Don-Thuan Le

CS 410: TIS

Progress Report

1. Progress made thus far: A lot of time has been invested into doing research for the project, specifically how to implement the chrome extension. I have gotten the overall gist of the chrome extension but am still exploring the capabilities as to how to use the chrome extension to scrape information from the google query and to aggregate the information into a data source. For the sentiment analysis, I decided to go with a pretrained model from the nltk library to free up some time. I have found a viable dataset for demographic based on text using the HappyDB dataset from Kaggle and have been working to clean the data for gender and age – I will probably decide to start with gender as it creates a simpler model.
2. Remaining tasks: I still have a lot to do. This includes fully implementing the chrome extension, and if this doesn’t work that at least creating a local example query dataset so that I may run test scripts on. The chrome extension seems to be the most extensive part. Then after I clean the HappyDB dataset, I need to train a model to predict gender. If time is taken too much by the chrome extension, I may opt to train using FastAI.
3. Challenges: My greatest challenge currently is figuring out how to scrape the URLs and information from the webpages returned from a query. Much of it means exploring chrome extension. I have found the fetch() function which seems to be along the lines of what I need. Then I need to be able to search each page for the specific main text that I need and pull that into a document so that I can run the python code on it.