Assignment 5

L245

Due Monday, March 27

1. Do question #7 in chapter 4 of the textbook (p. 122).
2. Do question #10 in chapter 4 of the textbook (p. 123).
3. Write a regular expression to capture the variants of the last name *Gaddafi*. See `http://blogs.abcnews.com/theworldnewser/2009/09/how-many-different-ways-can-you-spell-gaddafi.html` for insight on how this can vary.
4. **Bonus:** Do question #12 in chapter 4 of the textbook (p. 123–124).
5. (Based on a previous assignment by Jason Baldridge): Services like Twitter allow short, real-time commentary about whatever users feel like talking about, and there is often interest in automatically determining whether a given tweet is positive, negative or neutral toward a specific topic, company, etc. Here are old examples of positive & negative tweets regarding former president Obama & broccoli:

Positive	Negative
1a. Great goal to be set Mr. Obama thinking internationally the way to the 21st century meeting promises and goals of our people.	2b. Obama got crushed at town hall meeting today. His replies were terrible. He couldn't tell them the truth. It would kill Dems in Nov.
2a. It's not strange. I love cabbage and LOVE broccoli!	2b. I smell broccoli...oh how i hate that smell.... ehh...

These are all pretty straightforward for a document classification algorithm to pick up on because of the use of clear sentiment words that support the underlying sentiment, but in many cases it won't work so well. For example:

Positive	Negative
3a. All this President Obama backlash is terrible and is unjust. He inherited PROBLEMS, made promises, and was expected to walk on water	3b. America still needs to be focused on job creation. Not among Obama's great accomplishments since coming to office !!
4a. I hate ranch dressing, unless it's on my broccoli(like it is now)	4b. yeah i love broccoli HAHA

For example, 4a uses the discourse connective *unless* that contrasts the portion of the tweet with the sentiment word and in doing so, flips the polarity. 3a is positive about Obama, but its main focus is on characterizing the backlash. A document classifier would see the tweet mentions *Obama* and says a lot of negative words and would likely infer it to be negative about Obama. 3b involves negation (*Not*) which flips the polarity of the sentiment word but which is not immediately next to the sentiment word. That is harder to spot than things like *not great* and *do not love*. 4b is clearly sarcastic. Go to Twitter and, using its search functionality (`https://twitter.com/search-home`), find two positive and two negative tweets about Bloomington, IN, as follows:

- A positive tweet about Bloomington that is straightforward (as in the first group of tweets above).
- A negative tweet about Bloomington that is straightforward (as in the first group of tweets above).
- A positive tweet about Bloomington that would be trickier for a document classifier (as in the second group of tweets above).
- A negative tweet about Bloomington that would be trickier for a document classifier (as in the second group of tweets above).

Write all of these down in your homework submission and for each one, briefly state why it fits the description above.