Homework 3

Abhi Agarwal

Three Tables

General

1. I only filtered words that were in english - so I used a function to remove and normalize the datasets to only having english words.

Ebola dataset

1. The Ebola dataset contains 36,997 words, and 5,225 of the words are not english at all. Having removed the non-english words, the subset that remains has 25,122 words that are not in the dictionary. Therefore it leaves 6,650 unique words that exist in the english dataset.

2. I only did sentiment analysis on words that were within the english language or words that I could have corrected using similarity matching.

If They Gunned Me Down

1. The If They Gunned Me Down dataset contains 7,911 words, and 794 of the words are not english at all. Having removed the non-english words, the subset that remains has 3,525 words that are not in the dictionary. Therefore it leaves 3,592 unique words that exist in the english dataset.

US Top 10 Cities

1. The If They Gunned Me Down dataset contains 111,771 words, and 30,299 of the words are not english at all. Having removed the non-english words, the subset that remains has 75,434 words that are not in the dictionary. Therefore it leaves 6,038 unique words that exist in the english dataset.