Extraction, Transformation, and Load Technical Report

**SENTIMENT ANALYSIS OF JOKER MOVIE REVIEWS**

Hulester Bing

Antony Gikanga  
Subodh Khanal

Robert LaBelle

Aman Garg

Changler Dunklin

|  |  |
| --- | --- |
| **1.** | **INTRODUCTION** |

# 1.1 Summary

The objective of our ETL project was to scrape rottentomatoes.com for user reviews of the movie Joker so that we could upload them into a mongo database and conduct sentiment analysis to better understand how the movie was received by the public.

# 1.2 Scope

Our data was sourced from the 44,000 customer reviews of Joker from the Rotten Tomatoes website.

# 1.3 Technologies and resource contributions

*This section lists out the team members and their contributions towards the ETL initiative. Use this section to also outline (or list) the tech stack used to obtain the final outcome.*

* We used **Selenium** and **ChromeDriver** to scrape reviews from the Rotten Tomatoes website*.* Once we gathered the html responses, we used **Beautiful Soup** to parse through the data and gather the contents of each user review. Then we used **Text Blob** to analyze the subjectivity and polarity of each review and assign a set of scores. Lastly, we loaded the reviews, the polarity scores, and the subjectivity scores into our Mongo database using **PyMongo**.
  + Hu and Antony focused on scraping the data from Rotten Tomatoes
  + Amman and Subodh were responsible for parsing the HTML with Beautiful Soup and writing the code for the sentiment analysis.
  + Chandler and Robert were responsible for loading the data in MongoDB.

\*While we had individual responsibilities, all team members helped each other on all parts of the project.

|  |  |
| --- | --- |
| **2.** | **ETL DETAILS** |

*This section outlines a more detailed description of the processes utilized/proposed to achieve the objectives of this initiative.*

# 2.1 Data Import/Extract Sources and Method

We used Selinium and ChromeDriver to extract html responses from multiple webpages so that we could gather a large range of user reviews. Then we used PyMongo to load the data into our Mongo database.

# 2.2 Data Acquisition

The Joker is expected to leave theatres in 2 weeks after which we would not expect many additional reviews to be submitted. If someone wanted to use our code to analyze sentiment analysis on a different movie on Rotten Tomatoes website they could just substitute the URL and our code would work on any given movie that they selected.

# 2.3 Data Transform

As stated above Beautiful Soup and Text Blob were our primary tools for transforming the data so we could gather sentiment analysis. Beautiful Soup was utilized to parse through the contents of our html response so that we could gather each user review. TextBlob was then utilized to asses the polarity and subjectivity of each review and assign a set of corresponding scores.

# 2.4 Data Integrity

Our data was sourced from consumer reviews, and it is possible to falsify a review or for someone to provide a review who hasn’t seen a movie. However, given that we gathered over 44,000 reviews we’re confident that the dataset is large enough to outweigh any outliers.

# 2.5 Data Refresh Frequency

Again, given that the Joker is leaving theatres soon, we expect that the number of new reviews that will be added will fall precipitously. It would be more useful for someone to use our data to gather sentiment analysis from a new movie.

# 2.6 Data Security

Our data only scraped the content of the review on a public forum and no PII (personally identifiable information) was scraped.

# 2.7 Data Loading and Availability

All of the data was loaded to PyMongo, which would be accessible to users. We wrote code to reload the MongoDB database into a Pandas Dataframe for further manipulation by a client.

|  |  |
| --- | --- |
| **3.** | **DATA QUALITY** |

Address in this section success criteria for this project. Summarize the parameter KPIs such as Totals and expected counts. What user acceptance testing was performed and what were the outcomes. What is the recommended site acceptance testing that your client can perform to ensure the expected outcomes meets their expectations?

We were successful in scraping the reviews for Jocker movie, clean it, format the way we want it to be a and load into mongo database. We performed the analysis for only small section of the data because loading those data took a lot of time. We were able to scrape the reviews, add serial numbers for each review and ran sentiment analysis on each review. We were satisfied with the number of results, and expected outcomes. We pulled the mongo database to panda’s data frame for further analysis. We can now do averages on polarity and sensitivity for the entire movie.