Análise Exploratória de Dados:

Ranqueamento de interesse dos usuários da plataforma Goodreads utilizando Phyton.

Julia Santos Silva, Caio Ferraz Martins dos Santos

Bacharelado em Técnologia da Informação – Universidade Anhembi Morumbi (UAM) – São Paulo – SP – Brasil

Abstract. *The present article aims to analyze the data extracted from the Goodreads reading platform with the objective of gaining knowledge about the interests of which authors, books, and languages provided on the site are most frequently occurring among users of the literary community. This analysis will be based on the ranking of information, utilizing the Python programming language.*

Resumo. *O presente artigo tem a finalidade de analisar os dados extraidos da plataforma de leitura Goodreads com o objetivo de ter conhecimento dos interesses de quais autores, livros e linguagens fornecidos no site são de maior ocorrência entre os usuários da comunidade literária a partir do rankeamento das informações utiliizando a linguagem de programação phyton.*

Introdução

Com o aumento do número de usuários e da capacidade da Internet, a mesma tornou-se peça chave na publicação e acesso de livros online. Como solução para atender a esse tipo de demanda surgiram as bibliotecas digitais, quefacilitam o encontro entre a obra desejada e o usuário. Porém, com as plataformas de leitura digitais sendo também um meio de negócio e entretenimento é preciso ter conhecimento e se adequar aos interesses dos usuários para promover uma melhor experiênca.

O objetivo deste trabalho é através da extração e análise de dados da plataforma Goodreads (fundada em dezembro de 2006, lançada em janeiro de 2007 por Otis Chandler engenheiro de software e empresário, e Elizabeth Chandler. Adquirida pela Amazon em 28 de março de 2013), obter a visualização gráfica dos dados disponibilizados por um usuário do Kaggle a partir da mineração dos dados utilizando a linguagem de programação phyton, para organizar as informações geradas e posssibilitar a visualização gráfica necessária para explorar a partir do rankeamento das maiores ocorrências do site, os maiores interesses dos usuários entre os tópicos; autores, livros e linguagens.

1. Metodologia

Os dados são importados das bibliotecas que auxiliam no processo de funcionalidades para lidar com dados, visualização, clusterização e pré-processamento para manipulação e análise.

É definido um caminho para a base de dados no diretório "Goodreads" e em seguida é realizada a importação dos dados de um arquivo CSV chamado "books.csv" para o dataframe denominado "base_gd". O parâmetro "error_bad_lines" é definido como False para lidar com possíveis linhas com erros no arquivo, O dataframe "base_gd" é exibido usando a função "display". Isso permite visualizar os dados importados da base de dados "books.csv".

A função info() é chamada para exibir informações sobre a base de dados base_gd. Esse método fornece um resumo conciso, incluindo o número total de linhas, o índice das linhas, o número de colunas e o tipo de dado de cada coluna, cada uma delas é listada com seu nome e as informações:

Column: Nome da coluna.

Non-Null Count: O número de valores não nulos presentes na coluna.

Dtype: O tipo de dado da coluna.

o código gera um heatmap que mostra a correlação entre as variáveis numéricas da base de dados base_gd. Isso permite identificar quais variáveis estão mais correlacionadas umas com as outras, o que pode ser útil para análises e modelagem posteriormente.

Os dados são visualizados; a contagem dos títulos de livros é realizada na coluna 'title' do dataframe base_gd. Os 10 livros com o maior número de ocorrências são selecionados usando a função value_counts() e armazenados na variável 'books'. Além disso, a coluna 'average_rating' é selecionada e os 10 valores correspondentes são armazenados na variável 'rating'.

Um gráfico de barras é criado usando a função barplot() da biblioteca seaborn. Os eixos x e y são configurados como 'books' e 'books.index', respectivamente, para exibir a contagem de ocorrências dos livros no eixo x e os títulos dos livros no eixo y.

Para visualizar as línguas mais recorrentes na base de dados, a função value_counts() é aplicada à coluna 'language_code' do dataframe base_gd para contar a frequência de cada valor. Os resultados são exibidos com o comando print, mostrando a contagem de ocorrências para cada código de idioma após esse processo foi criada uma tabela chamada tabela_tipo_language que armazena a contagem de ocorrências de cada código de idioma. Em seguida, é definida uma lista vazia chamada colunas_agrupar.

A seleção dos 10 livros com maior número de avaliações o dataframe base_gd é ordenado de forma decrescente com base na coluna 'ratings_count', que representa o número de avaliações. Os 10 primeiros registros são selecionados usando a função head(10) e o índice do dataframe é definido como 'title'. O resultado é armazenado na variável most_rated e a contagem dos títulos de livros é realizada na coluna 'title' do dataframe base_gd. Os 10 livros com o maior número de ocorrências são selecionados usando a função value_counts() e armazenados na variável 'books'. Além disso, a coluna 'ratings_count' é selecionada e os 10 valores correspondentes são armazenados na variável 'rating'.

A organização dos 10 livros mais lidos pelos usuários é realizado a partir do dataframe base_gd que é agrupado pela coluna 'authors' e a contagem de títulos de livros é realizada na coluna 'title', os resultados são armazenados em um novo dataframe com as colunas 'authors' e 'title', e ordenados de forma decrescente com base no número de títulos usando a função sort_values(). Os 10 primeiros registros são selecionados usando a função head(10) e o índice do dataframe é definido como 'authors' e o resultado é armazenado na variável most_books.

Para organizar os 10 autores mais bem avaliados o dataframe base_gd é filtrado para incluir apenas os registros em que a coluna 'average_rating' (classificação média) é maior ou igual a 4.0 onde o resultado é armazenado na variável high_rated_author.

Para a contagem dos livros o dataframe high_rated_author é agrupado pela coluna 'authors' e a contagem de títulos de livros é realizada na coluna 'title', os resultados são armazenados em um novo dataframe com as colunas 'authors' e 'title', e ordenados de forma decrescente com base no número de títulos usando a função sort_values(). Os 10 primeiros registros são selecionados usando a função head(10) e o índice do dataframe é definido como 'authors' onde o resultado é armazenado na variável high_rated_author. Para obter a nota geral Conversão dos valores de média de avaliação: A coluna 'average_rating' do dataframe base_gd é convertida para o tipo de dados float utilizando a função astype(float). Os valores representam as médias de avaliação dos livros que pode ser visualizado a partir da função distplot() da biblioteca seaborn que é utilizada para exibir a distribuição dos valores de média de avaliação. O parâmetro rating é passado como os dados para o gráfico.

2. Análise e Consolidação

O ranking dos top 10 livros com maior número de leitores. A Figura 1 apresenta o resultado do processo de agrupamento de dados dos livros por ocorrência na plataforma.

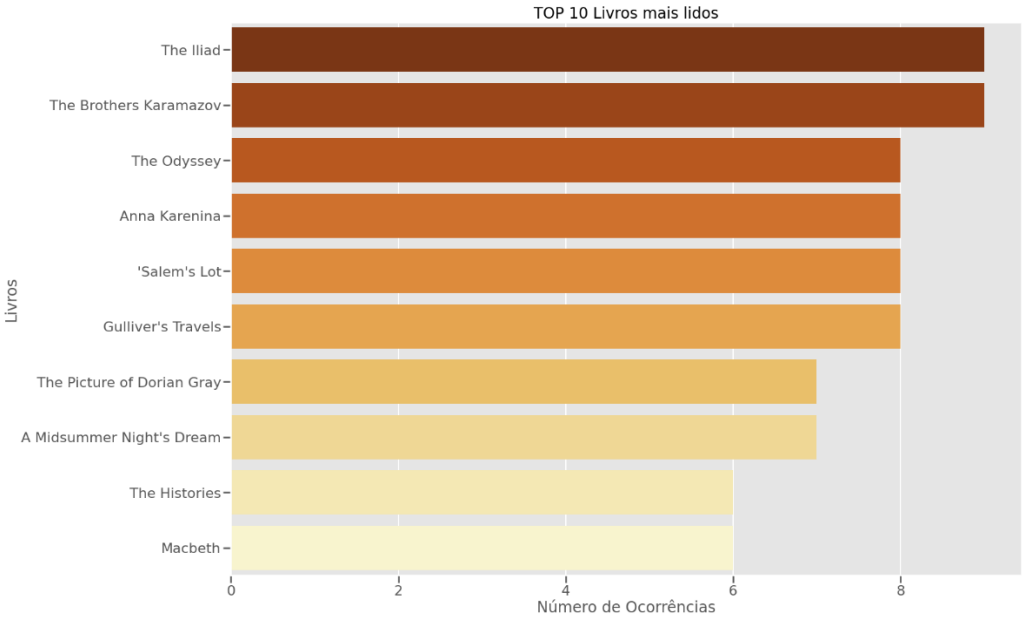

Figura 1. Processo de agrupamento de dados por ocorrência de pesquisa.

A primeira fase da análise consiste em agrupar as informações dos livros com maior número de leitores na plataforma. Essa etapa é essencial para identificar e classificar, de forma decrescente e através dos gráficos, o ranking de interesse do público. Na Figura 2, são apresentadas as linguagens mais recorrentes na base de dados analisada, bem como aquelas mais utilizadas nos livros. Essa visualização permite compreender quais idiomas predominam na leitura dos livros e quais são os mais frequentemente encontrados nas obras literárias analisadas.

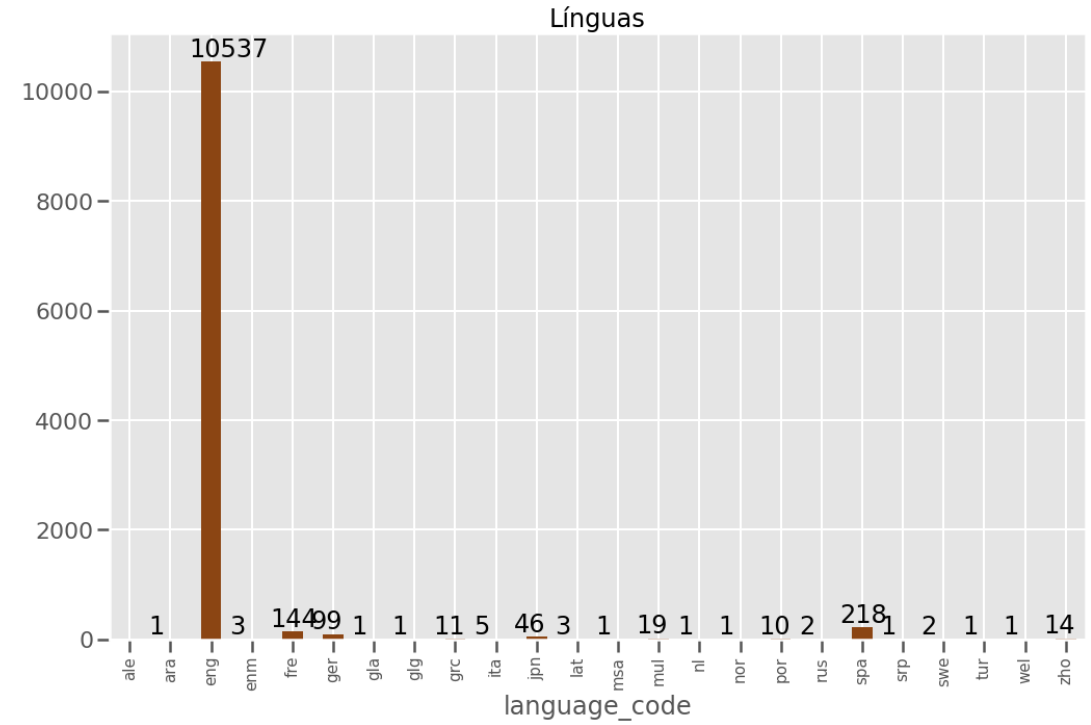

Figura 2. Processo de análise da linguagem mais utilizada nos livros da plataforma.

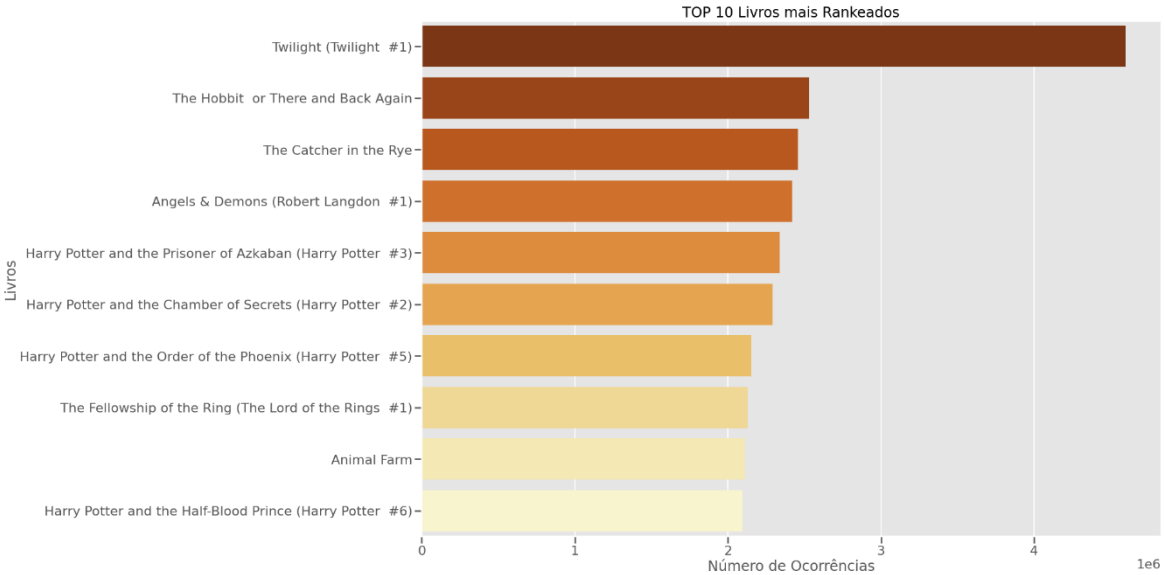

Figura 3. Apresentação do ranking dos livros mais avalidados.

Pode-se observar que os livros mais populares, conforme mostrado na primeira imagem, não necessariamente são os mais bem avaliados na plataforma. Além disso, é interessante notar que não há nenhum título presente em ambas as listas, indicando uma desconexão entre a popularidade de um livro e sua classificação pelos usuários. Isso ressalta a importância de considerar outros fatores, além do número de leituras, ao avaliar a qualidade e a preferência dos usuários.

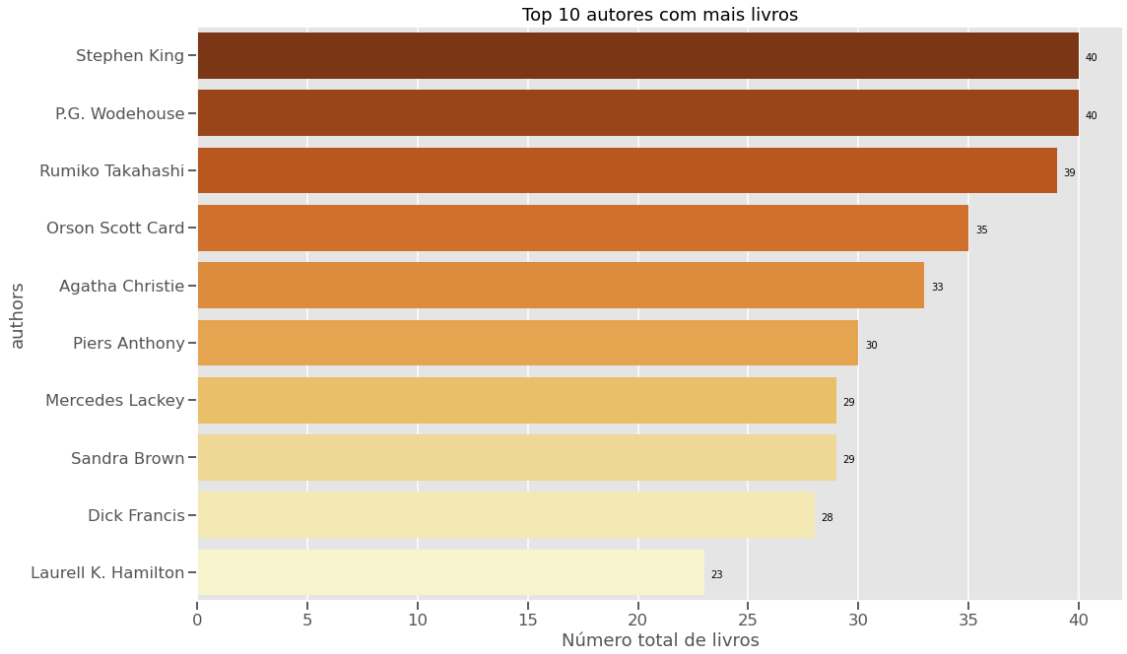

Figura 4. Apresentação do ranking dos autores com o maior número de livros.

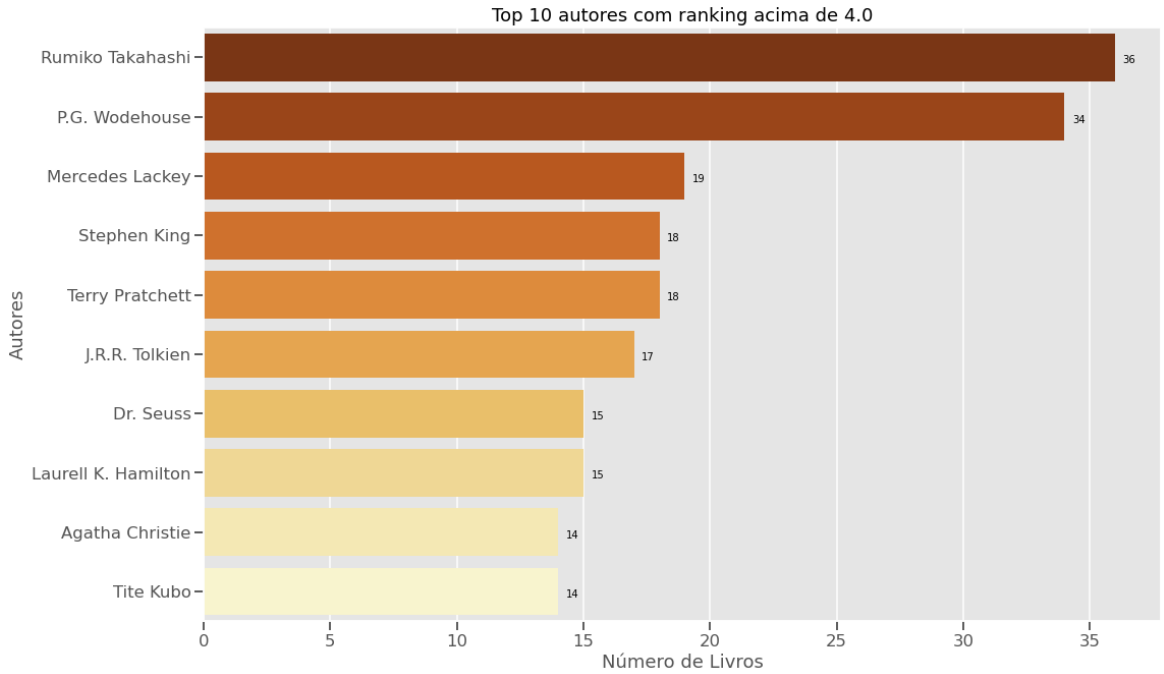

Figura 5. Apresentação do ranking dos melhor avaliados.

O gráfico resultante oferece uma visão abrangente dos 10 autores com as melhores avaliações, com base no número de livros que eles possuem, e com uma média de avaliação igual ou superior a 4.0. É interessante observar que há uma correlação entre os livros mais lidos, apresentados na Figura 1, e o ranking dos autores mais bem avaliados, mostrado na Figura 3 e Figura 5. Isso indica que os livros mais bem avaliados são, em sua maioria, escritos por autores que possuem uma alta média de avaliação no site. Essa relação entre a popularidade dos livros e a qualidade dos autores ressalta a importância de considerar tanto o número de leituras quanto as avaliações ao analisar o cenário literário da plataforma.

Conclusão

Através dos rankings apresentados, foi possível identificar tendências e preferências dos usuários. Essas informações são relevantes tanto para os próprios usuários, que podem descobrir novos livros e autores com base nas recomendações, quanto para editores, e profissionais do ramo editorial, que podem utilizar esses insights para direcionar suas estratégias de publicação e marketing. Além disso, a utilização da linguagem de programação Python para realizar a análise de dados e gerar gráficos permitiu uma abordagem mais eficiente e automatizada. Isso significa que o trabalho não apenas proporcionou uma compreensão mais profunda dos dados, mas também demonstrou o potencial e a aplicação prática da análise de dados e da programação na área literária. Em resumo, o trabalho realizado proporcionou uma visão detalhada sobre os interesses e preferências dos usuários da plataforma Goodreads, destacando informações valiosas para os envolvidos no mercado literário e demonstrando a importância da análise de dados e da programação nesse contexto.

Referências

Goodreads. Disponível em:

https://www.goodreads.com. Acesso em Junho, 2023.

Kaggle. Disponível em:

https://www.kaggle.com/datasets/jealousleopard/goodreadsbooks. Acesso em Junho,2023.

SBC Template. Disponível em:

https://www.sbc.org.br/documentos-da-sbc/summary/169-templates-para-artigos-e-capitulos-de-livros/878-modelosparapublicaodeartigos. Acesso em Junho, 2023.

Jupyter. Disponível em:

https://jupyter.org. Acesso em Junho, 2023.

Python Software Foundation. (2022). Python Programming Language. Disponível em: https://www.python.org. Acesso em Junho, 2023.

Github. Disponível em:
https://github.com/caioferraz/A3_Big_Data/blob/main/A3_de_big_data.ipynb. Acesso em junho,2023.