGULAR VALUE DECOMPOSITION (SVD)**
- **NON-NEGATIVE MATRIX FACTORIZATION (NMF)**

Tabela 2. NMAE dos modelos

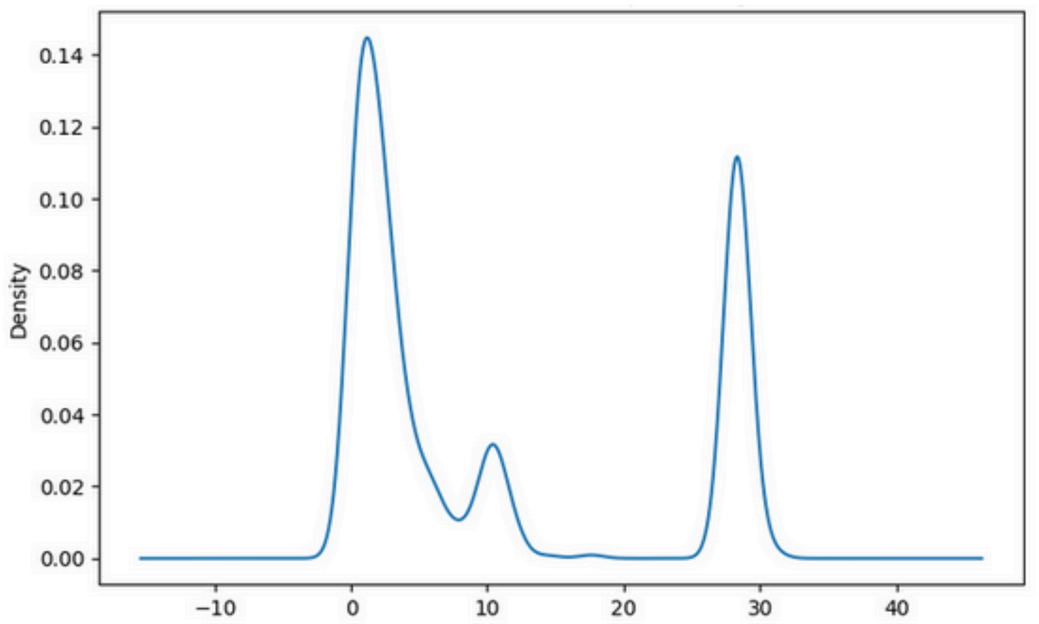
Modelo	NMAE
SVD	0.0085
NMF	0.0087
ALS	0.0081

AGRUPAMENTO DOS PERFIS

O clustering foi feito utilizando o algoritmo K-means e foi definido um número arbitrário de 20 clusters. Após o agrupamento dos usuários foi feita a contagem de consumo de cada categoria de produto dentro dos clusters e considerou-se como categorias mais consumidas aquelas que representam uma parcela maior de 20% das categorias consumidas pelo clusters. Como categorias menos consumidas, considerou-se aquelas que representam uma parcela inferior a 10%.

A média de avaliação das categorias foi de 4.057 para as mais consumidas e 4.061 para as menos consumidas.

Gráfico 2. Densidade de consumo por categoria em cada cluster

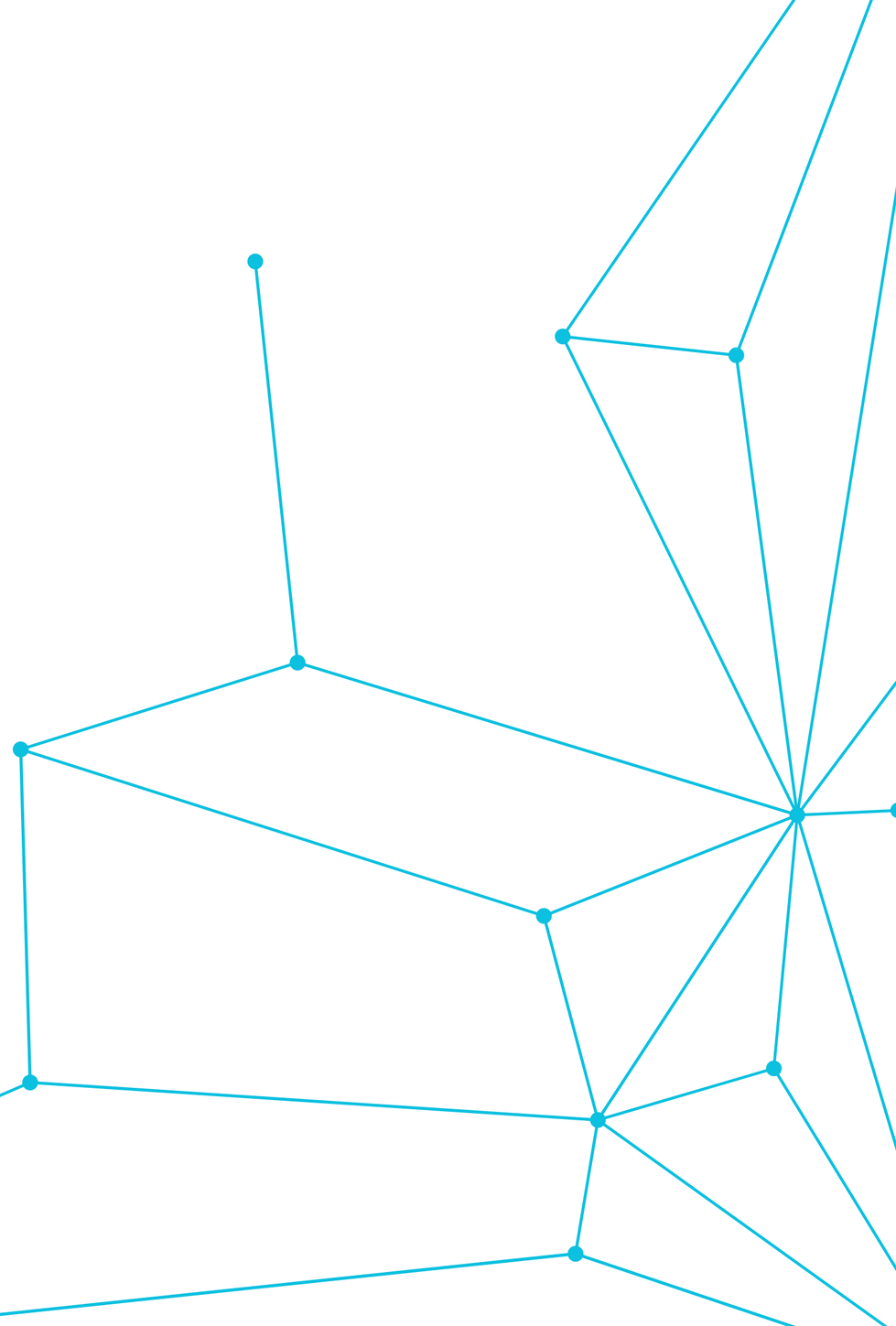


Hipotese	Resultado
Aspectos técnicos da mídia emergem como tópicos.	Confirmada
Gêneros emergem como tópicos	Confirmada
Os tópicos extraídos podem ser utilizados para predizer consumo futuro de outros produtos	Confirmada
Os tópicos extraídos podem ser utilizados para predizer consumo futuro de outros tópicos.	Confirmada
Usuários pertencentes a um determinado grupo avaliam pior produtos de categorias pouco consumidas pelo seu grupo.	Não confirmada

Perguntas	Repostas encontradas
Que tópicos emergem da descrição dos produtos?	A modelagem de tópicos foi capaz de encontrar temáticas relacionadas tanto a aspectos técnicos dos produtos, como de gênero. Além disso, foram encontrados tópicos ainda mais amplos, como exercícios físicos, música, história, cultura japonesa, desenhos animados e etc. Também foram encontrados tópicos mais específicos, como Scooby-Doo, Doctor Who e Star Trek
Esses tópicos são capazes de descrever relações com a compra de outros produtos?	Os tópicos encontrados foram capazes de de produzir perfis dos consumidores que foram utilizados para predizer consumo futuro com eficácia
Esses tópicos podem se relacionar com as categorias do produto e refletir aspectos de sua avaliação avaliação pelos consumidores?	O modelo usado não foi capaz de capturar relações entre os tópicos e as categorias, fazendo a resposta ser inconclusiva

LIMITAÇÕES E TRABALHOS FUTUROS

- O conjunto de dados utilizado na pesquisa é restrito a usuários com interações densas
- Houve uma demora em compreender o conjunto de dados, os modelos e métodos de avaliação utilizados
- As limitações técnicas do autor provocaram um longo atraso



Referências

NI, Jianmo; LI, Jiacheng; McAULEY, Julian. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In: CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING (EMNLP), 2019, Hong Kong. p. 188–197. DOI: 10.18653/v1/D19-1018

ALEKSEEV, Anton; NIKOLENKO, Sergey. Word Embeddings for User Profiling in Online Social Networks. Comp. y Sist., Ciudad de México , v. 21, n. 2, p. 203-226, jun. 2017 . Disponivel em: http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S1405-55462017000200203&lng=es&nrm=iso. Acesso em 24 de outubro de 2025. https://doi.org/10.13053/cys-21-2-2734.

GROOTENDORST, Maarten. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv preprint arXiv:2203.05794, 2022. Disponível em: https://arxiv.org/abs/2203.05794

GUAN, X.; LI, C.-T.; GUAN, Y. Matrix Factorization With Rating Completion: An Enhanced SVD Model for Collaborative Filtering Recommender Systems. IEEE Access, v. 5, p. 27668–27678, 2017. DOI: 10.1109/ACCESS.2017.2772226

SHAH, D.; SHAH, P.; BANERJEE, A. Similarity based regularization for online matrix-factorization problem: An application to course recommender systems. In: IEEE REGION 10 CONFERENCE (TENCON), 2017, Penang, Malaysia. p. 1874–1879. DOI: 10.1109/TENCON.2017.822816