DSI-US-5
September 9, 2018

Executive Summary

NLP Lab: r/todayilearned and r/AskReddit

The NLP lab focuses on which characteristics of a post on Reddit contribute most to what subreddit it belongs to. For this project I chose subreddits with a lot of subscribers and are somewhat related to each other. For the subreddit, r/todayilearned, people post interesting articles from the Internet to share facts and r/AskReddit is a question and answer forum. They are related due to similar topics they cover.

For this lab to be successful, I had to conduct a webscrape, sanitize texts using Beautifulsoup, and apply my own custom filters to remove certain keywords. For example, all of the posts from r/todayilearned start with a 'TIL' tag in the title. I had to remove this word because it would make the classification too easy. I wanted to increase the difficulty slightly.

After cleaning, I created a data frame and created an extra feature where I concatenated the thread and the title. Many posts did not have a thread so it was necessary to feature engineer. Once I had the feature column, I vectorized it, train/test split them, and ran it through three models (which were optimized using GridSearchCV).

The models I used were Gradient Boost, Random Forest, and Decision Tree. All of the model's test scores hovered around 92-93%. All of the models' training score overfit the test data's score. Random Forest's training score was 99% and Decision Tree's training score was 97%. Gradient Boost's training score was 93% and the testing score was 92%. I thought this was the best model since it did not exhibit overfitting to such an extreme degree.

The models also showed interesting patterns in feature importance. The most important words that determined which subreddit it belong to were: "What", "What's", and "How". This makes sense because these are question words and the r/AskReddit has a lot of questions. Besides from question words, I realized there were words pertaining to physical spaces such as, "Earth"

and “City”, and “Mexico”. There were other key words that pertained to historical and scientific questions such as, “WW2”, “Called”, and “Life”.

I thought it would be interesting to investigate what the accuracy scores would be like if I removed key question words. I decided against removing them in the first iteration because I thought they were valid words to leave in.

Finally, I believe this lab would be useful for receiving Reddit Gold. Even though it is not a perfect metric, I think the top 100 words selected by feature importance hint at what kind of questions/posts are popular. I would like to see if there is a correlation between popular posts and gilded posts.