Assignment 1: First Vectorized Representation

Sonny Desai

January 2020

Introduction & Problem Statement

The purpose of this assignment is to find words that are important and relevant and test them out using *tf-idf* and eventually Word2Vec and Doc2Vec output. First I evaluated words that I manually pulled from the seven movie review documents I got from the internet. I chose the last seven Academy Award Best Picture winners as my chosen seven films. From these seven documents, I chose 20 words manually that were a combination of relevant words to the film industry, and words my colleagues had chosen as relevant to their film reviews which all happened to be different films than the seven I chose. From here we ran external text analyzers on the words, with the top words becoming the corpus from which we trained, tested and found words for our model. I then ran the models that were provided by Paul in the Dimension rection, and compared the words I selected to the results of the two models' predictions.

Data

The data that was used in this assignment came from the seven movie reviews I provided as well as the other movie reviews my colleagues provided. This is where the data used in the models was formed. However, prior to this, I manually selected 20 words which I felt were relevant to the film industry. Some of these were common or similar in theme, but others were more tied to emotion. I also took into my colleague's suggestions in their own choices of 20 words. I then ran the data through two online text analyzers, Termine and FiveFilter, which provided results about which were the top words chosen and had the highest repetition value. Most of the data seemed to have low scores across the board with the text analyzers. However a few words had some scores in the analyzer. There were only a couple words that had similarity amongst all of the reviews, which is more likely to be caused by the vast differences in the themes and subject matter of the films that were chosen. Part of my choosing of the manual words were terms that I felt were similar along movie reviews and pertinent and relevant to the film industry. I also looked to guidance from the words chosen by my colleagues. IN ultimately choosing the three words I chose, I focused on the words that had the most repetition amongst the documents and those that had the highest text evaluation scores. This was pretty easy since a large amount iof the words I chose did not have consistency along all of the documents, and only a few had text analysis scores. Having determined these words were important (due to the original choice) in the manually formed words document, I decided that these were also prevalent, and in my words case the only ones that were prevalent. The words I chose are: film, story, and love. This was then compared to the results of the data of the words from the corpus which were formed from the models that were mentioned above.

Research Design and Modeling Method(s)

The two models that were chosen to test the corpus of data are Word2Vec and Doc2Vec which come from the genism data library. Word2vec is a technique for natural language processing. The word2vec algorithm uses a neural network model to learn word associations from a large corpus of text. Once trained, such a model can detect synonymous words or suggest additional words for a partial sentence. As the name implies, word2vec represents each distinct word with a particular list of numbers

called a vector. The vectors are chosen carefully such that a simple mathematical function (the cosine similarity between the vectors) indicates the level of semantic similarity between the words represented by those vectors.

Doc2Vec model, as opposite to Word2Vec model, is used to create a vectorized representation of a group of words taken collectively as a single unit.

To predict and further test the words we used: TfidfVectorizer and k means vector. The TfidfVectorizer uses a in-memory vocabulary (a python dictionary) to map the most frequent words to features indices and hence compute a word occurrence frequency (sparse) matrix. The word frequencies are then reweighted using the Inverse Document Frequency (IDF) vector collected feature-wise over the corpus.. The K means clustering put the documents into separate clusters for analysis.

Results

Of the words that I tested, film was the highest rated amongst all the seven film reviews which had been broken down into cluster 1 from the k means tlf dlf analysis. The main words that were gathered from there were: external, links, spotlight, family, boston, chiron, about, friend, inarritu, years. None of the words I chose were included in this final prediction. Story was found in only two reviews and love was only found in one.

Analysis and Interpretation

The results from the tlf vectorizer are not surprising. The seven films that won best picture the last seven years have very different themes, actors, and directors. I tried to choose words that would be similar enough that they are widely held in the film industry, but only one was universal, the word film itself. The top words that were chosen by the analyzer seem to be very specific to certain reviews and not others. This tends to suggest that certain words were more important and used more in certain reviews than other words were used in other reviews. This can be gleaned rom the words like Spotlight, which is the name of the film, Inarritu, the name of the film director of one of the films, and years, which is part of the name of one of the films. This suggest that in these individual reviews, these words played heavily, and might have done more so than in other words. Another issue was the different authors writing the different reviews. Some authors have certain styles and themes which they focus on, but having seven different authors write reviews over seven different years will likely lead to difficulty finding words in common over seven reviews. Still the word film, although not highly used in the reviews, was found in nearly all of the reviews, and the word story was also found in a lot of reviews, which suggest that there are some similarities amongst reviews and words used.

Conclusions

It is very difficult to find similar words for seven different move reviews of seven different films written by seven different authors. However, although none of the words that were chosen as the top words were the ones I selected in the manual review, the words film and story were still highly relevant in the vast majority of the reviews. Lastly, certain words are highly emphasized in certain reviews, like the

name of the film, the actor's name, or director's name which heavily skews the text analysis towards the overall tlf analysis, however they might be more focused in their singular reviews rather than being an overall analysis.