Mini-Project 1 – Multicore Programming

Due: Monday, April 25th at 11:59PM EST

The goal of this project is to use your understanding of distributed computing to create a MapReduce program to compute the relative n-gram frequencies (for n=1, n=2, n=3) in a set of text files.

For n=1, i.e. f(wave) = N("wave")/N("*")

$$f(w_j) = \frac{N(w_j)}{\sum_{w'} N(w')}$$

For n=2, i.e. f(wave|big) = N("big wave")/N("big *")

$$f(w_j|w_i) = \frac{N(w_i, w_j)}{\sum_{w'} N(w_i, w')}$$

For n=3, i.e. f(wave|the big) = N("the big wave")/N("the big *")

$$f(w_j|w_i, w_{i+1}) = \frac{N(w_i, w_{i+1}, w_j)}{\sum N(w_i, w_{i+1}, w')}$$

The output file(s) must contain n-gram frequencies for (n=1, n=2, n=3) for each unique n-gram that appears in the input data.

i.e.:

```
...
waves       0.0020625
...
big waves   4.26942350E-3
big ...     3.84654807E-4
...
foaming waves     8.69316215E-6
...
the big waves     9.34161350E-5
...
```

The techniques you may want to consider are introduced in Lectures 12, 13 and 14.

Grading Criteria

- 60% - Correctness
- 30% - Write up - For three optimizations explored, describe:
  - How the speed up works
  - What is the expected speed up?
  - What is the observed speed up?
  - An explanation of any difference between the expected and observed speed ups
- 10% - Code quality - Good coding practices and well commented code

Submit your final version of `NgramCount.java` to **Mini-Project 3 - Code** on gradescope
- https://www.gradescope.com/courses/357643/assignments/1948305

Submit your team project writeup to:
- https://www.gradescope.com/courses/357643/assignments/1948289