BDA1 -SPARK -Exercises

Wuhao Wang(wuhwa469)

Mucahit Sahin(mucsa806)

Q1: What are thelowestandhighesttemperaturesmeasuredeachyearfortheperiod1950-2014. Provide the lists sorted in the descending order with respect to the maximum temperature. In this exercise you will use the temperature-readings.csv file. The output should at least contain the following information (You can also include a Station column so that you may find multiple stations that record the highest (lowest) temperature.):

(only part of data here)

```
================= FINAL OUTPUT ============================================

(u'1975', (36.1, -37.0))

(u'1992', (35.4, -36.1))

(u'1994', (34.7, -40.5))

(u'2014', (34.4, -42.5))

(u'2010', (34.4, -41.7))

(u'1989', (33.9, -38.2))

(u'1982', (33.8, -42.2))

(u'1968', (33.7, -42.0))

(u'1966', (33.5, -49.4))

(u'1983', (33.3, -38.2))

(u'2002', (33.3, -42.2))

(u'1970', (33.2, -39.6))

(u'1986', (33.2, -44.2))

(u'1956', (33.0, -45.0))

(u'2000', (33.0, -37.6))

(u'1959', (32.8, -43.6))

(u'1991', (32.7, -39.3))

(u'2006', (32.7, -40.6))

(u'1988', (32.6, -39.9))

(u'2011', (32.5, -42.0))

(u'1999', (32.4, -49.0))
```

Q2: Count the number of readings for each month in the period of 1950-2014 which are higher than10degrees.Repeattheexercise,this time taking only distinct readings from each station. That is, if a station reported a reading above 10 degrees in some month, then it appears only once in the count for that month. In this exercise you will use the temperature-readings.csv file. The output should contain the following information:

```
================= FINAL OUTPUT ============================================

((u'1972', u'10'), 378)

((u'1973', u'05'), 377)

((u'1973', u'06'), 377)

((u'1972', u'08'), 376)

((u'1973', u'09'), 376)

((u'1972', u'05'), 376)

((u'1972', u'06'), 375)

((u'1971', u'08'), 375)

((u'1972', u'09'), 375)

((u'1971', u'09'), 374)

((u'1971', u'06'), 374)

((u'1972', u'07'), 374)
```

Q3: Find the average monthly temperature for each available station in Sweden. Your result should include average temperature for each station for each month in the period of 1960- 2014. Bear in mind that not every station has the readings for each month in this timeframe. In this exercise you will use the temperature-readings.csv file. The output should contain the following information: (only part of results)

```
================= FINAL OUTPUT ============================================
((u'2014', u'04', u'191720'), -2.06)
((u'2014', u'03', u'103410'), 0.69)
((u'2014', u'09', u'62130'), 14.01)
((u'2014', u'08', u'133180'), 13.17)
((u'2014', u'11', u'122430'), -3.31)
((u'2014', u'03', u'128390'), 2.81)
((u'2014', u'02', u'122430'), -4.55)
((u'2014', u'10', u'64130'), 10.51)
((u'2014', u'02', u'62180'), 3.61)
((u'2014', u'07', u'74180'), 17.26)
((u'2014', u'09', u'149340'), 8.67)
((u'2014', u'11', u'63590'), 5.51)
((u'2014', u'10', u'83230'), 9.14)
((u'2014', u'08', u'192840'), 11.57)
((u'2014', u'03', u'66420'), 5.16)
((u'2014', u'06', u'86340'), 14.3)
```

```
((u'2014', u'09', u'85490'), 10.7)
((u'2014', u'01', u'66400'), 0.58)
((u'2014', u'07', u'113420'), 17.29)
((u'2014', u'09', u'114140'), 9.85)
((u'2014', u'05', u'161940'), 7.38)
((u'2014', u'09', u'89230'), 14.18)
((u'2014', u'04', u'188850'), -2.64)
((u'2014', u'10', u'128390'), 6.39)
((u'2014', u'05', u'82110'), 12.94)
((u'2014', u'03', u'158740'), -2.15)
((u'2014', u'11', u'143440'), -1.34)
((u'2014', u'07', u'98490'), 17.9)
((u'2014', u'06', u'167990'), 10.02)
((u'2014', u'03', u'71420'), 6.12)
((u'2014', u'08', u'96040'), 15.43)
```

Q4: Provide a list of stations with their associated maximum measured temperatures and maximum measured daily precipitation. Show only those stations where the maximum temperature is between 25 and 30 degrees and maximum daily precipitation is between 100 mm and 200mm. In this exercise you will use the temperature-readings.csv and precipitation-readings.csv files. The output should contain the following information:

```
================= FINAL OUTPUT ==========================================

(There is no result matching the requirement)
```

Q5:) Calculate the average monthly precipitation for the Östergotland region (list of stations is provided in the separate file) for the period 1993-2016. In orderto dothis, you willfirstneed to calculate the total monthly precipitation for each station before calculating the monthly average (by averaging over stations). In this exercise you will use the precipitation-readings.csv and stations-Ostergotland.csv files. HINT (not for the SparkSQL lab): Avoid using joins here! stations-Ostergotland.csv is small and if distributed will cause a number of unnecessary shuffles when joined with precipitationRDD. If you distribute precipitation-readings.csv then either repartition your stations RDD to 1 partition or make use of the collect function to acquire a python list and broadcast function to broadcast the list to all nodes. The output should contain the following information:

```
================= FINAL OUTPUT ==========================================
(2016, 1, 22.33)
(2016, 7, 0.0)
(2016, 4, 26.9)
(2016, 6, 47.66)
(2016, 5, 29.25)
(2016, 2, 21.56)
(2016, 3, 19.96)
(2015, 11, 63.89)
(2015, 6, 78.66)
(2015, 10, 2.26)
(2015, 8, 26.99)
```

```
(2015, 5, 93.22)
(2015, 7, 119.1)
(2015, 12, 28.93)
(2015, 1, 59.11)
(2015, 4, 15.34)
(2015, 3, 42.61)
(2015, 2, 24.82)
(2015, 9, 101.3)
(2014, 9, 48.45)
(2014, 7, 22.99)
(2014, 12, 35.46)
```