Fall Semester 2016
Iowa State University

ENGL 520 - Computational Analysis of English

Assignment 1
Submission Deadline: 10 SEP 2016, end of the day

Instructions: This assignment consists of three questions. Each question carries 5 marks. First question does not involve programming. Second and third question involve programming and you can use any programming language you want. Create a zip file with three components (a pdf file for first question, 2 perl or python files). If any of the programs does not run and throws errors, you cannot get credit for that. Late submissions are allowed, but will not be awarded full credit.

Question 1

Describe a method to detect tense errors in English learner essays automatically. You don't have to write a program. Explain your approach to solving this problem either as a step by step procedure, or using a flowchart, and briefly mention what kind of information/resources will you need to write a program to detect tense errors based on your procedure description. Please write clearly, and assuming that someone will have to write code after reading your algorithm.

Question 2

Write a perl or python program that takes the path to a .txt file as input, and does the following:

- Replaces all citations in the file with the word CITATION, using regular expressions (Citations come in various forms e.g., [12], ABC (2009), (ABC,2009), ABC and XYZ (2009), (A & B, 1984; X & Y, 2004), (XX et al., 2010) and so on.)

Note: You do not have to make it a perfect program, but try to include atleast 5 such formats in your regex (and mention what patterns does your program cover by writing a comment at the start of the program)

Question 3

Write a perl or python program that takes a path to a .txt file as input, and does the following:

1. Removes punctuation markers and splits the text into words by using white space seperator.

2. Counts all character trigrams in the text (character trigrams refer to sequences of three characters. So, if your text has a single sentence: "this is a sentence", character trigrams are: thi, his, is , s i, is , s a, a , a s, sen, ent, nte, ten, enc, nce)

3. Prints them in descending order (trigram which occurred most number of times appears first) along with their counts.