Ben Prescott
Third Research/Programming Assignment
MSDS 453 2021SP
5/15/2021

Reddit Corpus Exploration Using Unsupervised Clustering Methods

Abstract

This paper involves exploring a portion of the Reddit corpus using k-means clustering, hierarchical cIustering, topic modeling, and spectral biclustering. Rather than using the full corpus with nearly 300,000 documents across 100 subreddits, I chose to define my working ‘full corpus’ as 3,000 documents sampled from eight subreddits, and my ‘reduced corpus’ as 300 documents across eight subreddits. My specific research involves clustering and topic modeling across many different subreddits, where more specific results may be obtained by limiting modeling efforts to describe a single subreddit’s documents.

Results from the k-means cluster are compared with those from hierarchical clustering, Latent Dirichlet Allocation is used for topic modeling, and spectral biclustering is compared with clustering and topic modeling. An broad ontology is also defined to represent the Reddit corpus for further study. In a previous study it was determined that Doc2Vec word embeddings performed the best in comparison to TF-IDF and analyst judgement. The Doc2Vec embeddings are used for this research, with the exception of TF-IDF results used for topic modeling.

Introduction

This research is being conducted to help develop some understanding with the underlying Reddit corpus documents. While the corpus has the document and the subreddit from where the document originated, I may be able to determine similarity in documents that exist from different subreddits. This study may be used for initial exploration prior to conducting sentiment analysis by identifying topics across documents and groups of similar documents. These results are used to further describe the corpus’ features and can be used as a basis in evaluating sentiment analysis for cyberbullying cases.

Literature Review

Based on my research, there are others leveraging Reddit data specifically to explore topic modeling using Latent Dirichlet Allocation. One article titled “Topical Classification and Divergence on Reddit” by Chow and Hong focuses on topic divergence, operating under the assumption that the first comment on a Reddit thread is more related than its subcomments.

There are also many posts on sites such as TowardsDataScience of topic modeling efforts and document clustering using Reddit data. The overall data used seems to vary from involving multiple subreddits (such as in Chow and Hong’s paper), to focusing on identifying topics within a single targeted subreddit.

Methods

The methods of this research were predefined, focusing on k-means clustering, dimensionality reducing using t-SNE, hierarchical clustering, cluster visualizations and comparisons, topic modeling using Latent Dirichlet Allocation, spectral biclustering, and developing an ontology visualization for further research.

To start, I selected a subset of the overall corpus to act as a limited ‘full corpus’, as if the other data were not collected. This is to help reduce processing requirements and ease exploration, as the original corpus size is nearly 300,000 documents. The adapted ‘full corpus’ is 3,000 documents across eight subreddits, with the reduced corpus having 300 documents across eight subreddits.

Using my Doc2Vec matrix as my best performing, I started by evaluating k-means performance using the elbow method and silhouette scores. Based on my data, it was determined the best cluster count for this use case is eight. I then performed k-means using using a fitted model from the original Doc2Vec matrix, as well as a second model using a two-dimensional matrix created by t-SNE. The subreddits were plotted by cluster and the results were then compared. The same process was then repeated for a hierarchical clustering method using the same number of clusters (eight). The method used for hierarchical clustering was Scikit-Learn’s *AgglomerativeClustering*. These results were then visualized and compared with k-means.

Using the TF-IDF matrix, Latent Dirichlet Allocation was used to identify topics within the various documents. I chose to use five LDA components to create five topics, and represented each with the top 30 words relating to each topic in an effort to attempt to identify which subreddit the topic may belong to.

Spectral biclustering was then used to cluster data on both rows and columns simultaneously, then visualized to compare with k-means and hierarchical clustering. Due to my Doc2Vec matrix having four vectors per document, I had to use four (or less) clusters for modeling.

Results

Using eight clusters for both k-means and agglomerative clustering, it was determined that both methods provided similar outcomes, with k-means providing slightly more distinct clusters. Both methods improved by training their respective algorithm on matrices with two dimensions, using t-SNE prior to training. Based on the results, my preferred method would be k-means for presenting findings, as it identified slightly better groups. When plotting the ground truth data points by unique label color, and plotting the subreddit labels by cluster, it can be determined that many documents from all subreddits have similarity. An interesting find is that one specific k-means cluster, cluster #2, seemed to have a majority of documents from the “programming” subreddit.

The results of modeling five topics with the 30 most common terms provides little insight, at least in this use case. However, given the subreddits sampled and based on the results, the first topic (Topic #0) could be related to the subreddit “Economics”, as common terms such as “wages”, “jobs”, and “minimum” are all common. Topic #1 is likely to be related to the subreddit “MovieDetails”, as common terms are “cool”, “movie”, and “source”. Topic #2 and Topic #3 could not be related to a specific subreddit. However, Topic #4 is likely related to the subreddit “programming”, as common terms are “experience”, “version” , and “money”.

Spectral biclustering results in the most distinguishable clusters yet, but was limited due to having to be less than or equal to the number of dimensions in my Doc2Vec matrix, which was four. However, when using four clusters the overlap in clusters was very minimal or non-existent. Based on the visual results, and given the Doc2Vec matrix were recreated using more dimensions, the best approach for clustering seems to be spectral biclustering. This approach also seemed to provide more consistent labeling for the different subreddits.

Conclusions

Ultimately, more refinement on the corpus is likely needed to process/clean text prior to clustering and topic modeling. However, we can leverage the current clusters to find document similarity and their respective subreddit, which can provide reference for reviewing sentiment analysis results. While not entirely surprising, these results also help to describe how similar comments (utterances) are between many different subreddits. This helps provide information to management to describe similarity in the corpus and which documents are grouped.

References

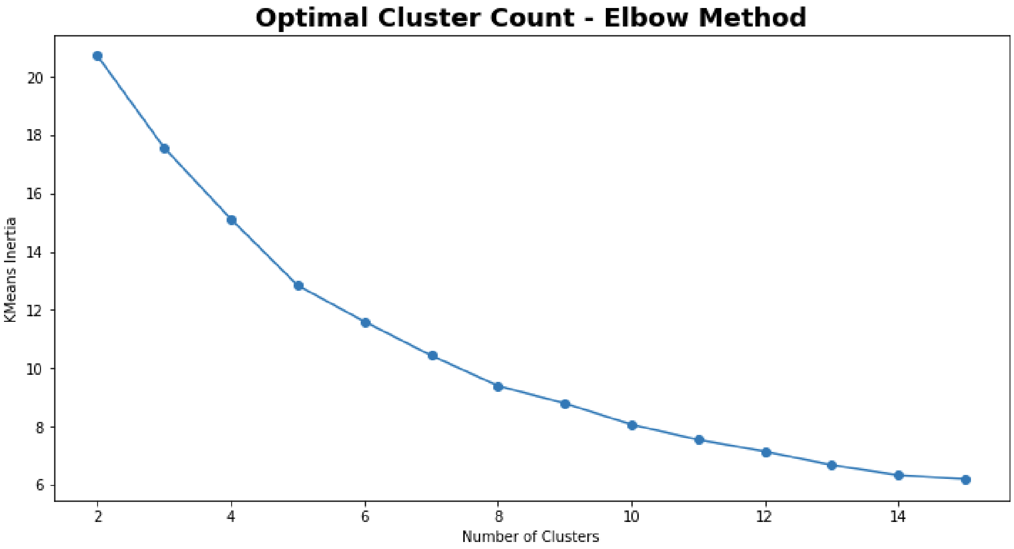
Brownlee, J. 2019. Deep Learning for Natural Language Processing: Develop Deep Learning Models for Natural Language in Python.

Chow, A., & Hong, J. (2016). Topical Classification and Divergence on Reddit.

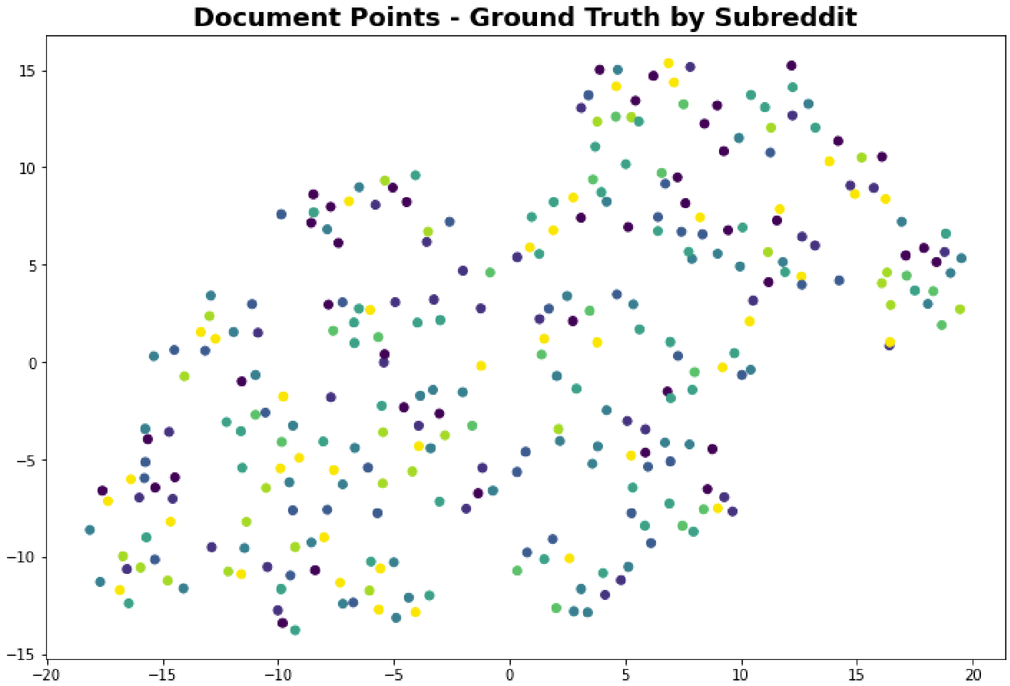
Cohen, J. (2019, September 27). *Understanding Fortnite's Reddit Community using Unsupervised Topic Modeling*. Medium. https://towardsdatascience.com/understanding-fortnites-reddit-community-using-unsupervised-topic-modeling-30f984f58129.

Appendix

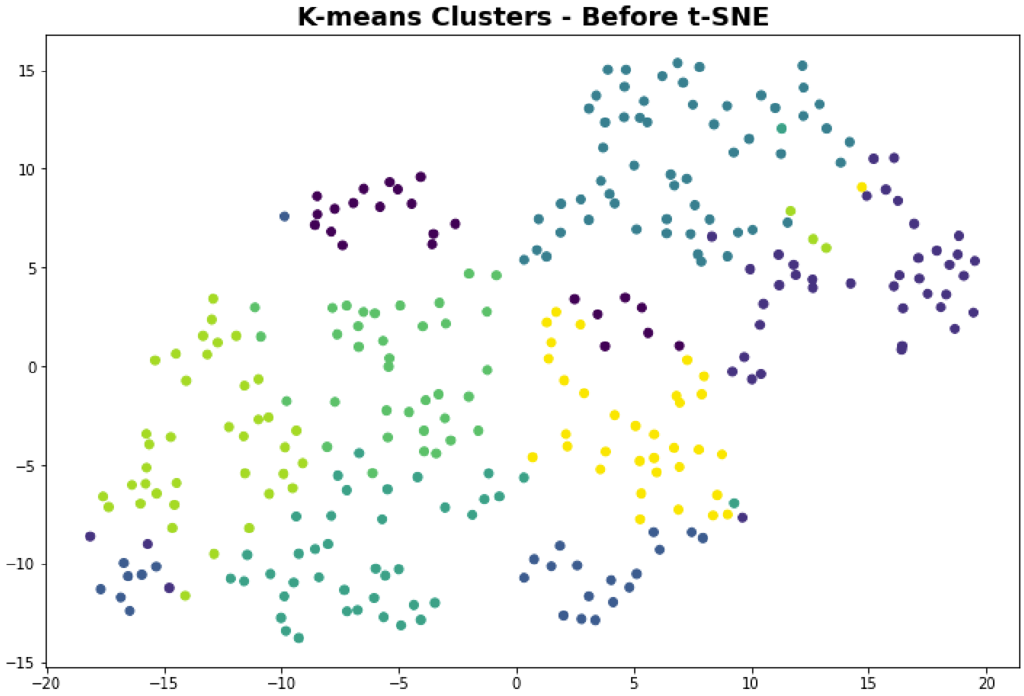
K-means Elbow Method



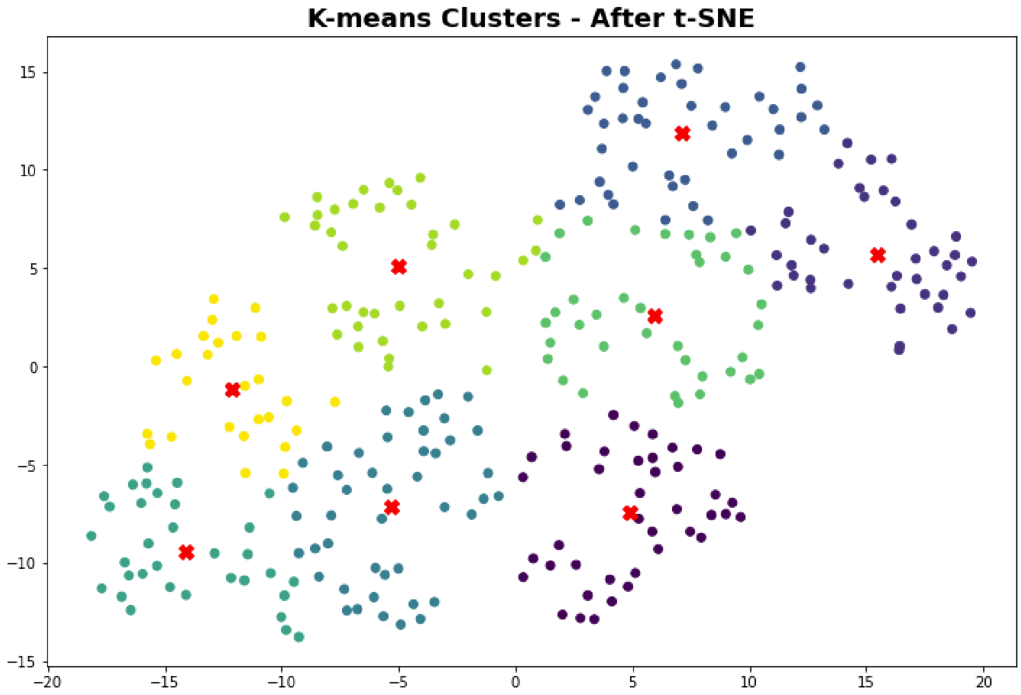
Document Ground Truth by Subreddit



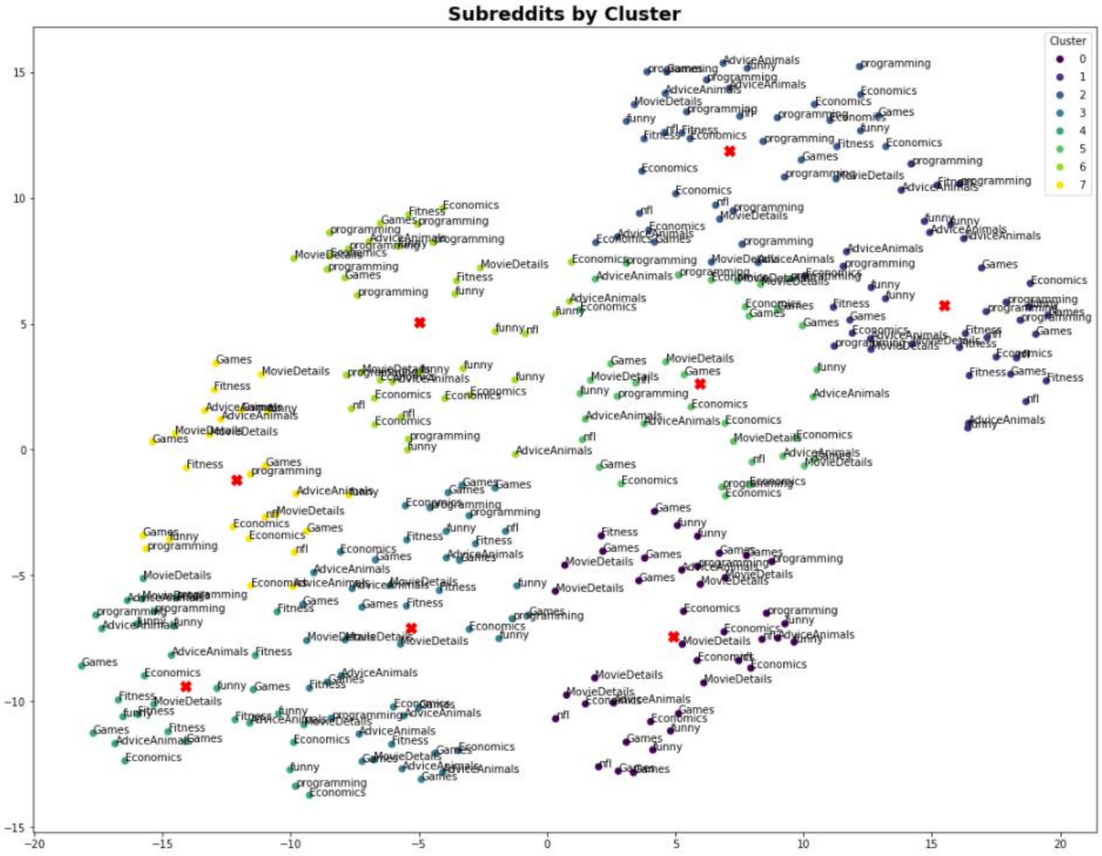
K-means Clustering Before t-SNE



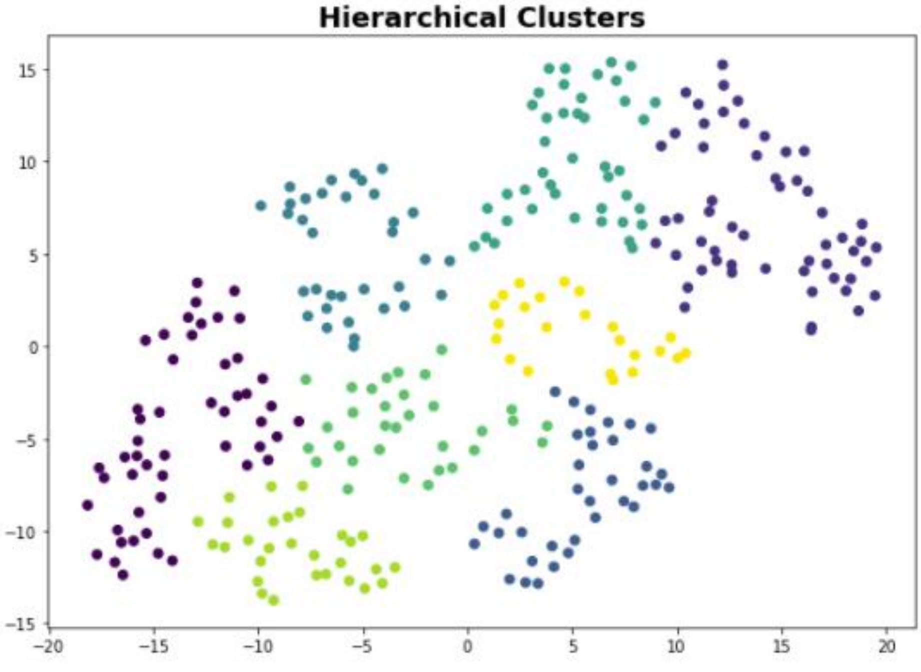
K-means Clustering After t-SNE



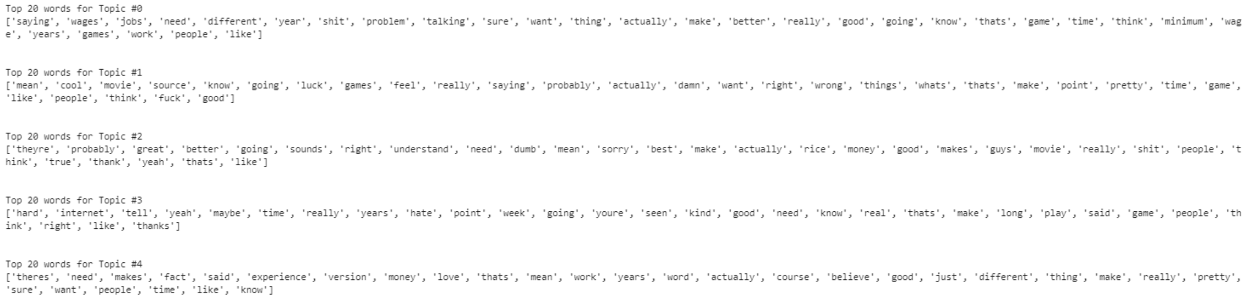
Subreddit by Cluster



Hierarchical Clusters After t-SNE

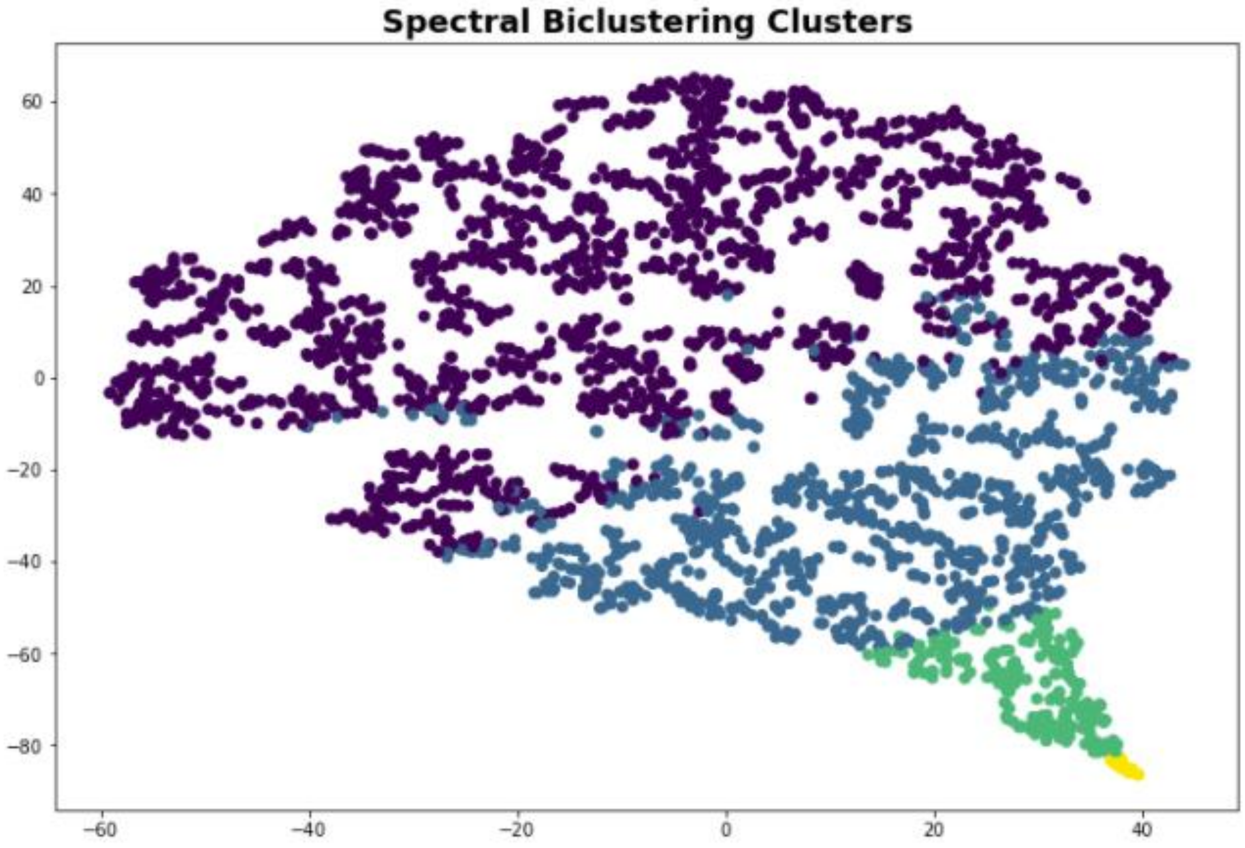


Topic Modeling

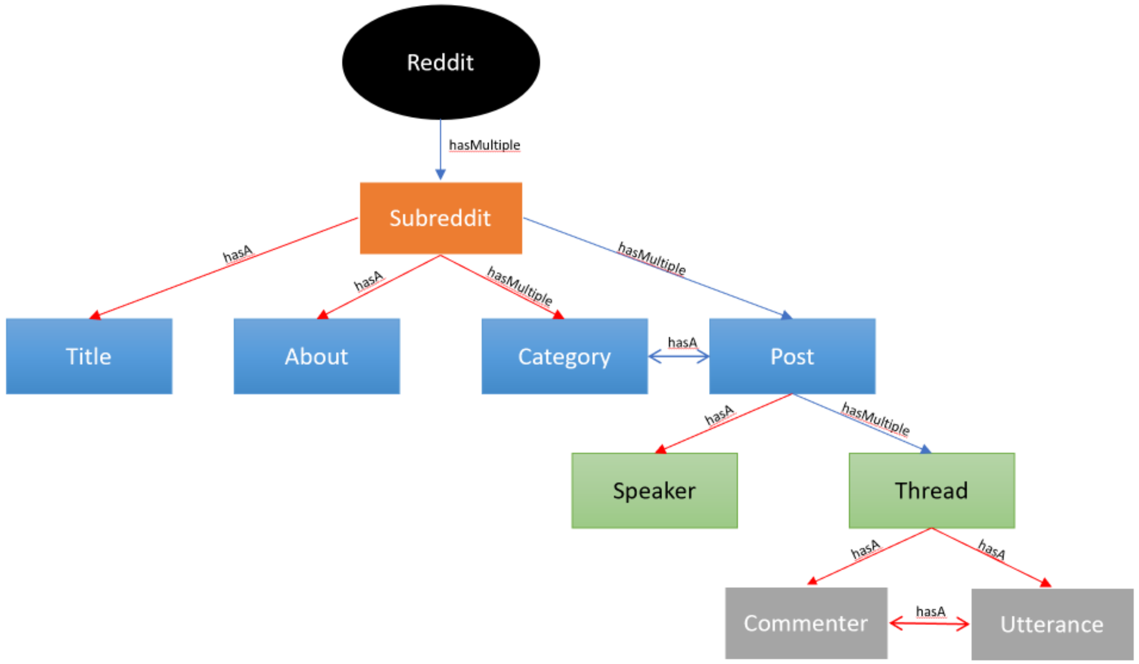


```
Top 20 words for Topic #0
['saying', 'wages', 'jobs', 'need', 'different', 'year', 'shit', 'problem', 'talking', 'sure', 'want', 'thing', 'actually', 'make', 'better', 'really', 'good', 'going', 'know', 'thats', 'game', 'time', 'think', 'minimum', 'wag
e', 'years', 'games', 'work', 'people', 'like']


Top 20 words for Topic #1
['mean', 'cool', 'movie', 'source', 'know', 'going', 'luck', 'games', 'feel', 'really', 'saying', 'probably', 'actually', 'damn', 'want', 'right', 'wrong', 'things', 'whats', 'thats', 'make', 'point', 'pretty', 'time', 'game',
'like', 'people', 'think', 'fuck', 'good']


Top 20 words for Topic #2
['theyre', 'probably', 'great', 'better', 'going', 'sounds', 'right', 'understand', 'need', 'dumb', 'mean', 'sorry', 'best', 'make', 'actually', 'rice', 'money', 'good', 'makes', 'guys', 'movie', 'really', 'shit', 'people', 't
hink', 'true', 'thank', 'yeah', 'thats', 'like']


Top 20 words for Topic #3
['hard', 'internet', 'tell', 'yeah', 'maybe', 'time', 'really', 'years', 'hate', 'point', 'week', 'going', 'youre', 'seen', 'kind', 'good', 'need', 'know', 'real', 'thats', 'make', 'long', 'play', 'said', 'game', 'people', 'th
ink', 'right', 'like', 'thanks']


Top 20 words for Topic #4
['theres', 'need', 'makes', 'fact', 'said', 'experience', 'version', 'money', 'love', 'thats', 'mean', 'work', 'years', 'word', 'actually', 'course', 'believe', 'good', 'just', 'different', 'thing', 'make', 'really', 'pretty',
'sure', 'want', 'people', 'time', 'like', 'know']
```

Spectral Biclustering



Reddit Ontology


Reddit
hasMultiple
Subreddit
hasA
hasA
hasMultiple
hasMultiple
Title
About
Category
hasA
Post
hasA
hasMultiple
Speaker
Thread
hasA
hasA
Commenter
hasA
Utterance