DSMA Individual Assignment

Jhanvi Sharma – 12120086

The assignment is based on scraping a site and performing keyword analysis. Feel free to work on the suggested problem stated below or extend your work on any project from the previous terms or any other problem of your choice and change the objective accordingly.

Suggested/Sample Problem: This problem is taken from the perspective of a movie theater owner (PVR cinemas, for example). They want to understand how movie goers select and search for movies on Google. Your objective is to generate keywords that people use to search for movies that they want to watch.

Assignment Instructions:

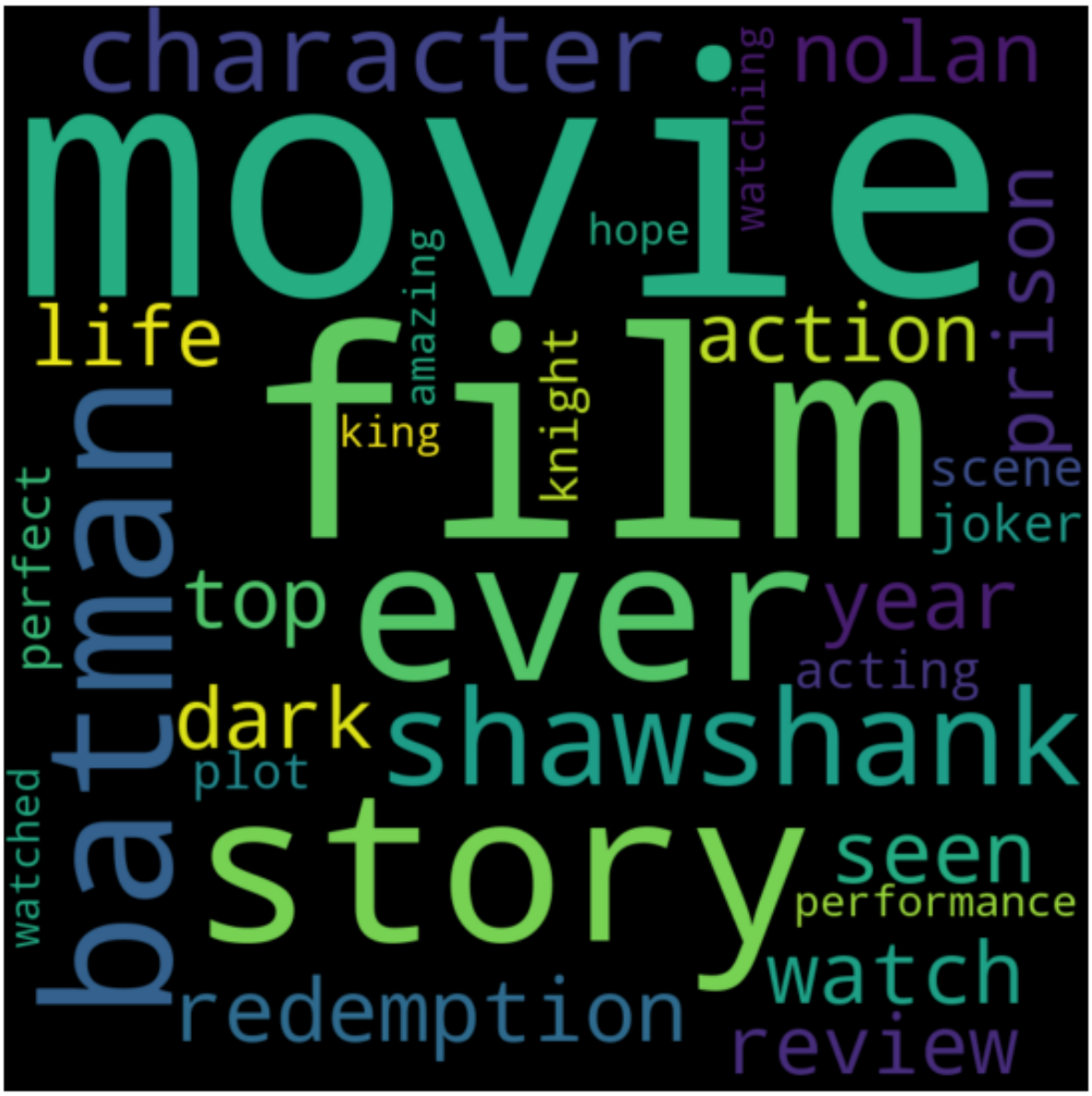
1. From IMDb (or any other site relevant to your chosen problem statement), extract user reviews of any 20 movies (or other relevant products/services). Preferably consider a diverse set of products/services.

20 movie reviews were scrapped from IMDB website. The CSV file was generated and names 'IMDB review' for further analysis.

2. Using either Python or R, perform keyword extraction and attempt to achieve a high diversity score.

Using the csv file generated, we found out mean, moving average and lexical diversity of the words –

```
The Lexical Scores are as follows:
 Simple_TTR: 0.1104954883647301
 Root_TTR: 17.564264957985756
 Log_TTR: 0.78270526860196
 Mass_TTR: 0.04935632790524033
 Mean Segment: 0.7805148514851474
 Moving Average: 0.7805075538284323
```

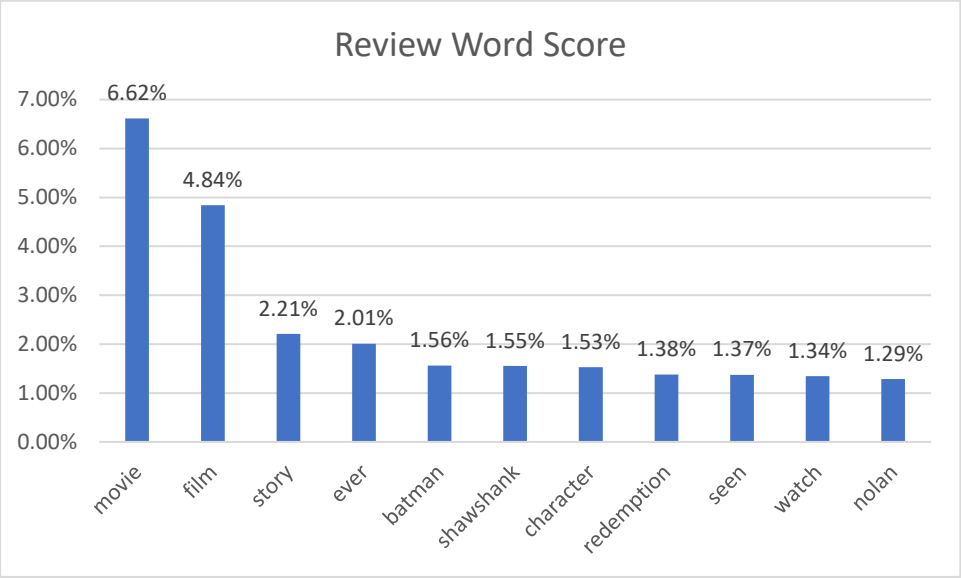


3. Explain the reasoning behind the generated keywords.

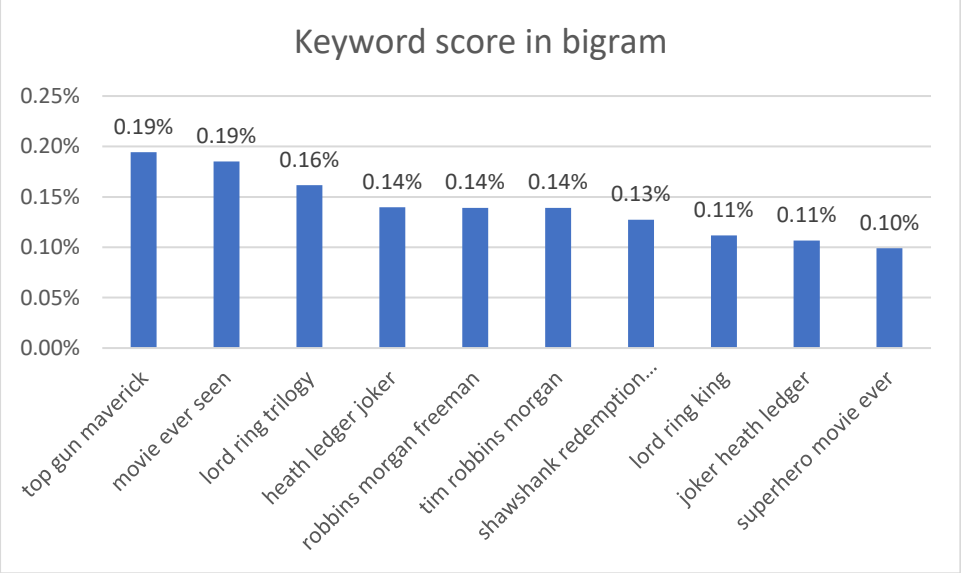
The following keywords have been generated as they are the most common words used by movie watchers to define any review of a movie.

Any person reviews a movie in simple understandable words so that as many possible people can understand the review and relate to it and add in their comments or views and support the review.

Words like – movie, film, story are very common and have the highest % of usage on IMDB as it's a movie review dedicated website and is one of the most searched review website all over the world for any review before watching a movie or even to post a review after a movie has been released.

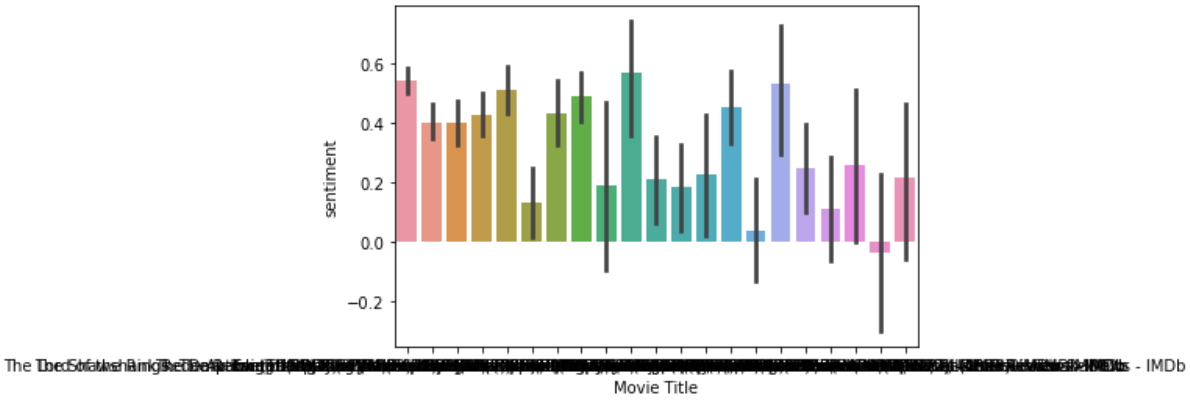


Keyword extracted using bigram –

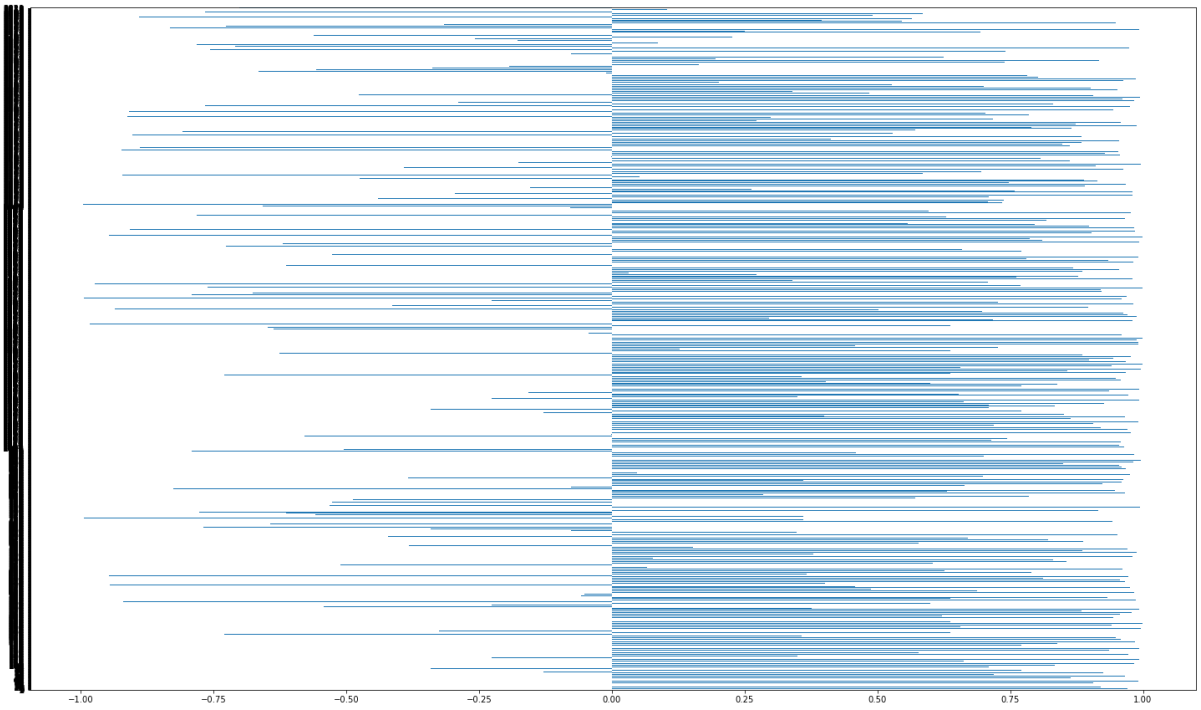


4. Perform sentiment analysis and LDA on the given data and explain the results.

Sentiment analysis on the 20 movie names represented by a boxplot –



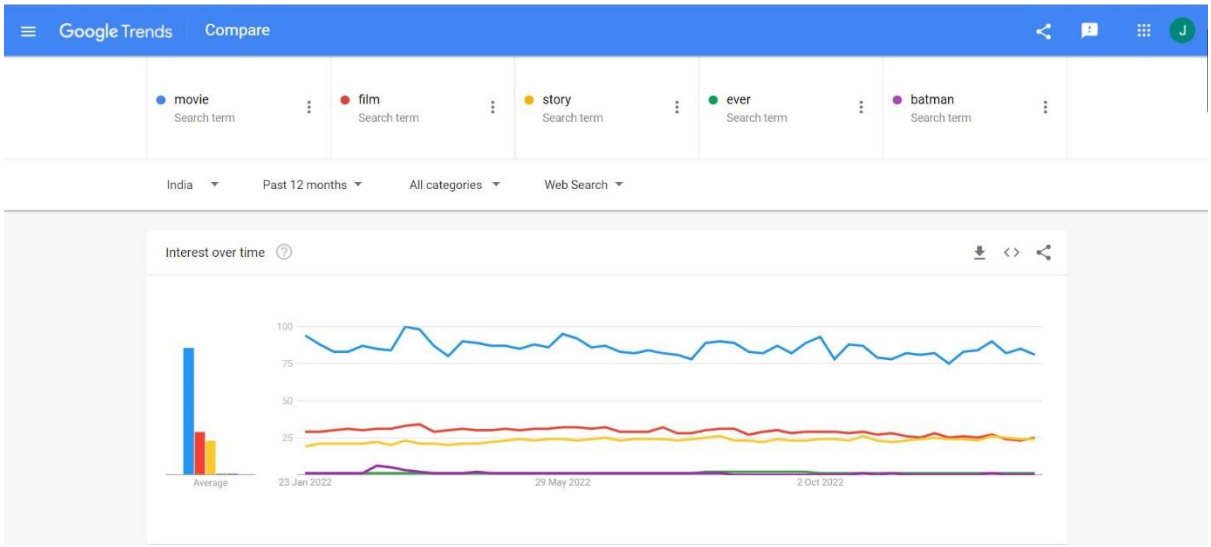
Sentiment analysis represented by subplot –

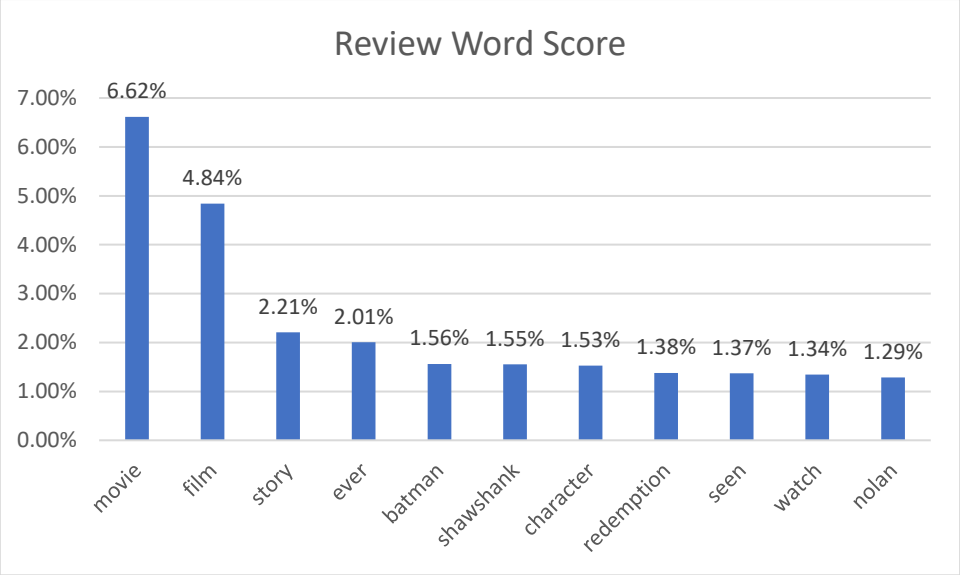


Word Tokenization is represented by –

5. Find the associated keywords or terms with the extracted keywords (You can use keywordtools.io, Google trends or Google ads).

Using google trends, I put in top 5 keywords and analysed their usage –




Review Word Score
7.00%
6.00%
5.00%
4.00%
3.00%
2.00%
1.00%
0.00%
6.62%
4.84%
2.21%
2.01%
1.56%
1.55%
1.53%
1.38%
1.37%
1.34%
1.29%
movie
film
story
ever
batman
shawshank
character
redemption
seen
watch
nolan