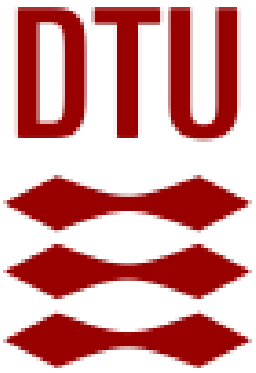


02807 - Computational Tools for Data Science
Final Report

Group 32

Name	Student Number
FOO YAN RONG	S231651
KAI JIE, JARED LIANG	S231648
NELSON JIA WEI CHOO	S231657
LEE JING TING	S231586

Problem/Motivation:

A crucial indicator to understanding whether a movie is well perceived by the general audience is to observe and analyse reviews. For instance, one aspect is understanding if the nature of these reviews tend towards a positive or negative sentiment. Another aspect is identifying any common themes, if any, across multiple reviews. All these can help provide valuable insights into the collective perception of the audience. However, with thousands of reviews on the internet, manually completing these tasks can exhaust many resources.

Marvel, once the reigning blockbuster, is currently undergoing a decline in popularity. The ultimate aim of conducting sentiment analysis and similarity analysis is to furnish Marvel movie directors and filmmakers with enhanced insights into reviews and critiques concerning their cinematic endeavors. This endeavor is designed to facilitate the production of films that align more effectively with the broad interests of the general audience, potentially resulting in heightened box office returns in subsequent productions. This analytical approach proves particularly advantageous for filmmakers involved in projects within a similar genre or crafting sequels to specific movies, allowing them to address and mitigate concerns raised in initial reviews for subsequent cinematic offerings.

In our subsequent analysis, we started off with sentiment analysis to evaluate the overall viewer reception of Marvel movies by discerning whether reviews tend towards positive, neutral, or negative sentiments, and identify any trends over time. To identify common themes/topics within reviews, we use similar items to see if the general public has similar opinions on the movies. From there, we examined our results from both fields and identified any interesting or similar findings to delve into further.

Data Exploration and Preprocessing

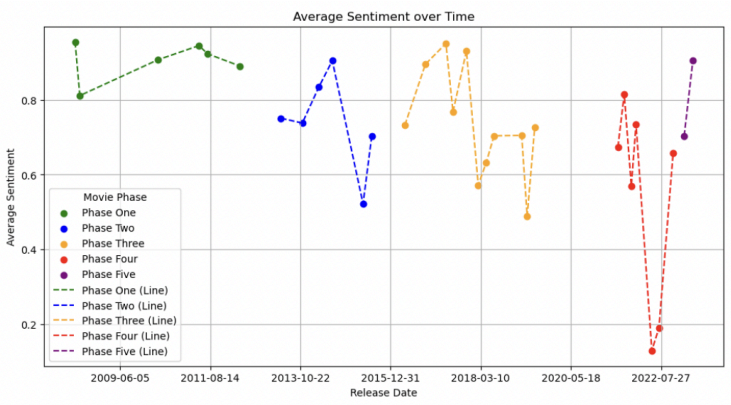
For the present undertaking, the Python package Cinemagoer, formerly known as IMDbPY, was employed to extract data from the IMDb movie database. Prior to this, the initial step involved compiling a roster of existing Marvel movies. To accomplish this, the MCU Movies dataset from Kaggle was employed. Subsequently, by systematically iterating through the list encompassing 32 Marvel movies within the Cinemagoer package, a total of 25 reviews were successfully extracted for each movie title.

For text data analysis, the reduction in noise helps to ensure the cleanliness of the text for subsequent analysis. To this end, a series of preprocessing steps were executed. Firstly, all words within the reviews were converted to lowercase, with both trailing and leading punctuation marks removed. The removal of stop words was executed through the use of the NLTK package so that only important and relevant words are captured in the analysis. Furthermore, to maintain analytical conciseness while capturing the core ideas, each word underwent lemmatization to reduce it to its base form.

Sentiment Analysis

Leveraging the VADER Lexicon through the nltk library and its Sentiment Intensity Analyzer (SIA) module, we scored reviews to gauge the emotional tone expressed in each movie's feedback. The process involved concatenating reviews for each movie and assigning sentiment scores to individual words based on the lexicon. Aggregating these scores determined the overall sentiment, representing the average sentiment value for each movie. Our analysis unveiled trends within each phase and across phases (Figure 1).

Examining the first four phases individually (Phase Five is incomplete), we noted a pattern where early movies exhibited a robust start, followed by a dip towards the phase's end - especially evident in MCU Phases Three and Four. Over the entire timeline, a gradual decrease in sentiment scores transpired. Intriguingly, the onset of each new phase witnessed an uptick in sentiment scores, reflecting heightened anticipation. Phase Four, however, stood out with a distinct dip in average sentiment compared to earlier phases.




```
Average Sentiments for Each Phase:
Phase One: 0.9058099999999999
Phase Two: 0.7426206666666665
Phase Three: 0.7370934545454546
Phase Four: 0.5387617142857143
Phase Five: 0.804656
```

Figure 1. Plot of Average Sentiment Scores for each movie across time (Left), and Average Sentiment Scores for each Phase (Right)

Evaluation of method

Our sentiment analysis insights serve as a valuable proxy for Marvel filmmakers, offering a nuanced understanding of audience reception trends. Pinpointing key timeframes and specific movies for detailed review provides a strategic lens to comprehend sentiments and identify areas for improvement. This alternative internal evaluation method goes beyond subjective movie rating scores, guiding filmmakers in addressing potential shortcomings and enhancing the overall quality of future cinematic endeavors.

Similarity Document Search

After conducting sentiment analysis, the movies are categorized into positive, negative, or neutral sentiments. The next step is to identify any common themes or topics within reviews in each movie. This exploration aims to ascertain whether the general audience shares common or disparate opinions. The objective is to discern whether specific strengths or weaknesses are concentrated in a singular aspect of the movie or if they are attributed to diverse facets of the cinematic experience.

Approach 1 - Shingling, minhashing, locality-sensitive hashing

The approach for searching similar documents is by performing shingling, minhashing and locality-sensitive hashing. The reason for using such an approach instead of the bag-of-words approach is because order is important. For example, comparing these 2 documents: “The plot was great but the ending was terrible.” and “The ending was great but the plot was terrible”, it can be observed that the two documents contain the same words but have completely different meanings.

Shingling

Shingling involves grouping consecutive elements together where overlapping subsets of shingles are created to identify patterns or similarities in different texts. Considering that the reviews are not near-duplicates, it makes more sense to consider words, rather than characters, as elements to form the shingles. In this case, k=3 words were chosen to form shingles. The choice of k helps to capture impactful phrases.

MinHashing

Minhashing helps to efficiently estimate the similarity between 2 sets of documents by condensing the set of shingles into a more manageable size, represented by signatures. The hash functions are generated using SHA256 with salting. The salt values are randomly generated to give each hash function its uniqueness. A total of 300 hash functions were used, resulting in a signature of length 300 for each document.

Locality-Sensitive Hashing (LSH)

Lastly, locality-sensitive hashing is performed to efficiently find approximate nearest neighbors between the documents. The signature matrix is divided into b=100 bands. Using this approach, candidate pairs are

generated when 2 documents are hashed into the same bucket for each band. For each candidate pair, the respective signature vectors are extracted and checked for similarity. This is done using Jaccard similarity with a threshold of 0.2. The threshold is determined by intuition and observing the total vocabulary size.

Evaluation of method

It was observed that the identification of similar reviews remained limited despite the hypertuning of parameters. This limitation arises from the fact that the similar items algorithm focuses on capturing the lexical similarity but not the semantic similarity. Hence, its efficacy is higher when applied to documents that exhibit near-identical content, a characteristic not representative of the reviews. For instance, diverse reviews conveying the sentiment that the movie is good or great, albeit using different terminology, share a common meaning. However, the algorithm fails to discern such semantic nuances, making the identification of similar reviews difficult even with a reduced threshold in Jaccard Similarity.

Approach 2 - Word Embeddings and Cosine Similarity

The poor performance of minhashing and LSH can be attributed to the fact that this keyword-based approach is a lexical search that relies on exact matches. However, it is clear that two texts can contain different words but have the same meaning. Thus, there is a need to capture the semantic relationships between words.

Addressing this challenge involves representing words using word embeddings accomplished through vectors. Word embeddings can help to enhance semantic search by representing words in a continuous vector space, where the positioning of words reflects their semantic relationships. For this reason, the Word2Vec python package is utilized to convert words into word embeddings, enabling the measurement of semantic similarity among movie reviews. Word2Vec converts words to embeddings by learning distributed representations of words based on their context within a given corpus. Opting for cosine similarity as the metric of choice, as opposed to Jaccard similarity, facilitates the assessment of vector orientation in a multi-dimensional space, accounting for both direction and magnitude of word vectors. A threshold of 0.8 has been selected for this purpose.

$$cos(\theta) = \frac{A \cdot B}{||A||\,||B||}$$

Evaluation of method

The outcomes of this approach of similarity document search reveal a pattern aligned with our expectations. Several movie titles exhibit a substantial number of similar reviews, while others have few to none. Notably, movies such as The Incredible Hulk, Iron Man 3, and Captain America: The First Avenger have large counts of similar reviews. In contrast, movies like Avengers: Endgame, Spider-Man: Far From Home, Eternals, Thor: Love and Thunder, and Shang-Chi and the Legend of the Ten Rings show no similar reviews through this approach.

The sentiment analysis and similar document search results indicate a noteworthy trend. Movies with the highest count of similar reviews predominantly belong to the franchise's initial phase and tend to exhibit higher sentiment values. This likely suggests a shared viewer consensus regarding the effectiveness of the early movies in laying the foundation for subsequent phases of the franchise. Conversely, in later phases, movies may encounter more diverse critiques, resulting in greater variance in reviews and comparatively lower sentiment scores.

TF-IDF

Following an assessment of the overall audience perception of the movies, our focus shifts to a more granular exploration aimed at discerning the specific elements that contribute to a movie's positive or negative reception. To achieve this, the TF-IDF approach is employed to extract the most crucial keywords within the reviews for each movie. This method allows us to pinpoint and prioritize the terms that carry the most significance in capturing the essence of audience sentiments toward individual movies.

In our project, our original intent was to use a weighted sum of TF-IDF scores to incorporate both term frequency and movie-level significance for a nuanced sentiment analysis. This approach intended to give additional importance to terms that were not only frequent in individual reviews but also held significance across all reviews for a movie.

A distinctive emphasis was placed on adjectives during the entity extraction phase using spaCy, with the intention of highlighting words that carry emotional tone and sentiment. A consolidation of all reviews for each movie facilitated the computation of TF-IDF scores at the movie level. TF-IDF scores were then calculated, revealing the most significant terms and offering insights into key words shaping the overall sentiment across all reviews for each movie.

Fig 2. Word Clouds of adjectives in 'Iron Man 3' reviews (Left) and 'The Incredible Hulk' reviews (Right) based on TFIDF values

Evaluation of method

Our analysis uncovered limitations in the effectiveness of TF-IDF, particularly in capturing sentiment nuances related to movie reviews. Despite the emphasis on adjectives in the entity extraction, TF-IDF's intrinsic nature posed challenges for sentiment analysis in the context of evaluating movies. TF-IDF, relying on word frequency and uniqueness, often emphasised adjectives related to specific movie elements such as settings, characters instead of describing their feelings towards the movie itself. Consequently, it struggled to distinctly capture the emotional aspects of viewer opinions, which are crucial in understanding audience sentiments toward a movie. This is apparent from how TF-IDF at the review level revealed more nuanced sentiments compared to its application at the movie level. Recognizing these limitations, we opted not to use TF-IDF as a complement to sentiment analysis but instead incorporated dedicated sentiment analysis techniques to directly analyse and interpret emotional tones in the reviews.

Limitations

From our analysis, we realise the phenomena of superhero fatigue, which refers to the later movies in the franchise having fewer enthusiasts. Ideally, the reviews that were given by people should be independent between movies, without the influence from having watched the other movies. However, it is likely the case where most people who reviewed the later films have already watched the preceding films too.

Conclusion

Generally, we found that sentiment analysis was effective in capturing the sentiments of individual words. However, due to the many adjectives that are in the reviews, the sentiment analysis was not as beneficial at capturing the sentiments of opinions in the review, because adjectives themselves have sentimental value. This also affects TF-IDF because of the nature of the dataset.