NTMI 2016: Project Exercises
Part A (Step I)

This project has four steps. The steps are gradual and allow you to build the project in blocks that you can re-use in subsequent steps.

1 STEP 1: Extracting ngram statistics (Due date: 4th Feb. 2016

Write a program which takes as input a natural number n and a large text file (a corpus). The program should construct a table of all ngrams – word sequences of length n – in the corpus together with the number of times each sequence appears in the corpus (the ngram corpus frequency of the sequence). The word sequences should be exactly as they appear in the corpus, so 'The' $\neq$ 'the'.

Extend your program to take an additional argument m, and output the m most frequent sequences (together with their frequencies and in decreasing frequency order).

Test your program with the AUSTEN TRAIN corpus (`http://www-nlp.stanford.edu/fsnlp/statest/austen.txt`) as the input corpus. In this exercise we consider the whole corpus as one long sequence of words (in subsequent exercises we will split the corpus into sentences).

Submit the following:

1. For $n = 1$, $n = 2$ and $n = 3$, submit the list of the 10 most frequent sequences.
2. For $n = 1$, $n = 2$ and $n = 3$, submit the sum of all frequencies of all sequences for that n.
3. Submit your program with clear instructions as to how it should be compiled and run. Use the following command line format.

Command line format:
`./a1-step1 -corpus [path] -n [value] -m [value]`

`a1-step1` can be a bash/csh script file invoking a Java virtual machine or a python interpreter. If you are not familiar with scripts, it is sufficient if your java class accepts the specified parameters. Then you command line may look like: "`java -cp . A1Step1 -corpus [path] -n [value] -m [value]`." This applies to all assignments.