CIE 427

Prepared For
Dr. Elsayed Hemayed

Prepared By

Omar Gaballah	201801697
Ahmed AbdelSalam	201801597
Mazen Hassan	201801897
Ibrahim Hamada	201800739

1. Idea

Our Idea is to analyze the IMDB reviews and predict whether a movie review is helpful using machine learning models and do sentiment analysis on reviews.

2. Scope

We want to do extensive data analysis on the movies dataset, which should give us some crucial insights that would help us learn plenty of information about the data and apply machine learning models to predict how much help a review is and sentiment analysis on the reviews.

3. Problem Statement

Movie recommender systems help people in narrowing their ranges of which movie to watch. However, people still look for other people's reviews to determine whether they would like to watch this movie or not. We believe that ranking people's reviews can be of great help for websites and viewers.

4. Proposed Solution

We will be analyzing the data using Spark to obtain the following insights:

- The correlation between average review rating and date of review and investigate if there are certain weekdays/months/holidays that correlate to changes in average ratings.
- The list of movies that get most of the review activity.
- The list of movies that get least of the review activity.
- The IMDB reviewers' "highest ranked" movies by average rating.
- The overlap between fandoms through reviewers who rate multiple titles.
- The most active reviewers on IMDB by determining the percentage of all reviews are posted by these reviewers.
- The distinctions that distinguish a helpful review from one that gets voted unhelpful.
- The effect of controversial reviews on the amount of votes.

5. Dataset

The dataset is obtained from IMDb Review Dataset - ebD | Kaggle. It is collected from IMDB with total records of 5, 571, 499 reviews for 453, 528 total shows. The reviews are written by 1, 699, 310 users. The data is split into 6 json files each of which has the following elements:

Content	Details
review_id	It is generated by IMBb and unique to each review
reviewer	Public identity or username of the reviewer
movie	It represents the name of the show (can be - movie, tv-series, etc.)
rating	Rating of movie out of 10, can be None for older reviews
review_summary	Plain summary of the review
review_date	Date of the posted review
spoiler_tag	If 1 = spoiler & 0 = not spoiler
review_detail	Details of the review
helpful	List of people find the review helpful.