Marvel versus DC

5 October 2023

Team 3: Allison Kinnan, Jeremy Moynihan, Connor Paszkiewicz

Introduction and Objectives:

People have always debated between Marvel and DC, comparing and contrasting the two and ultimately deciding one is better than the other. While deciding which one is better is completely objective, finding the true similarities and differences between Marvel and DC movies is achievable through analysis. This can lead to conclusions on what makes both movies successful and what differentiates the two movies in consumers' minds.

To do this analysis, we chose two Marvel movies- Thor Ragnarok and Spider-Man (2002)- and two DC movies- Wonder Woman and Dark Knight Rises- to analyze the scripts of. This gives us enough data to truly compare Marvel versus DC movies. We will perform a sentiment analysis, character frequency analysis, and fight scene analysis. It is theorized that DC movies tend to be darker than Marvel movies. By performing a sentiment analysis, we will test this theory to see if there is a difference in the number of positive and negative scenes in the four movies. Then, we will perform a character frequency analysis. We will look at each time the hero, villain, and love interest appear or are mentioned in each script. Finally, we will use topic modeling to identify what is and isn't a fight scene, comparing the frequency and timing of fight scenes across all four scripts.

Data Set:

We found our data on IMSDb, a website containing the original scripts from a variety of movies. We chose to focus on Thor Ragnarok, Spider-Man (2002), Wonder Woman, and Dark Knight Rises as they represent the range of action movies that Marvel and DC produce. Therefore, if our results are true for these four movies, they should be an accurate representation of the overall Marvel and DC franchises. These scripts include stage directions, scene description, and character lines.

Data Preparation:

In order to perform our analysis, we knew it would be important to break these long scripts into smaller parts. Therefore we took the scripts off the websites and loaded them into Excel, manually marking each scene. This left us with a table showing the Scene Number and the Script. We kept the stage directions, scene descriptions, and character lines as we thought they were all important for use in the three analyses we were performing and that separating them would not be beneficial or harmful to the analytical process. We then loaded this data into Python using the read_excel() function and the lower() function to ensure all words in the script would be in the same format and ready to analyze. Wonder Woman had 113 scenes, Spider-Man had 213 scenes, Thor Ragnarok had 132 scenes, and Dark Knight Rises had 166 scenes.

We additionally loaded each script into a .txt file for when we want to do analyses using the entire script. Here we again kept all stage directions, scene descriptions, and character lines.

This was then loaded into Python using the open().read().lower() functions to ensure all of the words in the script would be in the same format and ready to analyze.

Sentiment Analysis:

We analyzed the sentiments of each scene in all four movies to see if there was a possible difference between Marvel and DC movies in terms of overall mood.

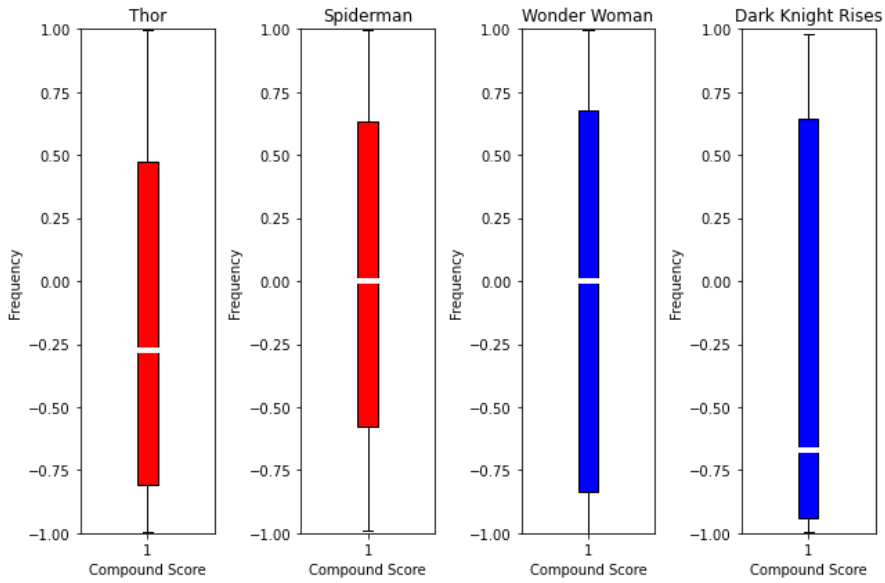
Methods Used:

We did not have labeled data and there were several hundred scenes to label by hand, so we opted for a rule-based approach using the VADER (Valence Aware Dictionary and sEntiment Reasoner) model. We accessed this SentimentIntensityAnalyzer from the NLTK VADER library. This allowed us to very quickly and easily evaluate the sentiment of each scene based on the seven-thousand term lexicon with crowdsourced sentiment scores. We stored the four output values - positive, negative, neutral, and compound - from the sentiment analyzer in the dataframe of scenes. We then created another column that was assigned a value of "negative," "neutral," or "positive" based on the compound score with a neutral cutoff of -0.05 to +0.05. With this data, we calculated the percentage of scenes in each film that were negative, neutral, and positive.

Insights Gathered:

		Negative	Neutral	Positive
Marvel	Thor	53.8%	9.8%	36.4%
	Spider-Man	41.3%	21.1%	37.6%
DC	Wonder Woman	48.2%	7.0%	44.7%
	Dark Knight Rises	59.0%	0.6%	40.4%

We found that Marvel and DC movies were relatively close, with notable observations including Marvel's 12% higher neutrality rate than DC, and DC's 6% higher positive and negative rates. When looking at the individual films, we see that Dark Knight Rises has the highest rate of negative scenes, which makes sense given its far darker tones than the other films examined. We see that Spider-Man has the lowest negative rate, which can be attributed to its kid-friendly themes, represented by the nickname "friendly neighborhood Spider-Man." We can plot the distribution of compound scores via box plots, with each film's color correlating to DC or Marvel, to clearly see these differences between films and how little correlation there is within studios.



One observation of note from this visualization is that all of the movies are more negative than positive, which we believe can be explained by the nature of superhero movies. There needs to be a lot of buildup to the climactic final hero scene, meaning that there will be more negative scenes throughout the film to establish the villain and give the superhero motivation for the final scenes.

Another thing we noticed with the sentiment analysis was that in looking at some of the scenes that were given strong positive or negative compound scores, some of the flaws with using a rule-based approach like VADER came to light. For example, a scene describing a pilot drowning in a plane and then being rescued right at the end of the scene received a strong positive score of 0.9, despite being overwhelmingly negative for the majority of the scene. Overall, however, we felt that the VADER model was fairly accurate and provided a good representation of the films' overall sentiments.

Conclusion:

We expected to see DC movies being more somber than Marvel movies because Marvel markets themselves as more light-hearted and family-friendly, especially with their association with Disney. Instead, we found an insignificant difference between studios, but large differences between movies, indicating that the superheroes' stories play a bigger role in the sentiment of the film than the studio producing the film. We would also need to analyze more films per studio to make a definitive statement about the overall sentiments of the studios, since there is such large variation between the movies we analyzed.

Fight Scene Identification:

We created a model to tag every time there was a fight scene in the four movie scripts.

Methods Used:

To determine which scenes were fight scenes, a hybrid method with keyword matching,sentiment analysis, and character name and recognition. We made sure to use a wide list of keywords for the matching, some more direct (like “fight” or “battle”) but also more indirect terms (like “defend”).

Inclusion of these words, however, doesn’t guarantee that a scene is a fight scene, so another method was used for sentiment analysis. Fight scenes have heightened emotion, so we would expect to see a high sentiment analysis value.

Furthermore, by leveraging Name Entity Recognition, character names were identified. As it takes two characters to have a fight scene, we incorporated this as a factor into determining a fight scene.

These were the libraries used:
Pandas, re, textblob, nltk (ngrams), collections (Counter)

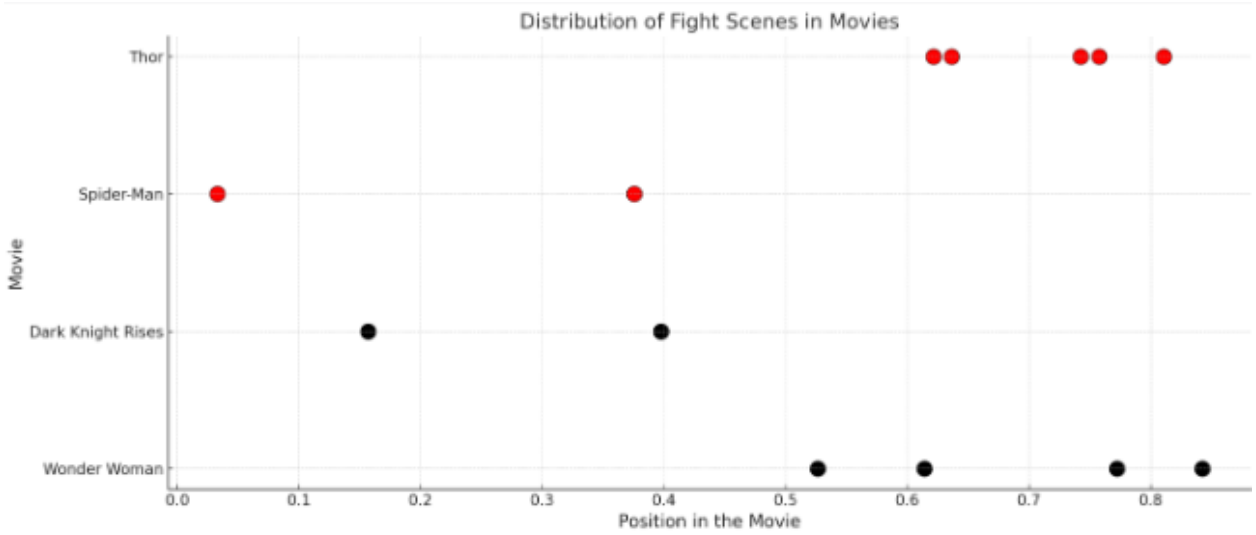
Insights Gathered:

The fight scene counts were very similar:

Movie	Fight Scenes
Wonder woman	4
Dark Knight Rises	2
Spider-Man	2
Thor	5

Dark Knight Rises and Spider-Man had the same amount of fight scnes (2). On the other hand, Wonder Woman and Thor had about double that (4 and 5, respectively).

The fight scene plots between DC and Marvel virtually mirrored each other:



Thor and Wonder Woman have heavy fighting at the end, whereas Dark Knight Rises and Spider-Man has light fighting in the beginning. This analysis shows a higher inter-company similarity than it does intra-company.

Conclusion:

There are fewer scenes that are considered "fighting" than would make sense for superhero movies. I think that in the future, a LLM would much better serve to determine which scenes are and are not fight scenes. However, I would maintain that the rules that guided my rules-based approach (intensity, violence-related words, multiple-characters) should also

Character Frequency Analysis:

We analyzed the amount of times a character was mentioned or appeared throughout the script. We focused on the hero, villain, and love interest for each movie, which are identified as follows:

	Wonder Woman	Spider-Man	Thor Ragnarok	Dark Knight Rises
Hero	Diana	Peter/Spider-man	Thor	Batman/Bruce
Villain	Ares	Ock	Hela	Bane
Love Interest	Steve	Liz	Valkyrie	Selina

Methods Used:

In order to accomplish this task, we first use nltk.tokenize.word_tokenize() to split the data into tokens. We converted this to a text using nltk.text.Text() and then used FreqDist() to

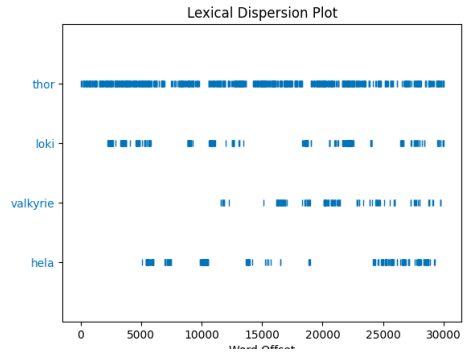
show the overall appearance of the characters in each script. We also calculated the tf and idf of each character throughout their scenes in each script so that we can compare the amount of times they appear and the timing of when they appear. We did both of these by defining our own functions.

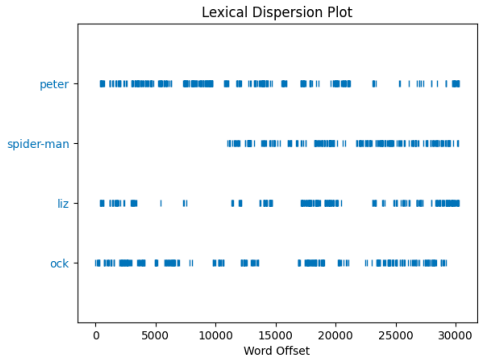
Insights Gathered:

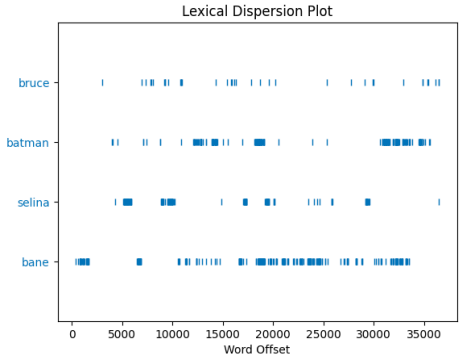
TF, IDF	Wonder Woman	Spider-Man	Thor Ragnarok	Dark Knight Rises
Hero	0.83, 0.18	Peter: 0.36, 1.01 Spider-man: 0.32,1.15	0.62, 0.48	Bruce: 0.16, 1.87 Batman: 0.29, 1.22
Villain	0.58, 1.36	0.33, 1.09	0.31, 1.18	0.17, 0.75
Love Interest	0.25, 0.53	0.25, 1.36	0.23, 1.47	0.46, 1.72

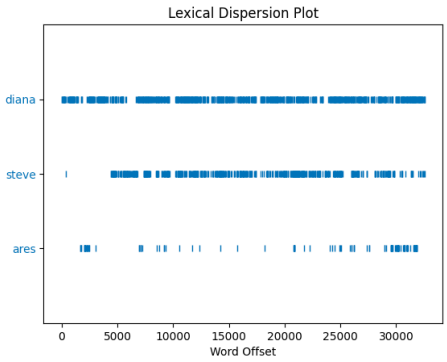
Here we can see that the overall distribution for Marvel movies tends to be the same, after denoting that Spider-man's hero will be split up into two phrases. DC doesn't appear to have a set distribution, it changes from movie to movie. Therefore, there are no clear similarities or differences between DC and Marvel in this analysis.

Then, we moved to an analysis of the timing of the character's appearances.









We can draw a few more conclusions from this. The hero consistently appears throughout the duration of all four films. However, in Spider-man and Batman, they are referred to by their

human name more at the beginning of the movie, and their superhero name more at the end of the movie. Across all four movies, the appearances of the love interest are drastically different. Additionally, the villain shows up at the end of all four movies for the final battle, but otherwise has varying appearances across all four movies. Again, we don't see clear similarities or differences between DC and Marvel.

Conclusion:

We can conclude that with regards to character frequency, there is no clear difference between DC and Marvel, but there are differences from movie to movie. A difficulty with performing this analysis is that some characters are referred to by multiple names, and that some characters are hard to define what archetype they are. Therefore, it is important to have clear definitions and decide on which characters are important early in the analysis.

Overall Conclusion:

From our analyses, we can conclude that DC and Marvel films are not as different as we expected. The true differences lie between individual movies, not production companies. In order to find stronger trends between DC and Marvel movies, we would need to analyze more movies for each production company. We have also noticed that classification models are not always accurate, but can still provide useful insights to support our conclusions.