Clustering Creative Writing
Project Fletcher
Harmeet Hora

Project Design:

The concept behind this project was to use responses on the writing prompts subreddit as a source of creative writing that would be used for NLP analysis. The idea was to see if there were common themes that were visible within the prompt responses.

Tools:

- PRAW (Python Reddit API Wrapper)
- Sci kit learn
- Pandas
- Seaborn
- NLTK

Data:

The data was scaped from reddit via the Reddit API. I was able to scrape approximately 5000 unique comments based on a few conditions. The comments were only pulled from the top-rated posts in the subreddit and had to have at least 100 upvotes in order to be scraped. The data was quite limited due to restrictions from reddit, who do not allow more than 1000 posts to be returned for a specific query.

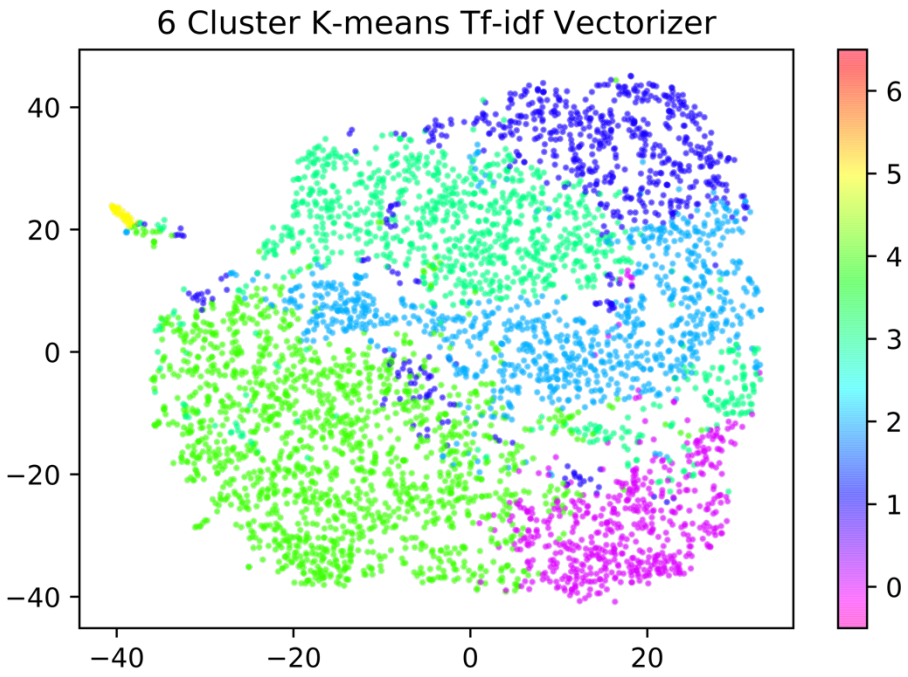
Algorithm:

The following data extra process was performed on the data in order to cluster the comments I scraped:

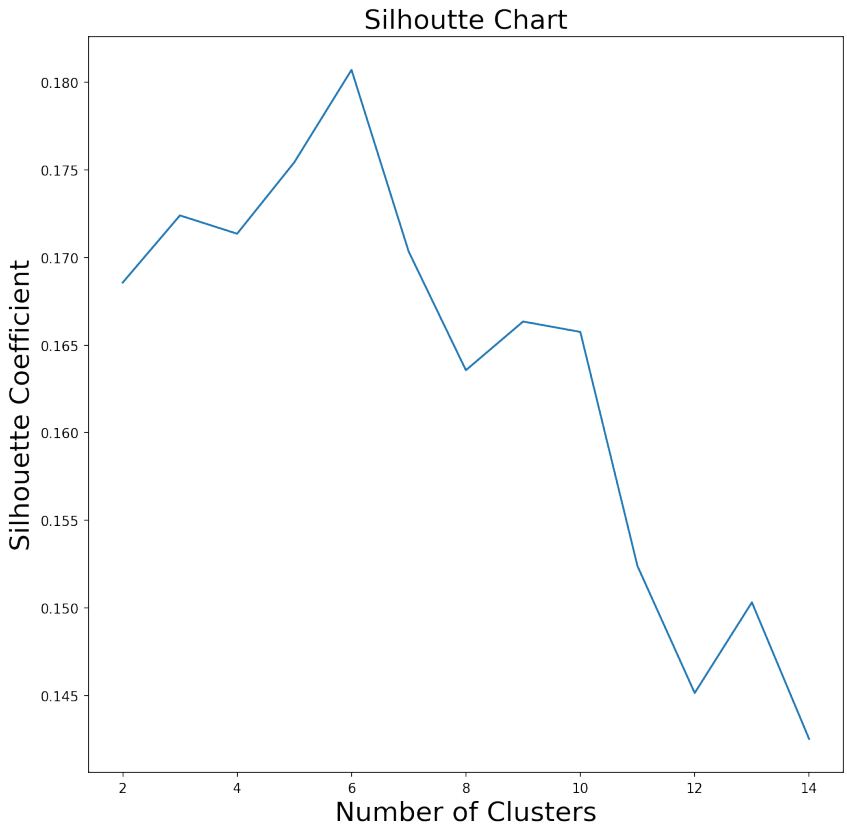
- TFIDF Vectorizer with Lemma Tokenizer
- LSA Dimensionality reduction
- Standard Scaler
- Kmeans clustering algorithim (6 clusters)
- tSNe plotting to display clusters

Results:

After clustering my data using Kmeans, I ended up having a bit of an amorphous shape with not as much separation between clusters as desired.



I settled on 6 clusters after generating a silhouette plot to find the optimal number of clusters.

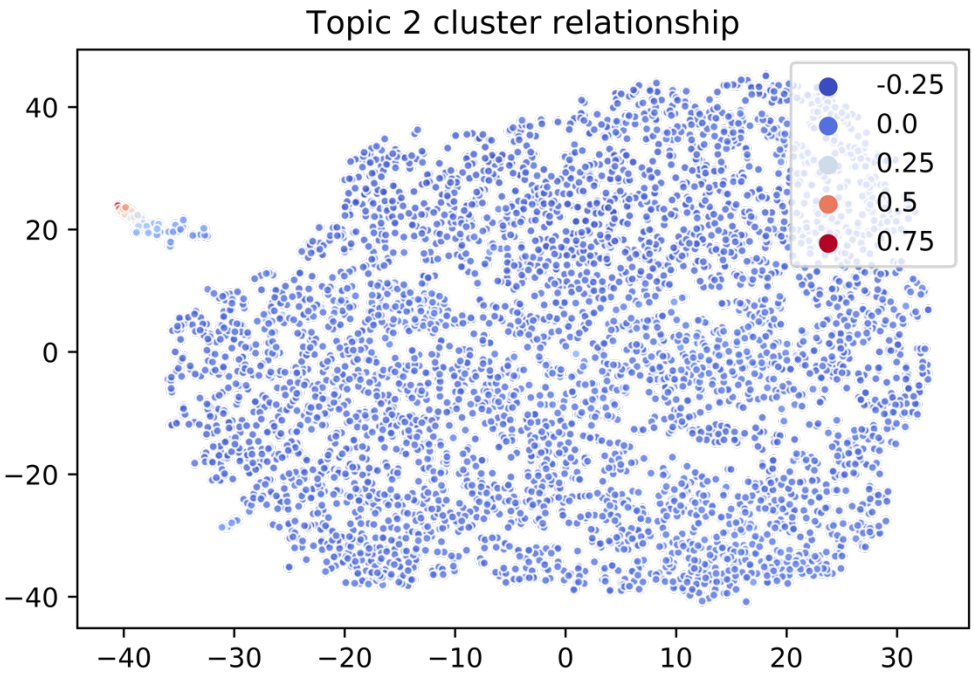


After examining posts from the centroids of the clusters, the themes tended revolve specifically around the writing prompt. For example, cluster 1 centered around comments related to immortality. After looking at the prompts for these comments, it was clear that the posts clustered together were written for the same prompt. In retrospect, this seems intuitive. This issue of clustering comments by prompt was not the desired effect, and would likely have been helped with significantly more data. The lack of data, I believe, was a big detriment to my results.

These are the 6 themes I observed from the clusters:

1. Immortality/Mortality
2. Old Age/Slow
3. Music
4. Working/Profession
5. Evil Forces
6. Harry Potter

I dug into these themes by looking at topic modeling outputs from the different vectorizers. Below is an example of the clustering of topic 2, which was related to a Harry Potter prompt. I was able to directly tie this back to cluster number 5 in my clusters shown above, as this seemed like a recurring theme in cluster 5 posts near the centroid.



What I would do differently:

If I were to redo this project, I would probably choose a different topic altogether. I would want to choose a project with a large number of records that were easier to obtain without restrictions and work arounds. I would also be interested in trying some of the other clustering methods, as well dimensionality reduction methods. Unfortunately, I spent a large chunk of my time on this project trying to figure out how to work around the reddit post restrictions.

I am still very much interested in literary analysis, as literature is something I very much appreciate, so I would probably do something more along the lines of analyzing historical works (Project Gutenburg, for example). Overall, this project was a good learning experience, even though I was not too satisified with my results.