Computer Science 4140
Midterm Question 7

Ethics policy reminder and computing environment reminder are the same as for **Midterm Question 1**.

Problem Overview

Text summarization is an application area where an automated system generates a summary of the main content of a document. The simplest of these tools generates the summary by extracting excerpts from the document. In this problem you will write a very simple summarization tool that is based on printing the sentences of a document that contain the highest total word frequency. You will use *FreqDist()* to count word frequencies and *sum* to sum the frequencies of the words in each sentence. Rank the sentences according to their score. Finally, print the *n* highest-scoring sentences in document order. Carefully review the design of your program, especially this approach to this double sorting. Coding style is important and particularly elegant solutions can earn bonus points.

Program Requirements

- Name your program *mt7.py*. A starter template is provided on the class WWW page. The template is designed to use books from the Gutenberg corpus. There will also be a couple of sample results that you can use as test cases. However, do not assume that these test cases cover all situations.
- Because this approach is biased in favor of sentence length, we will also use a parameter *maxLength*. The *n* highest-scoring sentences should be constrained to contain no more than *maxLength* words.
- Per the computing environment requirement, I must be able to execute your program on our Linux server with a command line. An example command line is the following. Suppose I wanted to generate as a summary the top 10 sentences from *Moby Dick* whose length does not exceed 12 words. The following command line would be used.

  python3 mt7.py melville-moby_dick 10 12

- Further technical details are provided in the program template.

Program Deadlines

March 20, 09:00 A.M.

Deadline for submitting completed program. The <*assignment_number*> should be *mt7*.

Program Evaluation

This question is worth 30 points. Your grade on this question will be weighted as follows:

Code Correctness	70%
Coding Style and Documentation	30%