Assignment 2

Sonny Desai

February 2020

Identify the problem, describe data, clarify

Clustering is an essential part of machine learning and natural language processing. The main point of clustering is to group similar items together so that each group or cluster, contains items that are similar to one another. In this corpus, movie reviews have been compiled from different films into a large text document. The issue is to find of all of the nearly 100 films, which can be grouped together based upon similarities in their movie reviews. The data for this experiment is interesting, because on its face, the data has very little similarities between each other. The corpus is compiled of movie reviews of each film which have been written by different writers, with different editors, along a large period of time (nearly forty years!). The data is compiled of the text words that are written in each film review. The objective then, is to use the words in the movie reviews to try to find clusters of similar movies to group the films into. By finding similar words and topics, the algorithm finds the films that are most similar to each other and places them into corresponding clusters.

Summarize algorithm/key features of code

This assignment uses a k-means clustering algorithm to cluster the films into groups. K-means clustering is an algorithm that groups similar data points together and discover underlying patterns. To achieve this objective, K-means looks for a fixed number of clusters in a dataset. K must be defined for each algorithm, and for this algorithm k was defined as 8 and thus the algorithm broke the clusters into 8 different groups. This algorithm is performing the clustering algorithm on thedoc2vec model. The doc2vec model creates a numeric representation of a document, regardless of its length. A word vector is generated for each

word, and a document vector is generated for each document. This model took k defined as 8 and broke the films into 8 different clusters.

Review results

The second K means algorithm clustered the films based on the doc2vec score and clustered the films based upon the text data from each of their reviews. The outcome looked like this with the algorithm functioning correctly and breaking down the groups into 8 different clusters:

```
Cluster 0:
 KS_Doc1_2001 Space Odyssey.txt
 KS_Doc2_Wargames.txt
 KS_Doc4_I_Robot.txt
 VPD_Doc2_I_wanted_to_explore.txt
 YF_Doc2_The_Imitation_Game.txt
 YF_Doc3_Argo.txt
Cluster 1:
 KS_Doc6_Her.txt
 KS_Doc7_The Terminator.txt
 SW_DOC7_FUGITIVE.txt
 VPD_Doc5_Existing_somewhere_between.txt
 YF_Doc1_Hidden_Figures.txt
 SD_12 years a slave.docx
 SD_Shape Of Water.docx
Cluster 2:
 PP_Doc2_Cinderella_Review_Straight-Faced.txt
 SW_DOC3_CRIMSONTIDE.txt
 VPD_Doc1_Paris_a_city_of_light.txt
 YF_Doc6_Lincoln.txt
 SD_Birdman.docx
 SD_Moonlight.docx
Cluster 3:
 SW_DOC2_CLEARANDPRESENTDANGER.txt
 SW_DOC5_INTHELINEOFFIRE.txt
 VPD_Doc6_The_Bucket_List.txt
 YF_Doc7_The_Fifth_Estate.txt
 SD_Parasite.docx
Cluster 4:
 KS_Doc3_The_Matrix.txt
 KS_Doc5_Ex_Machina.txt
 PP_Doc4_Review_Beauty_And.txt
 PP_Doc5_Film_Review_Tim.txt
 PP_Doc6_The_Jungle_Book.txt
 SW_DOC6_PATRIOTGAMES.txt
 VPD_Doc4_Feel_Good_Movie_Set_in_Rajasthan_India.
 YF_Doc4_Snowden.txt
 SD_Green Book.docx
 SD_Spotlight.docx
Cluster 5:
 PP_Doc1_Disneys_New_Mulan.txt
 SW_DOC4_HUNTFORREDOCTOBER.txt
 VPD_Doc3_As_beguiling_as_a_stroll.txt
Cluster 6:
 PP_Doc3_Aladdin_Review_This.txt
 VPD_Doc7_Frisky_and_frivolous.txt
Cluster 7:
 PP_Doc7_Whats_A_Nice.txt
 SW_DOC1_AIRFORCEONE.txt
 YF_Doc5_The_Revenant.txt
```

Interpret results

They k-means algorithm produced interesting results with some of the clusters producing results that seemed a bit out of place. Some of the clusters seemed to correctly place all of their films correctly into the right clusters, however this was the rarity. Cluster 0 all have films with science fiction themed films. The Imitation Game seems tenuous as it's a historical drama, but it is based upon the science of code breaking. The reality is the difference of the films are heightened because these movies have been reviewed by different writers, of differing ages and vantage points, along a large period of time. The age range of the films in cluster 0 is nearly forty years. Writers have inherent biases based upon taste and discretion, and so comparing different reviews of different films is likely to lead to varying conclusion because of these human biases. However, the algorithm did a fairly good job with Cluster 0. Cluster 1 is an example of a mixed cluster with some confusing results that produces more questions than answers. Some of the films (Shape of Water, The Terminator, Her) deal with science fiction whilst other films (12 Years a Slave, Hidden Figures) are historical dramas with African American protagonists. In going back to the vector inputs its possible there are not enough sufficient terms and term frequency strengths. The doc2vec model took 100 vectors, a window of 2, and a min count of 1. This could have possible been too low across the board which led to inconsistent results. In addition, the k-means cluster took a random state of 89 which could have affected the results as well.

Some films in other clusters seemed to be wholly random and not correlated to the other films in the cluster, pointing back tuning needed of the vectors in the doc2vec model. Cluster 4 is by far the largest of the clusters with 10 films. In comparison the next largest cluster

is Cluster 1 with 7 films. Cluster 4 has a broad range of films, some which are correlated to one another, but have little to no correlation with other films in the cluster. In a sense, this cluster has mini-clusters within the cluster, and had there been a better tuning of the model, or more likely the cause, more text data, the outcome would have properly separated the mini-clusters into their own separate clusters. For example, there is a swath of Disney animated films like beauty and the Beast and the Jungle Book, which are correlated to one another but have no correlation to The Matrix or Ex machina which are both sci/fi, action films. In addition, there are films to seem not to correlate to either of these two mini-clusters in Green Book, Spotlight, and Snowden which are all based upon true stories. The analysis from this points to a more refined model and more extensive data needed to create better results.

Summarize insights and findings

The model correctly performed as designed, with 8 separate clusters being create – the k input was 8. However, there were some inconsistent and confusing results with the output pointing to a more refined model, and possibly, more data needed. Some clusters were correctly created with films corresponding correctly, but the majority had films included that seemed to be mini-clusters, or related to a few other films in the cluster, but not to others. The size and scope of the different clusters points to a review of the vectors and other tuning measurements being corrected to produce better results.