Assignment 5

Cryptography
Document classification

Due Thursday, October 31

2. Do problem #4 from the same website.
3. (Based on a previous assignment by Jason Baldridge):
   Services like Twitter allow short, real-time commentary about whatever users feel like talking about, and it is one of the services that is creating data of great interest for sentiment analysis. Often, there is interest in automatically determining whether a given tweet is positive, negative or neutral toward a specific topic, person, company, etc. Here are examples of positive and negative tweets with regard to President Obama and broccoli:

<table>
<thead>
<tr>
<th>Positive</th>
<th>Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>1a. Great goal to be set Mr. Obama thinking internationally the way to the 21st century meeting promises and goals of our people.</td>
<td>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.</td>
</tr>
<tr>
<td>2a. It’s not strange. I love cabbage and LOVE broccoli!</td>
<td>2b. I smell broccoli...oh how i hate that smell.... ehh...</td>
</tr>
</tbody>
</table>

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:

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

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 in a classifier 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.

(It is okay if your examples use swear words, though please do avoid using especially offensive tweets.)

4. Select two articles in the Indiana Daily Student (http://www.idsnews.com/) written by different authors.

(a) Give the title, author’s name, and date of each article.
(b) What are some prominent stylistic differences between the two authors? Describe at least 3 differences.
(c) Could these differences be detected automatically? How?