Assignment 1: First Vectorized Representation

Vivian Xia

Northwestern University

MSDS453: Natural Language Processing

Syamala Srinivasan

January 23, 2022

Introduction & Problem Statement

Natural language processing utilizes text to create a representation or understanding of the words for computers. Because computers have no semantic understanding, the text must be processed and encoded as vectors to carry that meaning with them. The goal is to develop a corpus vocabulary that encapsulates the meaningful and important terms that would best represent the documents, so that they can be clustered with their corresponding genre and sentiment.

To accomplish this goal, different data wrangling methods will be experimented to see which best processes the texts for the intended purpose. Word2Vec and Doc2Vec models will be used to create word and document embeddings, in which the similarity between tokens and documents can be observed. The similarity will show if the tokens or documents are able to be clustered and classified with one another.

Data

The corpus includes reviews from 25 movies, consisting of six action, horror, sci-fi and seven comedy movies. There were five positive and five negative reviews collected for each of the movies, resulting in a total of 250 documents in the corpus. The tokens within the document were extracted. But needed to be preprocessed before adding it to the vocabulary.

The libraries pandas and re were used in preprocessing the text by removing punctuation, non-alphabetic characters, short tokens, and changing the words to lower case. Using the Python library NLTK, the text was preprocessed using methods such as tokenization, stemming, lemmatization, and stop words. A list of stop words was downloaded from NLTK. The stop words were removed from the text before using the other preprocessing methods. Tokenization used Python's split() function to separate the words by white space. Python's string.punctuation,

a list of punctuation characters !"#$%&'()*+,-./:;<=>?@[\]^_`{|}~, was referenced as a list of punctuation that needed to be removed. Regex was used to replace those characters with nothing ("Clean text," 2018).

Stemming is a method to reduce the number of tokens in the vocabulary by removing ending word affixes such as plural and future tenses. PorterStemmer from the NLTK library is used for stemming. This method, however, results in tokens that are not real English words anymore and there can also be errors due to the over stemming and under stemming. Lemmatization is another way to approach reducing the number of tokens using roots. This method converts the words to a meaningful real English base term. It is also from the NLTK library. These libraries are used to preprocess and manipulate the text to reduce unimportant tokens to the corpus vocabulary (Srinivasan, 2022b).

These tokens are then represented numerically through term frequency-inverse document frequency, td-idf. This numerical score vectorizes the frequency of the word in a document (Srinivasan, 2022b). The td-idf scores are calculated using the sklearn library TfidVectorizer. Each document is exported with the token and its score.

Using the Genism library, a Word2Vec model can be built using the continuous bag-of-words method. This model uses the one-hot encoded vectors of the words to create continuous and dense vectors that can predict the target word using context to give semantic meaning to the words. Those with similar embedding vectors have similar meanings. Using TSNE the sklearn library, the dimensions of the embeddings are reduced to 2 to plot on a t-distribution curve. The package for cosine similarity from the sklearn library is used to visualize a heatmap based on the cosine similarity between the embedding vectors (Srinivasan, 2022a).

Another type of embedding was done with the documents rather than the words. The Doc2Vec package from Genism was used to create a distributed bag of words model. Doc2Vec is the average of all the features from word embeddings to get document vectorizations. The Doc2Vec model outputs embedding vectors for each document. These embedding vectors are plotted on a t-distribution curve using TSNE and also on a heatmap for cosine similarity (Srinivasan, 2022a).

Research Design and Modeling Method

Before performing any processing among the documents, a qualitative approach was taken to analyze and identify terms in the ten documents I collected that I thought would be a good addition to the vocabulary. My ten documents were reviews on *Guardians of the Galaxy* in the science fiction genre. I paid attention to the entities within my documents. I also considered whether the terms were too specific for my own documents or too broad that they would not be significant in helping identify which cluster the documents belong to.

Looking at those documents, I picked the terms “Star Wars,” “planet,” “alien,” “space,” and “spaceship.” These five terms appear a few times in the ten documents, indicating that they are both important and prevalent. “Star Wars” refers to a major science fiction film that is often mentioned or brought up when it comes to the science fiction genre, so it could be helpful in identifying the film’s genre. The term “Star Wars” with its surrounding terms could contribute to identifying the sentiment analysis of whether it is a positive or negative review. The other terms are related to the science fiction genre and can be used to identify such genre.

To analyze the tokens from different methods of preprocessing, three data wrangling methods were used to clean the data. The data wrangling used in Method 1 consisted of removing punctuation and short tokens with three characters or less and normalize the tokens to

lowercase. Method 2 used Method 1's methods as well as stemming and removed stop words. Similarly, Method 3 also used Method 1's methods but used it with lemmatization, removal of stop words, and removal of non-alphabetic tokens.

For each method, the tf-idf was calculated. The tf-idf scores for each of the qualitative approach terms as well as tokens that were among the highest tf-idf mean scores were recorded.

To create a Word2Vec model, each word in the text is processed into dense vectors. Each method experiments using 100, 200, and 300 embedding dimensions. For visualization, a subset of one hundred random tokens were picked and their dimensions reduced to two, so they can be plotted in a t-distribution curve. A heatmap is also plotted to see the cosine similarity between the embeddings.

Like the Word2Vec model, the Doc2Vec model produces dense vectors for the 250 documents. But these vectors represent documents. The vectors can be plotted in a t-distribution curve and a heatmap to see how similar the documents are to each other.

Results

Method 1

The number of tokens was over 890 thousand after preprocessing. There were some punctuations that were missed in the cleaning of the text. There are a lot of tokens but not all of them are meaningful or important such as "it's" and "didnt" and "2003."

The tf-idf scores are calculated for each of the tokens in each document. The tf-idf scores for the terms identified in the qualitative approach are shown in Figure 1. The scores were recorded for each term in each document to compute the average score of the term. Figure 2 shows the ten terms that had the highest tf-idf average scores for the ten documents I gathered. The score of each term in each document was taken and, for the same terms, the scores were

averaged together. Then the scores for each token was divided by ten to get the mean score of each term for the ten documents.

Figure 3, 5, 6 shows a visualization of a subset of tokens that have been embedded from the Word2Vec model of 100, 200, and 300 embedding dimensions in a t-distribution curve, respectively. In addition, Figure 4 shows the Word2Vec with 100 embedding dimensions heatmap for cosine similarity. The heatmap for 200 and 300 embedding dimensions look very similar to Figure 4 as well. There is no clear clustering from these plots and heatmaps.

The embedding vectors of Doc2Vec with 100, 200, and 300 embedding dimensions are visualized using TSNE to reduce the dimensions to 2 and plotted on a t-distribution curve. A heatmap for cosine similarity of these embedded vectors is also created for each of the embedding dimensions of 100, 200, and 300. There are no clusters seen from the plots. The heatmaps show the only red areas are the diagonal, which show no similarity between other sets of documents other than themselves. The two sets of *Red Notice* documents are the exception.

Method 2

The preprocessing of the text in Method 2 tokenized, normalized, and removed short characters and punctuation like Method 1's as well as applied stemming of the text and removed stop words. These additional preprocessing methods reduced the tokens from over 890 thousand to 666 thousand. The td-idf scores are recorded for the qualitative approach terms and the highest scoring terms in Figure 1 and 2. The visualization results of the embeddings show that there are no clusters between the tokens or between the documents that makes sense.

Method 3

This data wrangling combination of methods include tokenization and normalization as Method 1 as well as lemmatization, removal of stop words and non-alphabetic tokens. The

number of tokens decreased to 751,000. The total is less than that of Method 1's preprocessing but greater than Method 2's preprocessing. The plot for the Word2Vec and Doc2Vec model was successful in showing some clear clusters. For the embedding dimensions of 200 and 300, there were clusters present. The heatmap did not show any similarities between the documents.

Analysis and Interpretation

Method 1

With this method, there were tokens such as "thatll" and "isnt" that were contractions before the cleaning. However, these tokens and their original combination of words do not hold much information anyways. Some apostrophes were also not removed, so there are tokens such as "there's" and "o'brien" and there are also quotation marks like "'pacific" where there is one set of quotation marks at the beginning of the token. There are many tokens such as "13bn" and "2021all" that do not have any actual meaning as well. Hyphens between words have been removed, combining the two terms into one token such as "reallife." Some hyphens or em dashes were missed during preprocessing.

Guardians of the Galaxy Documents

While looking for the td-idf scores of the terms identified in the qualitative approach, there were a few variations for each term. Because of preprocessing, "Star Wars" turned to "star" and "wars." To find the td-idf score for the term "Star Wars" in the documents, the individual td-idf scores for "star" and "wars" was found and added together for each document. The overall average score for the term using Method 1 was computed using the sum of the scores of each document. Documents 3 and 10 had a tf-idf score for "star" but used it to refer to "Star Trek." The context of "Star Trek" is that of another reputed and well-known science fiction film, as recognizable as "Star Wars." Because "Star Trek" represents the same meaning as "Star Wars,"

the term “trek” was also considered. Therefore, the scores for “star,” “war,” and “trek” were considered as having the same meaning as “Star Wars.” There is also the instance that “Star Wars” was present in Document 6 but so were terms such as “star child” and “real-life star.” The “star” td-idf score reflected not only “Star Wars” but also other contexts. “Star Wars” was only occurred in the document once, so the “star” score was divided by three to account for only the term in “Star Wars.” The preprocessing eliminated the need to account for terms such as “Star-Lord” as the removal of punctuation changed the turn to “starlord.” For instances where the document had “Star Lord,” the document did not have “Star Wars,” so a score of zero was given for those documents.

Generally, the term “planet” was used only in the intended context within the documents. There was one document with the term “planet” but referring to “Planet of the Apes,” so the term and its score was not considered for that document. A variation was observed for the term “planet” in “planetindanger” from the original term of “planet-in-danger.” Because “planetindanger” refers to a similar context as “planet,” the term and its score was considered for the term “planet.” There was a term “planetary” in a document that was in the context of “planetary danger” which seemed similar to “planet-in-danger” from the other document. Because “planet-in-danger” has a similar meaning to the context of “planetary,” “planetary” was also taken into account and its score was summed with that of the “planet” term also present in that document.

For the term “alien,” the plural “aliens” and singular “alien” scores were considered and added together when it occurred in the same document. Similarly, “spaceship” had both plural and singular versions that were considered as well as a synonym “saucer.” The term “space” was used as is for all the documents. For one document, it was used as an attributive noun to describe

multiple other nouns such as “space opera” or “space pirate,” but was under the same intended context as “space,” so the score was left unchanged.

The terms that had the highest tf-idf average score included the terms “guardians,” “galaxy,” “marvel,” “quill.” The terms “guardians” and “galaxy” are in the film title, so it was expected that it would get a high tf-idf score. The other two terms “marvel” and “quill” are used quite often to refer to the comics and main character, respectively. However, “guardians,” “marvel,” and “quill” may be too specific to the film. There is no connection that can be made to Marvel Comics within the science fiction genre movies in the corpus. The character “quill” does not appear in the other movies.” The other terms “galaxy,” “ideological,” “formulaic,” and “flying,” are terms that may be able to be related to other science fiction movies. The terms “prologue,” “ikea,” and “values” seem too broad to make a connection with other movies.

The qualitative approach terms do seem relevant and important in the corpus vocabulary. The tf-idf scores for each of the terms all average out a score larger than the top ranked tokens in the ten documents, representing the weight or importance of those terms in the ten documents. This is due to the semantic understanding of gathering the scores of the terms manually because there were many variations including synonyms and plural of the terms that had separate scores. The terms picked do seem important enough to be used for clustering and classification, especially “alien,” “star wars,” and “space.”

Word2Vec with 100 Embedding Dimensions

In Figure 3, the plot for the subset of tokens from the Word2Vec shows which tokens are similar to each other in meaning. The more similar the word, the more similar the vector. For example, the terms “incapable,” “putdowns,” “gullible,” and “shrugging” are close to each other in the feature space. The terms “story” and “director” are very close to each other as well but far

from the terms “churn” and “chop,” which make sense as those pairs of words are not as similar to the other in meaning. Clusters represent certain features that the words have in common, but there are no clusters that can be identified in the plot. Figure 4 shows the heatmap for cosine similarity. There are no clear clusters. The coloring of the heat map varies throughout.

Word2Vec with 200 Embedding Dimensions

This plot, shown in Figure 5, looks very similar to the plot with 100 embedding dimensions. There are no clusters in this plot except a small one at the lower right side. That one small cluster includes terms such as “dodging,” “seeds,” “gullible,” and “methods,” which are not similar in meaning. The heatmap for cosine similarity also do not have any evident clusters.

Word2Vec with 300 Embedding Dimensions

There are no clear clusters in this plot, Figure 6. There are a few separations between chains of terms. Like the 100 and 200 embedding dimension plots, this plot is very linear where one term is close to one or two terms but not a group of terms and no clear clusters.

Doc2Vec with 100 Embedding Dimensions

The plot with the embedding vectors for each document shows no clear clusters. The documents are distributed all over the plot. In general, the plot is very noisy. The heatmap shows that there are also no clusters or similarities between other documents other than themselves. The only red parts of the heatmap are along the diagonal. The only discrepancy is that there were two sets of ten reviews of the same movie, *Red Notice.* This resulted in similarity detected between the two which can be seen in Figure 7. Otherwise, the heatmap shows no similarity between other sets of documents.

Doc2Vec with 200 Embedding Dimensions

This plot does not show any clear clusters of documents. Three of the documents with the movie *Red Notice* make up the only cluster in this plot, but that is expected as reviews on the same movie. The heatmap looks very similar to the preceding heatmap. There are no similarities between other sets of movie documents other than themselves. Again, the only similarity is between the two sets of *Red Notes* documents.

Doc2Vec with 300 Embedding Dimensions

There are no clear clusters in the plot. The documents are distributed all throughout the plot. Similarly, the heatmap also does not show any similarities between other sets of documents, apart from the two sets of documents for *Red Notice*.

Method 2

The use of stemming changed the terms to tokens that are not actual English words such as “arriv” and “qualifi.” Tokens that contain information are left in the vocabulary after removing the stop words. The plural versions of terms were reduced to its singular form. There are no missed removal of punctuation for this method.

Guardians of the Galaxy Documents

The td-idf scores for each of the qualitative approach terms was recorded and averaged between the ten documents. The resulting average scores for each term was smaller compared to the corresponding average scores using Method 1. The same pattern of smaller values using Method 2 is also seen for the terms for each document. For example, Document 1 had a score of 2.88 for “alien” using Method 1 and a score of 2.67 using Method 2. These smaller values were a result of adding stemming and removing stop words.

Stemming made it easier to look for each term because it changed the plural terms to its singular form. This preprocessing method made it easier to look up the scores for “alien” and

"spaceship." The total tf-idf score corresponded to the singular form, so the plural forms did not have to also be looked up and added together with its singular forms' scores. For "planet," the variations that were included as discussed in Method 1 changed from "planetary" to "planetari" and "planetindanger" to "planetindang." These variations continued to be considered like Method 1.

For "Star Wars," the term "star" still existed within the documents that it was present in, but "wars" which was preprocessed to become "war" was given the score of zero for every document. This score of zero greatly reduced the scores of "Star Wars" in the documents and overall average score in the ten documents. The term "trek" still received a score, though, like the others, it was a lower score than its corresponding Method 1 score.

The terms with the highest tf-idf scores using Method 2 are shown in Figure 2. The tokens are similar to the tokens from Method 1. There are some changes in the ranking order and two of the tokens. The tf-idf score of "ideological" gained weight among the ten documents as well as "guardian." The tokens "galaxy" and "prologue" became "galaxi" and "prologu" but kept the same score. On the other hand, "formulaic" became "formula" and also decreased in its score. Similar to Method 1, the terms "guardian," "marvel," and "quill" are too specific to the film and do not seem like they will be helpful in clustering and classification with other film reviews. On the other hand, the tokens "galaxi," "ideological," "formula" can potentially be useful. The other terms "adult," "prologu," "ikea," "hypocrisy" are too vague to be used to connect to other film reviews.

Despite the smaller scores compared to Method 1, the average scores for the qualitative terms are still larger than the scores of the ranked top ten. This shows that these identified terms are useful and important in the documents to be used for clustering and classification.

Word2Vec with 100 Embedding Dimensions

There is some clustering in this plot as shown in Figure 8. There is an evident small cluster at the top right of the plot with tokens such as “name,” “begin,” and “almost,” There is a feature that clustered those tokens close together, although they do not have similar meanings. There is a cluster in the middle of the plot containing “sham,” “sarah,” “substitut,” “seal,” “frontier,” “openly.” These tokens also do not seem to have very similar meanings, so the clustering, where there are some, is not great. The heatmap for the cosine similarity also does not show clusters, representing that the tokens are not very similar to each other in meaning.

Word2Vec with 200 Embedding Dimensions

This plot, Figure 9, is similar to the preceding plot. There is not much evident clear clustering. There is one cluster in the middle. It contains tokens such as “redempt,” “network,” “superflu,” “outcome,” “dodg,” “vibrant,” and “fide,” which do not have similar meanings. The clustering is not great. The heatmap for cosine similarity also do not have any clear clustering of any of the terms.

Word2Vec with 300 Embedding Dimensions

There are no evident clusters between the tokens in the plot or in the heatmap. The plot has no separations between tokens and forms a long chain of tokens. The heatmap also does not show any darker coloring in one area, so there is no similarity between the tokens.

Doc2Vec with 100 Embedding Dimensions

There is no clear clustering from the documents except for *Red Notice*. The documents are all scattered and distributed throughout the plot. The heatmap for cosine similarity, Figure 10, is similar to Method 1’s for 100 embedding dimensions for Doc2Vec. There is, however,

some lightening of the surrounding cells other than the diagonal. This shows that there is a bit of similarity that can be seen from the documents with other documents compared to Method 1.

Doc2Vec with 200 Embedding Dimensions

This plot also has no clear clustering of the documents. The heatmap is similar to the preceding heatmap. There seems to be a little similarity that can be found between some of the documents due to the red color sneaking through the black.

Doc2Vec with 300 Embedding Dimensions

Figure 11 shows some little clusters forming in the plot. There is a cluster at the top of the plot with four of the *Pirates of the Caribbean: The Curse of the Black Pearl* documents, one *Pacific Rim* document. It makes sense that the *Pirates of the Caribbean* documents would be clustered together. *Pacific Rim* is a science fiction and action movie, so the clustering of these movies does make sense. There are a few movies documents very close to each other in groupings of two, so there are no big clusters. The heatmap looks similar to the prior two from Method 2, so there are no clear similar documents.

Method 3

The lemmatization, removal of stop words and non-alphabetic characters did reduce the number of tokens in the vocabulary. Like Method 2, all the stop words and punctuation were removed to produce tokens such as “yearold” rather than “year-old” and “illmannered” instead of “ill-mannered.” Otherwise, the tokens look like actual English words with the use of lemmatization rather than the result seen in Method 2 with the use of stemming.

Guardians of the Galaxy Documents

The average td-idf scores for the qualitative approach terms were smaller than that of their corresponding Method 1 scores. The terms “space” and “spaceship” had the same average

score for Method 2 and 3 among the ten documents. “Star Wars” and “alien” scored higher while “planet” scored lower in this method compared to Method 2’s corresponding scores. Unlike Method 2, “wars” in “Star Wars” was given a score greater than 0, which factored into the average score of “Star Wars” in the documents. This resulted in a larger score in Method 3 than Method 2 for “Star Wars.”

In Method 1, the plural and singular form of “alien” and “spaceship” were separated and given separate scores. Method 2 and 3’s preprocessing removes that extra step of having to look and record both and instead just have the singular form of those two terms to look for. The term “planet” also included the term “planet-in-danger” and “planetary.” Method 2 changed “planet-in-danger” to “planetindang” and “planetary” to “planetari,” which made it harder to connect it back to the original word at first glance. Method 3, however, still maintained the familiarity of the original word after preprocessing, as Method 1 did.

The top ten highest scoring terms were the same ten as Method 1’s. There is, however, a difference in the ranked number one and two in compared to Method 1’s. Method 1 has “guardians” as the ranked one and “ideological” as the ranked two, but Method 3 has it the other way around. Method 2 had “ideological” and “guardian” as ranked one and two as well, respectively. The scores for Method 3’s highest ten terms have a smaller range compared to that of Method 2’s. The range of scores for this method are similar to Method 1’s range of scores. Method 1’s largest score is 2.06 and lowest is 1.17 while Method 3’s largest score is 2.05 and lowest is 1.17. Method’s largest score is 2.48 and lowest score is 1.08, so there is more variance when it its scores.

Like the conclusions from the other methods, “ideological,” “galaxy,” “formulaic,” “flying” can be useful terms that help with identifying similar movie reviews. The tokens

“guardians,” “marvel,” and “quill” are too specific and “adult,” “prologue,” and “ikea” are too common. The scores for the qualitative approach terms were still greater than that of the top ten terms in this method, the same as it was in Method 1 and 2. These terms are important and meaningful to be used for clustering and classification.

Word2Vec with 100 Embedding Dimensions

The tokens that make up the graph are plotted in a different shape than Method 1 or 2’s. Method 1 and 2’s tokens have been plotted in a very linear fashion while Method 3’s curves at the end. At the bottom right of the plot shown in Figure 12, there are a few small clusters that form. One cluster includes the tokens “might,” “scott,” “away,” and “hard.” These words do not have similar meanings. Another little cluster has “deal,” “fair,” “cartoon,” “particular,” and “jaegar.” These words together also do not make sense except for “deal” and “fair.” The heatmap does not show any clusters. Therefore, there are no similar meanings that were found between tokens.

Word2Vec with 200 Embedding Dimensions

Like the preceding plot, the shape of the tokens plotted in this graph is not linear but as a curve at one end. This plot, Figure 13, has a clear cluster at the bottom left of the plot. The cluster consists of tokens including “preventing,” “misshapen,” “dour,” “personified,” “frenzy,” “chased,” “spending,” “enthusiastically,” “accountability,” “outweigh,” “rachels.” Some of these tokens make sense together. The tokens “frenzy,” “chased,” “enthusiastically” have similar meanings. Another set that makes sense is “misshapen,” “dour,” “personified.” On the other side of the plot, there are two more small clusters, although one of the clusters seems to only have two terms. The other cluster has terms such as “track,” “usual,” none,” “pace” that do have similar meanings. This model has been the most successful at showing clear clusters that make

some sense compared to the other plots. The heatmap, however, does not show any clusters within it.

Word2Vec with 300 Embedding Dimensions

This plot, Figure 14, also takes on a curve at one its ends. This curve is a little sharper than the other two plots seen for this Method. A small cluster does appear at the end of the tail with the terms "production," "interesting," "early," "shot," "always," "script." These terms together do make sense because they do have similar meanings. This Word2Vec model did a good job clustering similar terms together, despite it only being one small cluster. The heatmap shows no clusters in this model as well.

Doc2Vec with 100 Embedding Dimensions

There are no clusters formed in this plot. There are two documents that are little farther away at the top than the other documents. Those two documents are for *Pacific Rim* and *King of Staten Island* which are sci-fi/action and drama/comedy respectively. These two documents do not make much sense together. The heatmap also supports the lack of clusters in the plot. It looks similar to Method 2's heatmaps where there is a little red on the outer parts of the map other than the diagonal.

Doc2Vec with 200 Embedding Dimensions

There are some clusters present in this plot, Figure 15. There is a cluster at the top and also at the bottom. The top cluster includes two *Interstellar* documents as well as one *Pirates of the Caribbean, The Matrix Resurrection, The Ring, Hereditary,* and *Spider Man 3* document. This cluster makes sense. These films all have an action theme to them. *The Ring* and *Hereditary* are horror thriller films. *Spider Man 3* and *Pirates of the Caribbean* are action films. Although

The Matrix Resurrection and *Interstellar* are manly science fiction, they both do have an action or adventure element to them. This cluster makes some sense but is not great.

The other cluster consists of three *Red Notice*, two *Interstellar,* one *Groundhog Day,* one *Lamb*, one *Arrival,* and one *Us* documents. This cluster is more of a chain than it is a grouping. The film *Us* and *Lamb* are horror films. *Arrival* has a bit of a mystery theme that can be related to horror, but otherwise is a science fiction movie. *Interstellar* and *Arrival* make sense together, but not *Groundhog Day and Interstellar* since *Groundhog Day* is a comedy. *Groundhog Day* is similar to *Red Notice* since *Red Notice* is an action comedy. This cluster is not great because it does not all relate to the other films in the cluster very well. With no other red cells except for the diagonal, the heatmap confirms those conclusions as well.

Doc2Vec with 300 Embedding Dimensions

This plot, Figure 16, clearly shows clusters at the top and bottom of the plot. The top cluster includes a document from *Pirates of the Caribbean: The Curse of the Black Pearl, Spider Man 3, Encanto,* and *Cruella.* This clustering makes sense. All these films have action in them as well as a comedy element. The bottom cluster has two *Interstellar,* one *Cruella,* one *Groundhog Day,* one *King of Staten Island,* two *Red Notice,* and one *Lamb* document. The outliers from this cluster are *Interstellar* and *Lamb.* The other films are comedies. This plot did a good job clustering documents compared to the others. The clusters, for the most part, make sense and are clearly separated from the rest of the plot. Despite that, the heatmap still shows no similarities between other documents except for themselves.

Conclusions

Data wrangling plays a big role in the output and results of text processing. The different tf-idf scores from each preprocessing method shows that data wrangling can affect the weight of

each token in each document, which can lead to different results. Each method and its processed text also contributed to the resulting embedding vectors created from using Word2Vec and Doc2Vec.

The third method that used tokenization, normalization, lemmatization, and removed stop words processed text that was used in Word2Vec and Doc2Vec to create the only plots among the other methods that had clustering and those clusters made sense. Method 3's data wrangling methods used in the Word2Vec with 200 embedding dimensions produced a plot that had two clusters whose tokens had similar meanings to one another. The same method was used in Doc2Vec with 300 embedding dimensions. The plot showed clusters of documents that were similar to one another. The trial of using different hyperparameters for each model was also an important step to find similar tokens and documents.

References

Clean text. Foundations of AI & ML. (2018, October 16). Retrieved January 2022, from https://mylearningsinaiml.wordpress.com/nlp/data-preparation/clean-text/

Srinivasan, S. (2022a). *MSDS 453 Natural Language Processing: Dimensionality Reduction in NLP (Weeks 3 & 4)* [Zoom Cloud Recordings].

Srinivasan, S. (2022b). *MSDS 453 Natural Language Processing: Sync Session #1* [Zoom Cloud Recordings].

Appendix

Figure 1. Average tf-idf scores for qualitative approach terms.

Document/Terms	Star Wars	planet	alien	space	spaceship
Method 1					
1	0.00	6.24	2.88	0.00	0.00
2	5.99	0.00	8.65	5.39	4.22
3	7.14	0.00	0.00	2.67	16.79
4	5.99	3.12	8.65	2.69	4.22
5	0.00	12.07	15.22	5.39	4.22
6	5.99	0.00	5.77	2.69	0.00
7	0.00	0.00	2.88	5.39	0.00
8	9.56	5.83	11.06	16.60	0.00
9	5.99	3.12	0.00	8.08	0.00
10	19.12	0.00	6.57	2.69	0.00
Average Scores	**5.98**	**3.04**	**6.17**	**5.16**	**2.94**
Method 2					
1	0.00	6.24	2.67	0.00	0.00
2	2.07	0.00	8.01	5.14	3.96
3	7.14	0.00	0.00	2.57	14.80
4	2.07	3.02	8.01	2.57	3.96
5	0.00	11.88	13.36	5.14	3.96
6	2.07	0.00	5.34	2.57	0.00
7	0.00	0.00	2.67	5.14	0.00
8	2.07	5.83	8.01	16.11	0.00
9	2.07	3.02	0.00	7.71	0.00
10	10.93	0.00	5.34	2.57	0.00
Average Scores	**2.84**	**3.00**	**5.34**	**4.95**	**2.67**
Method 3					
1	0.00	6.05	2.74	0.00	0.00
2	5.76	0.00	8.21	5.14	3.96
3	6.86	0.00	0.00	2.57	14.80
4	5.76	3.02	8.21	2.57	3.96
5	0.00	11.88	13.69	5.14	3.96
6	5.76	0.00	5.47	2.57	0.00
7	0.00	0.00	2.74	5.14	0.00
8	9.39	5.83	8.21	16.11	0.00
9	5.76	3.02	0.00	7.71	0.00
10	18.37	0.00	5.47	2.57	0.00
Average Scores	**5.76**	**2.98**	**5.47**	**4.95**	**2.67**

Figure 2. Top ten tf-idf terms.

Top	Token	tf-idf average score
Method 1		
1	guardians	2.06
2	ideological	2.05
3	galaxy	1.50
4	formulaic	1.47
5	marvel	1.46
6	quill	1.36
7	prologue	1.27
8	flying	1.27
9	values	1.24
10	ikea	1.17
Method 2		
1	ideological	2.48
2	guardian	2.10
3	galaxi	1.50
4	marvel	1.41
5	quill	1.36
6	adult	1.35
7	prologu	1.27
8	formula	1.19
9	ikea	1.17
10	hypocrisi	1.08
Method 3		
1	ideological	2.05
2	guardians	1.82
3	galaxy	1.50
4	formulaic	1.47
5	marvel	1.46
6	quill	1.36
7	adult	1.35
8	prologue	1.27
9	flying	1.27
10	ikea	1.17

Figure 3. Method 1 Word2Vec plot with 100 embedding dimensions.

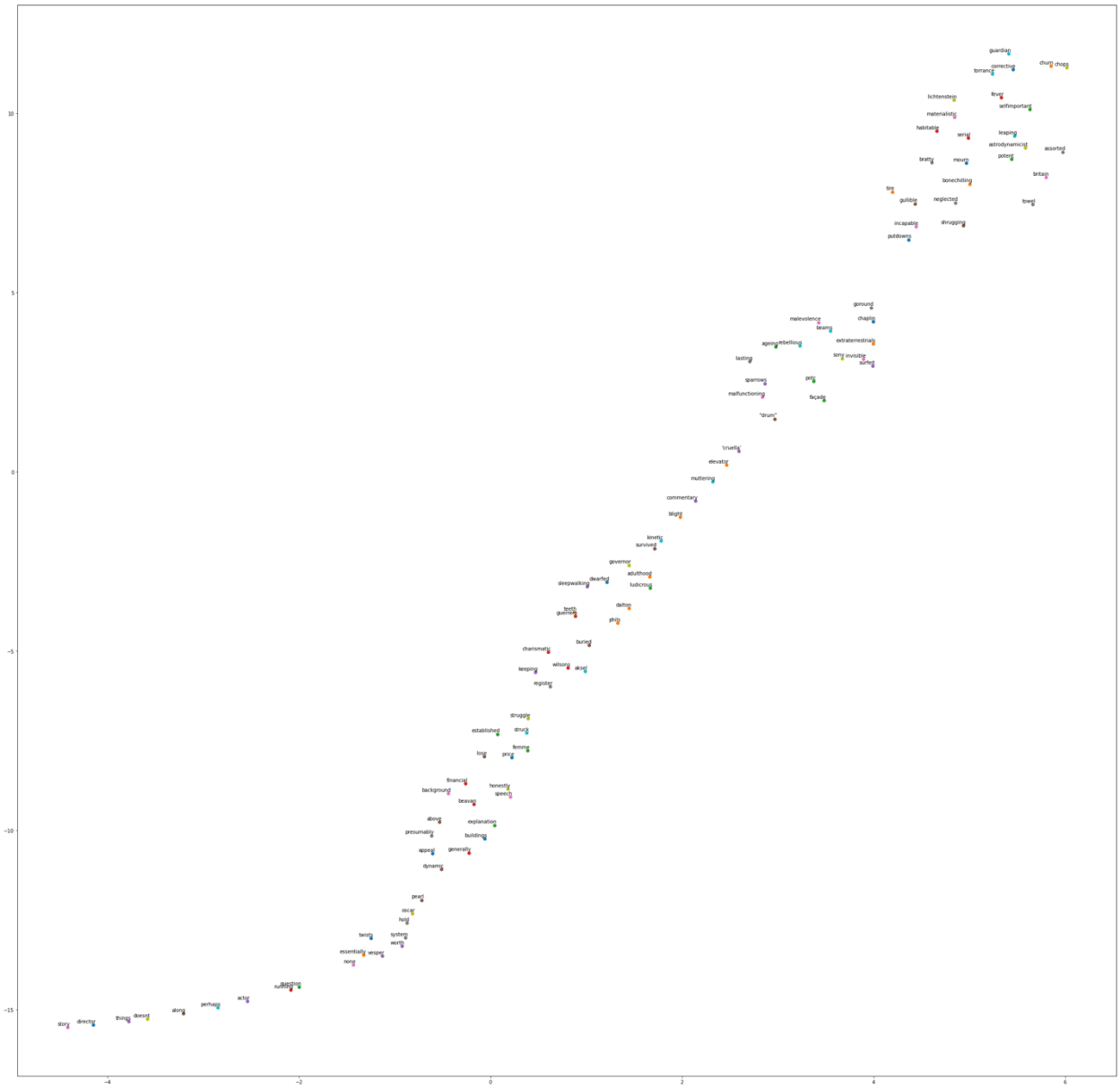

Figure 4. Method 1 Word2Vec heatmap with 100 embedding dimensions.

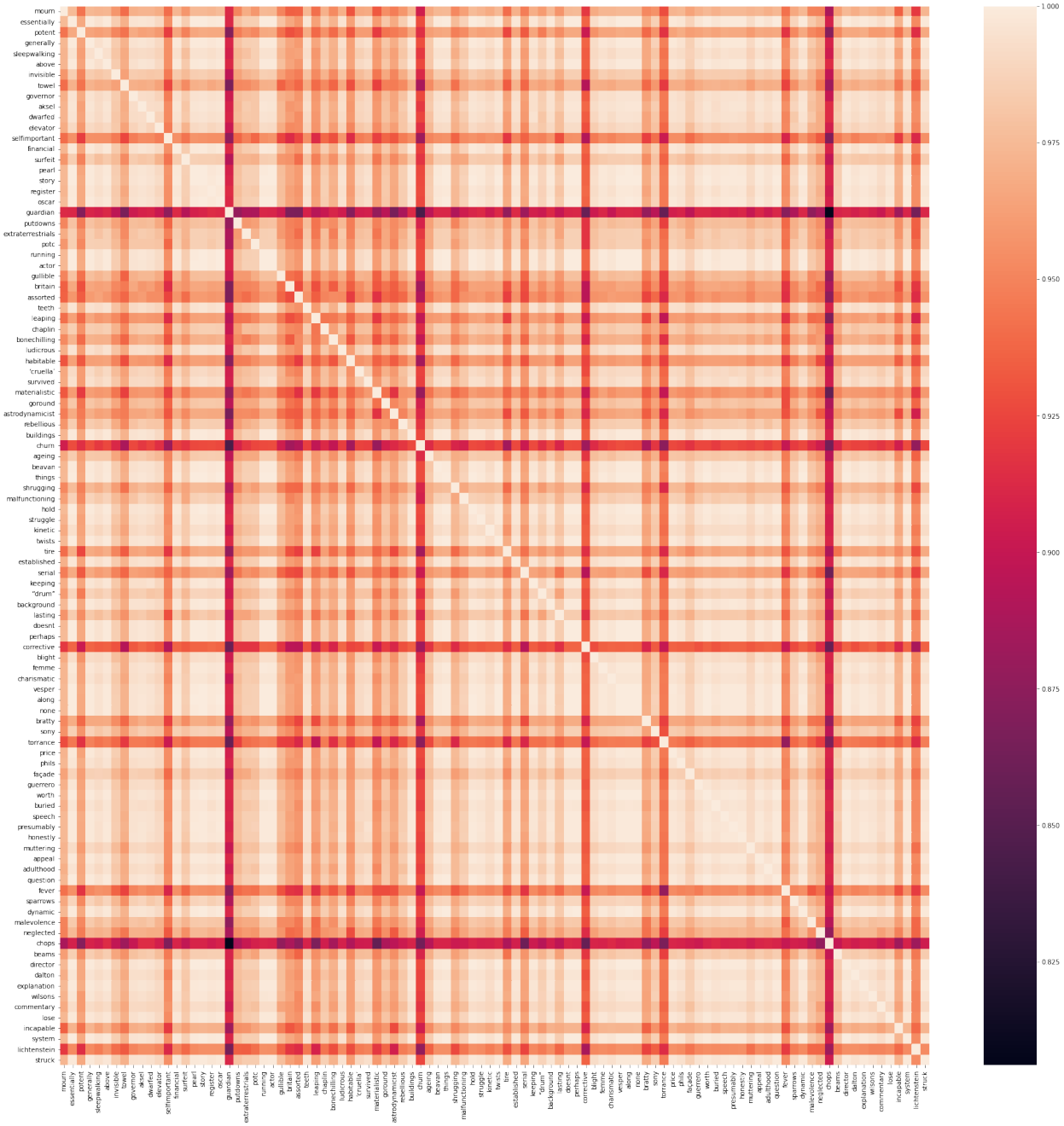

Figure 5. Method 1 Word2Vec plot with 200 embedding dimensions.

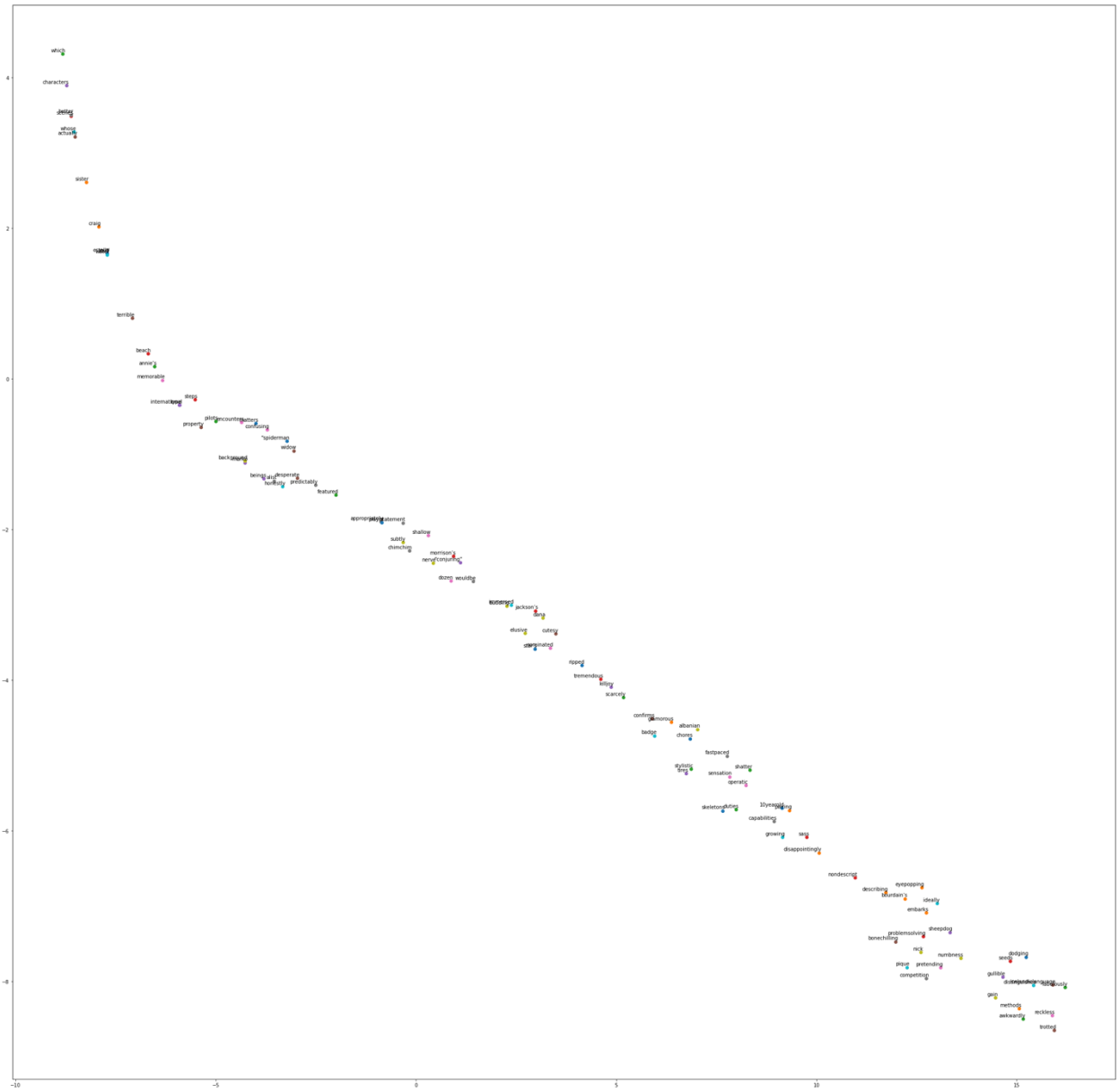

Figure 6. Method 1 Word2Vec plot with 300 embedding dimensions.

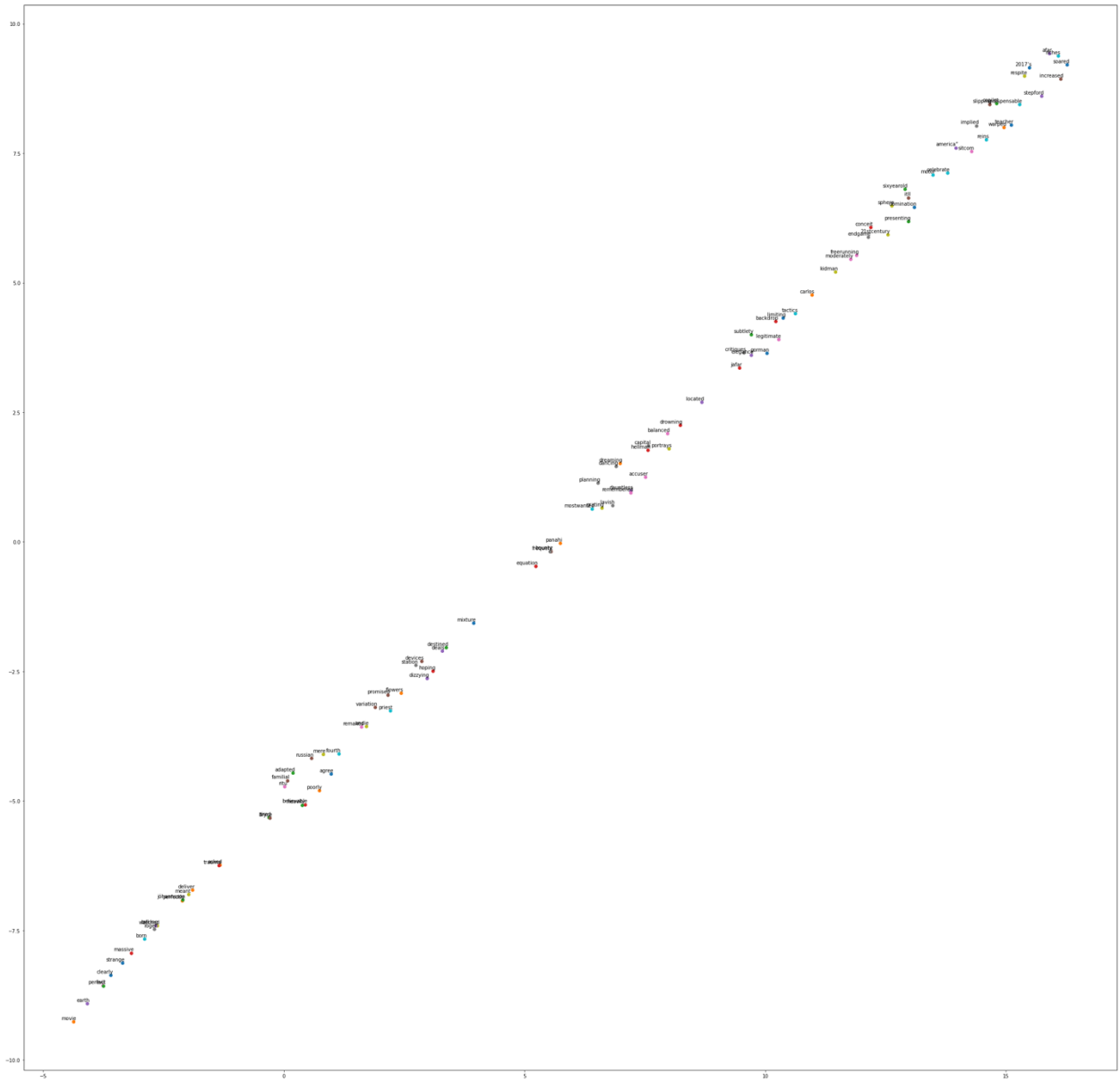

Figure 7. Method 1 Doc2Vec heatmap with 100 embedding dimensions.

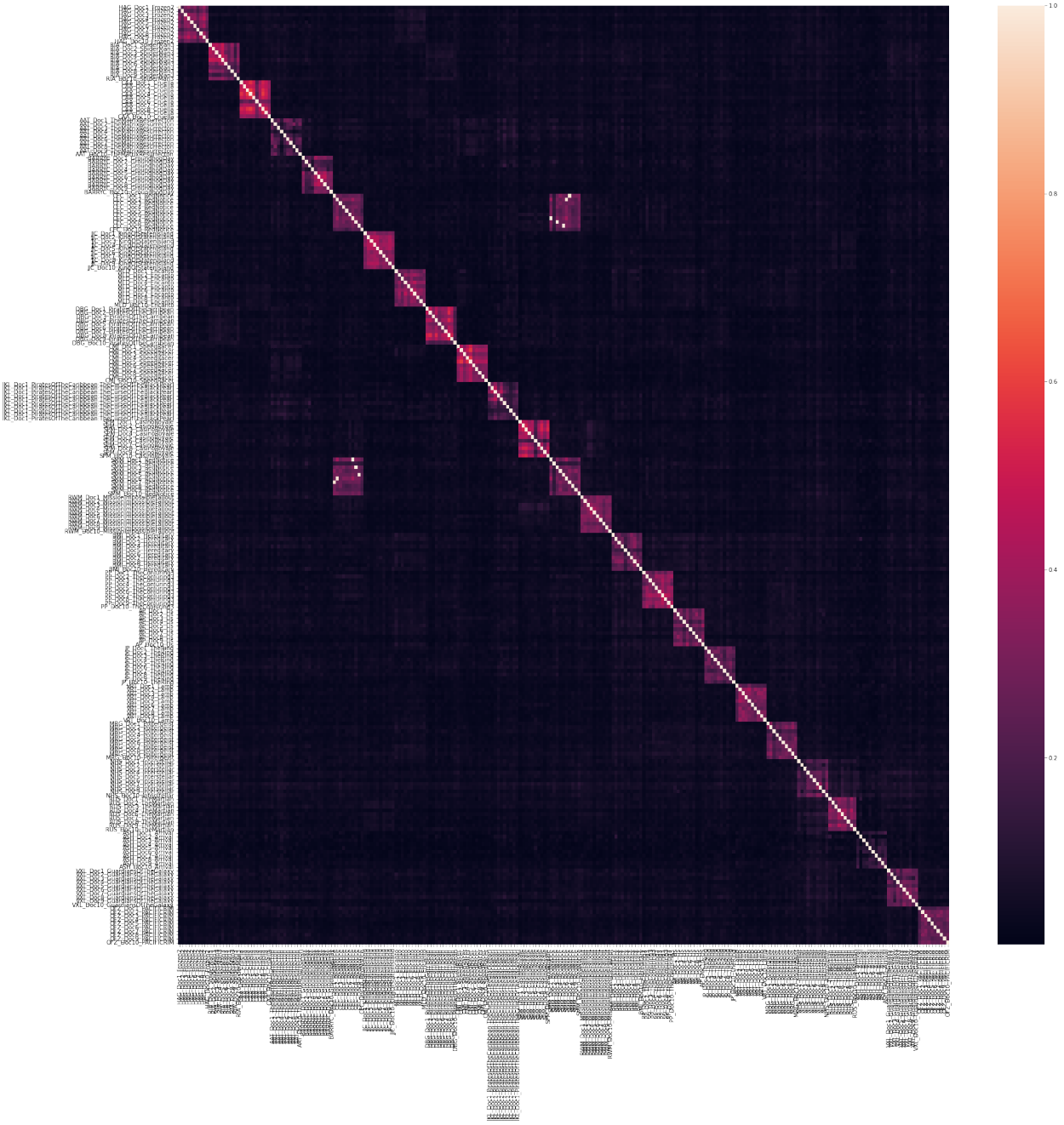

Figure 8. Method 2 Word2Vec plot with 100 embedding dimensions.

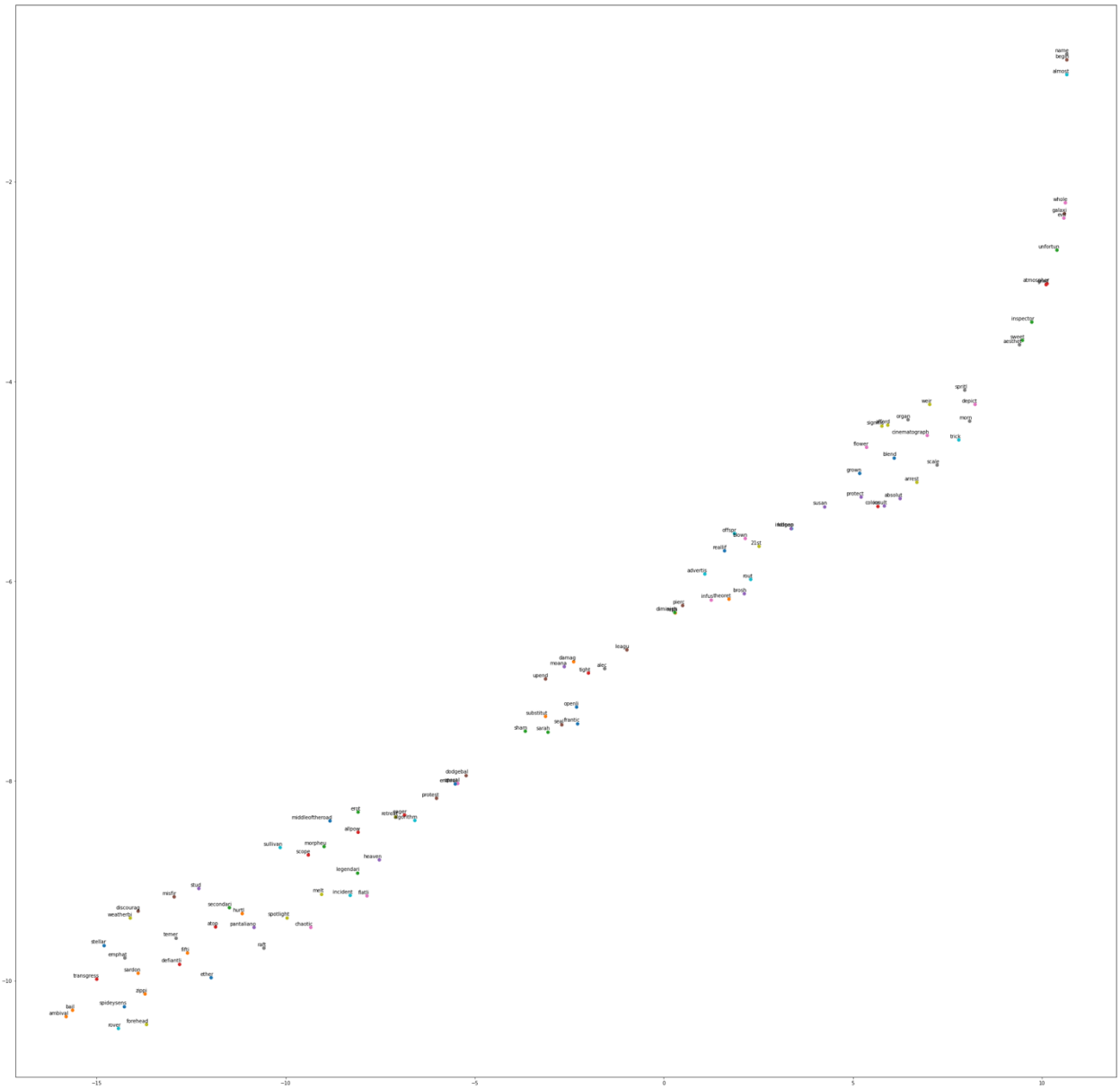

Figure 9. Method 2 Word2Vec plot with 200 embedding dimensions.

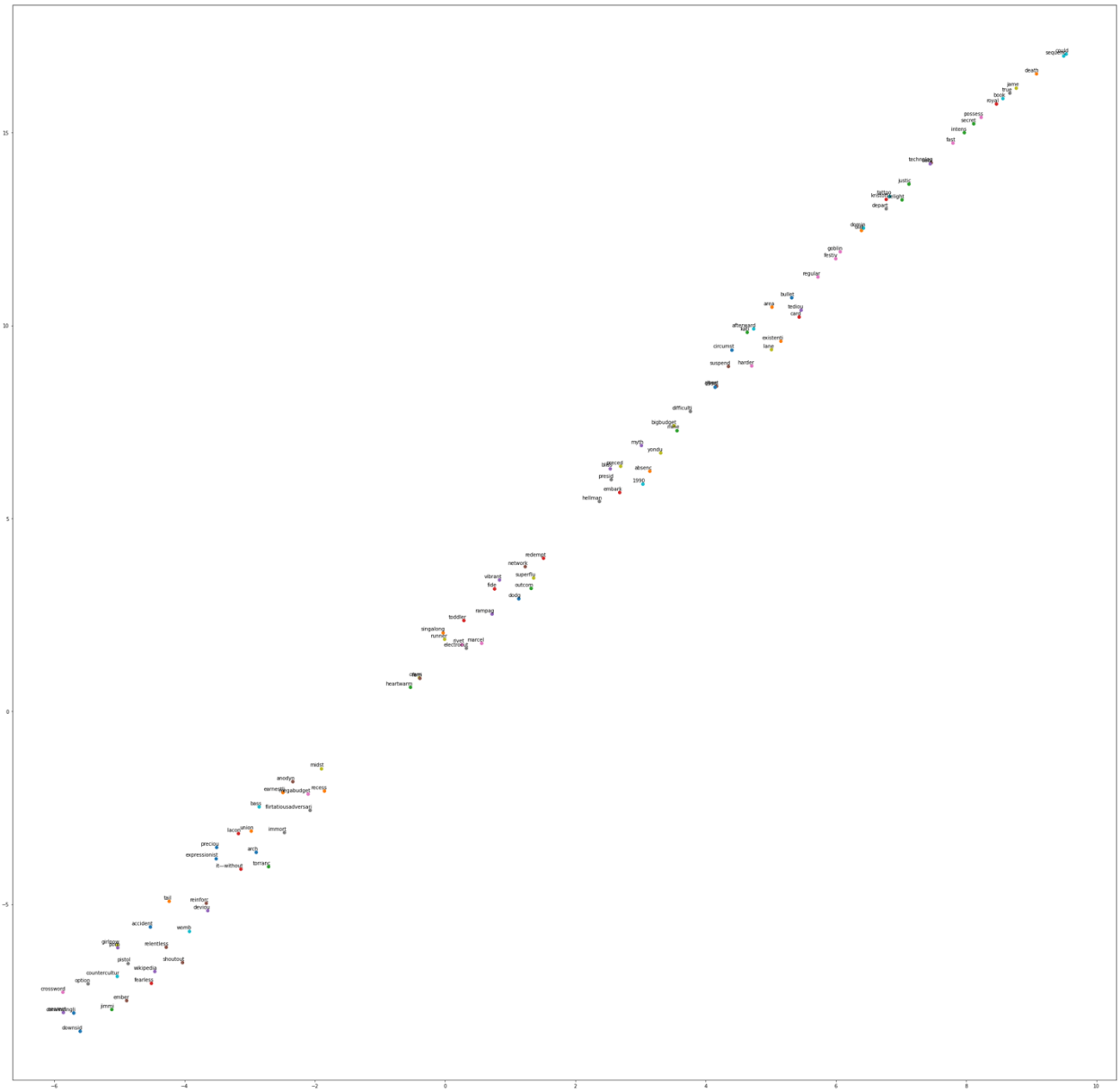

Figure 10. Method 2 Doc2Vec heatmap with 100 embedding dimensions.

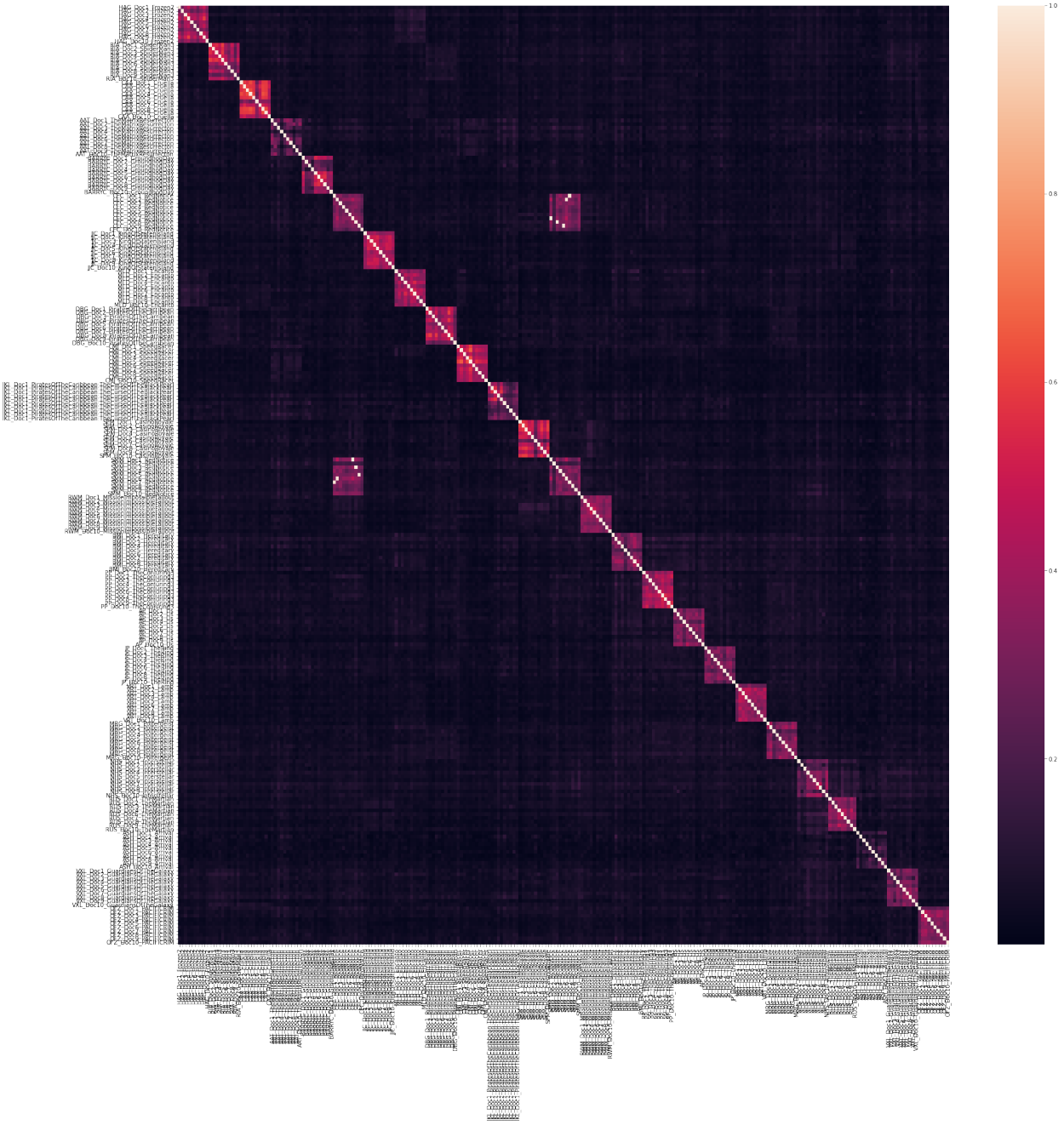

Figure 11. Method 2 Doc2Vec with 300 embedding dimensions plot.

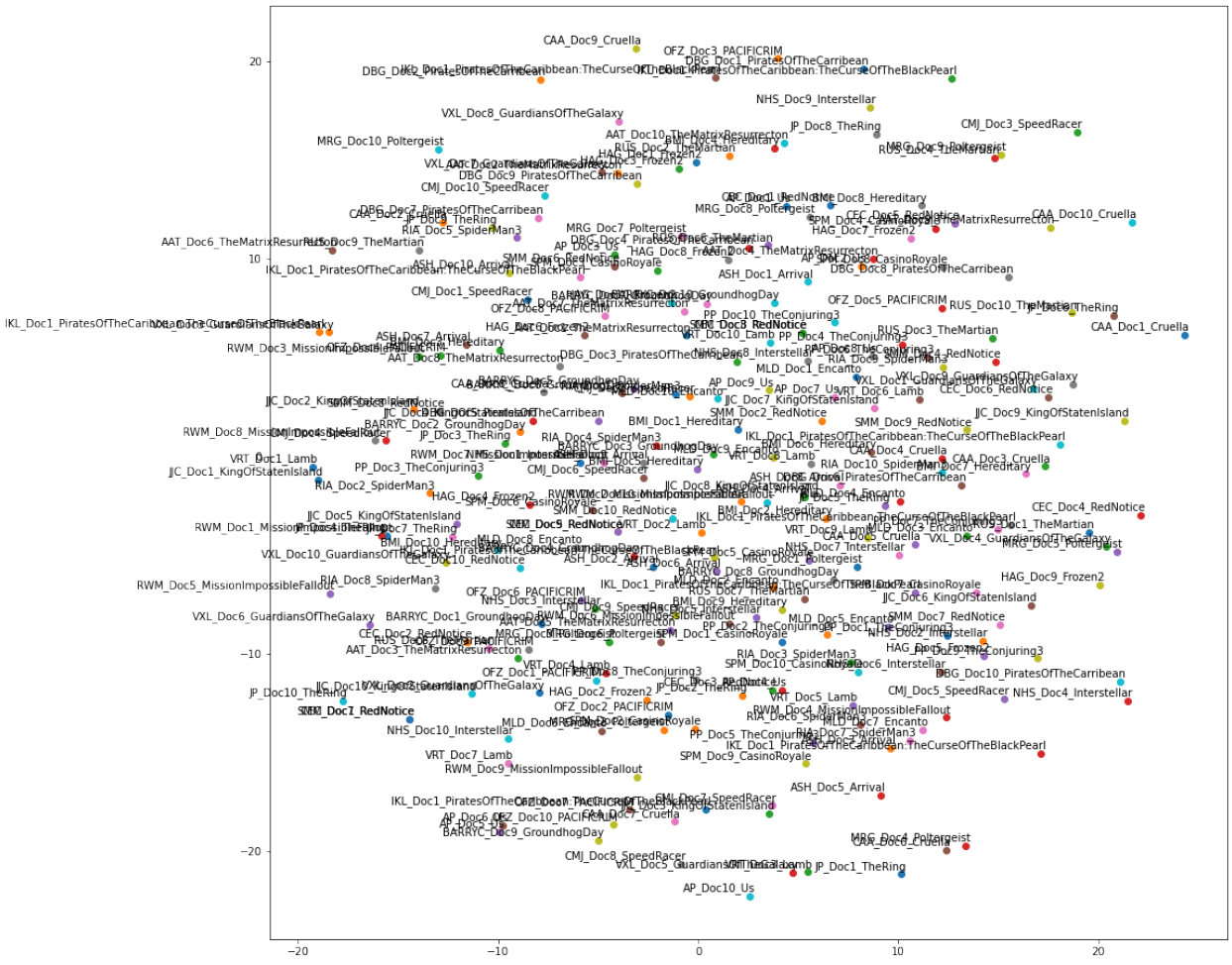

Figure 12. Method 3 Word2Vec plot with 100 embedding dimensions.

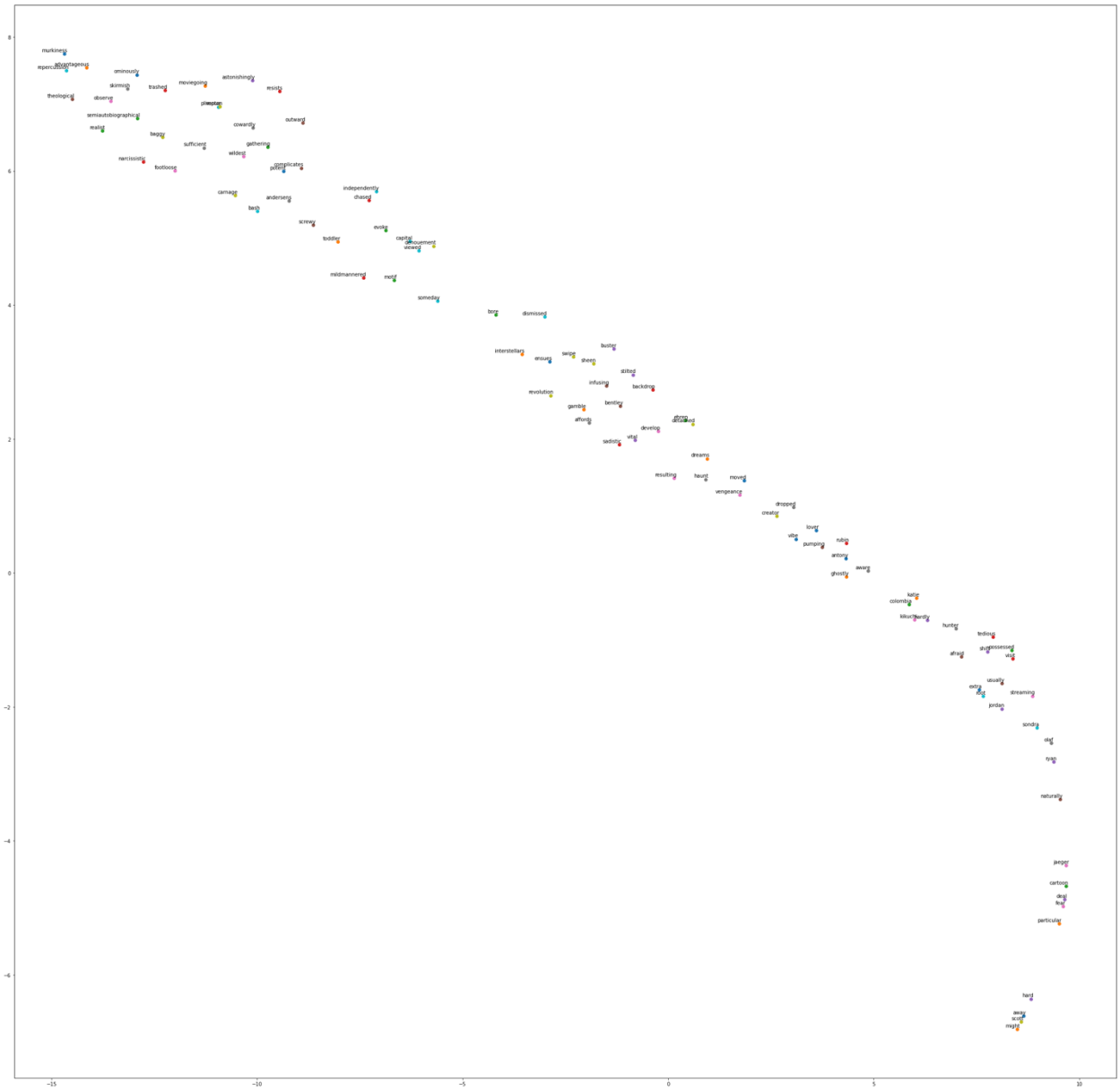

Figure 13. Method 3 Word2Vec plot with 200 embedding dimensions.

Figure 14. Method 3 Word2Vec plot with 300 embedding dimensions.

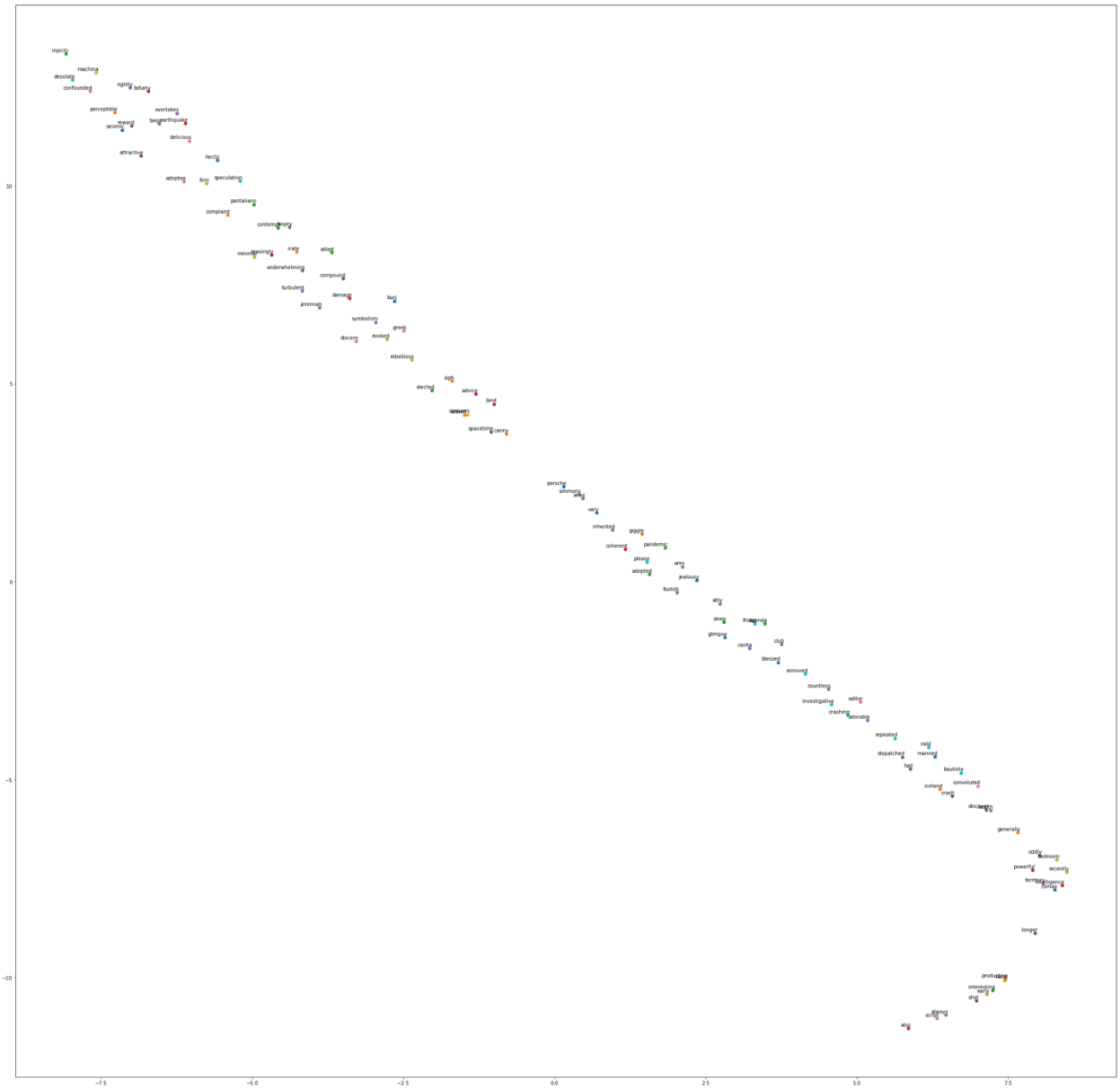

Figure 15. Method 3 Doc2Vec with 200 embedding dimensions plot.

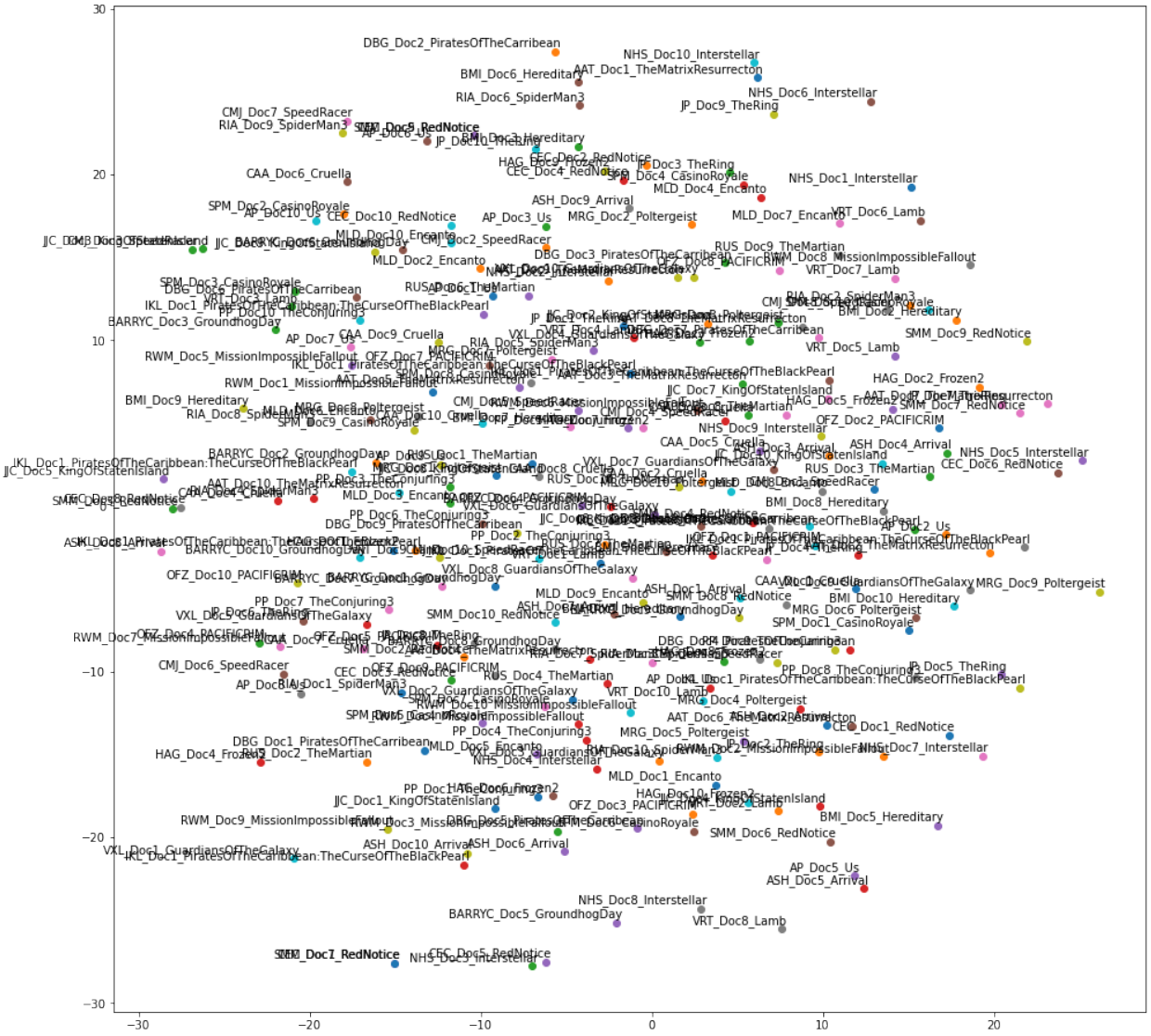

Figure 16. Method 3 Doc2Vec with 300 embedding dimensions plot.