COMP 6781 **Assignment 2** **Summer 2013**

Due date: Thursday May 30, 2013
Late submission : 25% per day.
Teams: The assignments can be done individually or in teams of 2.

Word Sense Disambiguation

Use one of the machine learning techniques seen in class for word sense disambiguation to disambiguate the senses of 2 different words in a corpus.

For the training, testing and gold standard corpora, you can use any of the standard corpora available from any previous edition of the Senseval or SemEval shared tasks (see http://en.wikipedia.org/wiki/SemEval). You should take only the Word Sense Disambiguation (Lexical Sample) data sets, but feel free to do the experiment in language other than English.

For the implementation of the machine learning algorithm, you can:

- use the freely available package called WEKA (see http://www.cs.waikato.ac.nz/ml/weka/). WEKA includes several machine learning algorithms such as Naïve Bayes, decision trees… A nice tutorial on Weka is available at the URL: http://weka.wikispaces.com/Primer
- or simply implement your own code

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

- The basic experimental setup (~1-2 pages):
  - Describe which data set and which words you used and why you chose them (number of senses, number of training/test samples, …)
  - Describe the algorithm that you chose to experiment with (do not re-explain the theory, but just the parameters you used; window size, stop-word filtering…)
  - Describe any additional code that you may have written.
  - Indicate the instructions necessary to run your code (files, commands, …)
- Analysis of the results (~2 pages)
  - Analyse your results with the gold standard. For example, does one method perform the same way for both words? What if you change some parameters (window size, stop words…)? What effect has the size of the training set on the quality of the disambiguation?
- Future Work
  - Indicate what you would improve if you had the time and energy.
- References
  - Properly indicate all external sources that you used to do your assignment.

Again, 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 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/