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 using a subset of each of the dataset. The first subset is all the words that for each letter of those words it exists in the english alphabet, and then the second subset I'm comparing is all the english words that are given provided they exist in the dictionary. The technique of sentimental analysis was to try and find the sentimental value of a particular word, multiplying it by how many times it occurs, then adding the result. It's a little basic, but it was difficult to figure out how to find the sentimental value for each particular word.

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

Here we can see interesting results. It's a little clear just looking at each of these particular examples that the words that are only in the dictionary improve the sentimental value in either way (either it improves it more positively or more negatively). I also can't accurately say which one is better as I don't have the tweets themselves. It's interesting to see this result because it shows how much non-dictionary words change the sentimental value of the dataset. It's hard to find a sentimental value for words that don't exist in the dictionary because it's hard to try and find their meaning without the user going through and labeling them.

2. The second table I concentrated on the most occurring words, and tried to give them some contextual information. I was trying to tag them depending on what they could mean. The first dataset I used was a dataset where all the letters of each word are in the english alphabet, and the second dataset was a dataset where all the english words were in the dictionary. The difference here becomes that we would know how to differentiate between locations/individuals & words in the dictionary using this - it's simple, but powerful in my opinion.

   The REFERENCE#1 in the table points to the word emabiggestfansjustinbieber (LaTeX had lots of problems with me putting it into the table). In the table I have put the most occurring words for each of the datasets I described above as rows.

Dataset	2014-10-20	2014-10-27	2014-10-06	2014-10-13
Ebola with only english words	vinson klain williamsburg	vinson klain othertech	braintree trembles paparazzo	klain vinson bridal
Ebola with only dictionary words	evoking unenforceable disorganization	exclusions sleuths intercepts	trembles paparazzo canines	bridal faulted widened
If They Gunned Me Down with only english words	pumpkins recount nh	shawshooting condemns flier	detention interrupt deadspin	shawshooting vonderritmyers anatomy
If They Gunned Me Down with only dictionary words	pumpkins recount antiquated	condemns flier reignite	detention interrupt evacuation	anatomy molester interventions
US Top 10 Cities with only english words	nashsnewvideo followmenash michaelbrown	followmecam sanayairani nashvschad	REFERENCE#1 cmaawards helpfindthem	REFERENCE#1 funfearlesslife helpfindthem
US Top 10 Cities with only dictionary words	crewed juries reclassified	emulsion biopics foamed	galactic ghouls reelection	flog dozier sleepwalkers

3. The third table

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

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.