CS4371

Introduction to Big Data Management and Analytics

Fall 2025

Homework 1

Submission Deadline: October 1st, 2025, 11:59 pm

In this homework, you will be using hadoop/mapreduce to analyze data.

**Q1.** Use the provided file *q1_dataset.txt* as the input to your Hadoop MapReduce programs. Implement the following:

A. **Word count:** Compute the number of occurrences of every word.

B. **Target words:** Report the counts for the specific words **"whale"**, **"sea"**, and **"ship"**.

C. **Unique words by pattern:** For each pair *(word length, first character)*, compute the number of unique words that have that length and start with that character.

   Example:

   Text: "All students are data scientists, and all are passionate about solving complex problems."

   For words with length **3** and first character **'a'**, the output is:

   (key, 3) — where key represents *(length = 3, first character = 'a')*.

   There are three unique words: **"all"**, **"are"**, **"and."**

**Q2.** Write a Hadoop MapReduce program to perform the following:

**A. Inverted Index Construction**

- Use the input file *q2_dataset.txt*.
- The **map function** should parse each line and emit a sequence of <word, line number> pairs.
- The **reduce function** should take all pairs for a given word, sort the corresponding line numbers, and emit a pair <word, list (line numbers)>.
- The collection of all such output pairs will form a simple inverted index.

**B. Maximum Occurrence Words (with Combiner)**

- Using the same output from Part A, apply a **Combiner** to identify the word(s) with the highest number of occurrences in the inverted index.
- Output Format**:**

  <Word><TAB><Number of occurrences>

**Submission Instructions**

Please submit the following:

1. The JAR files — one for each problem.
2. The Java source files.
3. The output file generated by your program.
4. A README text file that explains how to run your JAR files, including the necessary commands.

Failure to follow the submission instructions will result in a deduction of marks.

**Submission Guidelines**

- Combine all required files into a single ZIP archive.
- Use the following naming convention for your ZIP file:
  <NetID>-<FirstName>-<LastName>.zip
- Example: For student John Smith with NetID jxs220000, the ZIP file should be named:
  jxs220000-John-Smith.zip

Good luck!