Install_Apache_Spark_PySpark_Linux

September 4, 2016

Install PySpark on Linux
Youtube tutorial available at: https://www.youtube.com/watch?v=uhVYTNEe_-A
Download Spark

1) Go to the Apache Spark website.
http://spark.apache.org/downloads.html

```
<li> a) Choose a Spark release (I prefer 2.0.0)</li>
<li> b) Choose a package type: (this installation prefers "Pre-built for Hadoop 2.7 and later")</li>
<li> c) Choose a download type: (Direct Download) </li>
<li> d) Download Spark: http://d3kbcqa49mib13.cloudfront.net/spark-2.0.0-bin-hadoop2.7.tgz
</li> (you can click on this or go to http://spark.apache.org/downloads.html to choose your own Spark V
```

2) Make sure you have java installed on your machine. If you don't, I found the link below useful, but feel free to use something else.

http://tecadmin.net/install-oracle-java-8-jdk-8-ubuntu-via-ppa/

3) Go to your home directory. (You can use the command in red)

cd ~

4) Unzip the folder in your home directory using the following command.

tar -zxvf spark-2.0.0-bin-hadoop2.7.tgz

5) Use the following command to see that you have a .bashrc

ls -a

6) Next, we will edit our .bashrc so we can open a spark notebook in any directory

nano .bashrc

7) Don't remove anything in your .bashrc file. Add the following to the bottom of your .bashrc file

function snotebook () { #Spark path (based on your computer) SPARK_PATH=~/spark-2.0.0-bin-hadoop2.7

```
export PYSPARK_DRIVER_PYTHON="jupyter"
export PYSPARK_DRIVER_PYTHON_OPTS="notebook"

$SPARK_PATH/bin/pyspark --master local[2]
```

}

8) Save and exit out of your .bashrc file. Either close the terminal and open a new one or in your terminal type:

source .bashrc

9) Done! To test out PySpark please continue to the next tutorial or continue to step 10

Word Count Youtube: https://www.youtube.com/watch?v=jg7Z8ctKpEs Word Count Code: https://github.com/mGalarnyk/Python_Tutorials/blob/master/PySpark_Basics/PySpark_Part1_Word_Count_Removing_Pu

Optional: Word Count of Pride and Prejudice Text

```
In [1]: import re, string

        text_file = sc.textFile('Data/Pride_and_Prejudice.txt')

In [2]: punc = '!"#$%&\'()*+,./:;<=>?@[\\]^_`{|}~'

In [3]: def uni_to_clean_str(x):
            converted = x.encode('utf-8')
            lowercased_str = converted.lower()
            # for more difficult cases use re.split(' A|B')
            lowercased_str = lowercased_str.replace('--',' ')
            clean_str = lowercased_str.translate(None, punc) #Change 1
            return clean_str

In [4]: one_RDD = text_file.flatMap(lambda x: uni_to_clean_str(x).split())
        one_RDD = one_RDD.map(lambda x: (x,1))
        one_RDD = one_RDD.reduceByKey(lambda x,y: x + y)
        one_RDD = one_RDD.map(lambda x:(x[1],x[0]))
        one_RDD.sortByKey(False).take(15)

Out[4]: [(4331, 'the'),
         (4138, 'to'),
         (3611, 'of'),
         (3578, 'and'),
         (2225, 'her'),
         (2069, 'i'),
         (1947, 'a'),
         (1866, 'in'),
         (1846, 'was'),
         (1710, 'she'),
         (1579, 'that'),
         (1535, 'it'),
         (1429, 'not'),
         (1357, 'you'),
         (1336, 'he')]
```

Notes

The PYSPARK_DRIVER_PYTHON parameter and the PYSPARK_DRIVER_PYTHON_OPTS parameter are used to launch the PySpark shell in Jupyter Notebook. The –master parameter is used for setting the master node address. Here we launch Spark locally on 2 cores for local testing.

For Python 3 Users

You have to add the line in red before you use alias snotebook='$SPARK_PATH/bin/pyspark –master local[2]' line or you will get the error in the image above. export PYSPARK_PYTHON=python3

Other useful PySpark tutorials

https://www.dataquest.io/blog/installing-pyspark/