First Notebook: Virtual machine test and assignment submission

This notebook will test that the virtual machine (VM) is functioning properly and will show you how to submit an assignment to the autograder. To move through the notebook just run each of the cells. You will not need to solve any problems to complete this lab. You can run a cell by pressing "shift-enter", which will compute the current cell and advance to the next cell, or by clicking in a cell and pressing "control-enter", which will compute the current cell and remain in that cell. At the end of the notebook you will export / download the notebook and submit it to the autograder.

** This notebook covers: **

Part 1: Test Spark functionality

Part 2: Check class testing library

Part 3: Check plotting

Part 4: Check MathJax formulas

Part 5: Export / download and submit

**Part 0: Setup

```
%cd '/content'
```

```
/content
```

```
!mkdir test_helper
%cd 'test_helper'

!mkdir init
!mkdir test_helper
!wget https://raw.githubusercontent.com/Walkisible/Big_Data_Analytics/main/test_helper/MANIFEST.in
!wget https://raw.githubusercontent.com/Walkisible/Big_Data_Analytics/main/test_helper/README.md
!wget https://raw.githubusercontent.com/Walkisible/Big_Data_Analytics/main/test_helper/setup.cfg
!wget https://raw.githubusercontent.com/Walkisible/Big_Data_Analytics/main/test_helper/setup.py
%cd 'test_helper'

!wget https://raw.githubusercontent.com/Walkisible/Big_Data_Analytics/main/test_helper/test_helper/test_hel
import test_helper

%cd ..

%cd 'init'
!wget https://raw.githubusercontent.com/Walkisible/Big_Data_Analytics/main/test_helper/init/__init__.py
%cd '/content'
```

```
/content/test_helper
--2024-01-13 07:48:35--  https://raw.githubusercontent.com/Walkisible/Big_Data_Analytics/main/test
Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 185.199.108.133, 185.199.109.
Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|185.199.108.133|:443... conr
HTTP request sent, awaiting response... 200 OK
Length: 65 [text/plain]
Saving to: 'MANIFEST.in'

MANIFEST.in         100%[===================>]      65  --.-KB/s    in 0s

2024-01-13 07:48:35 (2.95 MB/s) - 'MANIFEST.in' saved [65/65]

--2024-01-13 07:48:35--  https://raw.githubusercontent.com/Walkisible/Big_Data_Analytics/main/test
Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 185.199.108.133, 185.199.109.
Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|185.199.108.133|:443... conr
HTTP request sent, awaiting response... 200 OK
Length: 295 [text/plain]
Saving to: 'README.md'

README.md           100%[===================>]     295  --.-KB/s    in 0s

2024-01-13 07:48:35 (12.4 MB/s) - 'README.md' saved [295/295]

--2024-01-13 07:48:35--  https://raw.githubusercontent.com/Walkisible/Big_Data_Analytics/main/test
Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 185.199.108.133, 185.199.109.
Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|185.199.108.133|:443... conr
HTTP request sent, awaiting response... 200 OK
Length: 40 [text/plain]
Saving to: 'setup.cfg'

setup.cfg           100%[===================>]      40  --.-KB/s    in 0s

2024-01-13 07:48:35 (2.39 MB/s) - 'setup.cfg' saved [40/40]

--2024-01-13 07:48:35--  https://raw.githubusercontent.com/Walkisible/Big_Data_Analytics/main/test
Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 185.199.108.133, 185.199.109.
Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|185.199.108.133|:443... conr
HTTP request sent, awaiting response... 200 OK
```

```
Length: 519 [text/plain]
Saving to: ‘setup.py’

setup.py            100%[===================>]     519  --.-KB/s    in 0s

2024-01-13 07:48:35 (31.9 MB/s) - ‘setup.py’ saved [519/519]

/content/test_helper/test_helper
--2024-01-13 07:48:35--  https://raw.githubusercontent.com/Walkisible/Big_Data_Analytics/main/test
Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 185.199.108.133, 185.199.109.
Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|185.199.108.133|:443... conn
HTTP request sent, awaiting response... 200 OK
Length: 30093 (29K) [text/plain]
Saving to: ‘test_helper.py’

test_helper.py      100%[===================>]  29.39K  --.-KB/s    in 0.004s

2024-01-13 07:48:36 (6.59 MB/s) - ‘test_helper.py’ saved [30093/30093]
```

```
# import lib tools
import numpy as np
import pandas as pd
```

** Part 1: Test Spark functionality **

```
'''
Spark installation source code

* Disclaimer: Spark is frequently updated their version, when Spark is updated,
              some line cannot be executed.

              To fix this, change the Saprk to the latest version and try to again

Tutorial: https://github.com/Walkisible/Big_Data_Analytics/blob/main/Spark_install.py
'''

# check for upgradable packages
!apt update

# install java
!apt-get install openjdk-11-jdk-headless -qq > /dev/null

!wget -q https://dlcdn.apache.org/spark/spark-3.5.0/spark-3.5.0-bin-hadoop3.tgz

# unzip the spark file to the current folder
!tar -xf spark-3.5.0-bin-hadoop3.tgz

# to remove Spark.tgz file
!rm -rf spark-3.5.0-bin-hadoop3.tgz

# install findspark using pip
!pip install -q findspark
!pip install -q pyspark

# set your spark folder to your system path environment.
import os
os.environ["JAVA_HOME"] = "/usr/lib/jvm/java-11-openjdk-amd64"
os.environ["SPARK_HOME"] = "/content/spark-3.5.0-bin-hadoop3"

# create SparkContext
from pyspark import SparkContext
import findspark

findspark.init()
sc = SparkContext.getOrCreate()
sc
```

```
Hit:1 http://archive.ubuntu.com/ubuntu jammy InRelease
Get:2 https://cloud.r-project.org/bin/linux/ubuntu jammy-cran40/ InRelease [3,626 B]
Get:3 http://archive.ubuntu.com/ubuntu jammy-updates InRelease [119 kB]
Hit:4 https://developer.download.nvidia.com/compute/cuda/repos/ubuntu2204/x86_64  InRelease
Get:5 http://security.ubuntu.com/ubuntu jammy-security InRelease [110 kB]
Hit:6 http://archive.ubuntu.com/ubuntu jammy-backports InRelease
Hit:7 https://ppa.launchpadcontent.net/c2d4u.team/c2d4u4.0+/ubuntu jammy InRelease
Hit:8 https://ppa.launchpadcontent.net/deadsnakes/ppa/ubuntu jammy InRelease
Hit:9 https://ppa.launchpadcontent.net/graphics-drivers/ppa/ubuntu jammy InRelease
Get:10 http://archive.ubuntu.com/ubuntu jammy-updates/universe amd64 Packages [1,309 kB]
Hit:11 https://ppa.launchpadcontent.net/ubuntugis/ppa/ubuntu jammy InRelease
Get:12 http://archive.ubuntu.com/ubuntu jammy-updates/main amd64 Packages [1,617 kB]
Fetched 3,159 kB in 2s (2,103 kB/s)
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
24 packages can be upgraded. Run 'apt list --upgradable' to see them.
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 316.9/316.
  Preparing metadata (setup.py) ... done
  Building wheel for pyspark (setup.py) ... done
```

SparkContext

Spark UI

Version
    v3.5.0
Master
    local[*]
AppName
    pyspark-shell

** (1a) Parallelize, filter, and reduce **

```
# Check that Spark is working
largeRange = sc.parallelize(range(100000))
reduceTest = largeRange.reduce(lambda a, b: a + b)
filterReduceTest = largeRange.filter(lambda x: x % 7 == 0).sum()

print(reduceTest)
print(filterReduceTest)

# If the Spark jobs don't work properly these will raise an AssertionError
assert reduceTest == 4999950000
assert filterReduceTest == 714264285
```

```
4999950000
714264285
```

** (1b) Loading a text file **

```
# download data
!mkdir data
%cd 'data'
!wget https://github.com/Walkisible/Big_Data_Analytics/raw/main/dataset/shakespeare.txt
%cd '/content'
```

```
/content/data
--2024-01-13 07:58:31--  https://github.com/Walkisible/Big_Data_Analytics/raw/main/dataset/shakespear
Resolving github.com (github.com)... 140.82.113.3
Connecting to github.com (github.com)|140.82.113.3|:443... connected.
HTTP request sent, awaiting response... 302 Found
Location: https://raw.githubusercontent.com/Walkisible/Big_Data_Analytics/main/dataset/shakespeare.txt
--2024-01-13 07:58:32--  https://raw.githubusercontent.com/Walkisible/Big_Data_Analytics/main/dataset/
Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 185.199.108.133, 185.199.109.133,
Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|185.199.108.133|:443... connecte
HTTP request sent, awaiting response... 200 OK
Length: 5327978 (5.1M) [text/plain]
Saving to: ‘shakespeare.txt’

shakespeare.txt     100%[===================>]   5.08M  --.-KB/s    in 0.08s

2024-01-13 07:58:32 (65.4 MB/s) - ‘shakespeare.txt’ saved [5327978/5327978]

/content
```

```
# Check loading data with sc.textFile
rawData = sc.textFile("/content/data/shakespeare.txt")

# Due to lazy computing, caching is necessary to reduce read time from disk to RAM
rawData.cache()

shakespeareCount = rawData.count()

print(shakespeareCount)

# If the text file didn't load properly an AssertionError will be raised
assert shakespeareCount == 122395
```

```
122395
```

** Part 2: Check class testing library **

** (2a) Compare with hash **

```

```

```
# TEST Compare with hash (2a)
# Check our testing library/package
# This should print '1 test passed.' on two lines
from test_helper import Test

twelve = 12
Test.assertEquals(twelve, 12, 'twelve should equal 12')
Test.assertEqualsHashed(twelve, '7b52009b64fd0a2a49e6d8a939753077792b0554',
                        'twelve, once hashed, should equal the hashed value of 12')
```

```
1 test passed.
1 test passed.
```

** (2b) Compare lists **

```
# TEST Compare lists (2b)
# This should print '1 test passed.'
unsortedList = [(5, 'b'), (5, 'a'), (4, 'c'), (3, 'a')]
Test.assertEquals(sorted(unsortedList), [(3, 'a'), (4, 'c'), (5, 'a'), (5, 'b')],
                  "unsortedList does not sort properly")
```

```
1 test passed.
```

** Part 3: Check plotting **

** (3a) Our first plot **

After executing the code cell below, you should see a plot with 50 blue circles. The circles should start at the bottom left and end at the top right.

```
# Check matplotlib plotting
import matplotlib.pyplot as plt
import matplotlib.cm as cm
from math import log

# function for generating plot layout
def preparePlot(xticks, yticks, figsize=(10.5, 6), hideLabels=False, gridColor='#999999', gridWidth=1.0):
    plt.close()
    fig, ax = plt.subplots(figsize=figsize, facecolor='white', edgecolor='white')
    ax.axes.tick_params(labelcolor='#999999', labelsize='10')
    for axis, ticks in [(ax.get_xaxis(), xticks), (ax.get_yaxis(), yticks)]:
        axis.set_ticks_position('none')
        axis.set_ticks(ticks)
        axis.label.set_color('#999999')
        if hideLabels: axis.set_ticklabels([])
    plt.grid(color=gridColor, linewidth=gridWidth, linestyle='-')
    map(lambda position: ax.spines[position].set_visible(False), ['bottom', 'top', 'left', 'right'])
    return fig, ax

# generate layout and plot data
x = range(1, 50)
y = [log(x1 ** 2) for x1 in x]
fig, ax = preparePlot(range(5, 60, 10), range(0, 12, 1))
plt.scatter(x, y, s=14**2, c='#d6ebf2', edgecolors='#8cbfd0', alpha=0.75)
ax.set_xlabel(r'$range(1, 50)$'), ax.set_ylabel(r'$\log_e(x^2)$')
pass
```

** Part 4: Check MathJax Formulas **

** (4a) Gradient descent formula **

You should see a formula on the line below this one:

$$\mathbf{w}_{i+1} = \mathbf{w}_i - \alpha_i \sum (\mathbf{w}_i^\top \mathbf{x}_j - y_j) \mathbf{x}_j \ .$$