Introduction

In this assignment, you have to write a series of Apache Flink programs to analyze a *flow cytometry* data set that was collected to study the immune reaction of an organism to a virus infection. The analysis is structured into three separate tasks.

Input Data Set Description

The input data can be found in the HDFS directory /share/cytometry in our cluster. There are several csv files all with names like measurements_arcsin200_pX.csv. All of them have the same comma-delimited format, with each line representing the measurements for a single cell:
sample, FSC-A, SSC-A, CD48, Ly6G, CD117, SCA1, CD11b, CD150, CD11c, B220, Ly6C,...

Each sample consists of a number of individual cells that have been analysed (such as a blood sample). Each row in the measurement files corresponds to one cell. Note that there is a header row that just gives the meaning of each column:
- The sample column specifies the specific sample to which a measurement belongs to.
- FSC-A and SSC-A are quality metrics about the measurements that reflect how much laser light was scattered from a measured cell (FSC: ”forward-scatter”; SSC: ”side-scattered”).
- The remaining 14 columns contain the measurements for various cell markers. The values have been already normalised by us into the same value space (using arcsin). Each marker corresponds to a specific antigen on the cell surface that was investigated in an experiment. These antigens correspond to different functions or development stages of a cell, and the more an antigen is expressed by a cell, the higher is the measured value. For the purpose of this assignment, think of them as coordinates in a 14-dimensional (normalised) value space where each dimension has a name that corresponds to one of the antigens.

The meta-data for the experiments are stored in the csv file /share/cytometry/experiments.csv Each line of this file contains a record with the following comma-delimited information:
sample, date, experiment, day, subject, kind, instrument, researchers

The sample column specifies to which specific flow cytometry sample a measurement belongs to. The date column specifies when an experiment was conducted. In our use case, the different samples are from an investigation of the immune reaction of an organism to an infection with the West Nile Virus (WNV). Day gives the number of days since the infection until a sample was taken and analysed. A value of 0 corresponds to a healthy cell before the infection, a value of, eg., day 3 would be a sample taken at the third day of an infection. The samples are taken from different subjects and are of a certain kind. In our case, the samples are taken from the bone marrow of different individual mice. The last column lists one or multiple researchers who were conducting the different experiments, with the names of the researchers separated by semicolon.

Analysis Task Descriptions

1. **Task 1: Number of (valid) measurements conducted *per researcher*.**

The first task is an explorative analysis of the flow cytometry data.
You shall write a Flink program that finds the number of valid measurements done per individual researcher. We consider a measurement valid if its FSC-A and SSC-A values are in the range of 1 to 150,000. Note that there might be more than one researcher involved in analysing a certain sample. In this case, you can count all measurements from the same sample to each researcher who was involved.

The output file should have the following format, ordered by number of measurements in descending order (researchers with most measurements first); researchers with the same number of measurements should be listed alphabetically:
researcher \t numberOfMeasurements

2. **Task 2:** k**-means clustering of the measurements.**

In the second task, you shall implement the iterative k-means algorithm and then use this algorithm to cluster all valid cell measurements into k clusters (k as input parameter). You can restrict your clustering to three dimensions, with the default: Ly6C, CD11b, and SCA1
Filter out all measurements with FSC-A or SSC-A values outside the range 1 to 150,000. Then start with k random centroids and conduct k-means clustering for a number of iterations, with a default value of 10 iterations (configurable at runtime). To determine the distance between a point p and a cluster c use the standard Euclidean distance:
distance(p, c) = sqrt(sqr(c.x - p.x) + sqr(c.y – p.y) + sqr(c.z – p.z))
The output file should have the following tab-delimited format (ordered by cluster ID):
clusterID \t number_of_measurements \t Ly6C \t CD11b \t SCA1
The cluster ID is an integer value between 1 and k. For each cluster, give the number of valid measurements which have been associated to this cluster (i.e. the size of the cluster). The last three values specify the centroid of the cluster for the three dimensions which we consider (default: Ly6C, CD11b and SCA1).

3. **Task 3: Outlier removal and reclustering.**

In the third task, you shall identify and remove outliers from the dataset: For each cluster, remove the 10% of measurements with the highest residual error, where residual error is defined as the Euclidean distance from a measurement to its assigned cluster centroid. Create a new dataset with these identified outliers removed. Then use this new dataset to produce a new k-means clustering with the same values for k and the same number of iterations. The output should have the same format than for Task 2.