Simja649 Danko376

# Lab report 1 - Spark

## Assignment 1

For this assignment, we had to provide a list of temperatures measured for the period 1950-2014. The list outputs the highest and lowest temperature measured for each year and is sorted in descending order from the year with the highest temperature.

First we extracted the data into a new map where data was split using “;” as the input file was of type .csv. Then we extracted the values that we needed to complete the assignment, year (filtered for the period) and temperature. Then we made two new maps, one for max temperatures and one for the min temperatures by using the reduceByKey function. This gave us the maximum and minimum temperatures for each year and then we just used a join function to get both values for each key (year) in a map. Sort by descending for our first value (maxtemp) and then save the file. See result below.

```
================ 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'1986', (33.2, -44.2))
(u'1970', (33.2, -39.6))
(u'1956', (33.0, -45.0))
(u'2000', (33.0, -37.6))
(u'1959', (32.8, -43.6))
(u'2006', (32.7, -40.6))
(u'1991', (32.7, -39.3))
(u'1988', (32.6, -39.9))
(u'2011', (32.5, -42.0))
(u'1999', (32.4, -49.0))
(u'1953', (32.2, -38.4))
(u'1955', (32.2, -41.2))
(u'1973', (32.2, -39.3))
(u'2003', (32.2, -41.5))
(u'2007', (32.2, -40.7))
(u'2008', (32.2, -39.3))
(u'2005', (32.1, -39.4))
(u'1969', (32.0, -41.5))
```

## Assignment 2

For assignment 2, we were to count the number of readings for each month in the period of 1950-2014 which was higher than 10 degrees for task a. In task b we were to repeat by only taking distinct readings from each station.

To complete both tasks. We extracted key value pairs as station, year, month (key) and temperature (value). We first filtered between 1950 and 2014 and degrees over 10. In task a, we only took year and month as key and value as 1. Then reducing the key to get a count on the value. In task b we did nearly the same, but instead to the distinct stations and then reduced year and month by key and added them up by count. See results below.

Simja649 Danko376

2a

```
================= FINAL OUTPUT ===
((u'2008', u'11'), 1663)
((u'1970', u'01'), 1)
((u'1971', u'06'), 47448)
((u'1989', u'12'), 32)
((u'1966', u'11'), 213)
((u'1956', u'05'), 12696)
((u'1998', u'07'), 120912)
((u'1975', u'02'), 33)
((u'1983', u'10'), 12074)
((u'1992', u'05'), 34325)
((u'1969', u'10'), 14305)
((u'1994', u'10'), 7348)
((u'1988', u'06'), 64141)
((u'1976', u'07'), 64960)
((u'1963', u'12'), 1)
((u'1989', u'01'), 90)
((u'2007', u'10'), 18619)
((u'1970', u'12'), 2)
((u'1959', u'08'), 23997)
((u'1996', u'09'), 45183)
((u'2008', u'02'), 101)
((u'1952', u'09'), 5667)
((u'2011', u'08'), 133483)
((u'1975', u'11'), 771)
```

2b

```
================ FINAL OUTPUT =
((u'2008', u'11'), 107)
((u'1970', u'01'), 1)
((u'1992', u'05'), 311)
((u'1971', u'06'), 374)
((u'1989', u'12'), 10)
((u'1966', u'11'), 76)
((u'1956', u'05'), 125)
((u'1998', u'07'), 326)
((u'1975', u'02'), 19)
((u'1983', u'10'), 246)
((u'1978', u'09'), 358)
((u'1969', u'10'), 346)
((u'1994', u'10'), 265)
((u'1988', u'06'), 322)
((u'1963', u'12'), 1)
((u'1976', u'07'), 356)
((u'1965', u'07'), 349)
((u'2007', u'10'), 269)
((u'1970', u'12'), 2)
((u'1996', u'09'), 340)
((u'2008', u'02'), 27)
((u'1952', u'09'), 114)
((u'1989', u'01'), 29)
((u'1975', u'11'), 190)
((u'1950', u'10'), 46)
((u'2013', u'06'), 302)
((u'2014', u'07'), 297)
```

# Assignment 3

For this assignment, we had to find the monthly temperature for each station in Sweden for the period 1960-2014. The result had to show what the average monthly temperature was for each station a given year and month.

Started by extracting relevant data and putting this into a map and the dates filtered for our period:
Key: Station, date
Value: Temperature

We extracted the max and min temperatures for each day by using reduceByKey two times so we got two different maps. These were then combined so we had a map with (station, date: max_temp, min_temp). We then added these temperatures together, added a count and also made a new key consisting of only the year and month for each station. To get the monthly average we once again used reduceByKey to sum all of the temperatures, sum the count. Then we filtered the period for years and finally we calculated the monthly average by dividing the sum with the count multiplied by 2 because for each day we have two temperatures (the min and the max).

This gave us the following output:
((station, year, month), average monthly temperature)

```
================= FINAL OUTPUT ====================
((u'62560', u'1969', u'06'), 17.22833333333333)
((u'86420', u'2004', u'08'), 17.591935483870966)
((u'103090', u'1983', u'03'), -1.1129032258064515)
((u'180730', u'1976', u'03'), -8.301612903225807)
((u'83420', u'2013', u'03'), -2.520967741935484)
((u'105450', u'1994', u'12'), -0.9354838709677419)
((u'123250', u'1981', u'05'), 9.633870967741935)
((u'62040', u'2007', u'01'), 3.7806451612903222)
((u'77180', u'1969', u'02'), -3.189285714285714)
((u'127380', u'1961', u'09'), 11.086666666666668)
((u'54550', u'1993', u'03'), 2.6451612903225805)
((u'96350', u'1970', u'04'), 2.2)
((u'86420', u'2012', u'10'), 5.859677419354839)
((u'116430', u'1989', u'01'), -0.596774193548387)
((u'103090', u'2014', u'07'), 18.19838709677419)
((u'62190', u'1970', u'12'), 2.0629032258064517)
((u'147090', u'2004', u'02'), -10.067241379310344)
((u'68550', u'2010', u'09'), 13.240000000000002)
((u'63450', u'1987', u'01'), -9.191935483870969)
((u'172770', u'2001', u'08'), 12.998387096774195)
((u'173810', u'2005', u'02'), -9.691071428571432)
```

# Assignment 4

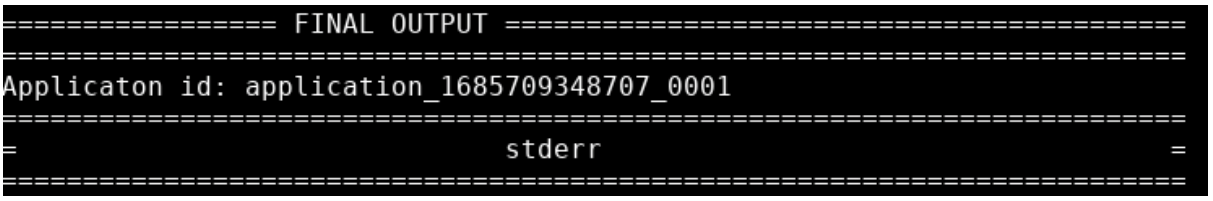
In task 4 we were to get stations with maximum measured temperatures and maximum measured daily precipitation. With temps between 25 and 30 degrees and precipitation between 100 and 200mm.

To complete this task we first created help-functions max and sum which took the maximum and the sum. We are using them later when reducing by key.

We first retrieved maximum temp by getting the station (key) and temperature (value). Then we reduce by key to get the maximum for each station. Then filtering on temperatures between 25 and 30.

After that we retrieved the daily precipitation by getting the station and full date (key) and the precipitation (value). We first got the sum of each day by using reduced by key with the sum

help-function. Then filtering between 100 and 200mm of the sum of each day. Then we create a key with station and value as precipitation. Lastly, we combine both temperature and precipitation by using a join to get the stations we desire. See result below (no stations with these requirements).




# Assignment 5

For the last assignment we calculate the average monthly precipitation for the whole of Östergötland, the period was from 1993 to 2016.

Same as for all of the previous assignments we extract the data that we want to use, from the stations-ostergotland.csv-file we wanted to get the stations numbers. From the precipitation-readings.csv-file we wanted to get the station numbers, year, month and also the precipitation readings. When this data was extracted we made a new map using all the values from the precipitation-readings but filtered the stations so we only got station readings from the Östergötland region in the period 1993-2016.

Now we have a map with the following format:
Keys: Station (only Östergötland stations), year, month
Value: Precipitation, 1

Using this map we can sum the monthly total precipitation and count the days for each month for the whole of Östergötland. This gave us a new map with the year and month for keys and the values are total precipitation for that month together with the number of days. Again, dividing the values with each other to get the average. The resulting map gives us the average precipitation for each year and month for the whole of Östergötland.

Simja649 Danko376

```
================ FINAL OUTPUT ==========
((u'2012', u'09'), 72.75)
((u'1995', u'05'), 26.00000000000002)
((u'2011', u'08'), 86.26666666666667)
((u'2007', u'04'), 21.249999999999996)
((u'2007', u'06'), 108.95)
((u'1993', u'04'), 0.0)
((u'2011', u'10'), 43.75)
((u'2014', u'10'), 72.13749999999999)
((u'1996', u'09'), 57.466666666666676)
((u'1995', u'07'), 43.6)
((u'2002', u'05'), 72.13333333333334)
((u'2010', u'04'), 23.783333333333335)
((u'1999', u'01'), 61.933333333333394)
((u'2013', u'11'), 46.37500000000002)
((u'2010', u'03'), 23.88333333333334)
((u'1999', u'10'), 18.549999999999997)
((u'2003', u'11'), 54.45000000000001)
((u'2014', u'04'), 31.762500000000006)
((u'2006', u'09'), 19.26666666666667)
((u'2016', u'02'), 21.5625)
((u'2013', u'09'), 26.18750000000001)
((u'2016', u'05'), 29.250000000000007)
((u'2015', u'01'), 59.112500000000026)
((u'2009', u'07'), 113.16666666666663)
((u'2008', u'05'), 23.133333333333336)
```