COMP 6781 **Assignment 1** **Winter 2016**

Due date: February 11, 2016
Late submission: 25% per day.
Teams: No teams allowed. The assignment must be done individually.

Question 1 (85%) : Zipf's law

Verify Zipf's Law by computing the frequency of words in a corpus. Pick an electronic corpus of your choice (for example, look at http://digital.library.upenn.edu/books/ for a list of online books). Make sure that your corpus contains plain text only (no HTML, for example), then:

1. Compute the frequency of each word. For the top 50 most frequent words, print the word, its frequency, its rank, and the frequency times the rank.
2. Print this information for every 50 words in steps of 1000 (ie. 1-50; 1000-1050; 2000-2050; …).
3. Print the number of words in the corpus that have frequency count i for i = 1 to 50 (the frequency of frequencies).
4. Graph the log-log values of the frequency and the rank as in Figure 1.1 of Manning & Schütze (p. 26).
5. Experiment with your program with different corpora, different corpus sizes … Does the data basically confirm Zipf's law?

Notes:

- For steps 1-3, any programming language will do. If you know Perl, this might be a good choice.
- To remove HTML tags, the following can be useful http://www.mbayer.de/html2text
- The following Unix commands may be useful. Do `man <theCommand>` for more information on each:
  - `sort <file>` sorts lines of text files (place one word per line)
  - `uniq <file>` removes duplicate lines from a sorted file (the option –c prefixes lines by the number of occurrences)
  - `cut <file>` removes sections from each line of files (the option –f outputs only specific fields)
  - `nl <file>` counts the numbers of lines in a file
- For step 4, you can use any Unix graphing/math packages (ex. Matlab, Splus, GnuPlot); or, on Windows, you can even use Excel. In that case, plot a XY (Scatter); right click on a data point, and do "Add Trendline" and ask for a linear regression model.
- For step 5, let your imagination and intuition guide you. The point is to experiment with different situations, then report your findings.

The report:
Write a report (~5 pages) to describe your code and your experimentation. You report must describe:

- The program:
  - Describe your code itself (choice of language, data structures, …)
  - Indicate the instructions necessary to run your code (files, commands, …)
- The experiments:
  - Describe your corpus/corpora briefly (size, source, …)
  - Describe what you did, and why you did it (why you thought it would be interesting to test)
- The results:
  - Analyse your results (do your experiments always confirm Zipf's law? compare the results across corpora, …)

Question 2 (15%): Linguistic Essentials

Do exercises 3.1 (only the first 2 sentences), 3.2, 3.3, 3.4, 3.9 and 3.12 and 3.13 on pp. 114-115 of Manning & Schütze.
Briefly justify or discuss any controversial points.

Submission:

You must submit your assignment electronically through the Electronic Submission Form (https://fis.encs.concordia.ca/eas/). Please submit :

- For question 1, the code of your programs, results and an electronic version of your report.
- For question 2, your typed answer.

The assignment must be handed electronically by midnight on the due date.

1. Make sure that you have signed the expectation of originality form (available on the Web page; or at: http://www.encs.concordia.ca/documents/expectations.pdf) and given it to me.

2. In addition, write one of the following statements on your assignment:
   "I certify that this submission is my original work and meets the Faculty's Expectations of Originality"
   with your signature, I.D. #, and the date.

3. Upload your files:
   - Create one zip file, containing all files of your assignment.
   - Name your zip file *a1_studentID*, where *studentID* is your ID number.
   - Upload your zip file at: https://fis.encs.concordia.ca/eas/