Homework 3

Abhi Agarwal

Three Tables

1. The first table I have created is to show sentiment analysis across all three datasets. Specifically I am looking at the sentiment analysis result that comes out when I'm using all the english words that are given, and comparing them to all the english words that are given provided they exist in the dictionary.

Dataset	2014-10-20	2014-10-27	2014-10-06	2014-10-13
Ebola with only english words	0.00194	0.000158	0.00411	0.00120
Ebola with only dictionary words	0.00999	0.000756	0.0202	0.00671
If They Gunned Me Down with only english words	-0.00201	-0.0201	-0.00333	-0.00106
If They Gunned Me Down with only dictionary words	-0.00399	-0.0398	-0.00660	-0.00211
US Top 10 Cities with only english words	0.000135	-0.0000689	-0.000104	0.0000601
US Top 10 Cities with only dictionary words	0.00162	-0.000615	-0.00111	-0.000861

2. The second table

x	y
x	y

3. The third table

x	y
x	y

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.