Pig Workshop - Keys

Exercise 2:

```
stationsRaw = Load '/home/cloudera/BixiData/Stations_2017.csv' using
PigStorage(',') as
(code:chararray,name:chararray,latitude:double,longitude:double);

bixi07Raw = Load '/home/cloudera/BixiData/OD_2017-07.csv' Using
PigStorage(',') as (start_date:chararray,
start_station_code:chararray, end_date:chararray,
end_station_code:chararray, duration_sec:int, is_member:int );

stations = Filter stationsRaw By not (code == 'code');

bixi07 = Filter bixi07Raw By not (start_date == 'start_date');

-- st5 = Limit stations 5;

-- Dump st5 ;

-- bx5 = Limit bixi07 5;

-- Dump bx5 ;


bixi07grp = Group bixi07 by start_station_code;

bixi07sc = Foreach bixi07grp Generate group as start_station_code,
COUNT(bixi07.start_station_code) as count;

joinedss = Join bixi07sc by start_station_code, stations by code;

sstcounts = Foreach joinedss Generate name, count;

sstcountsdesc = Order sstcounts by count desc;

sstop5 = Limit sstcountsdesc 5;

dump sstop5;
```

Exercise 3:

```
dataset = LOAD '/home/cloudera/pig/tw.txt' AS (id: long, fr: long);

-- check if user IDs are valid (e.g. not null) and clean the dataset

SPLIT dataset INTO good_dataset IF id is not null and fr is not
null, bad_dataset OTHERWISE;

-- organize data such that each node ID is associated to a list of
neighbors
```

```
nodes = GROUP good_dataset BY id;

-- foreach node ID generate an output relation consisting of the
node ID and the number of "friends"

friends = FOREACH nodes GENERATE group,COUNT(good_dataset) AS
followers;

-- count the following

nodes2 = GROUP good_dataset BY fr;

followings = FOREACH nodes2 GENERATE group, COUNT(good_dataset);

-- find the outliers

outliers = FILTER friends BY followers<3;

STORE friends INTO '/home/cloudera/pig/tw/';

STORE followings INTO '/home/cloudera/pig/tw/';

STORE outliers INTO '/home/cloudera/pig/tw/';
```