CSCE 578 Final Project Proposal

Noemi Glaeser & Nick Quan
29 March 2019

We plan to conduct several analyses on the corpus of our shared friend group chat. They are as follows:

1 Clustering

We propose a continuation/improved version of Noemi's Assignment 02, which uses `tf-idf` values to group people based on what they talk about. Some problems to address are misspellings, dealing properly with URLs, possibly excluding emojis, and using an updated stopword list that excludes contextually meaningless words like "lol", "yeah", and "like".

2 Text Complexity

We also plan to run a continuation/improved version of Noemi's Assignment 03, which uses the Stanford POS tagger to determine each person's average sentence complexity (using element tree depth). An improvement could come from looking into the *Washington Post*'s political language analysis methods.

3 Sentiment Analysis

Another interesting possibility is conducting a sentiment analysis to see if the mood of the chat shifts based on the time of year. Two different sentiment analysis libraries (NLTK VADER and TextBlob) will be compared to observe which works the best to categorize messages into positive, negative and neutral categories. We will analyze the messages before and after exam periods to see if there is more negative sentiment around these times. We may also compare different seasons of the year to see if there is evidence for some sort of seasonal depression, i.e. more negative messages in the winter than in the summer.

4 Topic Modelling

We also hope to do a continued/improved version of Nick's Assignment 03. We propose to use `gensim` topic modelling on daily chat logs to look for overarching monthly topics. An improvement to be made is using NLTK POS tagging to identify nouns, stem them, and

use them as input tokens (instead of the naïve implementation which simply uses capitalized words).