

2023

[Mineração de Dados RELATÓRIO FINAL]

JOANA SANTOS, MARCOS RIBEIRO – MESTRADO EM HUMANIDADES DIGITAIS

**Índice**

[Introdução 3](#_bookmark0)

Ferramentas usadas 3

[Metedologia, organização 4](#_bookmark2)

Desenvolvimento 4

Visualização dos dados 8

[Conclusão 11](#_bookmark3)

[Bibliografia. 11](#_bookmark5)

# Introdução

Neste relatório, apresentaremos os resultados de um projeto que utilizou análise de sentimento das críticas dos usuários do IMDb em relação aos filmes nomeados para o Óscar de Melhor Filme. O nosso objetivo principal foi comparar o sentimento expresso pelos usuários em relação aos filmes indicados com os vencedores dessa categoria nos últimos 10 anos. Além disso, também exploramos a relação entre o sentimento expresso e a avaliação dada por cada usuário.

Os resultados deste projeto podem fornecer insights interessantes sobre as preferências e perceções dos usuários do IMDb em relação aos filmes nomeados para o Óscar de Melhor Filme. Além disso, podem revelar se existe uma relação entre o sentimento expresso nas críticas e a escolha final dos vencedores do Óscar. Essas descobertas podem ser úteis para entender como o público em geral percebe e reage aos filmes indicados, bem como para aprimorar os critérios de avaliação e seleção dos vencedores do Óscar.

No próximo capítulo, discutiremos a metodologia utilizada para realizar a análise de sentimento das críticas e como obtivemos os dados relevantes do IMDb. Em seguida, apresentaremos os resultados obtidos, seguidos de uma discussão e conclusões sobre as implicações desses resultados.

# Ferramentas usadas

# Python:

# NLTK (análise de sentimento);

# Pandas; Selenium; BeautifulSoup (webscrapping).

# OpenAI

# Metedologia e organização

Neste projeto, seguimos os seguintes passos para extrair os dados do IMDb, realizar a avaliação de sentimento das críticas dos usuários e apresentar os dados de forma adequada:

1. Extração de dados do IMDb: Utilizamos técnicas de web scraping para extrair informações relevantes do IMDb, como os filmes nomeados para o Óscar de Melhor Filme nos últimos 10 anos. Recolhemos os títulos dos filmes, os seus respetivos anos de indicação e outros detalhes relevantes.
2. Recolher críticas e avaliações dos usuários: Com base nos filmes selecionados, extraímos as críticas e avaliações dos usuários disponíveis no IMDb. Utilizamos técnicas de web scraping para aceder as páginas de cada filme e recolher as críticas publicadas pelos usuários. Ao todo, recolhemos um total de 134.842 críticas e avaliações.
3. Avaliação de sentimento: Para avaliar o sentimento expresso nas críticas dos usuários, aplicamos uma técnica de análise de sentimento. Utilizamos um modelo de processamento de linguagem natural capaz de classificar o sentimento das frases ou textos como positivo, negativo ou neutro. Esse modelo atribuiu um valor de sentimento a cada crítica, com base na polaridade do texto.
4. Apresentação de dados: Após a avaliação de sentimento das críticas, procedemos à junção de todos os dados extraídos num único dataset. Esse dataset foi organizado de forma a incluir as informações sobre cada filme (título, ano) juntamente com as críticas dos usuários, avaliações de sentimento e outras métricas relevantes.

**Desenvolvimento**

**1. Extração de Dados (Web Scraping):**

Nesta primeira etapa, o nosso principal objetivo foi coletar dados de avaliações de filmes do site IMDb. Para isso, desenvolvemos um script em Python que utilizou as bibliotecas Selenium e BeautifulSoup. Com essas ferramentas, navegamos pela página do filme, acessamos as opiniões dos usuários e retiramos as suas reviews, ou seja, as suas opiniões e avaliações sobre os filmes, que foram armazenadas num dataset através da biblioteca Pandas para posterior análise. Além disso, registramos também a avaliação dada pelo usuário ao filme, numa escala de 1 a 10, consoante descrita no imdb.

Breve descrição das partes mais importantes do script:

Começamos por colocar o url de um filme específico para remoção das opiniões dos usuários, e em seguida informamos ao Selenium que vamos utilizar o Google Chrome.

A screen shot of a computer

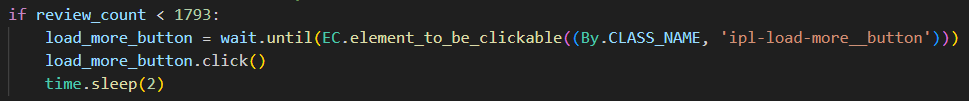
Description automatically generated with low confidence

Em seguida utilizamos um While loop com a condição ‘menor’ do que 1793 (que são todas as reviews existentes para este filme em específico). Como existiram alguns obstáculos na extração das reviews, tais como opiniões escondidas por causa de spoilers ou por chegarmos ao final da página onde existe um botão para carregar mais opiniões, tivemos que criar algumas condições para que os pudéssemos contornar. Utilizamos então um try/except para verificar em cada opinião se existe um botão de spoiler que neste caso teria a class de ‘show-more\_\_control’, se esta class existisse, o script pressionava o botão, se não existisse, seguia normalmente extraindo a opinião com a class ‘imdb-user-review’, e a avaliação com a class ‘rating-other-user-rating’.

A screen shot of a computer code

Description automatically generated with low confidence

Após a extração das opiniões da página, se ainda existiam mais opiniões, tivemos que pressionar o botão ‘Load More’, lendo então uma página nova com opiniões adicionais.



Finalmente criamos um dataframe com a biblioteca Pandas onde criamos duas colunas, uma para as ‘reviews’ e outra para os ‘ratings’, extraindo os dados para um ficheiro .xlsx.

A black screen with white text

Description automatically generated with low confidence

**2. Análise de Sentimento:**

Com os dados coletados na etapa anterior, passamos à análise de sentimentos. Para isso, utilizamos a biblioteca NLTK do Python, que nos permitiu analisar os textos das reviews retiradas na etapa de Web Scraping. Desta forma, conseguimos identificar se os comentários dos usuários eram, em geral, positivos, negativos ou neutros em relação a cada filme.

Primeiramente começamos por criar um diretório onde o script iria buscar os data sets das opiniões dos usuários, neste caso seriam todos os filmes nomeados para o Óscar no ano de 2014.



Primeiramente criamos uma função para a análise de sentimento através da biblioteca NLTK e em seguida criamos uma condição para que o script apenas leia ficheiros no formato .xlsx, e através do Pandas o script leu toda a coluna das opiniões de cada data set, atribuindo um valor de sentimento para cada review, e posteriormente calculou a média de todos os valores de sentimento para cada filme, que variam entre -1 (mais negativo), 0 (neutro) e 1 (mais positivo)

A picture containing text, screenshot, font

Description automatically generated

A screen shot of a computer code

Description automatically generated with low confidence

Como as avaliações dos usuários estavam em uma escala de 0 a 10, decidimos normalizar o valor dos dados de sentimento para que pudéssemos comparar ambos os valores de forma mais clara, para tal, somamos ‘1’ ao valor do sentimento e multiplicamos por ‘5’, por exemplo, se o sentimento de determinado filme fosse -0.5, ao somarmos 1 daria 0.5, e ao multiplicar por 5 daria 2.5/10, ou seja, o mesmo que -0,5/1, mas em uma escala bastante mais legível.



Após isso extraímos os dados e criamos um data set novo com a informação toda do data set inicial de Web Scraping juntamente com os dados da análise de sentimento.

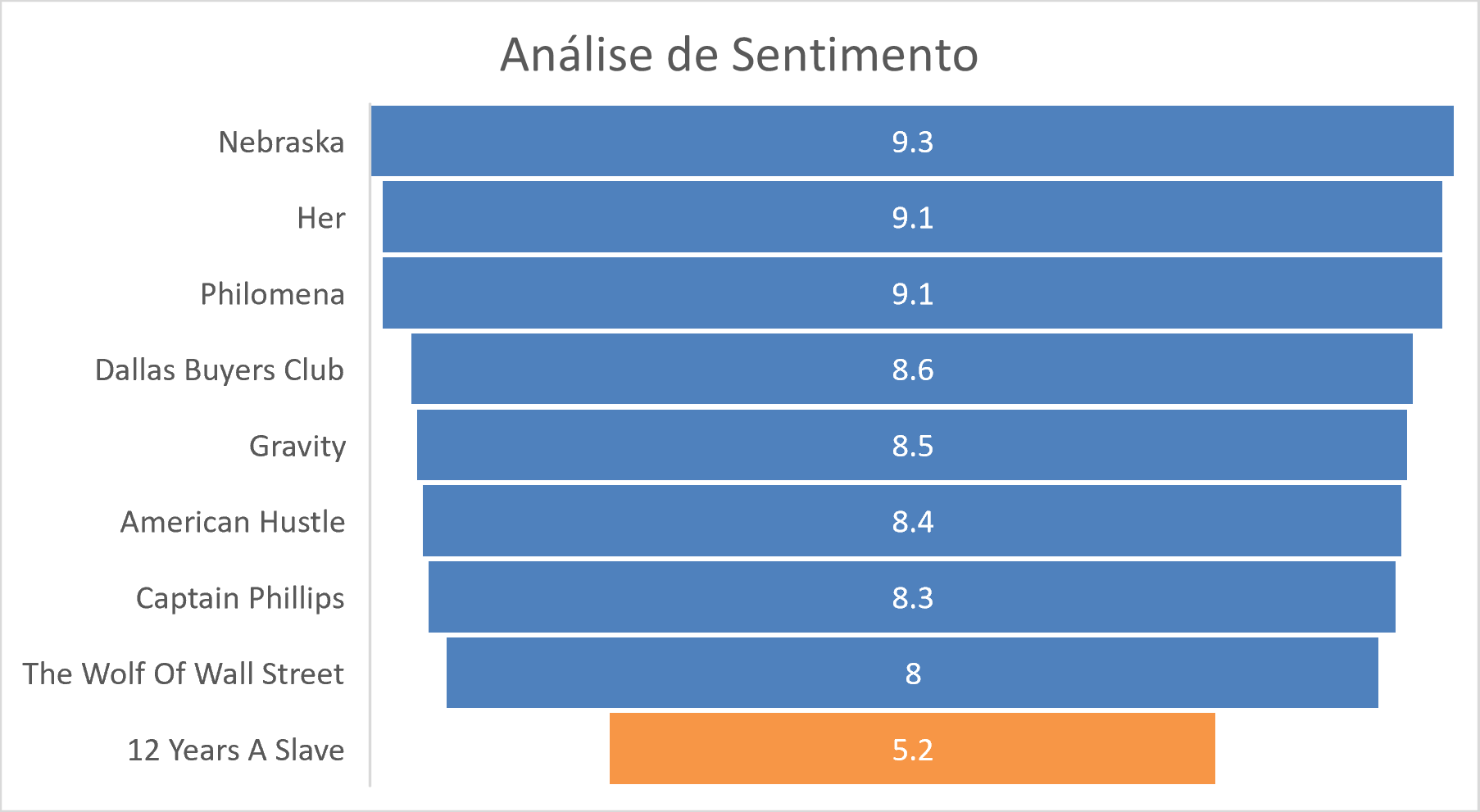
**3. Criação de um data set para análise final:**

Na última etapa, a partir do data set único que continha as informações obtidas tanto na etapa de Web Scraping como na de Análise de Sentimento, criámos um novo data set. Este novo conjunto de dados foi refinado para conter apenas as informações mais relevantes para a nossa análise: a média da análise de sentimento e a média da avaliação dos usuários para os filmes que foram nomeados para o Óscar num ano específico. Com este novo data set, conseguimos fazer uma comparação mais direta e significativa entre o sentimento médio dos usuários e a avaliação média dada aos filmes. Além disso, também permitiu analisar se os vencedores do Óscar estavam alinhados com o sentimento médio dos usuários, proporcionando uma nova visão sobre a relação entre a opinião do público e a premiação oficial do cinema.

**Visualização dos dados**

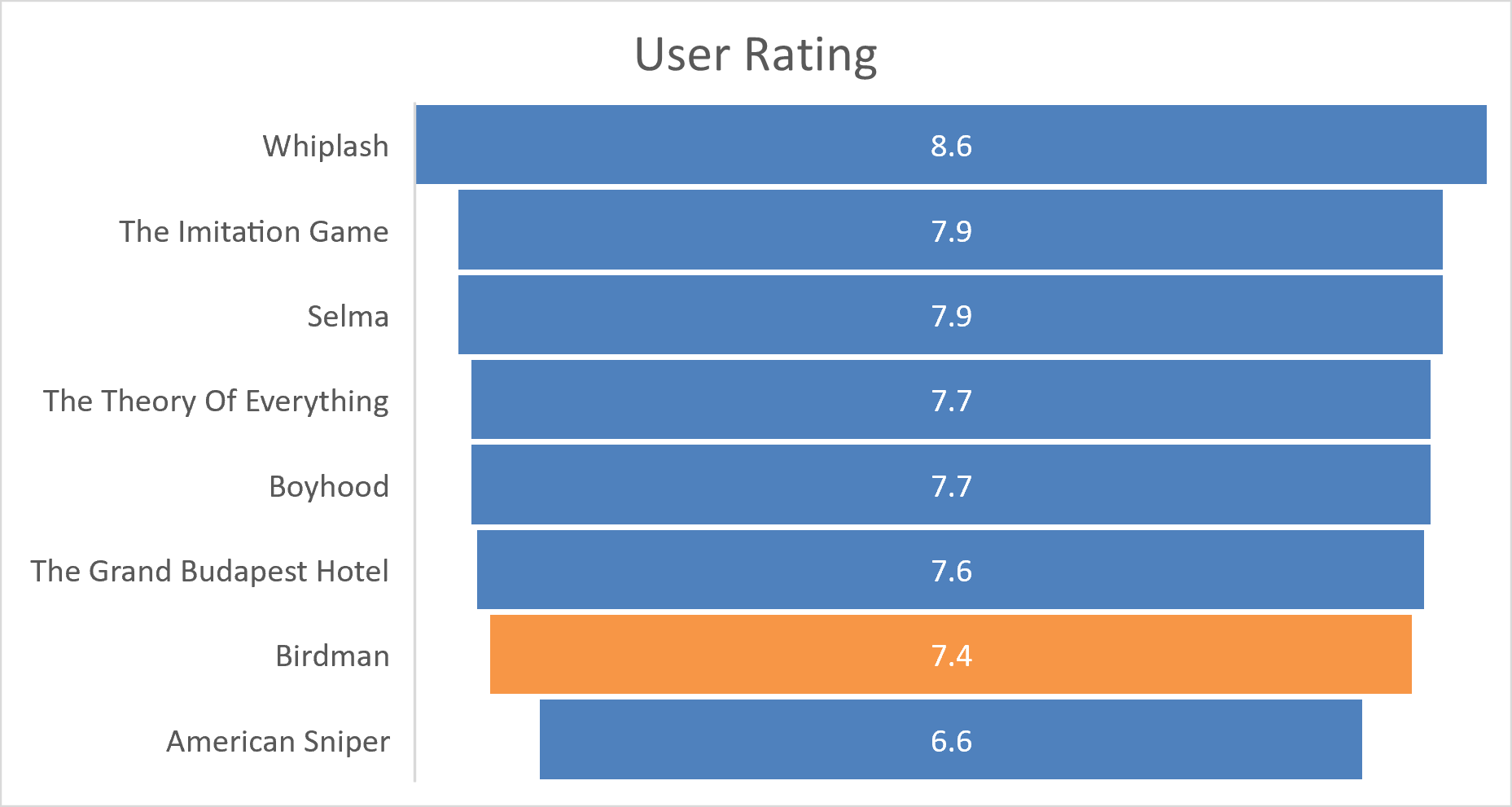
Em seguida iremos mostrar os gráficos em que houve mais disparidade entre os dados, em geral não foi detetada grande diferença entre os dados, no entanto quando o assunto abordado no filme fosse mais negativo, apercebemo-nos de uma descida na análise de sentimento.

* **2014**

**A picture containing text, screenshot, parallel, number

Description automatically generated**Nos gráficos abaixo podemos ver todos os filmes nomeados ao Óscar de melhor filme em 2014, e a laranja está indicado o vencedor. Neste caso, podemos ver a maior disparidade detetada, em que o filme vencedor, por retratar temas mais negativos como é o caso da escravatura, houve uma descida entre o rating e a análise de sentimento.

* **2015**

****Mais uma vez, um filme com temas mais negativos, no entanto não houve uma disparidade tão grande como foi o caso dos dados de 2014, pois a negatividade do tema não é tão predominante.

**A picture containing text, screenshot, parallel, line

Description automatically generated**

* **2016**

A picture containing text, screenshot, parallel, line

Description automatically generated**A picture containing text, screenshot, parallel, line

Description automatically generated**Provavelmente o caso mais dramático em termos de análise de sentimento. O filme em questão abordava milhares de casos verídicos de violação de crianças e o seu encobrimento por parte da igreja católica. No entanto podemos ver que ficou em primeiro em termos de user rating.

# Conclusão

Com base na comparação entre a análise de sentimento e a avaliação dos usuários, podemos concluir que, na maioria dos casos, os valores estão próximos. Isso indica uma tendência geral de concordância entre o sentimento expresso nas críticas e a avaliação dada pelos usuários. No entanto, observamos que a negatividade ou positividade do tema pode influenciar os resultados.

Embora a análise de sentimento e a avaliação dos usuários forneçam uma perspetiva valiosa sobre a perceção dos filmes nomeados para o Óscar de Melhor Filme, é importante ressaltar que nem sempre estão alinhadas com o vencedor. Isso sugere que outros fatores, além do sentimento expresso nas críticas e da avaliação dos usuários, podem desempenhar um papel importante na decisão final do vencedor do Óscar.

No entanto, a análise de sentimento e as avaliações dos usuários continuam a ser ferramentas valiosas para entender as preferências e opiniões do público em relação aos filmes indicados.

Em suma, embora a análise de sentimento e a avaliação dos usuários ofereçam perspetivas interessantes sobre os filmes indicados ao Óscar de Melhor Filme, é importante reconhecer que elas não estão necessariamente alinhadas com o vencedor.

# Bibliografia

*Academy Awards, USA (2023) - IMDb*. (n.d.). IMDb. <http://www.imdb.com/event/ev0000003/2023/1/>

Richardson, L. (2007). Beautiful soup documentation. *April*.

McKinney, W., & others. (2010). Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference* (Vol. 445, pp. 51–56).

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*. " O&#x27; Reilly Media, Inc."