SOCIAL MEDIA ANALYSIS FINAL PROJECT PROPOSAL: VOCABULARY ENTROPY AMONG REDDIT COMMUNITIES

NICHOLAS LINES

1. Introduction

In this document I present a proposal for my final project, which is focused on vocabulary entropy among Reddit communties, focused primarily on political discussion subreddits. It is not a draft of the paper, but is intended to help me explore the sections I'll be writing. This discussion is motivated by a desire to augment distance-reading methods like Topic Modeling with language-agnostic techniques for profiling a community of writers from a distance. In my final project I'll ask whether political discourse on Reddit is most similar to high-school level, graduate-school level, or non-native English writing, using as my metric the mean vocabulary entropy in samples taken using the Python Reddit API PRAW. My intended peer-review process is to submit this to KDD.

2. Background

Use of vocabulary to profile writers or groups of writers is commonplace, but fraught with difficulties. While it is generally acknowledged that vocabulary size is correlated with fluency and education levels, sub-exhaustive methods for approximating vocabulary size are difficult find. One possible solution is to instead measure vocabulary entropy in a document. Suppose we have a document that is n words long composed using V distinct words $X_1, X_2, ..., X_V$. These words's frequencies are observed and used to write the vocabulary's empirical probability distribution for the random variable for word choice, X. The Shannon entropy of the document is defined to be

$$H(X) = -\sum_{i=1}^{V} P(X_i) \log_{10} P(X_i).$$

This provides an excellent measure of the spread of vocabulary use in the document, and intuitively more entropy indicates broader vocabulary knowledge, which is associated with better education and language skills. However, vocabulary use in writing follows Heap's law, which states that the number of unique words observed V in a document of length n words is given by

$$V(n) = \alpha n^{\beta},$$

where α and β are parameters dependent on the author and possibly the language and context of the document. This means that document length will greatly affect the entropy measure. Suggestions have been made for how to resolve this problem so that comparisons can be made between documents of varying lengths. In [2] the authors put forward "a novel approach" for normalizing the vocabulary entropy

Date: April 23, 2021.

English as a second language	High school	Graduate School	US Politics
534	923	4260	20411

TABLE 1. Subreddit group sample sizes

term, in which they simply divide it by the maximum entropy of a document of the same size, $-\log_{10}(1/n)$, and call the result the vocabulary quotient. It turns out that this exact method was also advocated in [1, p. 551]. Unfortunately, it can easily be shown that the vocabulary quotient is unstable, and indeed that it strictly decreases as n increases. As a result, we will restrict our review to universal fixed lengths of $n = 100$ words. We'll continue to use vocabulary quotients rather than the unnormalized Shannon entropy since it provides a minimal amount of consistency.

3. Data

I have gathered posts and comments from Reddit in four main categories: a sample from Redditors for whom English is a second language, a sample from high school related subreddits (i.e. text written by high school students), a sample from graduate school related subreddits, and finally a much larger sample from a variety of US political discussion subreddits. From these I kept only the posts or comments that were 100 tokens long or more after removing punctuation and tokenizing. Each item is cut down to just the first 100 tokens, resulting in the sample sizes shown in Table 1.

4. Methodology

The research question to be tested is whether the the average vocabulary quotient of Reddit political discussions is more similar to that of graduate students or one of the less entropic groups. I've started some of this, but still have work to do. So far I've shown that the mean vocabulary quotient for each population is different, with the expected result that graduate students have higher entropy in their text than high school students, and the English learners have the least diverse vocabulary. However, I've been surprised to see that the US Politics vocabulary quotient is better even than graduate school text. These results are summarized in a box plot in Figure 1. I will need to show that these differences are statistically significant. Next I'd like to break the political data out by subreddit and see if there are some subreddits more prone to lower vocabulary quotients. If I have time, I'd also like to take examine a social network whose nodes are Redditors whose posts/comments in US Politics subreddits appeared in the dataset, but including edges from shorter messages as well; if possible I'd like to investigate the relation of node centrality and vocabulary quotient.

I am also a big believer in open research and reproducibility, so I want to provide all my code and relevant material in a GitHub repository that is world viewable.

5. Questions for the instructor

I have a couple areas I'd appreciate guidance or suggestions in. I'd like to provide a copy of the data that I used, but I also want it to be sanitized. Do you think it is sufficient for me to simply provide token index lists rather than the raw data? Admittedly this is just a substitution cipher using numbers for words, so a lot can

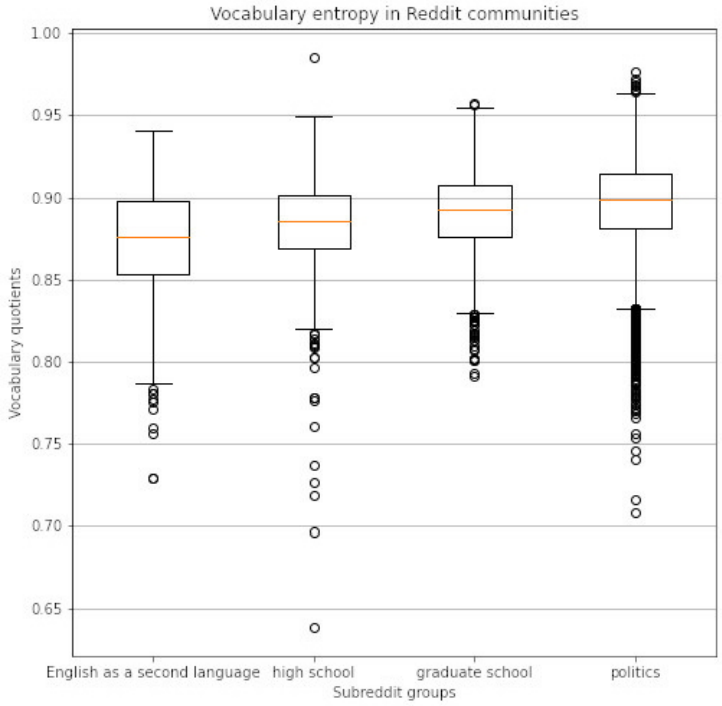

FIGURE 1. Vocabulary quotients in several subreddit groups.

be recovered based on frequency tables, and confirmed using public Reddit data. But the data is all public on Reddit anyway, so I don't think that is a big deal. I also am looking for suggestions on where to host the data. In a pinch I can keep it in the GitHub repository, but that seems less professional than hosting it somewhere standard.

REFERENCES

1. R. Dale, H. Moisl, and H. Somers, *Handbook of natural language processing*, Taylor & Francis, 2000.
2. Nikhil Kumar Rajput, Bhavya Ahuja, and Manoj Kumar Riyal, *A novel approach towards deriving vocabulary quotient*, Digital Scholarship in the Humanities **33** (2018), no. 4, 894–901.

E-mail address: nicholasalines@gmail.com