CS 6320 Natural Language Processing

Homework 1: Due 09/05/2014

Your name: ****

Please submit your homework on elearning.
Please name your attached file Homework1_firstname_lastname.[doc|pdf|txt].
Remember to put your name in your submission (not just the file name, but in the actual write up).

1. (10 points) Try a few sentences using the Babelfish MT system (http://babelfish.com/) or Google translator (http://translate.google.com/). You can use any language pairs supported.
   (a) List 4-5 problems with the translations. Explain the problems in a way that somebody not speaking these languages can understand.
   (b) Identify two of the most severe problems from those in (a), and suggest some probable cause.
   Note: If you don't know languages other than English, try to team up with another student in the class to finish this question. Another solution is to use the system to translate an English sentence to another language, and then back to English. Based on the errors, try to determine what went wrong in the translation process.

2. (10 points) You must have all used some search engines (Google or others). Please give one or more examples when the system didn't interpret your queries correctly and the returned top ranked documents are not very relevant to what you were looking for. Can you speculate the reasons for the imperfect system and do you have suggestions on how to improve performance?

3. (15 points) Find some review text (for movies, products, etc.) and explain **using your examples** how to classify them into positive and negative categories, and why such a task is challenging. You can also use blogs or tweets for this problem.

4. (20 points) We mentioned in the class that for many natural language processing tasks, the first thing often needed is to split the text into sentences and words. This sounds easy, but it is not that trivial! Your assignment for this problem is to:
   (a) read the tokenization paper available from the course webpage
   (b) find some real examples to show the ambiguity in sentence boundary detection and word tokenization. In your submission, please write down where your examples come from. Note that the paper is from more than a decade ago. You might consider whether there are new rules or examples.

5. (30 points) Use the text provided in the homework page to do the following. For each file:
   (a) Word count: for each word in the book, count the number of times it occurs in the document. Print the top 30 words with highest counts, in descending order.
   (b) How many distinct words are there in the book?
   (c) How many words have occurred once? twice? three times?
   There are two files:
   http://www.hlt.utdallas.edu/~yangl/cs6320/homework/pride_and_prejudice.txt
   http://www.hlt.utdallas.edu/~yangl/cs6320/homework/sherlock_holmes.txt

one is *Pride and Prejudice*, from http://www.gutenberg.org/etext/1342.
The other is *The Adventures of Sherlock Holmes*, from http://www.gutenberg.org/ebooks/1661.
You can see how your results differ for the two input files.

You can use any programming languages for this problem. You may find a lot of Unix commands (e.g., sort, wc) very handy for the kind of statistics asked in this problem.

Please include a short description of your preprocessing step (e.g, how you did word segmentation) or assumptions (e.g., upper and lower case words are same/different).

6. (15 points) Paper reading. Pick one paper from ACL 2014 to read, write a paper critique. The conference webpage is http://www.acl2014.org/, and the proceeding is available from ACL anthology (http://aclweb.org/anthology/P/P14/). Please provide the paper information in your submission.