Python MapReduce

Veeral Patel

Regis University

April 21st, 2021

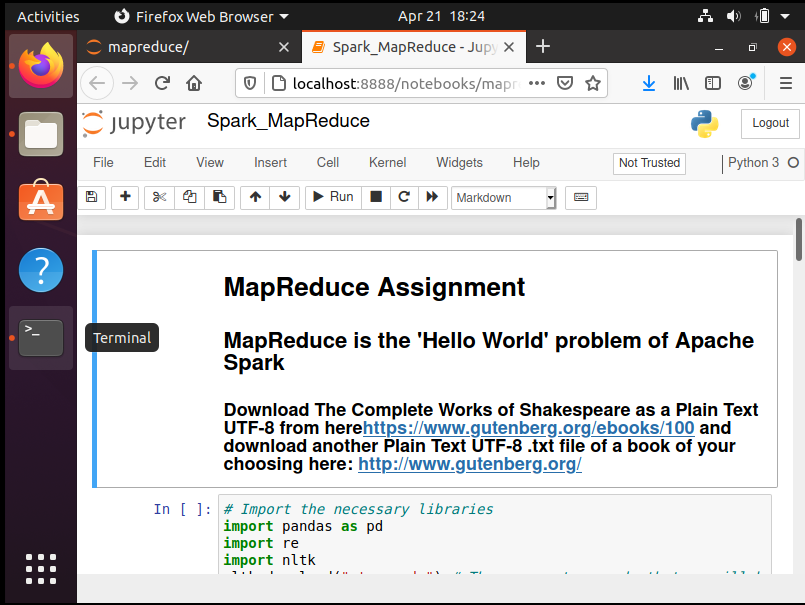
1. **Introduction**

For this project, we were to use a VM that we had installed the previous week. This week, we installed pyspark, learned how to use it and ran a python notebook through jupyterlab that was inside of the virtual box. Our goal was to learn how to load data as a RDD and then learn how to use MapRedce which is a part of Apache Spark. To do so, we downloaded book in Plain Text UTF-8 .txt file format. We performed operations on that text via python.

1. **Methods/Code/Screenshots**

I used a windows computer and I had to install pyspark. I used the PDF that was given to the class as well as some tutorials I found on YouTube. I had difficulty with getting the virtualbox to run. Later, I found out that in order to get it to open, I needed to disable “Boot in safe mode” from my system BIOS on my computer. This took a long time to figure out but finally, it was done. Then, I got it up and running and was able to install all of the necessary components in order to run this code. For the problem of the homework, I used the “anthem.txt” file that was available to download.

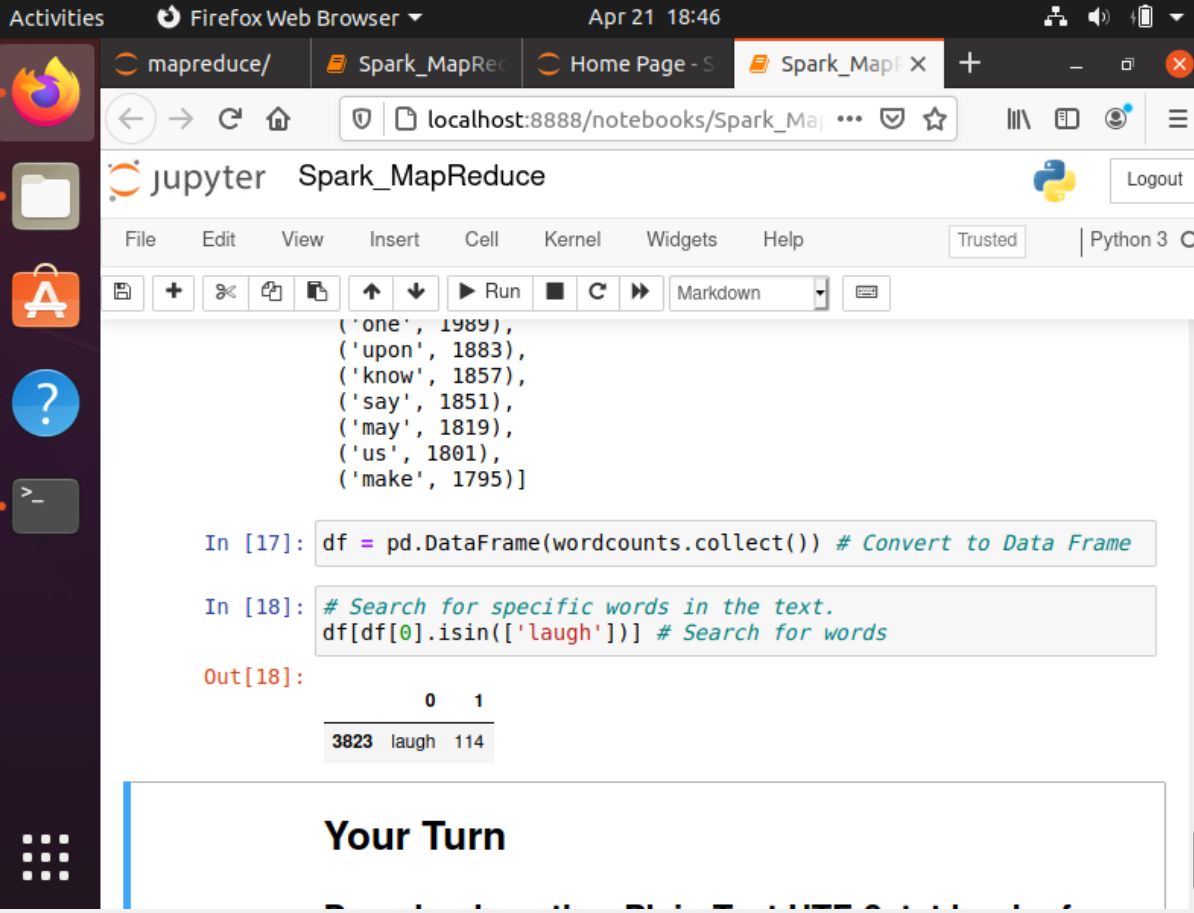
Examples of Python Notebook Running in VM:



Example of loading Shakespeare dataset:

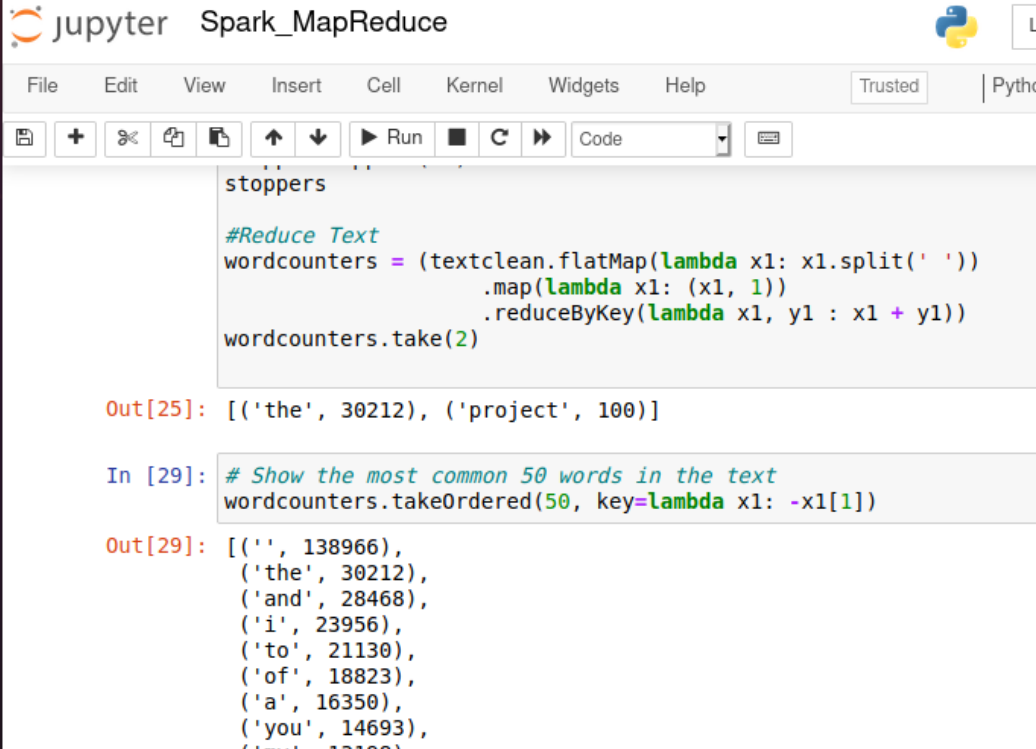


Running the code that was given for Shakespeare.txt file

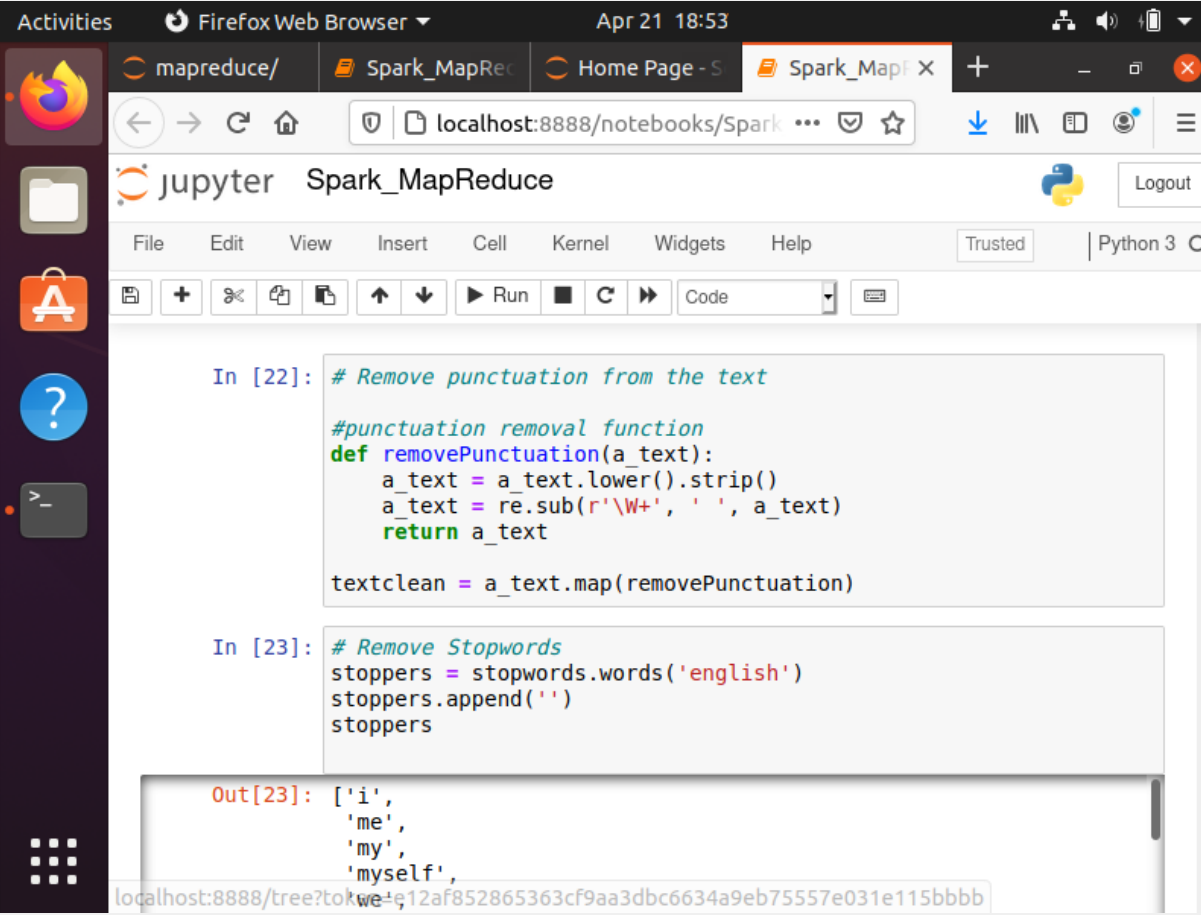


1. **Results**

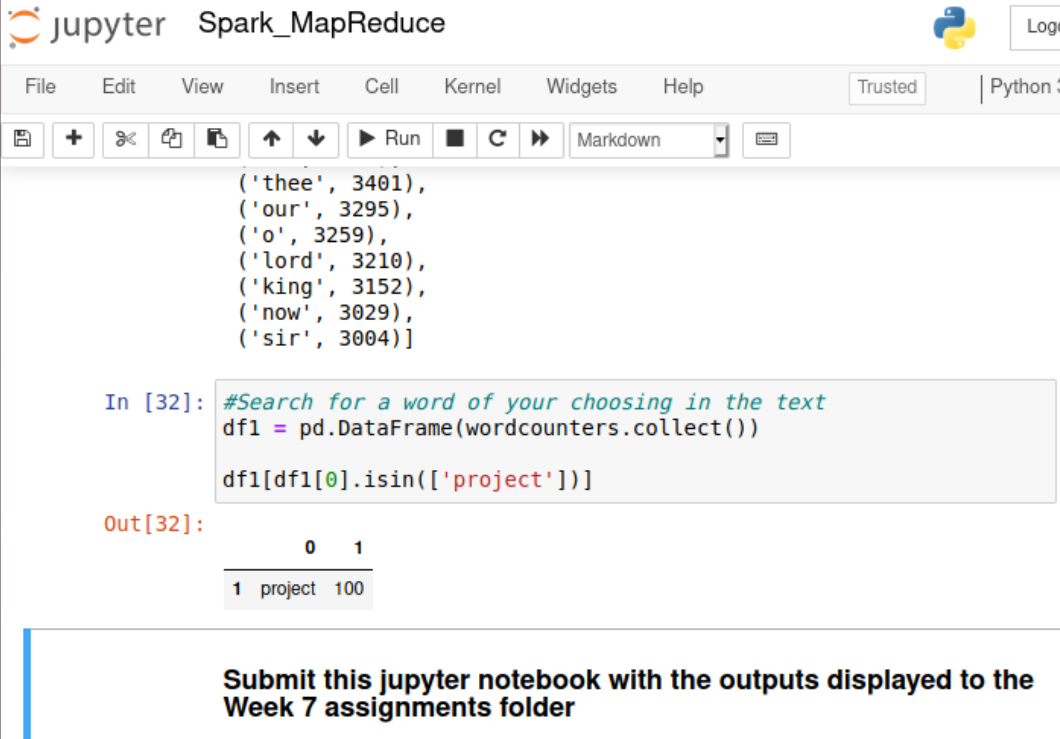
Reducing Common Words for problem assignment via Anthem.txt::



Removing Punctuation and stop words:



Searching for work in Anthem text:



1. **Analysis of Results**

In this lab, I thought that we learned a lot about how to use the mapreduce component of Apache Spark. Since spark is a widely used technology, I thought that it was really useful. What this means is that we are able to use RDD’s in order to quickly read and manipulate very large sets of data. I would think that this would be key when there are large streams of data coming into a large organization, which is pretty much always the case. This example, the HW, was a really good way to learn how easy it is to use map reduce and how fast it is, even when running inside of a VM.

1. **Conclusion**

I thought that this was a really good introduction to mapreduce and the pyspark module that can be used inside of Python. I think that the positive of this assignment was we learned how to use an RDD and how to break down text and how easy it is to clean text and manipulate it. However, I think that it would have been really helpful to learn what an RDD is and how it works. I also would like to dig deeper in what was RDD is different than a dataframe or a really long string and why it is so much faster. This technology would apply to my field, healthcare, because it allows us to put patient data into an RDD and be able to analyze it really fast. Really, I think that text analysis/cleaning/wrangling is pretty much useful in any field that has data. This was my first experience in the topic, I am still confused as to what an RDD is, I think that is because I am thinking of it as a data structure or something, and I think that it made it more confusing. I think that it was really helpful and I would certainly like to learn more about the topic.

1. **Reference(s):**

Gower, J. C. (1985). Properties of Euclidean and non-Euclidean distance matrices. *Linear Algebra and its Applications*, *67*, 81-97.

Verma, N., Mahajan, D., Sellamanickam, S., & Nair, V. (2012, June). Learning hierarchical similarity metrics. In *2012 IEEE conference on computer vision and pattern recognition* (pp. 2280-2287). IEEE.