Movie Reviews Under Review

Part 1: First Vectorized Representation

Aubrey Molitor

MSDS 453: Natural Language Processing

Introduction

Beginning the process to compare and classify movies based on their reviews, it is necessary to determine what the basis of that comparison should be. Even movie reviews for the same movie can have disparate views, word choices, and details. So, understanding which words are most important to identifying the entities spoken about, as well as the larger film context in which those movies occur, is key. There are two steps to this process: First is to identify key words that are able to both classify and relate documents. Second is to codify words, then documents, to relate the differences and similarities between documents, forming the basis of relational analysis. To do this, there are several methods worth comparing, including Word2Vec, Doc2Vec, and ELMo embedding. This document explores the advantages and disadvantages of each approach and draws conclusions based on these experiments.

The Data

The data set for this project is the class corpus data set, compiled of ten reviews for each student in the class. Each review has a minimum of 500 words, and half of all reviews are classified as positive whereas the other half are negative. For identifying keywords and embeddings, a process of data normalization was done that included removing punctuation, lower casing all words, removing tags and special characters, and stemming the words.

Research Design and Methods

Part 1: Qualitative Approach

The first step, as mentioned in the introduction, was to choose words that are necessary for both classifying the Toxic Avenger movie with others and distinguishing it from others. Based on a qualitative analysis of the data as completed in the previous weeks, the following were chosen as important words: fil, waste, superhero, classic, blood, remake, horror, satire, budget, illness, father, gore, funny, and Troma.

These words each capture one of two descriptors present in the review. One is the thematic focus of the movie – a superhero-esque, gory film with hints of climate change-related illness and a heartening father-son story line. The other is the general analysis of the film, describing it as a remake of a classic film in the Tromaverse, sitting on the line somewhere between out-of-line comedy and all-out gore.

When narrowing these down to prevalent terms, these main descriptors were the basis of the decision. Prevalent terms required to be present in multiple reviews and were chosen to be

most important for classification and differentiation between documents in the corpus So, these four words were chosen to capture the essence of the short descriptions above:

- Waste – an environment-focused film that encompasses the often-discussed "toxic waste"
- Superhero – a subgenre of film, great for classification, that links The Toxic Avenger to the genre
- Horror – a general descriptor for the genre of The Toxic Avenger that helps classify the movie
- Father – a specific word for a key relationship in the film, which could link to non-horror films

Part 2: Evaluate Code Results

The experiment design for the step two section of the research is to evaluate the differences between TF-IDF, Word2Vec, Doc2Vec, and ELMo embeddings in identifying similarities and differences between documents, as well as any discernible trends. For the TF-IDF, all documents will be analyzed to identify common terms and those with highest and lowest TF-IDF scores. Additionally, the TF-IDF scores for the prevalent words will be examined and evaluated accordingly.

For the Word2Vec and Doc2Vec experiments, the documents will be examined with an embedding size of 100 and then 200 to compare how results behave in general, and also how conclusions change with different embedding sizes.

For the ELMo embeddings, the texts will be processed in batches of 4 and 6 at a sample size of 20, and then at a batch of 4 with sample sizes of 40, 60, 80, and all documents to see how both batch and size differences change the t-SNE plots and embeddings heat maps.

The Results

TF-IDF

The highest TF-IDF for the corpus in general, unsurprisingly, has very general and indistinct words. "Film" is at the top, followed by similar general movie descriptors, as well as broad verbs like "has" (ha after stemming) and "like".

Since none of the chosen prevalent words were on here, I decided to make a separate table including these words and their TF-IDF scores. Additionally, I took the average of all of the mean TF-IDF scores to have a point of comparison for the prevalent terms:

It was expected that horror and superhero would have the highest TF-IDF of the prevalent terms though not necessarily in that order. The superhero genre is widespread in itself and among

references within other genres, so a higher TF-IDF was expected for this term. Horror is also a very general and widely used descriptor for the genre it refers to, so seeing this term's higher TF-IDF – especially where other horror films are considered – is unsurprising. Waste and father were not too far behind "superhero," though they did come out last as expected. The bottom two are very close to each other and the difference between them may not be statistically significant. It was expected that "father" would be higher than "waste," though the stemming of the latter term could have aided its higher ranking. In general, though, all of the prevalent terms had a higher average TF-IDF than all of the 7518 words combined, whose mean was .126. This indicates that the prevalent terms could have classification power by connecting some documents together while still leaving room for them to be distinct from others.

Word2Vec

For each iteration of the Word2Vec T-SNE, the words related to The Toxic Avenger tended to be towards the top or top right of the graph. They usually clustered together, with the word "brain" slightly lower, and the word "horror" higher or to the right. After adding "gore," this word tended to be closer to the middle and more similar to other words. The word "bloodi" tended to be somewhat close, but surprisingly "waste" and "environment" tended to be on opposite sides of the graph. The words farthest from the chosen words tended to be "modest," "affleck," "wardrobe," and "mytho." These were unsurprising, as thematically, these words were unrelated to both the prevalent and important words from The Toxic Avenger.

Changing the size of the Word2Vec iteration seemed to make the results slightly more accurate. With a size of 100, the chosen words were closely clustered, and there was not much differentiation relative to the other words besides their clustering in the top or top right. After changing the iteration size to 200, the chosen words spread out more amongst themselves, making the noted differences with "horror" and "brain" visible, and also had more traceable connections to the words such as "bloodi," "grotesq," and "disgust." As a result, the Word2Vec shows a commonality between the Toxic Avenger chosen words and those related to horror and violence.

Doc2Vec

The Doc2Vec is useful for examining which documents are related, suggesting that the reviews for the same movies would be closer on the graph. However, this was not necessarily the case. In several runs of the Doc2Vec experiment, The Toxic Avenger, Dirty Grandpa, 21 Jump

Street, and others all had documents scattered in opposite ends of the plot. Additionally, there was hardly any trend to be found among movies of the same type, suggesting that thematic specifics of the movies might link the documents more than the overarching genre. For example, The Toxic Avenger documents were commonly found to be near those of the action film No Time To Die, as well as the comedy film 21 Jump Street, despite the genre differences between the three. Additionally, The Toxic Avenger's documents, compared to the other movies, seemed to have more distance from the general cluster of the group, suggesting a more distinct theme to the movie. It is also worth noting that the experiment with an embedding size of 100 was generally more scattered around the plot, with a loose trend from one corner to the opposite visible among the entire group. However, with the 200 embedding size, there was a much sharper line from one corner to the opposite consisting of all the scattered documents and genres. This suggests a clearer relational mapping present with the 200 embedding size.

ELMO Embeddings

With each run of the ELMo experiment of sample size 20 and batch size 4, the t-SNE plot showed a very unrelated cluster of dots; there were hardly any overlapping dots, and each dot had roughly equal space from the other. With such a small sample of the documents, it seems much less likely to find ones similar enough to be perceived as the same genre, let alone the same movie. After increasing the sample size to 40, there started to form some clusters of dots. The points were still spread around the whole graph, but, for example, there could be seen a small cluster of dots to the mid right, and a couple of other clusters among the broader group. Additionally, a slight trend of corner-to-opposite-corner grouping started to emerge for the entire group of documents. Increasing the batch size from 40 to 60 did not have much of a change. Moving to a batch size of 80 seemed to make the general group more circularly clustered, and there were several pairings of points that began to overlap. Regarding the heatmap for each of these batch sizes, most documents had 80-90% cosine similarity with each other. However, there were a few standout documents, seen with dark streaks throughout the graph, that bore relatively little resemblance with the other documents, closer to 65-75%. 21 Jump Street, The Suicide Squad, and Dirty Grandpa seemed to have some of these darker lines. Interestingly, though, reviews from these movies were not necessarily similar to each other either. As an example, one text review for Dirty Grandpa had a very light intersection with one of the 21 Jump Street, but the same Dirty Grandpa review had an extremely dark intersection with a different 21 Jump

Street review. This could very well signal the differences in positive and negative reviews influencing the makeup of the heatmap.

The final run of the ELMo experiment, with all documents, was not much different than the sample size of 80. The plot seemed quite spread out, and still had some, though few, overlapping points. There was not a distinguishable pattern or even sections that might signal any particular relationship by genre or movie other than the closest clusters within the larger, spread-out group. The heatmap, with several more intersections, had quite a few lines signaling the differences in meanings for different reviews. Drag Me To Hell had a few stand-out reviews, along with Everything Everywhere All At Once and Oblivion. Noticeably, these are not all the same as some of the other stand-out reviews in the lower samples, suggesting that the smaller sample sizes did not always capture the extremes of the entire data set.

Conclusions

Overall, the prevalent words chosen for the Toxic Avenger movie – waste, superhero, father, and horror – seemed to have strong potential to both classify and distinguish the film and its reviews from others in the corpus. The TF-IDF for each of these was not too high, but above average and in a reasonable order for drawing limited document similarities. For the experiments, Word2Vec the words for The Toxic Avenger seemed to have a reasonable cluster and still had some outlying words like “gore” and “horror” that provide further distinction. However, Given that other words like “bloodi” were similar in theme but farther from the Toxic Avenger words, the Word2Vec did not have a great record for classifying words outside of the Toxic Avenger. For the Doc2Vec experiment, the document results seemed to be mixed. There were some clusters of documents for the same movie, but on different sides of the plot, which suggests the play of the positive and negative reviews in classifying the documents and movie similarities. Additionally, though each iteration had differences in where movie reviews were placed, some similarities could be seen, such as with The Toxic Avenger and 21 Jump Streets’ similar placements. However, this connectivity did not apply to movies of similar genres. Lastly, for the ELMo embeddings, changing the sample size did not influence the t-SNE plots to a great extent, and did not provide significant information for distinguishing movies. Only the heatmap gave some suggestions as to very stand-out reviews, but otherwise provided little information.