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 (loja do Kindle) para inferência de perfis de consumo a partir de análise de textos de avaliação dos produtos.

Perguntas de pesquisa

1. Quais palavras, expressões e tópicos são recorrentes nas avaliações dos produtos? Como isso varia entre as categorias de produto?
2. Que temas latentes podem ser extraídos da análise das avaliações dos produtos? Esses temas podem refletir aspectos de decisão de compra?
3. 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. Temas relacionados à originalidade, profundidade, criatividade, preço e facilidade de leitura são recorrentes em produtos de categorias relacionadas a entretenimento e ficção, enquanto utilidade, relevância, organização e confiabilidade das informações tratadas são temas mais recorrentes em produtos de categorias não ficcionais.
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, especificamente da loja do Kindle, 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.

Foi realizado um recorte no dataset para que o processamento fosse viável. Esse recorte foi feito a partir da seleção do produto de 30 maiores ranks da loja da Amazon por categoria, resultando em um total de 4536 produtos. Desses produtos, existem 77991 reviews no dataset. O recorte foi feito sobre um subset do dataset total, em que todos os produtos e usuários têm ao menos 5 avaliações.

Utilização de TF-IDF para identificação de termos recorrentes e relevantes nas avaliações por categoria (para responder a pergunta 1). Utilização de LDA para descoberta de temas latentes (para responder as perguntas 2 e 3). Clusterização (para responder a pergunta 3).

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