Distributed Computing Spark Installation Assignment

DIANE WOODBRIDGE, PH.D

Prerequisite:

- Use python version 2.7.12.
- Install Anaconda
 - This includes Ipython, jupyter
- Download JDK

brew update

brew cask install java

✓ Spark runs on Java 7+, Python 2.6+/3.4+ and R 3.1+. For the Scala API, Spark 2.0.0 uses Scala 2.11. You will need to use a compatible Scala version (2.11.x).

Install Spark

brew install apache-spark



```
Install Spark
  brew install apache-spark
```

```
[ML-ITS-603436:bin dwoodbridge$ brew info apache-spark
apache-spark: stable 2.0.2, HEAD
Engine for large-scale data processing
https://spark.apache.org/
/usr/local/Cellar/apache-spark/2.0.2 (1,206 files, 204.5M) *
Built from source on 2016-12-13 at 09:42:19
From: https://github.com/Homebrew/homebrew-core/blob/master/Formula/apache-spark
_rb
```

Spark-shell – uses scala or python.

- A program written in the shell is discarded after you exit the shell.
- → Testing and developing applications.

```
spark-shell
Pyspark
```

Spark-shell – uses scala or python.

spark-shell

```
[ML-ITS-603436:bin dwoodbridge$ spark-shell
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel).
16/12/13 09:58:46 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using buil
[tin-java classes where applicable
16/12/13 09:58:47 WARN SparkContext: Use an existing SparkContext, some configuration may not take effect.
Spark context Web UI available at http://192.168.1.9:4040
Spark context available as 'sc' (master = local[*], app id = local-1481651927379).
Spark session available as 'spark'.
Welcome to
Using Scala version 2.11.8 (Java HotSpot(TM) 64-Bit Server VM, Java 1.8.0_112)
Type in expressions to have them evaluated.
Type :help for more information.
```

Spark-shell – uses scala or python.

pyspark

>>>

```
[ML-ITS-603436:~ dwoodbridge$ pyspark
Python 2.7.12 | Anaconda 4.1.1 (x86_64) | (default, Jul 2 2016, 17:43:17)
[GCC 4.2.1 (Based on Apple Inc. build 5658) (LLVM build 2336.11.00)] on darwin
Type "help", "copyright", "credits" or "license" for more information.
Anaconda is brought to you by Continuum Analytics.
Please check out: http://continuum.io/thanks and https://anaconda.org
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel).
16/12/13 10:04:34 WARN NativeCodeLoader: Unable to load native-hadoop library for your pl
e applicable
Welcome to
   /__ / .__/\_,_/_/ version 2.0.2
Using Python version 2.7.12 (default, Jul 2 2016 17:43:17)
SparkSession available as 'spark'.
```

```
Jupyter set up (include in your ~/.profile or .bash_profile)
    export PYSPARK_DRIVER_PYTHON=jupyter
    export PYSPARK_DRIVER_PYTHON_OPTS="notebook"

Start pyspark (Python version spark shell)
    pyspark
```

When jupyter notebook starts, choose "new python"



Homework 1 (5%)

Change the title of the notebook file to "MSAN694_HW1_LastName_Firstname".

Run sc and show the output as below.

Submit the screenshot (The file name should be also "MSAN694_HW1_LastName_Firstname.png".) on Canvas.

