

**Assignment: Lab Assignment #2**

Quoc Tuong Dong

Analytics, College of Professional Studies, Northeastern University

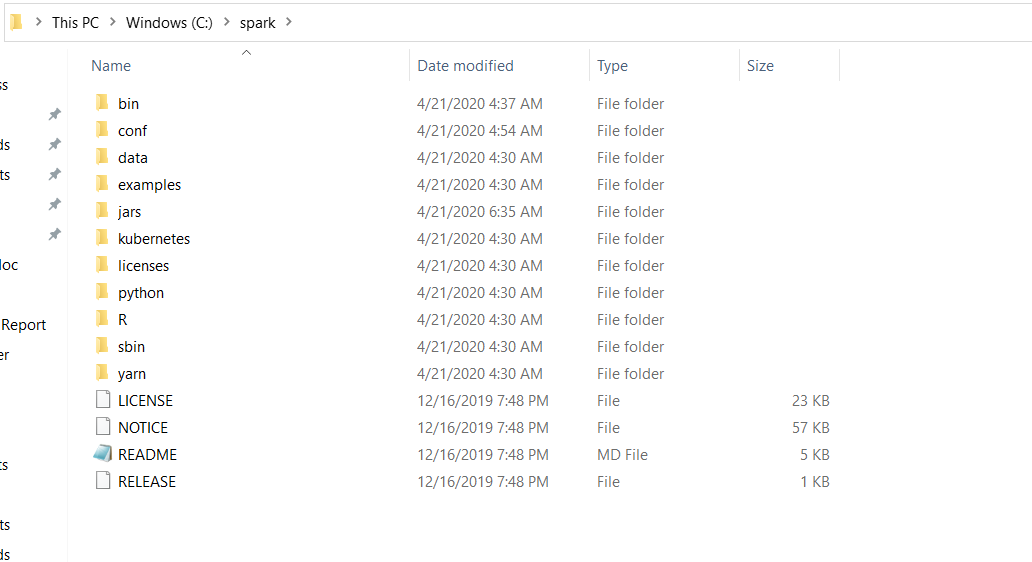
ALY6110 23240 Data Management & Big Data

Instructor: Doctor Reza Sadoddin

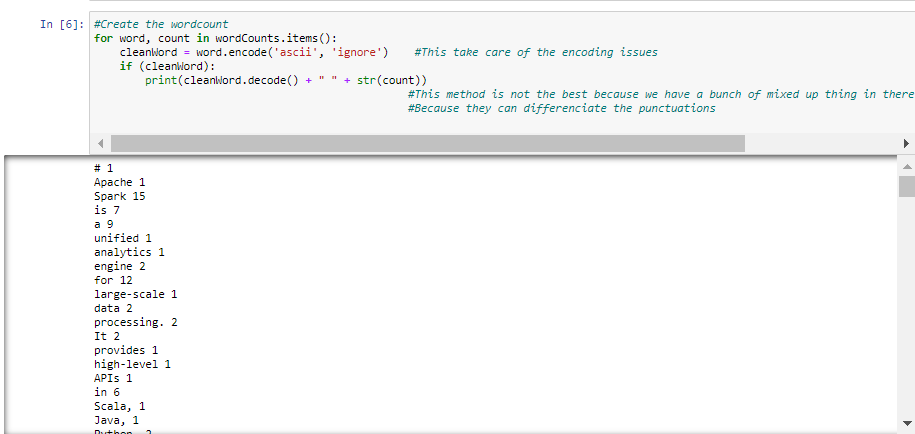
Due Date: April 9th, 2020

Submit the code, output and explanation of your steps from this assignment below.

I installed JDK and Spark earlier on my computer for another project and I mostly use PySpark to operate code command. I stored my Spark files in C://spark folder after downloading it from the system.



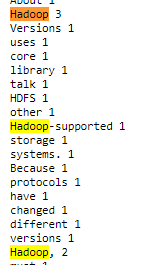
Then after that I ran the code on Jupyter notebook



**Question 3**

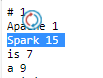
1. **How many times is the word "Hadoop" counted when the tutorial has printed out all the word counts?**

* Hadoop appears 3 times but there are some small variations of Hadoop words in there that should be acknowledged. So if we add all of them up, that would be 6



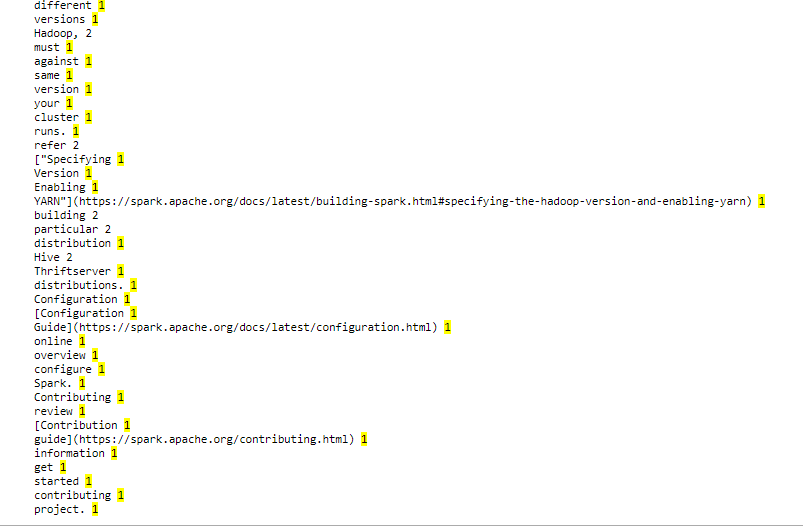
1. **Which is the most common word used in the file? How many times does the word occur?**

* The most common word is Spark, it appears 15 times



1. **Which word occurs the fewest times? How many time does the word occur?**

* There are plenty of uncommon words that happen only 1 time. Here are some of the examples



**Question 4**

It took 0.6 seconds for it to run properly on my computer.