

Homework 5

MapReduce using Hadoop, Due Thursday, Dec 6th.

Description

This homework requires implementing a Mapper/Reducer using Hadoop. The input text, *the The Project Gutenberg EBook of The Adventures of Sherlock Holmes* by Sir Arthur Conan Doyle is available at the URL <http://www.cs.umd.edu/class/fall2007/cm433/resources/advsh12.txt>

For each word in the input text, you must compute the occurrences of other words appear in every line that it appeared. For example, assume that we have the following line as an input text.

I had seen little of Holmes lately. My marriage had drifted us

With the word *I*,

- *had* occurs twice.
- *seen* occurs once.
- *little* occurs once.
- *and* so on.

With the word *had*,

- *I* occurs once.
- *had* occurs once. (not zero times!)
- and so on.

For this assignment, you are going to compute the occurrences of *every* word for *all* lines, not just for a single line.

Before You Begin

First of all, you should download Hadoop and configure it. You may use either Eclipse or command-line to run your program. Here is the link to the homepage of Hadoop project:

<http://lucene.apache.org/hadoop/>

Make sure you read *Quickstart* and try out *Map-Reduce tutorial*. They will be very useful for this homework problem. If you have trouble with downloading and running it, please contact us as soon as possible.

Guidelines

You must follow these guidelines to obtain correct result.

- Convert every string into lower-case.
- Get rid of all punctuation, i.e., any character other than a to z and space. It will be helpful to use a regular expression to replace them with empty strings. Of course, this will result in some non-sense situations (im, youre, and etc.), but you do not have to worry about this.
- Do not count the subject word. In other words, if you are computing the occurrence of words for a word *the*, you should NOT count the same *the*. Note, however, that you should count other *the*'s appear in the same line if there are any!
- The output of Mapper and Reducer is up to you. However, you must be able to answer the questions at the bottom of this assignment.
- Here are some sample results:
 - *sherlock* occurs 94 times with *holmes*
 - *holmes* occurs twice with *armchair*
 - *i* occurs 122 times with *see*
- If you follow these guidelines, you should get the same results as above.

Before You Begin

First of all, you should download Hadoop and configure it so that you can use it for the assignment. You may use either Eclipse or command-line to run your program. Here is the link to the homepage of Hadoop project:

<http://lucene.apache.org/hadoop/>

Make sure you read *Quickstart* and try out *Map-Reduce tutorial*. I found them very useful for this homework problem. In fact, you may use their line indexer example and modify it to do this assignment. (It basically gives you an idea of how to start).

What to Submit

- You should submit two files: `Mapper.java` and `Reducer.java`. Please print them out and hand-in on the due date.
- You should also submit the answers to the questions in the bottom of the assignment along with the code printouts.

Questions

1. How many times the word *asked* occur with the word you?
2. How many times the word *cut* occur with the word *hair*?