Application to analyse Marvel script dialogue to identify character similarities along and their sentiments

Group 37: Aimee Donsu, Dominique Rueda, Gargi Nandanpawar, Kornpol Chung

1 Introduction

The Marvel Cinematic Universe refers to the realm of superhero films produced by Marvel Studios. Due to its overwhelming popularity and recognition value, the films have expanded into a collection of over 27 films composed of different writers and genres, all sharing a common narrative. For this reason, not only do the Marvel Cinematic Universe films and their respective scripts provide a good representation of today's dominant works of fiction, but it is also one of the only examples of an organically growing corpus of fictional stories that are connected via a common theme and setting.

Given this, a common problem when screenwriting future projects is preserving continuity with regards to characterisation and relationships between existing characters in the MCU, simply due to the sheer number of characters and an even greater number of relationships they maintain between each other. The idea behind the application was to create a tool that could allow screenwriters to quickly view the sentiments of characters and their similarities to each other, based solely on scripts of existing films, in order to provide a holistic understanding of the existing themes and character dynamics in the MCU. Having such an understanding would thus aid them in keeping consistent the existing characterizations of characters while creating new and distinct future characters.

1.1 Research Question

To what extent can sentiment analysis, NER classification and similarity index of script dialogues help us understand the similarities between characters?

2 Dataset Preparation

2.1 Data Preperation and Inspection

The dataset that is being used for this text-mining project comes from a Kaggle repository (Hari, 2020) that is a combination of the original scripts and transcripts from all of the existing superhero films produced by Marvel Studios. These lines of spoken dialogue are organized by the character in a chronological manner, going from the beginning of the first MCU film (Iron Man 2008) to the most recent (Avengers: Endgame 2019).

The dataset consists of 15724 written lines of spoken dialogue and includes additional columns such as character, movie, year and number of words that appear in the line. Furthermore, the dataset also contains additional one-hot encoded columns which indicate for each column which writer contributed to the creation of the corresponding spoken line.

2.2 Data Cleaning and Pre-processing

The first step in data cleaning was the elimination of non-useful columns. Most apparent were the columns indicating writer names which were one-hot encoded therefore not in a format applicable to the text-mining project, and moreover, does not provide any contextual information as to the sentiment of a character. Moreover, the data set contained lines for every single character that appeared in the movies. This would also include minor rules with very few lines. To reach more accurate (sentiment) scores and results, it was decided to use only a small number of characters with enough lines to analyse. Of all the characters, the top five were selected with the most lines: Tony Stark, Thor, Steve Rogers, Peter Parker and Natasha Romanoff. Their lines would be extracted into a new data frame and further processed.

As part of the spaCy data cleaning step, We need to firstly dismiss stop-words such as ”You”, “I”, “the” etc. For the NER analysis, only the lemma of the words was used, for example, the verbs ”walks”, and ”walking” are forms of the same lemma ”walk” (Maynard et al., 2016). Taking the lemma of the possible words makes sure no word is overlooked.

Because each script line can consist of multiple sentences, it was necessary to tokenize each script line (Maynard et al., 2016). The tokenizer would split up the text into separate sentences and would then be fed into the sentiment analyser. This way, the script lines would hold more context and calculate a more accurate score.

3 Methodology

3.1 Sentiment Analysis

The first step in the application was to provide an accurate sentiment analysis of the MCU characters. The first thing that stood out when approaching this task was the length and brevity of each spoken line. Due to the fact that the corpus is a collection of dialogue lines (and likely also due to how the superhero movie scripts are traditionally written), the average length of each line was short, having a mean word length per line of 11.01 words. With the sentences being relatively short, the context is usually kept simple. Therefore, the analyser would have little difficulty calculating its sentiment.

For the sentiment analysis task itself, it was decided that the VADER sentiment analysis package be used (Hutto, 2022). This is because the VADER (Valence Aware Dictionary for sEntiment Reasoning) approach is an example of a lexical approach and in the context of the project, the obstacle of not having a dataset containing annotated sentiment labels per line can be circumvented (Vossen, 2022a) (Jurafsky and Martin, 2020a).

The VADER Sentiment Intensity Analyser was then applied to each line to extract a positive, neutral, and negative score. These scores would range between 0 and 1, and be summed up as a compound score. Then for each line of character dialogue, its average would be calculated from the compound scores and result in the final sentiment score/label ranging between - 1 being negative, 0 being neutral and +1 being positive. Following this, each character would have a collection of sentiment scores which could then be visualized in a stacked bar chart, giving a general view of the character’s sentiment. Finally and further, taking the sentiments scores from each character separately and presenting them over the course of each movie can display the character’s development (ritakalach, 2020).

3.2 Named Entity Recognition and Similarity Scores

The task of identifying the similarities between characters began first with performing Named Entity Recognition and Classification (Vossen, 2022b) (Zvornicanin, 2022) and POS-tagging (Pham, 2020) and then viewing the similarity scores between each character from the results of those two.

To accomplish this, the spacy module was used. Spacy (spaCy) requires its language model to be loaded, which is a collection of entities and POS identifications (spacy, unknown) (Pham, 2020). Following this, for each character, the most commonly used verbs, adverbs, nouns, adjectives as well as named entities were identified and their frequencies were recorded. A corresponding histogram plot was plotted for each character and for each category (Laura, 2019).

Following this, a similarity score was calculated based on a vocabulary consisting of the aforementioned categories per each character pair. The similarity score is calculated using spacy’s similarity() function, which calculates similarity based on comparing their word embedding representation (Jurafsky and Martin, 2020b). Finally, using this collection of paired similarity scores, a similarity matrix would then be plotted.

4 Results

With this, enough data had been gathered to come up with conclusions. We were able to use the data to extract sentiment scores for a chosen few characters as well as decipher the relations between the said characters. The diagrams we used to display the results are (stacked) bar charts, and a confusion matrix.

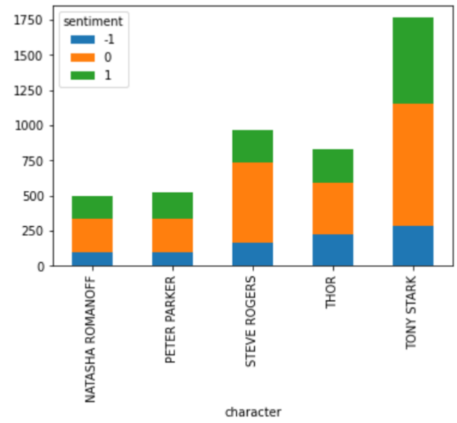


Figure 1: Sentiment Analysis

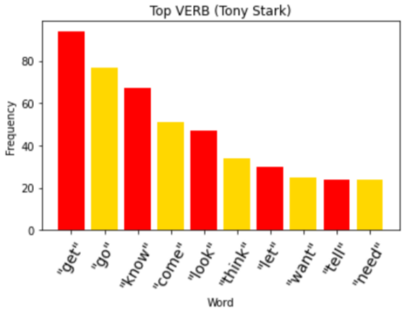


Figure 2: Verbs used by Tony Stark

We envision our application to allow the user to pick a character and the output would be their sentiments, their top used entities as well as the character who is the most similar to them. For the NER model, with the help of POS tagging,the information we were able to acquire is as follows: We got the frequencies of verbs, adjectives, nouns and adverbs, sorted by overall and by character. The top entities overall and by character which was used to produce a similarity score, a value between 0 and 1, to determine similarity between characters which was visually represented through a confusion matrix. For the sentiment analysis, we were able to get information in terms of a score: -1 (negative), 0 (neutral) and 1 (positive). The results were essentially put in one stacked bar graph, where the frequency of each sentiment score was given for each character. To describe our results more in-depth we will be using the example of the character, Tony Stark. The results for the remaining characters can be found in the appendix.

At first glance, the sentiment graph shows

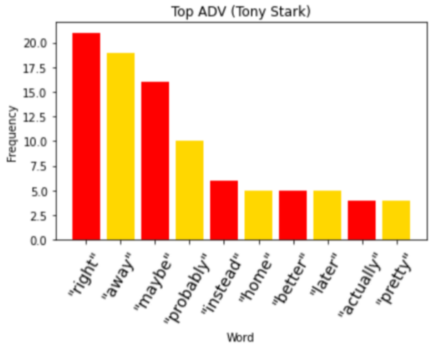


Figure 3: Adverbs used by Tony Stark

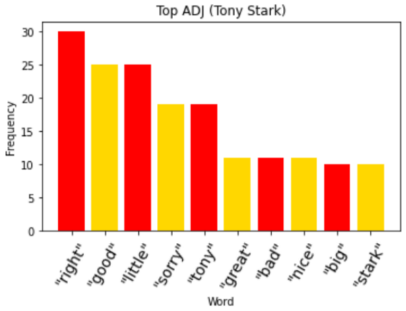


Figure 4: Adjectives used by Tony Stark

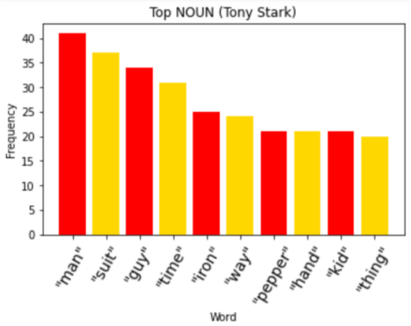


Figure 5: Nouns used by Tony Stark

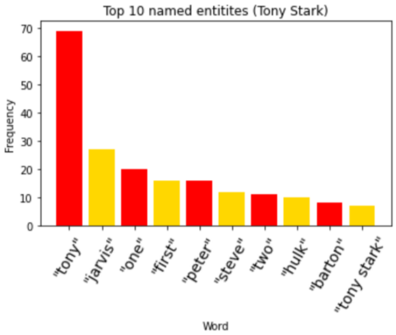


Figure 6: Top entities used by Tony Stark

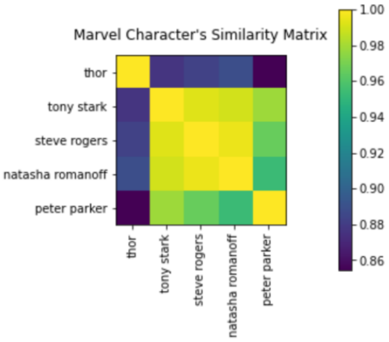


Figure 7: Similarity matrix

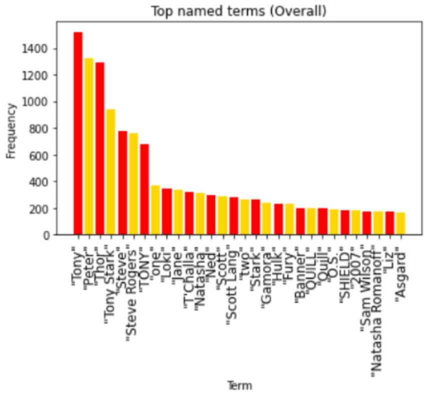


Figure 8: Top entities used overall

us that the characters have very neutral sentiments and their negative sentiments are relatively low. Their low negative sentiment is probably due to the fact that they are superheroes, who tend to be often optimistic and hopeful. Furthermore, we are assuming that the neutral sentiments stem from the fact that a majority of dialogues in movies are often casual with no definite sentiment.

Tony Stark was number one in the top five most script lines, which would result in him having the most sentiment scores and most obvious character development throughout the movies. In terms of the NER, we found it interesting that in the case of the nouns used by Tony Stark, he mentions other characters often such as “Pepper” and “kid”, “Jarvis”, “Steve” etc. This is also observable when looking at the nouns and top entities of Tony Stark. While the nouns and top entities were informative. The same could not be said for the verbs. The top spots were taken by words such as ”get”, “go”, “know”, “come” etc, which can be often found in commands.

This said, we further investigated the similarity between characters using their spoken lines. These results are somewhat expected. Considering that all the movies are part of the same ‘movie universe’ and have similar themes, it is easy to see why characters have a similarity score of 1 for most comparisons. We found it fascinating, however, to see that Thor was one character who has the lowest similarity score among all characters. This is also reflected in Thor’s sentiment score, where he has the second-highest negative sentiment. This likely stems from Thor’s dialogue writers writing his lines to speak in an old-fashioned inspired by Shakespeare plays, thus resulting in a distinct sentence phrasing and linguistic style, comparatively to other characters. Lastly, we took a look at the top named entities overall. In the first place, we had “Tony”, which was not unexpected considering he is one of the central figures of the movie series. Besides the Avengers crew, there are locations such as “Asgard”. As well, a time entity “2007” is mentioned and the organization “SHIELD” is also up there.

5 Conclusion and Evaluation

Given the aforementioned results, it can be concluded that the sentiment analysis scores and the vocabulary of each character provide a clear indication of how similar the characters are to each other as reflected by the similarity matrix. With regards to the needs of the end-user, screenwriters are able to quickly see that characters referencing similarly-named entities and engaging in certain vocabulary have a stronger similarity score.

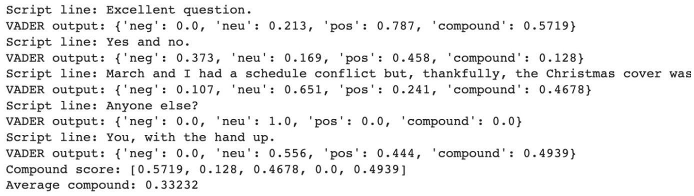
Within the scope of the current project, it did not make sense to create annotated labels for the working dataset, however, for future iterations, labels would allow for a supervised machine learning method to go alongside the existing lexical approach in parallel, for a more comprehensive overall approach to the project. Another improvement could be to include more characters for analysis as we thought it would be interesting to compare the nouns, verbs, and entities used by the villains with the heroes we already have and see how they fare in the similarity matrix.

6 Contribution

Our group has divided the work evenly throughout the labs as well as this project.We made sure that everyone had a part to do in each of the labs. We also had group meetings to discuss our parts as well as our findings. For the project, we started off with discussing the topic we all wanted to focus on and saw to it that each group member had something to do. We decided on the two NLP techniques and decided it would be best for two people to work on one technique. Aimee and Kornpol worked on sentiment analysis. Gargi and Dominique worked on NER. After this we still had the report to do so we decided to split up the sections of the essay equally. student 1 (Aimee): 1) 25%, 2) 25%, 3) 25% student 2 (Dominique): 1) 25%, 2) 25%, 3) 25% student 3 (Gargi): 1) 25%, 2) 25%, 3) 25% student 4 (Kornpol): 1) 25%, 2) 25%, 3) 25%

7 Appendix

7.1 Sentiment Analysis results



```
Script line: Excellent question.
VADER output: {'neg': 0.0, 'neu': 0.213, 'pos': 0.787, 'compound': 0.5719}
Script line: Yes and no.
VADER output: {'neg': 0.373, 'neu': 0.169, 'pos': 0.458, 'compound': 0.128}
Script line: March and I had a schedule conflict but, thankfully, the Christmas cover was
VADER output: {'neg': 0.107, 'neu': 0.651, 'pos': 0.241, 'compound': 0.4678}
Script line: Anyone else?
VADER output: {'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound': 0.0}
Script line: You, with the hand up.
VADER output: {'neg': 0.0, 'neu': 0.556, 'pos': 0.444, 'compound': 0.4939}
Compound score: [0.5719, 0.128, 0.4678, 0.0, 0.4939]
Average compound: 0.33232
```

Figure 9: Input, output and results from the VADER analyser

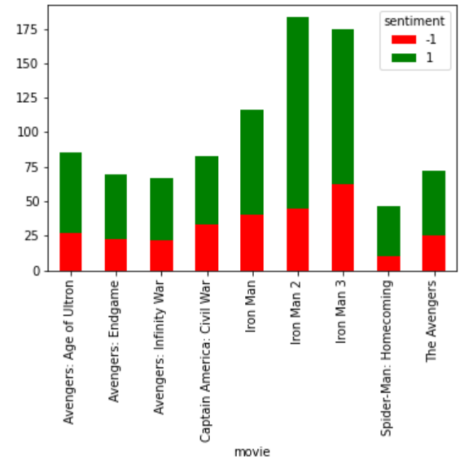


Figure 10: Tony Stark: Positive and negative sentiment frequency in the MCU movies

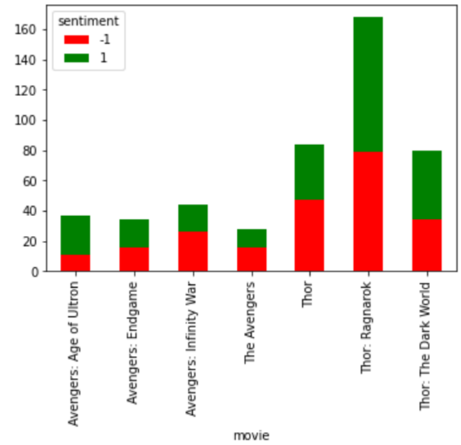


Figure 11: Thor:Positive and negative sentiment frequency in the MCU movies

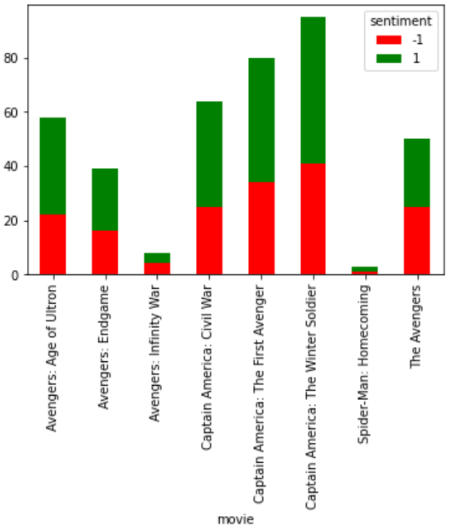


Figure 12: Steve Rogers:Positive and negative sentiment frequency in the MCU movies

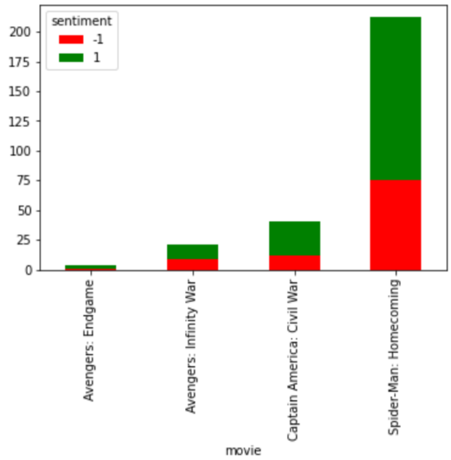


Figure 13: Peter Parker:Positive and negative sentiment frequency in the MCU movies

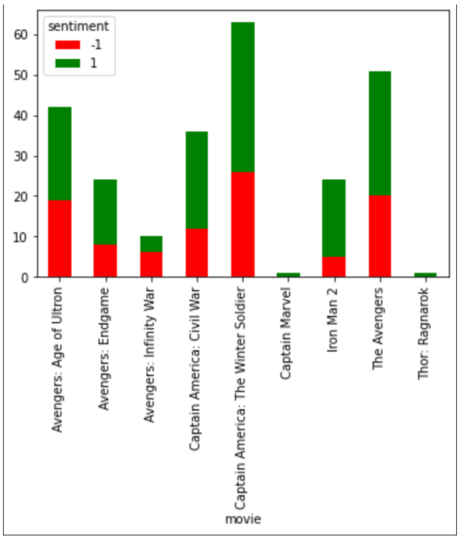


Figure 14: Natasha Romanoff:Positive and negative sentiment frequency in the MCU movies

7.2 NER results

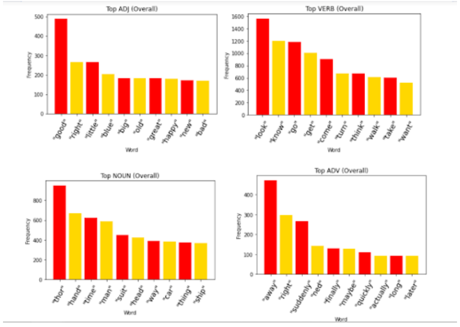


Figure 15: Entity analysis

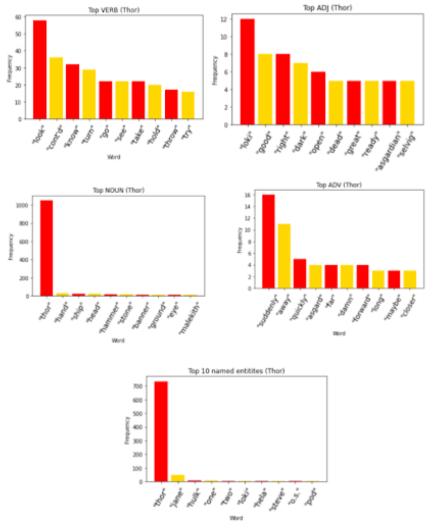


Figure 16: Thor analysis

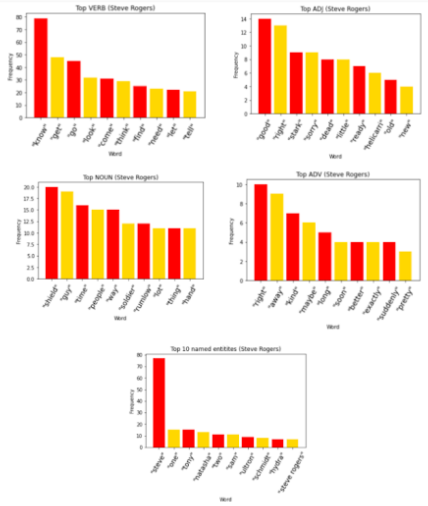


Figure 17: Steve Rogers analysis

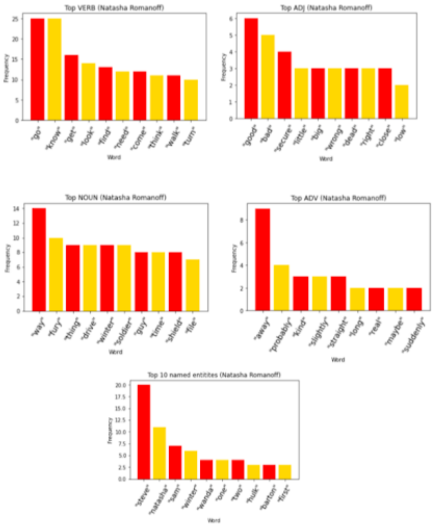


Figure 18: Natasha Romanoff analysis

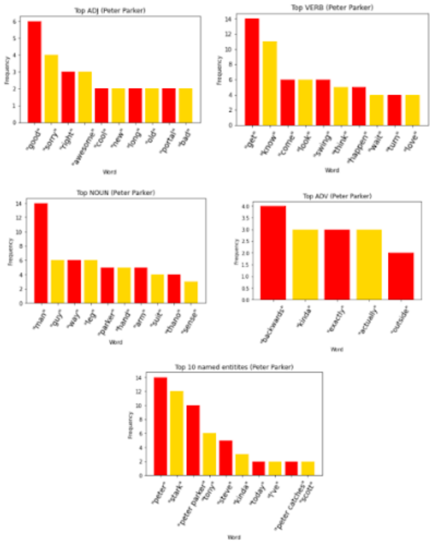


Figure 19: Peter Parker analysis

References

Pranav Hari. 2020. Marvel cinematic universe dialogue dataset.

et al. Hutto, C.J. 2022. Nltk vader documentation.

Daniel Jurafsky and H. James Martin, 2020a. *Speech and Language Processing*, chapter 20. not published - draft.

Daniel Jurafsky and H. James Martin, 2020b. *Speech and Language Processing*, chapter 6. not published - draft.

Laura. 2019. Predicting mbti with pos, nerc, sentiment with svm.

Diana Maynard, Kalina Bontcheva, and Isabelle Augenstein. 2016. Natural language processing for the semantic web. *Synthesis Lectures on the Semantic Web: Theory and Technology*, 6(2):1–194.

Bao Pham. 2020. Parts of speech tagging: Rule-based.

ritakalach. 2020. Using sentiment analysis to visualize character arcs in atla tv series.

spacy. unknown. spacy documentation.

Piek Vossen. 2022a. Lecture 4: Subjectivity mining.

Piek Vossen. 2022b. Lecture 5: Named entity detection and classification.

Enes Zvornicanin. 2022. Ner - named entity recognition tutorial.