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	Sentiment analysis of words on **2014-10-20**	Sentiment analysis of words on **2014-10-27**	Sentiment analysis of words on **2014-10-06**	Sentiment analysis of words on **2014-10-13**
Ebola with only english words	0.0194	0.00158	0.0411	0.0120
Ebola with only dictionary words	0.0999	0.00756	0.202	0.0671
If They Gunned Me Down with only english words	-0.0201	-0.201	-0.0333	-0.0106
If They Gunned Me Down with only dictionary words	-0.0399	-0.398	-0.0660	-0.0211
US Top 10 Cities with only english words	0.000135	-0.000689	-0.00104	0.000601
US Top 10 Cities with only dictionary words	0.00162	-0.00615	-0.0111	-0.00861

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.

The scale of the data basically goes from -1 being the most negative to +1 to be the most positive. It's quite interesting to see the results across all the platforms. It seems like the Ebola tweets were barely above average in sentiment so a little more positive than negative. This seems interesting because the conclusions we can make are that some people have a positive view of looking at it from a perspective of being hopeful while others had a negative view of looking at the sorrow.

For the If They Gunned Me Down it seems like they were both slightly on the more negative end of things, which means that people were more negative. It seems interesting because the sentiment of the title itself comes out to be negative because of the words Gunned and Down. In either of these cases it helps us realize if people were more hopeful or not.

The last one was a little confusing to me and the most interesting. It fluctuated between either being more towards the positive end and more towards the negative end. The data here is basically Tweets from random cities using geo-location. This makes it really hard to kind of understand what the data is about, but this fluctuation between positive and negative could basically come from the fact that the data is so diverse and not concrete on an issue.

Lastly, neither of the results above are either descending or ascending over time, but have a very little variance, which shows that time doesn't necessarily have to do with sentiment.

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.

Dataset	Most common words on **2014-10-20**	Most common words on **2014-10-27**	Most common words on **2014-10-06**	Most common words on **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	emabiggestfans cmaawards helpfindthem	emabiggestfans funfearlesslife helpfindthem
US Top 10 Cities with only dictionary words	crewed juries reclassified	emulsion biopics foamed	galactic ghouls reelection	flog dozier sleepwalkers

The emabiggestfans 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. The very interesting and clear details we're able to see here is that location/individuals are usually more common than english words, and that's actually very powerful. The fact that more individuals are mentioning ideas/people/concepts from each of these datasets is very powerful (definitely what relates these tweets together obviously). It's also very interesting to see the words that are used in those tweets that are also in the dictionary - such as bridal, faulted, intercepts for tweets related to ebola. It seems very bizarre, but these help give some contextual information about the events themselves as we see the changes in words that occur.

Also very interesting to words like evacuation, reignite, and recount in the If They Gunned Me Down dataset as this actually gives you an understanding of the events

without even knowing what the event was about. The word pumpkins also suggests some time frame of when it could have been (thanksgiving), and this is very cool to see because we managed to get both context and some description of the events.

As always there tweets that have Justin Bieber at that top! This exists mostly in the US Top 10 Cities dataset, which I'm hopeful for - because it should not have been high in the other two datasets (it's unrelated, and shows what people are tweeting/thinking about during these events).

From these events I think it's powerful to get some meaning out of them as removing words and fitting them into dictionary words shows you how many words there were, which were entities - location/individuals, etc.

This kind of shows that time is powerful when looking at words that people are using, and both context and the background information do change depending on the issues that occur. It was very interesting to see the progression of events that occurred for the US Top 10 Cities as we can see that there were big events with Michael Brown, Nash, some awards show, ema, and issues about life?

3. The third table is to look at the most occurring languages within each dataset. This is interesting to look at because we know that the tweets must have fallen through the twitter cracks because it was a mixture of both another language and english. So it'll give us some understanding of the different environments or places tweeting about this particular event. Now that we can get contextual information on location/people it's also interesting to see who the people are / what languages / what cultures are tweet about it. These languages are WITHOUT english, which is at the top for all of these.

Dataset	Most Occurring Languages across all dates
Ebola	Haitian Spanish French Portuguese Finnish
If They Gunned Me Down	Haitian Portuguese German Italian Swedish
US Top 10 Cities	Haitian Spanish Finnish French Italian

There are about 30 different languages for the Ebola dataset. I found this interesting because there were an incredible number of languages that I presume were embedded with the english language. I really was hoping to try and translate all these words to see if we could match any of them. This was a little too slow to process as was the language detection algorithm.

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 US Top 10 Cities 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.