COMP 6781 **Assignment 1** **Summer 2013**

Due date: Tuesday May 21, 2013
Late submission : 25% per day.
Teams: The assignments can be done individually or in teams of two.

Question 1 (10%): Linguistic Essentials

Do exercises 3.1 (only the first 2 sentences), 3.3, 3.4, 3.9 and 3.12 on pp. 114-115 of Manning & Schütze.

Briefly justify or discuss any controversial points.

Question 2 (90%): Collocations

Pick 2 electronic texts of your choice that describe topics in different domains (eg. sports, politics, cooking,…), download the texts and perform the following operations:

1. List the 50 most frequent bigrams in your corpus along with their frequency.

2. Use a part-of-speech tagger of your choice (see http://www-nlp.stanford.edu/links/statnlp.html#Taggers for free taggers) and develop tag patterns to filter your bigrams. Show the best collocations you find this way.

3. Develop and experiment with 2 other methods discussed in class to find collocations. Analyze your results and compare them to those you had in step 2 above.

Write a report (~4 pages) to describe your experiments. Your report must describe:

- The program:
  - Briefly describe your code (choice of language, data structures, …)
  - Indicate the instructions necessary to run your code (files, commands, …)
- The experiments:
  - Describe any assumptions that you made (definition of a word)
  - clearly indicate the part-of-speech tagger that you used, and the tag patterns you developed
  - Describe your corpora briefly (size, source, …)
  - Describe the 2 methods you chose to implement (do not re-explain the theory, but just the parameters you used; window size, …)
- The results:
  - Analyse your results. Explain what went right, and what went wrong. For example, which method seemed to give the best results? Do the methods perform the same way on both corpora?
  - Show some examples of the "best/most interesting" collocations that you extracted and show some examples of the "worst" collocations that you extracted. Analyse these. Why do you think you have the results that you have?
- Future Work
  - Indicate what you could improve if you had the time and energy.
- References
  - Properly indicate all external sources that you used to do your assignment.

Note that your report should not be a detailed explanation of your code (data structures …). It should be an analytical report describing your experiments and an analysis of the results.

Submission:
Submit your answers to question 1, the code, corpora (and anything else required to reproduce your experiments) and the report electronically through the Electronic Submission Form https://fis.encs.concordia.ca/eas/