Comparison of r/WFH and r/digitalnomads Using Natural Language Processing

Problem Identification

Reddit is one of the largest social platforms that is unlike Twitter, Facebook, or any other platform; it gets a constant flow of communities posting new discussions daily every minute. With the vast set of communities, it can be intimidating to join a subreddit that you don't know is right for you. By using r/WFH and r/digitalnomads, I want to see if I can create a model that could predict if I were to write a post, in which subreddit group would it be more likely to appear?

Data Extraction

Before drawing any insight from the social media platform, the first step is extracting the data; in doing so, we need to use PRAW. PRAW is a python Reddit API wrapper that scrapes active subreddits. To utilize PRAW, a Reddit account is required, and next, we would need to apply for credentials. The credentials (client_id and client_secret) are required to access the application. All the necessary websites to create credentials can be found in the first cells of my notebook. Never share your key, as it is your access to the API and should only be used by you.

After gathering the correct credentials, the next step is to write the function to collect the data from a certain number of subreddit posts. For this project, Max_docs shows the maximum number of documents for each subreddit, subreddit_text is the empty list of where the documents will be appended from submission.title (Add title to list of 'documents'), submission.selftext (Add text to list of 'documents'), submission.comments (Get list of all comments for a particular post), submission.comments.list (Iterate over comments in a specific post), and comment.body (Add comment to list of 'documents'). For our initial analysis, we picked r/WFH and r/digitalnomads.

Data Pre-processing

To ensure our data was usable, I used df.info() to evaluate the data and print(df['subreddit'].value_counts()) to make sure we had extracted the correct amount of documents from both subreddits. I cleaned any duplicate posts and the initial line of each subreddit used for the group's definition. I evaluated the process text using tokenizer, stemmer, and .lower(). Ultimately, I decided to use text and subreddit to continue the project.

Exploratory Analysis

This section goes through the data frames created for the post, titles, and comments from July 29 , 2022, to the last 5,000 documents from each subreddit. All text in the data frame is unique. I removed duplicate texts and the first column of each subreddit where it is the introduction.

Building the model
This section attempted several models, which showed two runs of different pipelines and grid search instances. However, I worked on other models, including knn, naive bayes, and random forest classifiers. My test data with the highest scores were with a count vectorizer and logistic regression.

The precision of this model was evaluated with a confusion_matrix. The graph shows that digitalnomads and WFH are predicted to be the same amount. The precision of WFH is .80 while digitalnomads is .83.

Data Visualization
Several visualizations were created for this project, one for the precision of the model, one for the common words in the subreddit digitalnomads, and a word cloud for the subreddit digitalnomads. If we observe the word clouds, we can see that the wordclouds from r/digitalnomads follow a pattern that we would be able to distinguish which one is a more adventurous work from home lifestyle while the other subreddit, r/WFH, is more serious about working from their literal home. The word correlation is Airbnb, visa, city, and nomad being the most prominent.

Future Work
Using Natural Language Processing methods, I analyzed both subreddits r/WFH and r/digitalnomads to train a Natural Language Processing model to identify what subreddit a test string is more likely to originate from (subreddit group). I could observe more specific subreddits or users with further effort, look at sentiment analysis, and add a view of how individual users change their subreddit digital fingerprint over time.

A good next step could be to use PRAW to gather posts from more groups such as r/WFH, r/workfromhome, and r/digitalnomands and analyze the intensity of the interaction connections between these three subreddits and visualize the suitable interconnections as the relationships extend to outer groups. Reddit is one of the only social media platforms that does not have follower counts; however, it utilizes karma which signifies a user's score, which is determined by totaling the number of upvotes against their downvotes. It would be interesting to see who are influencers in specific subreddits and see if they are active influencers in others. Are influencers in one group an influencer in another? What are the connections? If I were to continue this project, I would account for user comments contributing to narrowing down influencers.