**CIND 719 – Assignment 2**

**Brendan Dagys**

1) *Find hour of the day when highest number of tweets were generated by users on March 6, 2010*

a = load '/user/pig/full\_text.txt' as (id: chararray, ts: chararray, location: chararray, lat: float, lon: float, tweet: chararray);

b = filter a by ts MATCHES '2010-03-06.\*';

c = foreach b generate GetHour(ToDate(ts)) as hourofday;

d = group c by hourofday;

e = foreach d generate group, COUNT(c) as cnt;

f = order e by cnt desc;

g = limit f 1;

dump g;



2) *Find top 10 topics (#hashtags)*

a = load '/user/pig/full\_text.txt' as (id: chararray, ts: chararray, location: chararray, lat: float, lon: float, tweet: chararray);

i = foreach a generate FLATTEN(TOKENIZE(tweet)) as token;

j = filter i by STARTSWITH(token, '#');

k = foreach j generate REGEX\_EXTRACT(LOWER(token), ‘(#\\w+)’, 1) as tweet2;

l = group k by tweet2;

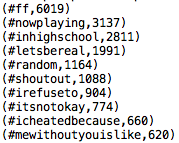
m = foreach l generate group, COUNT(k) as cnt2;

n = order m by cnt2 desc;

o = limit n 10;

dump o;

Hashtags on Twitter are NOT case-sensitive. Because of this, I chose to convert all tweets to lower-case before tokenizing and finding the most common 10 hashtags. As well, since hastags terminate when any special character is introduced, I used a REGEX\_EXTRACT function to extract the part of the hashtag that was NOT any special character. This way, I retrieved the true count of each hashtag by not treating them as unique when they contained special characters.



3) *Find top 10 mentions (@xxxxxxx)*

a = load '/user/pig/full\_text.txt' as (id: chararray, ts: chararray, location: chararray, lat: float, lon: float, tweet: chararray);

p = filter a by tweet MATCHES ‘^(?!.\*RT @USER\_).\*’;

q = foreach p generate FLATTEN(TOKENIZE(LOWER(tweet))) as token2;

r = filter q by token2 MATCHES '@user\_\\w{8}.\*';

s = group r by token2;

t = foreach s generate group, COUNT(r) as cnt3;

u = order t by cnt3 desc;

v = limit u 10;

dump v;

The second line uses a look-ahead assertion to filter OUT any rows that contain ‘RT @USER\_’. Tweets that contain this string are retweets, and are simply copies of other tweets. Because retweets are just copies of other tweets, the mentions within them are not original mentions, and shouldn’t be counted. On Twitter, when a tweet containing mentions is retweeted, it does not show up as a mention for the person contained in the tweet being retweeted. Hence, I removed all tweets that are retweets before tokenizing the tweets and summing the mentions.

