Course: Big Data

Lab 05

PySpark - DataFrame

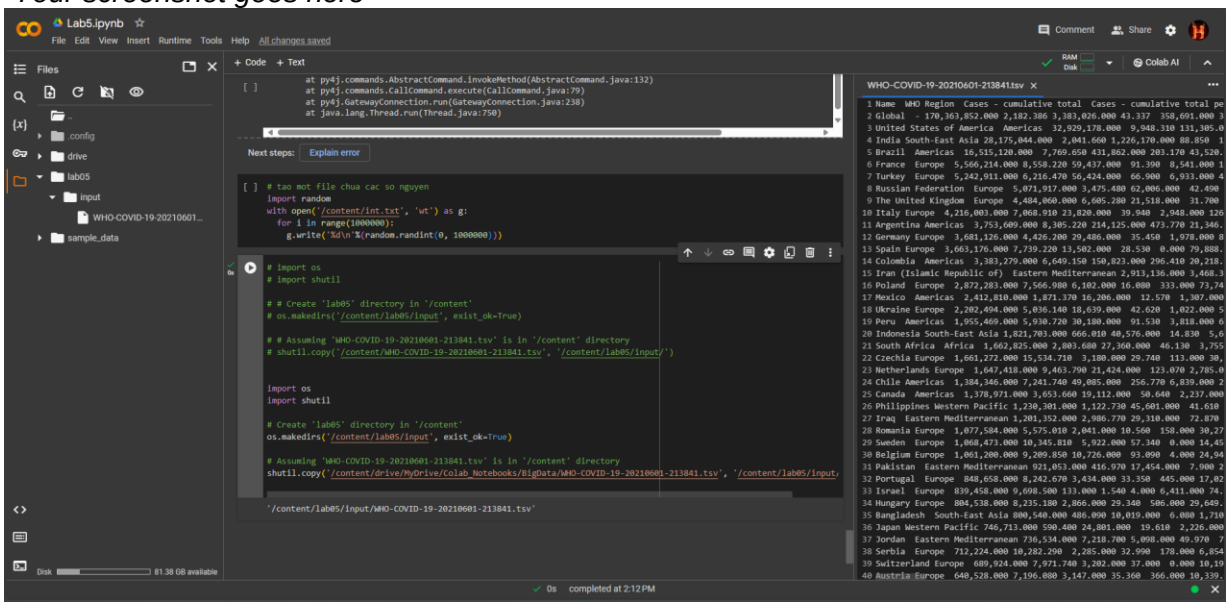
Question 1:

Given a tsv file [WHO-COVID-19-20210601-213841.tsv] which is corresponding to the [WHO Coronavirus (COVID-19) Dashboard].

Students are required to create a folder, named **lab05**, in **/content** directory of Google Colab and then copy the tsv to **/content/lab05/input/**

Take a screenshot to show your work.

Your screenshot goes here



Question 2:

Write a PySpark program, located in **ASEANCaseCount.py**, using DataFrames to

- to count the number of cumulative total cases among ASEAN countries (*South-East Asia Region in the given data table*)

- to find the country with the maximum number of cumulative total cases among ASEAN countries.
- to find the top 3 countries with the lowest number of cumulative cases among ASEAN countries.

- Insert your source code into the table below.

```
from pyspark.sql import SparkSession
from pyspark.sql.functions import col, max

spark = SparkSession.builder.appName('ASEANCaseCount').getOrCreate()
data = spark.read.csv('/content/drive/MyDrive/Colab_Notebooks/BigData/lab05/input/WHO-COVID-19-20210601-213841.tsv', header=True, sep='\t')
asean_countries = ['Brunei Darussalam', 'Cambodia', 'Indonesia', 'Lao People\'s Democratic Republic', 'Malaysia', 'Myanmar', 'Philippines', 'Singapore', 'Thailand', 'Viet Nam']
data = data.filter((data['WHO Region'] == 'South-East Asia') & (data['Name'].isin(asean_countries)))
data = data.withColumn('Cases - cumulative total', data['Cases - cumulative total'].cast('integer'))
total_cases = data.groupBy().sum('Cases - cumulative total').collect()[0][0]
print(f'Total cases in ASEAN countries: {total_cases}')
max_cases_country = data.orderBy(data['Name'].desc()).first()['Name']
print(f'Country with the maximum number of cases: {max_cases_country}')
lowest_cases_countries = data.orderBy(data['Cases - cumulative total']).select('Name').limit(3).collect()
lowest_cases_countries = [row['Name'] for row in lowest_cases_countries]
print(f'Top 3 countries with the lowest number of cases: {lowest_cases_countries}')
spark.stop()
```

- Take a screenshot of the terminal to visualize the program result.

Your screenshot goes here

Submission Notice

- Export your answer file as pdf
- Rename the pdf following the format:

  lab05_<student number>_HoTen.pdf

  E.g. lab05_123456_NguyenThanhAn.pdf

  If you have not been assigned a student number yet, then use 123456 instead.
- Careless mistakes in filename, format, question order, etc. are not accepted (0 pts).