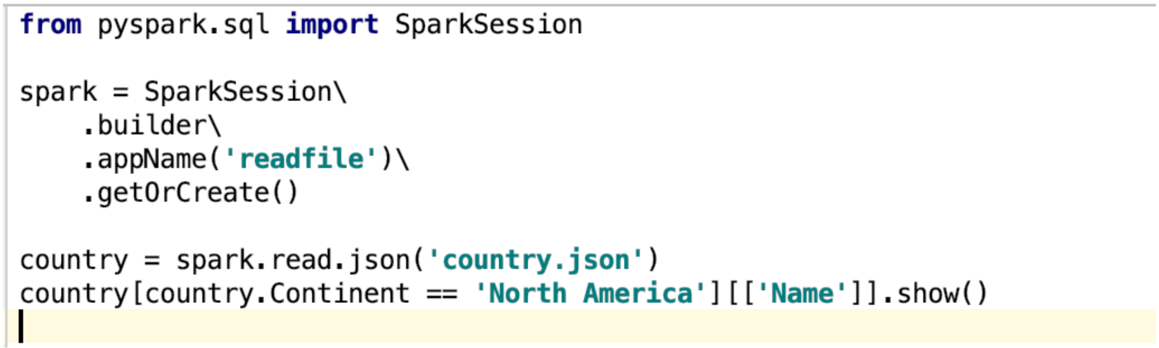
DSCI 551 Homework 5

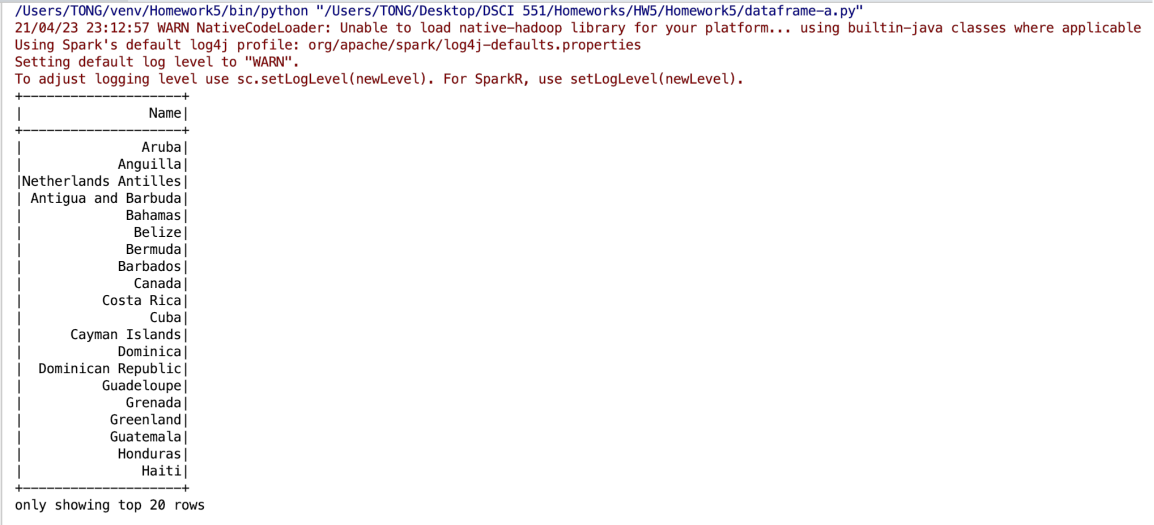
Question 2

Sitong Ju

a)

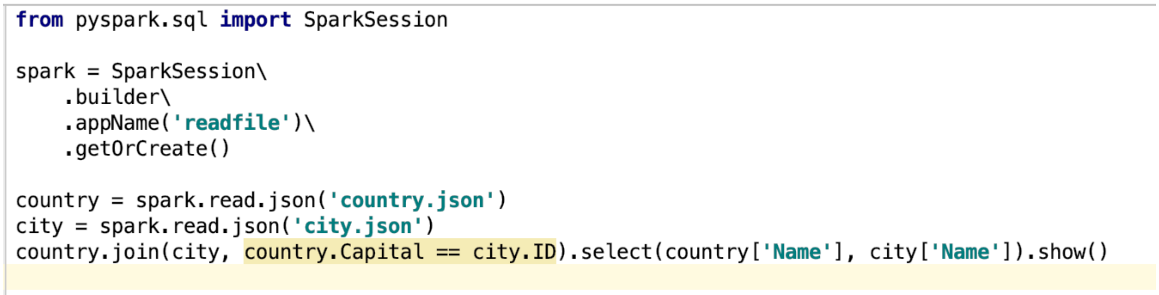


```
from pyspark.sql import SparkSession

spark = SparkSession\
    .builder\
    .appName('readfile')\
    .getOrCreate()

country = spark.read.json('country.json')
country[country.Continent == 'North America'][['Name']].show()
```

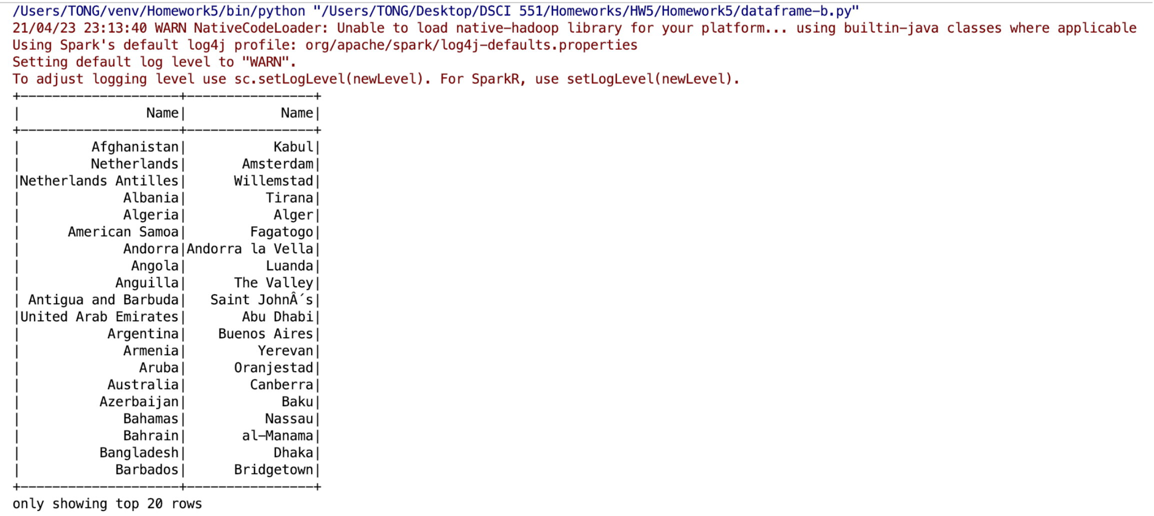


```
/Users/TONG/venv/Homework5/bin/python "/Users/TONG/Desktop/DSCI 551/Homeworks/HW5/Homework5/dataframe-a.py"
21/04/23 23:12:57 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
+--------------------+
|                Name|
+--------------------+
|               Aruba|
|            Anguilla|
|Netherlands Antilles|
| Antigua and Barbuda|
|             Bahamas|
|              Belize|
|             Bermuda|
|            Barbados|
|              Canada|
|          Costa Rica|
|                Cuba|
|      Cayman Islands|
|            Dominica|
|  Dominican Republic|
|          Guadeloupe|
|             Grenada|
|           Greenland|
|           Guatemala|
|            Honduras|
|               Haiti|
+--------------------+
only showing top 20 rows
```

b)

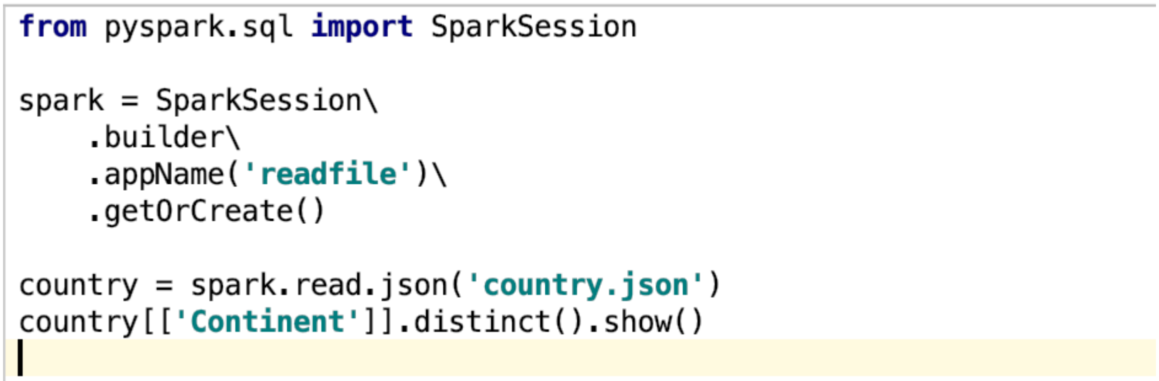


```
from pyspark.sql import SparkSession

spark = SparkSession\
    .builder\
    .appName('readfile')\
    .getOrCreate()

country = spark.read.json('country.json')
city = spark.read.json('city.json')
country.join(city, country.Capital == city.ID).select(country['Name'], city['Name']).show()
```

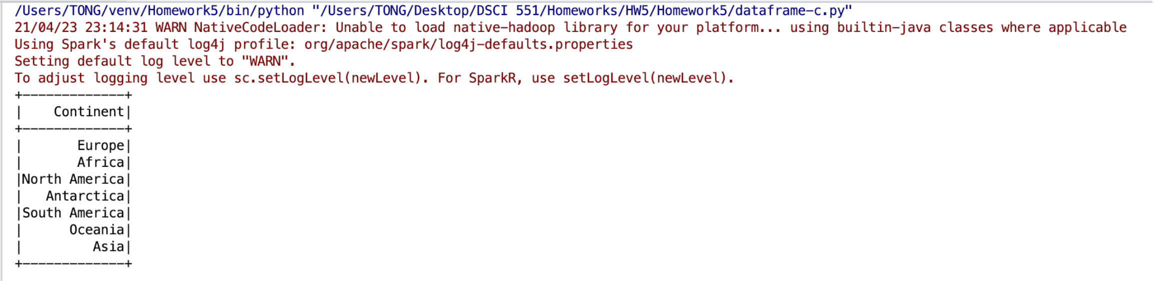


```
/Users/TONG/venv/Homework5/bin/python "/Users/TONG/Desktop/DSCI 551/Homeworks/HW5/Homework5/dataframe-b.py"
21/04/23 23:13:40 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
+--------------------+----------------+
|                Name|            Name|
+--------------------+----------------+
|         Afghanistan|           Kabul|
|         Netherlands|       Amsterdam|
|Netherlands Antilles|       Willemstad|
|             Albania|          Tirana|
|             Algeria|           Alger|
|      American Samoa|        Fagatogo|
|             Andorra|Andorra la Vella|
|              Angola|          Luanda|
|            Anguilla|      The Valley|
| Antigua and Barbuda|   Saint JohnÂ´s|
|United Arab Emirates|       Abu Dhabi|
|           Argentina|    Buenos Aires|
|             Armenia|         Yerevan|
|               Aruba|      Oranjestad|
|           Australia|        Canberra|
|          Azerbaijan|            Baku|
|             Bahamas|          Nassau|
|             Bahrain|       al-Manama|
|          Bangladesh|           Dhaka|
|            Barbados|      Bridgetown|
+--------------------+----------------+
only showing top 20 rows
```

c)

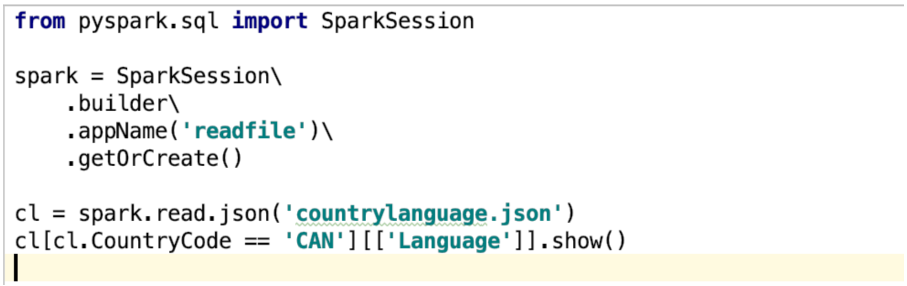


```python
from pyspark.sql import SparkSession

spark = SparkSession\
    .builder\
    .appName('readfile')\
    .getOrCreate()

country = spark.read.json('country.json')
country[['Continent']].distinct().show()
```

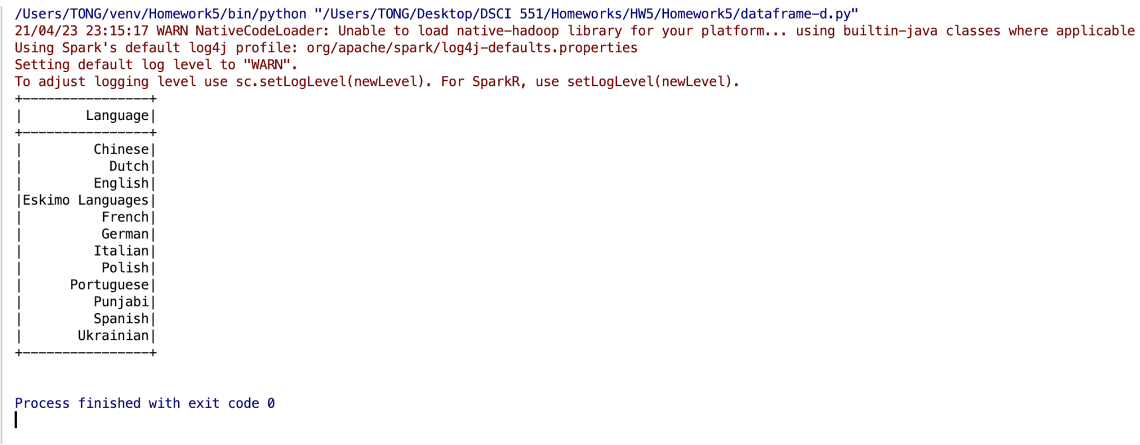


```
/Users/TONG/venv/Homework5/bin/python "/Users/TONG/Desktop/DSCI 551/Homeworks/HW5/Homework5/dataframe-c.py"
21/04/23 23:14:31 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
+-------------+
|    Continent|
+-------------+
|       Europe|
|       Africa|
|North America|
|   Antarctica|
|South America|
|      Oceania|
|         Asia|
+-------------+
```

d)

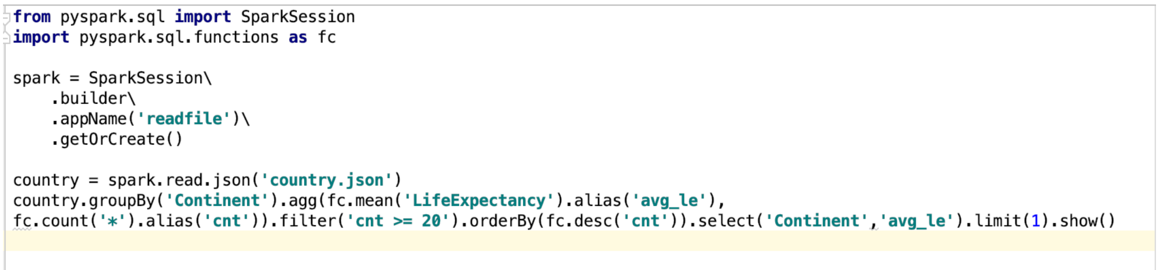


```python
from pyspark.sql import SparkSession

spark = SparkSession\
    .builder\
    .appName('readfile')\
    .getOrCreate()

cl = spark.read.json('countrylanguage.json')
cl[cl.CountryCode == 'CAN'][['Language']].show()
```

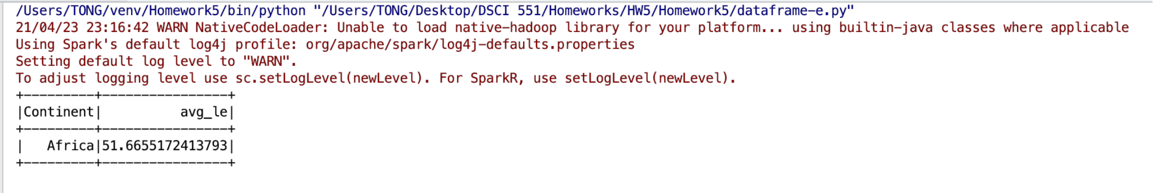


```
/Users/TONG/venv/Homework5/bin/python "/Users/TONG/Desktop/DSCI 551/Homeworks/HW5/Homework5/dataframe-d.py"
21/04/23 23:15:17 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
+----------------+
|        Language|
+----------------+
|         Chinese|
|           Dutch|
|         English|
|Eskimo Languages|
|          French|
|          German|
|         Italian|
|          Polish|
|      Portuguese|
|         Punjabi|
|         Spanish|
|       Ukrainian|
+----------------+


Process finished with exit code 0
```

e)

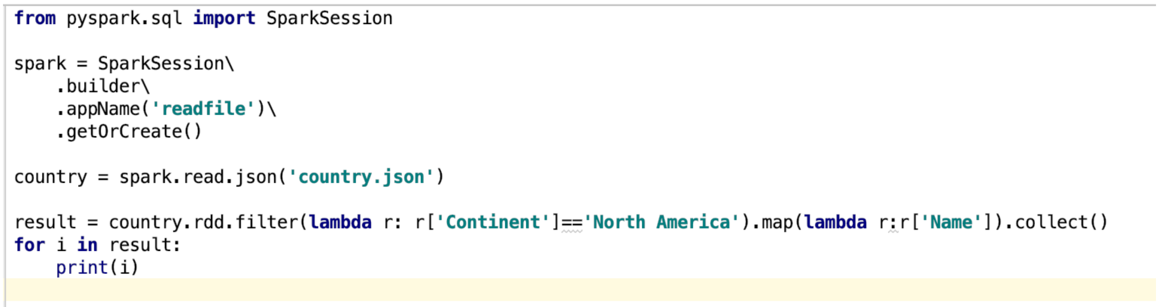


```python
from pyspark.sql import SparkSession
import pyspark.sql.functions as fc

spark = SparkSession\
    .builder\
    .appName('readfile')\
    .getOrCreate()

country = spark.read.json('country.json')
country.groupBy('Continent').agg(fc.mean('LifeExpectancy').alias('avg_le'),
fc.count('*').alias('cnt')).filter('cnt >= 20').orderBy(fc.desc('cnt')).select('Continent','avg_le').limit(1).show()
```

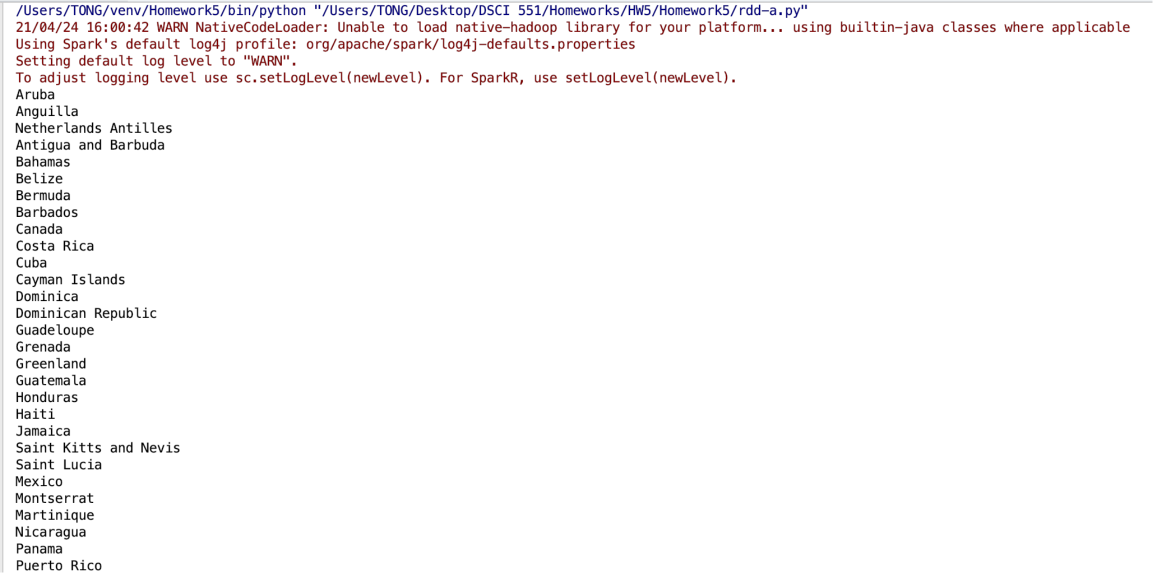


```
/Users/TONG/venv/Homework5/bin/python "/Users/TONG/Desktop/DSCI 551/Homeworks/HW5/Homework5/dataframe-e.py"
21/04/23 23:16:42 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
+---------+----------------+
|Continent|          avg_le|
+---------+----------------+
|   Africa|51.6655172413793|
+---------+----------------+
```

rdd_a)

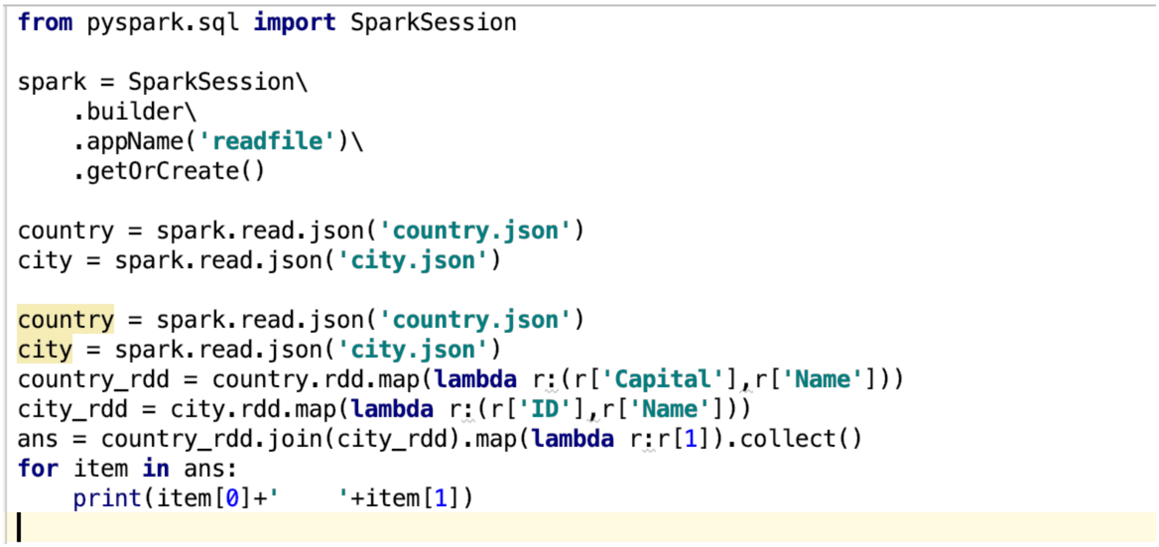


```python
from pyspark.sql import SparkSession

spark = SparkSession\
    .builder\
    .appName('readfile')\
    .getOrCreate()

country = spark.read.json('country.json')

result = country.rdd.filter(lambda r: r['Continent']=='North America').map(lambda r:r['Name']).collect()
for i in result:
    print(i)
```

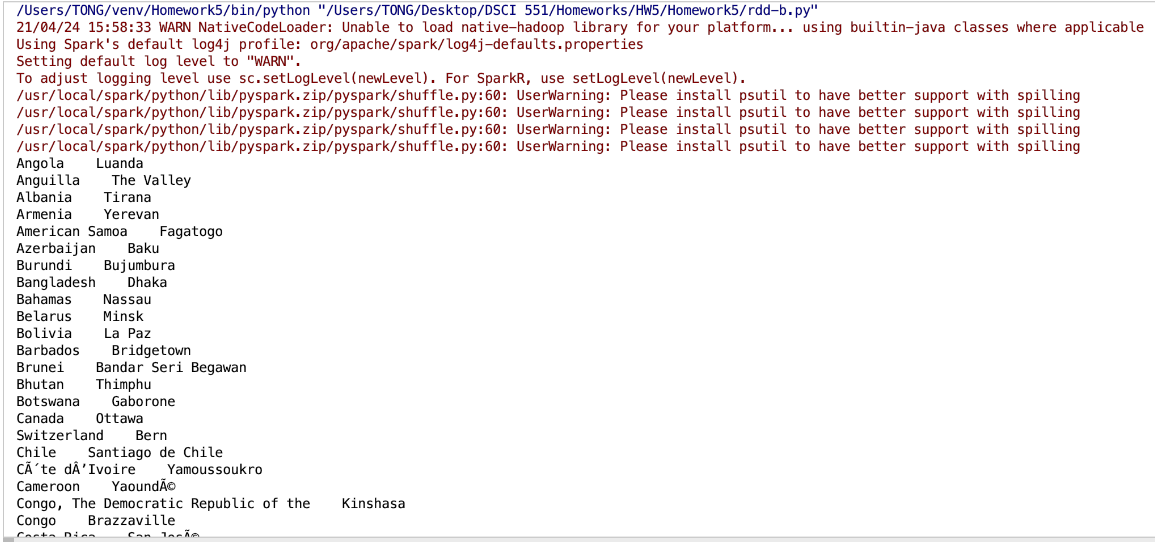


```
/Users/TONG/venv/Homework5/bin/python "/Users/TONG/Desktop/DSCI 551/Homeworks/HW5/Homework5/rdd-a.py"
21/04/24 16:00:42 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
Aruba
Anguilla
Netherlands Antilles
Antigua and Barbuda
Bahamas
Belize
Bermuda
Barbados
Canada
Costa Rica
Cuba
Cayman Islands
Dominica
Dominican Republic
Guadeloupe
Grenada
Greenland
Guatemala
Honduras
Haiti
Jamaica
Saint Kitts and Nevis
Saint Lucia
Mexico
Montserrat
Martinique
Nicaragua
Panama
Puerto Rico
```

rdd_b)



```python
from pyspark.sql import SparkSession

spark = SparkSession\
    .builder\
    .appName('readfile')\
    .getOrCreate()

country = spark.read.json('country.json')
city = spark.read.json('city.json')

country = spark.read.json('country.json')
city = spark.read.json('city.json')
country_rdd = country.rdd.map(lambda r:(r['Capital'],r['Name']))
city_rdd = city.rdd.map(lambda r:(r['ID'],r['Name']))
ans = country_rdd.join(city_rdd).map(lambda r:r[1]).collect()
for item in ans:
    print(item[0]+'    '+item[1])
```



```
/Users/TONG/venv/Homework5/bin/python "/Users/TONG/Desktop/DSCI 551/Homeworks/HW5/Homework5/rdd-b.py"
21/04/24 15:58:33 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
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). For SparkR, use setLogLevel(newLevel).
/usr/local/spark/python/lib/pyspark.zip/pyspark/shuffle.py:60: UserWarning: Please install psutil to have better support with spilling
/usr/local/spark/python/lib/pyspark.zip/pyspark/shuffle.py:60: UserWarning: Please install psutil to have better support with spilling
/usr/local/spark/python/lib/pyspark.zip/pyspark/shuffle.py:60: UserWarning: Please install psutil to have better support with spilling
/usr/local/spark/python/lib/pyspark.zip/pyspark/shuffle.py:60: UserWarning: Please install psutil to have better support with spilling
Angola    Luanda
Anguilla    The Valley
Albania    Tirana
Armenia    Yerevan
American Samoa    Fagatogo
Azerbaijan    Baku
Burundi    Bujumbura
Bangladesh    Dhaka
Bahamas    Nassau
Belarus    Minsk
Bolivia    La Paz
Barbados    Bridgetown
Brunei    Bandar Seri Begawan
Bhutan    Thimphu
Botswana    Gaborone
Canada    Ottawa
Switzerland    Bern
Chile    Santiago de Chile
CÃ´te dÂ’Ivoire    Yamoussoukro
Cameroon    YaoundÃ©
Congo, The Democratic Republic of the    Kinshasa
Congo    Brazzaville
Costa Rica    San JosÃ©
```