Name: Hii Nian Yu

Matric Number: 17007070/1 (WQD180038)

Github link: <https://github.com/hiinianyu/DataMining>

Video link: https://www.loom.com/share/9cce920834144fac9162e59de9c9d732

Title: Gold Price Prediction

**Milestone 1**

I am using selenium to crawl data from a website: <https://www.investing.com/commodities/gold-historical-data>

Below is the code:

Driver = webdriver.Chrome(executable\_path="/home/student/DM\_assignment/chromedriver\_linux64/chromedriver")

driver.set\_page\_load\_timeout(30)

driver.maximize\_window()

driver.get("https://www.investing.com/commodities/gold-historical-data")

for i in range(5):

dateRangeElements = driver.find\_elements\_by\_xpath("//div[@id='widgetFieldDateRange']")

if len(dateRangeElements) > 0:

time.sleep(3)

dateRangeElements[0].click()

break

else:

print(str(i + 1) + " trial, can't find element")

time.sleep(1)

i = i + 1

elementStartDate = driver.find\_element\_by\_xpath("//input[@id='startDate']")

elementStartDate.clear()

elementStartDate.send\_keys("01/01/2019")

driver.find\_element\_by\_xpath("//a[@id='applyBtn']").click()

time.sleep(3)

tableData = driver.find\_elements\_by\_xpath("//table[@id='curr\_table']/tbody/tr/td")

fileName = r"Result\goldprice\_" + time.strftime("%d-%b-%Y\_%H-%M-%S", time.localtime()) + ".csv"

resultFile = open(fileName, 'w')

resultFile.write("Date\_D,Price,Open,High,Low,Volume,Change\_%\n")

column = 1

for data in tableData:

if column != 7:

resultFile.write((data.text).replace(',', '') + ", ")

column = column + 1

else:

resultFile.write(data.text + "\n")

column = 1

resultFile.close()

After data saved, I will put the command below to save the file automatically into HDFS system.

def run\_cmd(args\_list):

"""

run linux commands

"""

# import subprocess

print('Running system command: {0}'.format(' '.join(args\_list)))

proc = subprocess.Popen(args\_list, stdout=subprocess.PIPE, stderr=subprocess.PIPE)

s\_output, s\_err = proc.communicate()

s\_return = proc.returncode

return s\_return, s\_output, s\_err

(ret, out, err)= run\_cmd(['hadoop', 'fs', '-put', '/home/student/DM\_assignment/' + fileName, '/user/hdfs/DataMining'])

driver.close()