ASSIGNMENT 4

COMP 550, Fall 2018

TA: Jad Kabbara, jad.kabbara@mail.mcgill.ca

Due: November 28^{th}, 2018 at 11:59 pm.

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:	40 points
Question 2:	60 points
	100 points total

Assignment

Question 1: Reading Assignment — Multi-document Summarization (40 points)

Read the following paper:

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`

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. The limitations of the approach. What do you suggest to address these limitations?
3. Discuss the advantages and disadvantages of using ROUGE as an evaluation measure.
4. 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.

Question 2: Multi-document Summarization (60 points)

For this question, you will implement the algorithm that you read about in the above paper, SumBasic:

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 paper, in order to generate 100-word summaries for each of your document clusters. Compare these three versions of the system:

1. `orig`: The original version, including the non-redundancy update of the word scores.
2. `best-avg`: A version of the system that picks the sentence that has the highest average probability in Step 2, skipping Step 3.
3. `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 fourth 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) Assess the quality of each of the methods. Does the constraint of always including the sentence with the best word help? 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.

What To Submit

Electronically: Submit the written portions of the assignment in a single pdf file called 'a4-written.pdf'. For the programming part of Question 2, you should submit one zip file called 'a4-q2.zip' with your source code, input document clusters, and output summaries. Both should be submitted to MyCourses under Assignment 4.