Atividade 03 - Projeto

Equipe

Nome: $sin(args)^0$
Repositório GitLab da equipe: *https://gitlab.com/antfer/pdi*
Membros:

- Antonio Carlos Farias Ferreira, 2413868, antfer, antfer@alunos.utfpr.edu.br, BSI, UTFPR

Tema

Descobrimento de temas latentes em reviews de produtos da Amazon (categorias de filmes e TV) para inferência de perfis de consumo a partir de análise de textos de avaliação e dos produtos.

Perguntas de pesquisa

1. Que termos são recorrentes na descrição dos produtos?
2. Quais palavras, expressões e tópicos são recorrentes nas avaliações dos produtos? Como isso varia entre as categorias de produto?
3. Que temas latentes podem ser extraídos da análise das avaliações dos produtos? Esses temas podem refletir aspectos de decisão de compra?
4. É possível agrupar consumidores com base nos temas extraídos de suas avaliações? Quais relações entre o comentário de avaliação, as notas dadas pelo consumidor e a compra de outros produtos podem emergir desse agrupamento?

Hipóteses

1. As expressões e tópicos recorrentes variam significativamente de acordo com as categorias dos produtos.
2. É possível extrair temas que representam o que é valorizado pelo consumidor e o interesse por produtos semelhantes
3. É possível fazer o agrupamentos dos consumidores com base nos temas extraídos de suas avaliações. Usuários pertencentes a um determinado grupo avaliam pior produtos de categorias pouco consumidas pelo seu grupo.

Dados e modelos

Dados de avaliação de produtos da Amazon, como nota dos produtos, metadados do consumidor e comentários de avaliação. Dados dos produtos avaliados, como nome do anúncio, metadados do vendedor, categoria e descrição do produto, autor e etc.

Será realizado um recorte de produtos consumidos por 1000 usuários em um subconjunto de dados em que cada usuário tem ao menos 5 avaliações realizadas.

Utilização de TF-IDF para identificação de termos recorrentes e relevantes nas avaliações por categoria e descrição. Utilização de LDA para descoberta de temas latentes de categorias e reviews. Embeddings e clusterização para agrupar usuários.

Cronograma

1. Construção de documentos por categoria e consumidores
2. Identificar expressões recorrentes nas avaliações
3. Extrair temas latentes aplicando LDA interpretá-los
4. Clusterizar consumidores de acordo com os temas encontrados
5. Avaliar e interpretar resultados