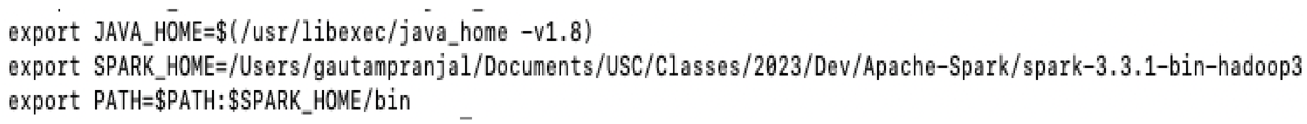
macOS

1. Install Java https://www.java.com/en/download/
2. Once installed, open the terminal and run the following commands:
    a. java -version
    b. javac -version

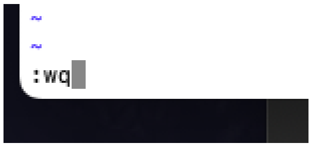
    If you are able to see the version (JDK) then Java is correctly installed
3. Download spark https://archive.apache.org/dist/spark/spark-3.1.2/ or the required version
4. Extract the contents and paste the folder to /Documents/DSCI553/Dev/Apache-Spark/ location.
5. Open a terminal and run the following command from the root location:
    a. vi ~/.bash_profile
    b. Move the pointer to the end of the file
    c. Press the key “i” on the keyboard. This will enable the “Insert” mode on the file
    d. Now type in the following commands in the file:



```
export JAVA_HOME=$(/usr/libexec/java_home -v1.8)
export SPARK_HOME=/Users/gautampranjal/Documents/USC/Classes/2023/Dev/Apache-Spark/spark-3.3.1-bin-hadoop3
export PATH=$PATH:$SPARK_HOME/bin
```

Note: SPARK_HOME is where you saved your downloaded Spark folder in Steps 3 and Step 4.

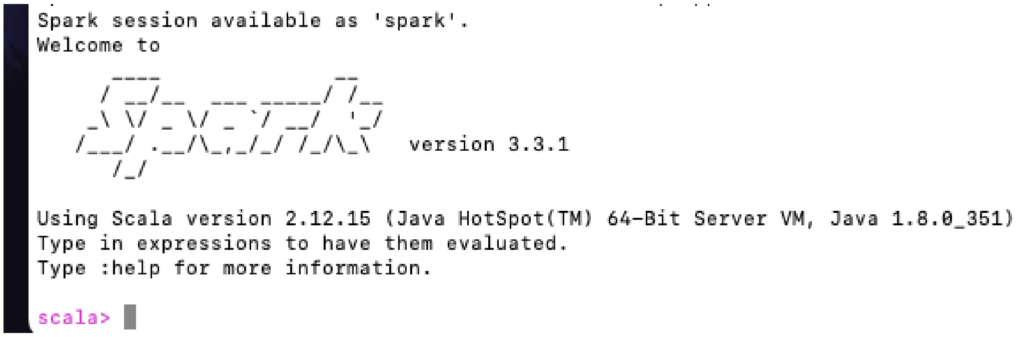
6. To exit the “Insert” mode, do the following:
    a. Press Escape Key
    b. Press Shift+[Key :] and then write ‘wq’ and press enter. This will appear the bottom left corner of your screen.



```
~
~
:wq
```

7. Once you are back in the terminal, type the following command:
    a. source ~/.bash_profile

8. Run the command “spark-shell” on the terminal. It will take some time to start spark. If you can see the following output, everything is working correctly till now.



```
Spark session available as 'spark'.
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /___/ .__/\_,_/_/ /_/\_\   version 3.3.1
      /_/

Using Scala version 2.12.15 (Java HotSpot(TM) 64-Bit Server VM, Java 1.8.0_351)
Type in expressions to have them evaluated.
Type :help for more information.

scala> 
```

9. In the editor’s terminal, type python and press enter. This will start the Python interpreter.
10. Type import pyspark
11. If the moduleNotFound error shows then come out of Interpreter by typing exit() and run the command:
    a. pip install pyspark
    b. Once the module is installed, repeat steps 9, 10, and 11 to see if it’s working.
12. If in step 9, in case you ran python3 to open the interpreter and the moduleNotFound error showed, then install pyspark using the following command:
    a. pip3 install pyspark
13. Go ahead in your Editor and write in the code for “word_count.py” mentioned in the slides.
14. Run your file as python word_count.py or python3 word_count.py depending on what you used in steps 9, 10, 11, and 12.
15. You should be able to see the results of your program. If in case you are not able to see the results or some error shows up, search it on google or contact a TA.

Python Spark Windows/Vocareum Install Guide

< Taken from Piazza https://piazza.com/class/lci8q5mczxb2xp/post/66 >

After a little bit of troubleshooting and messing around, I was able to get everything working both locally and on Vocareum, so below are the instructions that led me to this step, hope this is useful for some of you. For additional reference for software version and steps, @21, @58.

<Install JDK 1.8 (Skip If You Already Have)>

1. Download JDK 1.8 from the official website, look for JAVA 8 Windows, and select the desired file (x86 is 32-bit, x64 is 64-bit). Oracle will ask you for an account in order to download, which you can get here if you don't have one or don't wish to create one.

2. Run the installer and follow through to install JAVA 8

3. Press the Windows key and search for "Edit the system environment variables", and click on "Environment Variables" located at the bottom right corner

4. Look for "JAVA_HOME" in the lower "System variables" section. If it does not exist then create one. Set the value of "JAVA_HOME" to the folder in which you just installed JAVA 8, the default path is "C:\Program Files\Java\jdk1.8.0*" but please verify this yourself

5. Look for the double-click on "Path" under "System variables". Delete "C:\ProgramData\Oracle\Java\javapath" and "C:\Program Files (x86)\Common Files\Oracle\Java\javapath" if they exist. Add "%JAVA_HOME%\bin" and move it all the way to the top

6. Open the Windows command prompt, and type in "Java -version" to verify your JAVA version

<Install Spark v3.1.2 Windows>

1. Download Spark v3.1.2 w/ Hadoop v3.2 from the official website

2. Unzip the .tgz file twice and extract the final "spark-3.1.2-bin-hadoop3.2" folder

3. Download winutils.exe Hadoop v3.2.2 from GitHub

4. At your desired local location, create a folder for the setup, we will use "C:\Spark_Hadoop"

5. Within the master folder create two more subfolders each for Spark and wintils, we will use "C:\Spark_Hadoop\hadoop" and "C:\Spark_Hadoop\spark"

6. Copy and paste everything in the "spark-3.1.2-bin-hadoop3.2" folder (Step 2) into "C:\Spark_Hadoop\spark"

7. Create a subfolder under "C:\Spark_Hadoop\hadoop" called "bin", and place winutils.exe (Step 3) into "C:\Spark_Hadoop\hadoop\bin"

8. Press the windows key and search for "Edit the system environment variables", and click on "Environment Variables" located at the bottom right corner

9. Look for and click on "New" under "User variables". Add a new variable with the name "HADOOP_HOME" and value "C:\Spark_Hadoop\hadoop" (w\o the quotes). Then add another new variable with the name "SPARK_HOME" and value "C:\Spark_Hadoop\spark" (w\o the quotes)

10. Look for the double-click on "Path" under "User variables". Add two new lines, the first being "%SPARK_HOME%\bin", and the second being "%HADOOP_HOME%\bin" (w\o the quotes)

11. Open the Windows command prompt, and type in "spark-shell" (w\o the quotes), wait for a moment and you are done!

Note: Follow this youtube video if you need some visual guidance, steps are the same just different versions

<Install Python 3.6 (Skip If You Already Have)>

1. Download Python 3.6.8 from the official website, choose the correct version for yourself

2. Run the installer and follow through to install Python 3.6 (Add to PATH when asked by the installer)

3(a). If you do not use an IDE/editor that can choose the python version on a per-script basis, follow along to steps 3(b)-3(d), otherwise skip to step 4. I recommend using VSCodium

3(b). Press the Windows key and search for "Edit the system environment variables", and click on "Environment Variables" located at the bottom right corner

3(c). Look for the double-click on "Path" under "System variables" (Or "User variables" depending on configuration). Move the two lines with "Python36" to the very top to make Python 3.6 the default. You can do the same thing later to change it back to another version.

3(d). Open the Windows command prompt, and type in "py -V" or "python -V" to verify the version

4. Open the Windows command prompt to install relevant packages to Python 3.6. Do this by using "py -3.6 -m pip install packagename". (pyspark, pandas, numpy, matplotlib, seaborn, scikit-learn, imbalanced-learn, and tensorflow)

<Running "word_count.py" Locally>

1. Use the same code provided on the lecture slides

2. For the code to run locally, you will need to change the two "os.environ" lines to the correct local space. I found that this can be done automatically by simply using "sys.executable", provided you have your versions and permissions set up correctly.

3. Create a "text.txt" file in the same folder as "word_count.py" and fill it with random words

4. Run "word_count.py" by using "py -3.6 word_count.py" in the windows terminal, or any other way you like

Note: Attached script here: word_count.py

<Setting Up Vocareum>

1. You should have received an email titled "Welcome to Vocareum Cloud Labs" to signup

2. In the case that you did not, you can request a password reset with your USC email to obtain the account

3. If you still have no luck then consider opening a private post on Piazza to reach the instructors

4. Within Vocareum you should see "DSCI 553" and "Assignment 0", click on "My Work" to access the cloud lab

5(a). Inside the lab you should see Files (5(b)), Terminal (5(c)), and Source (5(d))

5(b). Located to the left, there are three categories. "resources" has a lock on it and are just system resources that you can't modify. "work" is where you can put all the files and work on them, click on "work" and you will see that you are able to upload files at the top. "Submission" is just a record of your submissions after you click on "Submit", this can't be modified

5(c). The terminal is towards the bottom, this is where you can execute code and debug errors. Vocareum is running on Ubuntu v16.04.4 LTS (Xenial Xerus), you can get this by entering "cat /etc/os-release" at the terminal (that is right, you can't seem to copy-paste anything in the terminal, someone let me know if they find a way that I missed)

5(d). Source is just the text editor, if you double-click on a file under "Files", it will show up here

<Running "word_count.py" In Vocareum>

1. Set the JAVA version to 1.8, I find that only doing the command in the slides was not working for me. However, entering "export JAVA_HOME=/usr/lib/jvm/java-1.8.0-openjdk-amd64" first, and then entering "export PATH=$JAVA_HOME/bin:$PATH" worked for me

2. Set the PySpark Python version to 3.6, do this by entering "export PYSPARK_PYTHON=python3.6"

3. Upload the "word_count_vocareum.py" script and a "text.txt" file into the "work" area. This should be an identical copy as that in the lecture slides

4. Run the "word_count_vocareum.py" script by entering "/opt/spark/spark-3.1.2-bin-hadoop3.2/bin/spark-submit --executor-memory 4G --driver-memory 4G word_count_vocareum.py"

5. If everything works properly you should get outputs and results in the terminal, after that you can click on "Submit" and turn in the assignment

Note: Attached script here: word_count_vocareum.py