Spring Semester 2018
Iowa State University

ENGL 516 - Formal Methods of Linguistic Analysis

Assignment 4
Submission Deadline: 10 MAR 2018

Instructions: This assignment consists of two questions. Each question carries 5 marks. Both the questions involve writing python code. Save your code as two .py files (one for each question, as yourname_question1.py and yourname_question2.py), and submit it as a zip file on Canvas. If the file does not run and throws errors, you cannot get a grade unless you defend it in the office hours. If I provide a test case for each question, use it as a means to check if your code works. For both the questions, you need to write one paragraph at the end of your program as a multi-line comment, explaining what your program does at each line.

Question 1

The first question involves writing a program that reads the file Ibsen.htm (attached with this assignment) and shows the following output:

1. Number of lines in the html file are: YY
2. Number of times "Peter Stockman" appears in html file is: ZZ
3. Number of words (after removing punctuation) in HTML file are: XX
4. What are the words between $< h3 >$ and $< /h3 >$ tags in this file? Note: The tags can appear as upper case on some computers i.e., H3. Also, what I want is only the text between the H3 tags, not text within the tags themselves i.e., if there is $< H3id = "blah > textinbetween < /H3 >$, I want only "text inbetween" to be extracted.
5. How many times does a Mr. or Mrs. or Dr. Stockman appear in this file compared to any other word followed by Stockman?

Note: XX, YY, ZZ - these are placeholders for the actual numbers I should be seeing later.

Question 2

You have to work with a file called Ibsen.txt (attached).

1. What are the 10 most frequent words in Ibsen.txt? Do the necessary pre-processing (removing punctuation, lowercasing etc)
2. What are the 10 most frequent words in Ibsen.txt, if we discard the words in stopwords.txt file?