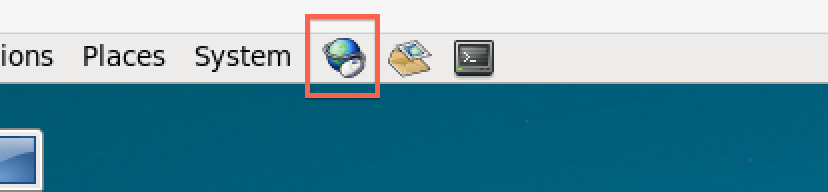
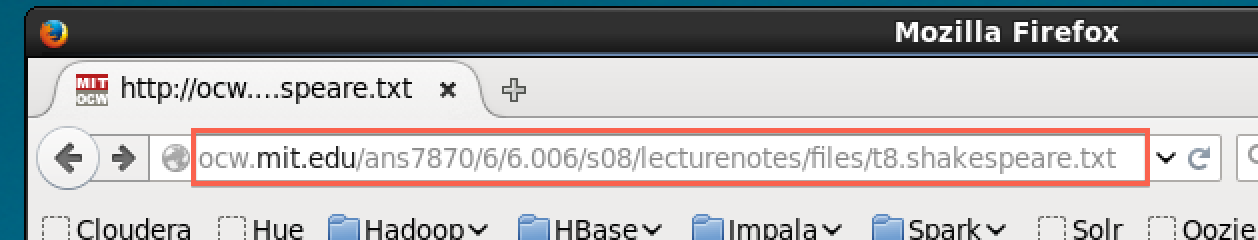
**Lab Practice 1: Hadoop Examples Week 5 11-17 Esfand**

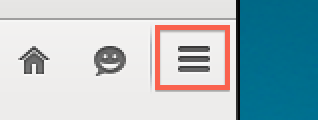
**Note: If you don’t use Cloudera change the provided path and instructions with the help of your TA.**

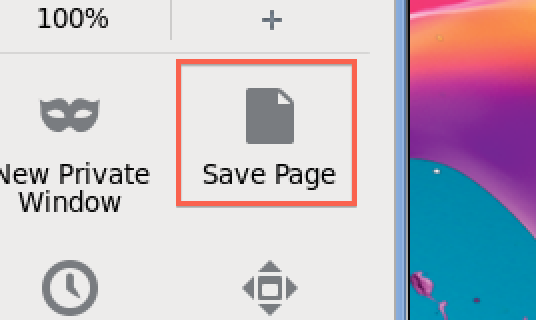
**Instructions for making an input file by downloading a text file to copy into HDFS:**

1.**Open a browser.**Open the browser by click on the browser icon on the top left of the screen.

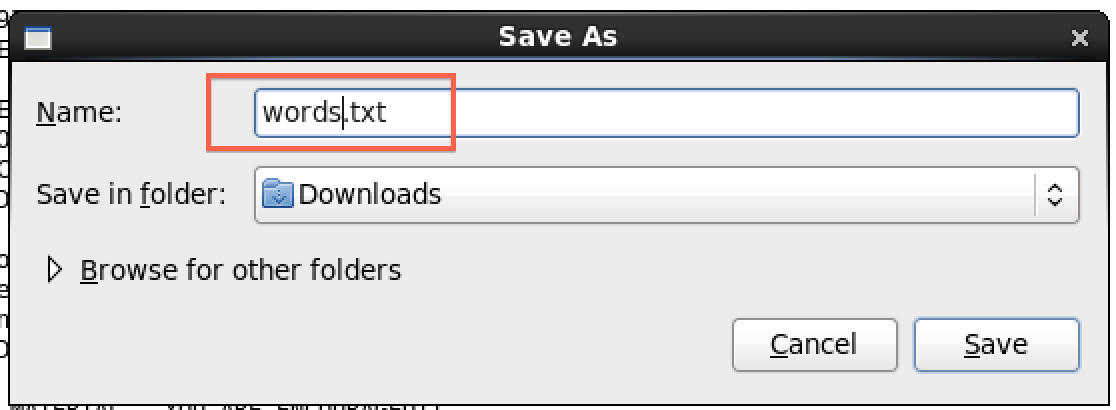
**2. Download the Shakespeare.**We are going to download a text file to copy into HDFS. Enter the following link in the browser: <http://ocw.mit.edu/ans7870/6/6.006/s08/lecturenotes/files/t8.shakespeare.txt>



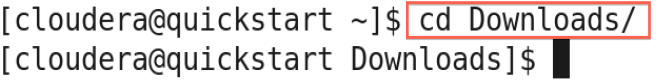


Once the page is loaded, click on the Open menu button.

Click on Save Page

Change the output to words.txt and click Save.

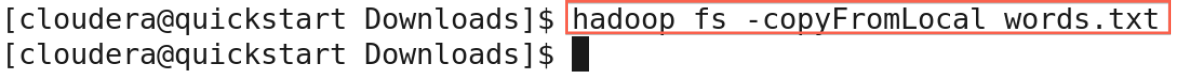


**2. Open a terminal shell.** Open a terminal shell by clicking on the square black box on the top left of the screen.

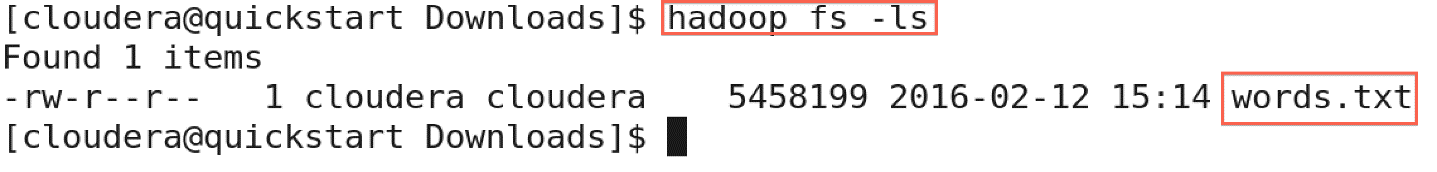
Run *cd Downloads* to change to the Downloads directory.

Run *ls* to see that words.txt was saved.

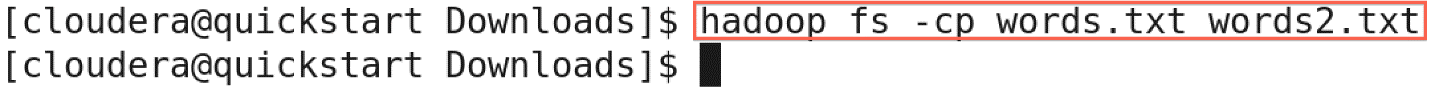
3. **Copy file to HDFS.** Run *hadoop fs -copyFromLocal words.txt* to copy the text file to HDFS.



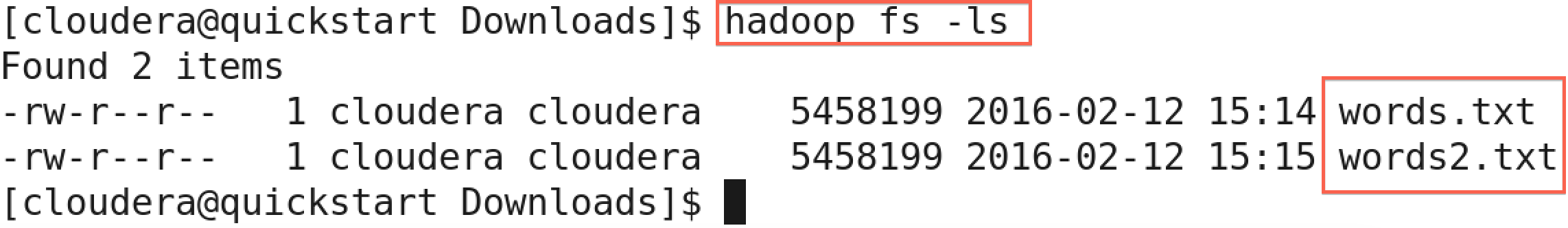
4. **Verify file was copied to HDFS.** Run *hadoop fs –ls* to verify the file was copied to HDFS.



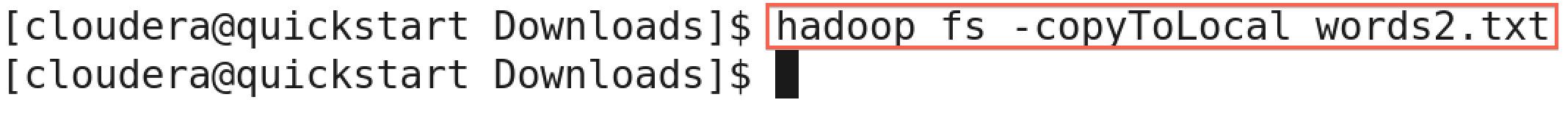
5. **Copy a file within HDFS**. You can make a copy of a file in HDFS. Run *hadoop fs -cp words.txt words2.txt* to make a copy of words.txt called words2.txt



We can see the new file by running *hadoop fs -ls*



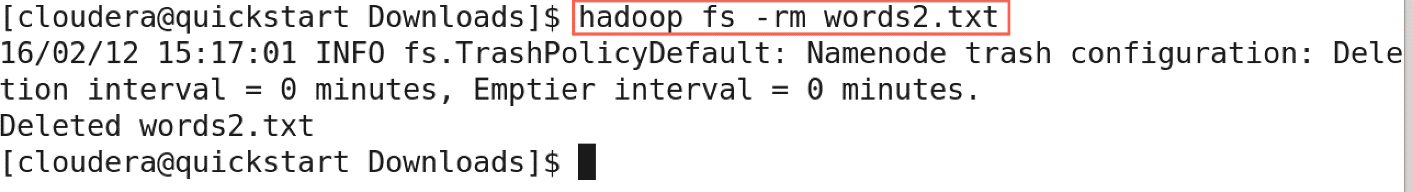
6. **Copy a file from HDFS.** We can also copy a file from HDFS to the local file system. Run *hadoop fs -copyToLocal words2.txt .*to copy words2.txt to the local directory.



Let's run *ls* to see that the file was copied to see that words2.txt is there.



7. **Delete a file in HDFS.** Let's the delete words2.txt in HDFS. Run *hadoop fs -rm words2.txt*



Run *hadoop fs -ls* to see that the file is gone.

**Running some of Hadoop Examples**

**Execute some of the example applications of Hadoop, mainly**

* wordcount: A map/reduce program that counts the words in the input files.
* wordmean: A map/reduce program that counts the average length of the words in the input files.
* grep: A map/reduce program that counts the matches of a regex in the input.
* And so on whatever you like …..

**Part1: Run the wordcount**

**Steps for WordCount:**

1. Start the Cloudera VM in VirtualBox and open a terminal shell.
2. As the first activity, you gonna running WordCount. The output says that WordCount takes the name of one or more input files and the name of the output directory. Note that these files are in HDFS, not the local file system.
3. Verify input file exists. (The complete works of Shakespeare which can be downloaded and copied into HDFS as mentioned).Make sure this file is still in HDFS so we can run WordCount on it. Run hadoop fs –ls
4. 4. See WordCount command line arguments. We can learn how to run WordCount by examining its command-line arguments. Run **hadoop jar /usr/jars/hadoop-examples.jar wordcount**.
5. 5. Run WordCount. Run WordCount for words.txt: **hadoop jar /usr/jars/hadoop-examples.jar wordcount words.txt out.**
   * As WordCount executes, the Hadoop prints the progress in terms of Map and Reduce. When the WordCount is complete, both will say 100%.
6. 6. See WordCount output directory. Once WordCount is finished, let's verify the output was created. First, let's see that the output directory, out, was created in HDFS by running hadoop fs –ls
   * We can see there are now two items in HDFS: words.txt is the text file that we previously created, and out is the directory created by WordCount.
7. Look inside output directory. The directory created by WordCount contains several files. Look inside the directory by running hadoop –fs ls out
   * The file part-r-00000 contains the results from WordCount. The file \_SUCCESS means WordCount executed successfully.
8. 8. Copy WordCount results to local file system. Copy part-r-00000 to the local file system by running hadoop fs –copyToLocal out/part-r-00000 local.txt

9. View the WordCount results. View the contents of the results: more local.txt.

* Each line of the results file shows the number of occurrences for a word in the input file. For example, Accuseappears four times in the input, but Accusing appears only once.

**Part2: Run the wordmean** : Take similar steps as above for the wordmean

**Part3- Run the grep:** Take similar steps for grep. Remember you need to give a regular expression as input as well. Try 3 different regular expression of your choice.

**Report Output Results:**

* For Wordcount choose 10 words by random an report their count.
* For wordmean, you report a number which is the mean of length of the all words in the input file.
* For grep, report your 3 regular expression and five matches for each (by random).

|  |
| --- |
| 1- Wordcount Result :   * Word1 count1 * … * Word10 count10   2- Wordmean result: a number  3-Grep results:   * Regexp1: result 1 to 5 * Regexp2: result 1 to 5 * Regexp3: result 1 to 5 |