Assignment 4 – DS8003

Amir Ghaderi – 500794236

November 10th, 2016

Question 1:

b = sample a 0.1;

a = load '/user/amir/pig/full\_text\_clean.txt';

store b into '/user/amir/pig/full\_text\_small.txt' using PigStorage('\t');

Question 2:

a = load '/user/amir/pig/full\_text\_clean.txt' AS (id:chararray, lat:float, lon:float, tweet:chararray, modified\_lat:float, modified\_lon:float);

b = foreach a generate flatten(TOKENIZE(tweet)) as token;

c = group b by token;

d = foreach c generate flatten(group) as word, COUNT(b) as cnt;

e = order d by cnt desc;

f = limit e 3;

dump f;

Question 3:

a = load '/user/amir/pig/cities\_clean.txt' AS (city\_name:chararray, lat:float, lon:float, modified\_lat:float, modified\_lon:float);

b = group a ALL;

c = foreach b generate COUNT\_STAR(a);

dump c;

Question 4:

a = load '/user/amir/pig/full\_text\_clean.txt' AS (id:chararray,lat:float, lon:float, tweet:chararray,modified\_lat:float,modified\_lon:float);

b = load '/user/amir/pig/cities\_clean.txt' AS (city\_name:chararray,lat:float, lon:float,modified\_lat:float,modified\_lon:float);

c= join a by (modified\_lat,modified\_lon),b by(modified\_lat,modified\_lon)

d= foreach c generate a::tweet as tweet,b::city\_name as city\_name,SQRT((a::lat - b::lat) \* (a::lat - b::lat) + (a::lon - b::lon) \* (a::lon - b::lon)) as distance;

e= group d by (tweet)

l= foreach e {

sortd = order d by distance asc;

g = limit sortd 1;

generate g;

};

dump l;