CMP 473.3 - Practical II **Instructor: Shailesh B. Pandey**

1 General Information

You can use programming language of your choice to complete this practical. Note that the lab computers may not have your language of choice installed. So it might be better to bring your laptop for the practical sessions. Alternatively you can choose to use the lab computer and programming languages installed in there. Please note that I may not be able to help with your code for languages that I am unfamiliar. I prefer to use `C, C++` and `JAVA`.

1.1 Submission

You will have to submit the practical report on `A4` sheets. You can prepare the report in a word processor and print it out. You have to include the source code in the `Appendix` section at the end of the report. Please do not print out the code (auto)generated by the programming platform. Make sure your code is pretty formatted and has comments to explain what is happening. Certain marks is allocated for good programming practices.

Submission deadline: Report is due next week. The submission will be followed by a 5-15 minute viva-voce. Performance on the viva-voce determines the marks you get for that practical. Late submission loses 20% marks for each late day. Please inform me ASAP in case you cannot make the submission deadline.

1.2 Tasks

For this practical you will need the text file (`shakespeare.txt`) containing complete works of William Shakespeare. You can either collect the text file from me or download it from `http://norvig.com/ngrams/shakespeare.txt`.

1.2.1 Part A

Your task is to read the contents of the file and produce:

1. A table containing 20 most frequent words. The table contains three columns: `rank, word` and `frequency`.
2. A table containing list of bottom frequencies. The table contains three columns: `frequency, word count` and `example words`. You are supposed to print word counts for frequencies 10 to 1. The rows in this table shows how many words have frequency 10,9,8 ...,1 with example of some of the words.
3. A table containing 20 most frequent word-pairs (bigrams). The table contains three columns: `rank, word pair` and `frequency`.

1.2.2 Part B

With the frequency counts of the word at our hand we calculate some basic probability estimates.

1. Calculate the relative frequency (probability estimate) of the words:
   (a) "the" (b) "become" (d) "brave" (e) "treason"
   [Note: $P(the) = \frac{count(the)}{N}$. Here, $count(the)$ is the frequency of "the" and "N" is the total word count.]

2. Calculate the following word conditional probabilities:
   (a) $P(court|the)$ (b) $P(word|his)$ (c) $P(qualities|rare)$ (d) $P(men|young)$
   [Read $P(B|A)$ as "the probability with which word B follows word A". Note: $P(B|A) = \frac{count(A,B)}{count(A)}$]

3. Calculate the probability:
   (a) $P(have, sent)$ (b) $P(will, look, upon)$ (c) $P(I, am, no, baby)$ (d) $P(wherefore, art, thou, Romeo)$
   Hint $\rightsquigarrow$ use the chain rule (multiplication rule):

   $$P(A, B, C, D) = P(A) * P(B|A) * P(C|A, B) * P(D|A, B, C)$$
   This is still difficult to calculate so we make the Markov assumption and end up with:
   $$P(A, B, C, D) = P(A) * P(B|A) * P(C|B) * P(D|C)$$

4. Calculate probabilities in Q3 assuming each word is independent of other words (independence assumption).

5. Find the most propoble word to follow this sequence of words:
   (a) I am no (b) wherefore art thou