Programming Assignment 4

1 Assignments

1. Calculate the relative frequencies of the co-occurrences of word pairs. Download the sample code from blackboard. Piece together the codes to calculate the relative frequencies of the co-occurrences of word pairs. This is the first implementation.
   - Do NOT change the hashCode() method in the wordpair class.
2. In the second implementation of the calculation of relative frequencies, remove the specific partitioner method from the mapreduce class. Instead, modify the wordpair class so that all the word pairs sharing the same left word go to the same reduce task.

2 Details

1. Use "pairs" approach in both implementations.
2. A word pair is a pair of words that are right next to each other.
   - For example, if <A B C> are three words in a sequence, then from B's point of view, both A and C are its neighbors. (B, A) and (B, C) are considered as two word pairs in this example when B is the focus.
3. The relative frequency of a word pair is defined as follows
$$f(w_j|w_i) = \frac{N(w_i, w_j)}{N(w_i, *)}. \quad (1)$$
$N(w_i, w_j)$ is the number of co-occurrences of the word pair (w_i, w_j). $N(w_i, *)$ is the number of co-occurrences of any word pair in which one word is w_i.
4. The input is a set of files. Each file contains multiple lines. Each line consists of a sequence of words that are separated by space.
5. The output should be ordered by the left word in a pair. For all the pairs with the same left word, order them using the right word.
6. Specify 3 reducers, i.e., generating 3 output files.

3 Submission

- Due date: February 25, 2019 at 11:59AM.
- Submission
  - For implementation 1, save the two class files into two files, name them wordpair1.java and relativefreq1.java; put these two files under folder "imp1".
  - For implementation 2, save the two class files into two files, name them wordpair2.java and relativefreq2.java; put these two files under folder "imp2".
  - tar -cvf pa4_<your last name>.tar imp1 imp2
  - Upload the tar file to blackboard before deadline