ASSIGNMENT 4

COMP 599, Fall 2015

Due: November 26^{th}, 2015 in class. No late assignments accepted.

You must do this assignment individually. You may consult with other students orally, but may not take notes or share code, and you must complete the final submission on your own.

Question 1:	60 points
Question 2:	40 points
	100 points total

Assignment

Question 1: Multi-document Summarization (60 points)

This question asks you to implement a simple, but surprising effective algorithm for multi-document summarization, SumBasic:

Ani Nenkova and Lucy Vanderwende. The Impact of Frequency on Summarization. *Microsoft Research, Redmond, Washington, Tech. Rep. MSR-TR-2005-101.* 2005. `http://www.cs.bgu.ac.il/~elhadad/nlp09/sumbasic.pdf`

a) Use a news aggregator tool such as Google News to find four clusters of articles on the same event or topic. Each cluster should contain at least three articles, and each article should be of sufficient length to generate an interesting summary from (at least 3–4 paragraphs).

You should clean the article text by removing all hyperlinks, formatting, titles and other items that are not the textual body of the articles. Use any method to do this (including by hand). You may have to deal with non-ASCII characters. You can handle them any way you like, including just replacing them by a similar-looking ASCII character. Save your input into text files called `docA-B.txt`, where A is an positive integer corresponding to the cluster number, and B is another positive integer corresponding to the article number within that cluster. For example `doc1-2.txt` is the second article in the first cluster. Put all of your documents inside a subfolder called `/docs`.

b) Implement SumBasic, as it is described in the lecture notes, in order to generate 100-word summaries for each of your document clusters. Compare these two versions of the system:

1. `orig`: The original version, including the non-redundancy update of the word scores.
2. `simplified`: A simplified version of the system that holds the word scores constant and does not incorporate the non-redundancy update.

Compare these versions against a third method, `leading`, which takes the leading sentences of one of the articles, up until the word length limit is reached. You may decide on how to select the article arbitrarily.

You should apply the standard preprocessing steps on your input documents, including sentence segmentation, lemmatization, ignoring stopwords and case distinctions. The main method that should run

your code should be in a file called `sumbasic.py`. Your code should be run using the following command structure:

```
python sumbasic.py <method_name> <file_n>*
```

And it should print the output summary to standard output.

For example, running

```
python ./sumbasic.py simplified ./docs/doc1-*.txt > simplified-1.txt
```

should run the simplified version of the summarizer on the first cluster, writing the output to a text file called `simplified-1.txt`.

c) Discuss quality of each of the three methods. Does the non-redundancy update work as expected? How are the methods successful or not successful? How would you order the summary sentences with the SumBasic methods, or another extractive summarization approach? Be sure to cover all aspects of summary quality that we discussed in class.

Question 2: Reading Assignment — Natural Language Generation (40 points)

Read the following paper:

Irene Langkilde and Kevin Knight. Generation that Exploits Corpus-Based Statistical Knowledge. *ACL 1998*. `http://www.aclweb.org/anthology/P/P98/P98-1116.pdf`

Write a max. one-page (c. 500 words) discussion on this paper, including the following points:

1. A brief summary of the contents of the paper, including the theoretical framework and the experiments.
2. Relate this paper to the following concepts that we have discussed throughout the term: *language modelling*, *underspecification*, and *morphological analysis*.
3. Three questions related to the paper. These can be clarification questions, or questions about potential extensions of the paper, or its relationship to other work.

What To Submit

On paper: Submit a hard copy of your response to Question 2, **separately from** the report part of Question 1 in class.

Electronically: For the programming part of Question 1, you should submit one zip file with your source code, input document clusters, and output summaries to MyCourses under Assignment 4.