Don-Thuan Le

CS 410: Text information Systems

Final Project Proposal

**Requirements**

If you choose this theme, please answer the following questions in your proposal:

1. What are the names and NetIDs of all your team members? Who is the captain? The captain will have more administrative duties than team members.
2. What topic have you chosen? Why is it a problem? How does it relate to the theme and to the class?
3. Briefly describe any datasets, algorithms or techniques you plan to use
4. How will you demonstrate that your approach will work as expected?
5. Which programming language do you plan to use?
6. Please justify that the workload of your topic is at least 20\*N hours, N being the total number of students in your team. You may list the main tasks to be completed, and the estimated time cost for each task.

**Team Members (Requirements: 1)**

1. (Captain) Don-Thuan Le – Net ID: dtle2

**Background and Motivation (Requirements: 2, 4)**

The topic that I chose was Intelligent Browsing. The specific task is the merging of two interests: data mining and analytics, and general research into how different ethnicities/cultures think. For example, if someone wanted to research the sentiment behind a certain topic, they would go to google and manually look through links to see what each person had to say about their topic. This project will take it one step further to automate the sentiment analysis of returned pages and try to create a way to categorize the sentiments by culture or even age.

The goal of the final product is a chrome extension (or maybe a different method of integration with google) that allows a user to perform sentiment analysis over their desired topic. From the desired topic, my product would first return the top most positive sentiment pages and the top most negative sentiment pages. The last step is to analyze each page to see if we can match the page to a certain culture or ethnicity. The output to the user would be a table showing what cultures or ethnicities are most common in having the respective sentiment about your chosen topic.

This is related to the course as it not only implements a text retrieval problem, but also adds sentiment analysis and probabilistic language modeling to find out what culture the page most likely refers to.

**Algorithms, Datasets, and Programming Language (Requirements: 3, 5)**

For sentiment analysis, algorithms that I will potentially use are logistic regression, SVMs, Naïve Bayes or maybe even random forests. To train these models for sentiment analysis I would potentially datasets such as: Sentiment140, Opin-Rank Review, and the IMDB Movie Reviews. For programming language, I will be using python to train the models. To create a final product linked to google chrome, I believe I will have to use Javascript.

**Workload and Justification (Requirements: 6)**

As a one-man team, I believe this project will take more than the 20-hour minimum requirement to fully complete. Below is a high-level summary of the tasks that will need to be completed (side note: considering how long it took me to wrap my head around MP3…I have my work cut out for me).

1. Data processing
   1. Tokenization, stop word removal, stemming, and lemmatization
2. Learning
   1. Feature extraction
   2. Model selection
3. Evaluation (sentiment)
4. From the above evaluations, mine the pages for more specific information using topic modeling
5. User interface

Future implementations of this project could be applied for users to do general research on opinions of different groups of people about certain topics. Presidential candidates could see what groups of people they need to be reaching out to more. Psychologists could conduct research with a hypothesis in mind when surveying how people think of a topic in person.