Pig assignment

Seongtaek Lim, School of Information
Date submitted: 09/14/12
Analyzing Big Data with Twitter in Fall 2012

Step 1: Learn the Tutorial

Question 1.a: Why do we do this? Under what circumstances would there be duplicates?

```
ngramed2 = DISTINCT ngramed1;
```

The reason why we do this is to limit the occurrence of each n-gram from a user up to once an hour before we count them all. By using *DISTINCT* command, we can get the unique records across the data we have. There are some cases in which the dataset *ngramed1* may include duplicates.

First, there would be duplicates if a user searched twice or more, with the same keywords within one hour. This happens because of the truncation of values of time field by *ExtractHour(time)*, which makes each record lose its uniqueness in the dataset.

Second, even though a user searched with different keywords within one hour, there would be duplicates if n-grams of different queries share some of their subsets. For example, if we have two records as follows,

BED75271605EBD0C	01	yahoo chat
BED75271605EBD0C	01	hawaii chat universe

each will generate n-grams as follows and we can still find a pair of duplicates from two different queries where *N=2*:

BED75271605EBD0C	01	yahoo
BED75271605EBD0C	**01**	**chat**
BED75271605EBD0C	01	yahoo chat
BED75271605EBD0C	01	hawaii
BED75271605EBD0C	**01**	**chat**
BED75271605EBD0C	01	universe
BED75271605EBD0C	01	hawaii chat
BED75271605EBD0C	01	chat universe

Like this, subsets of n-grams also can bring about unexpected duplicates.

Question1.b In English, what is this command doing?

```
hour_frequency1 = GROUP ngramed2 BY (ngram, hour);
```

What this statement does is regrouping the dataset with two criteria: n-gram used for search and when it was queried. Thus, a record of the result of this command shows an n-gram used for search, when it was used, and list (bag) of the users who queried, hour, and n-gram itself.

(edu,18)	{(B9E187FD56A5C322,18,edu 06878125BE78B42C,18,edu)}
n-gram *edu* was queried within 6-7 p.m.	The users B9E187FD56A5C322 and 06878125BE78B42C searched with the n-gram *edu* within 6-7 p.m.

Question 1.c Why is the FLATTEN command used here? What are we really counting with COUNT($1) mean? (Hint: see the GROUP documentation linked to above.)

```
hour_frequency2 = FOREACH hour_frequency1 GENERATE flatten($0),
                  COUNT($1) as count;
```

The reason why the *flatten($0)* command was used is to get the first field of *hour_frequency1*, or tuple of *ngram* and *hour*, in two atomic values.

What the command *count($1)* actually counts is the number of items within the bag type list of *user*, *hour*, and *ngram*. This number means the number of users tried to search with the n-gram per hour.

Question 1.d In English, what is this command doing?

```
uniq_frequency1 = GROUP hour_frequency2 BY group::ngram;
```

This statement regroups the records so that each unique n-gram has a list of tuples containing *ngram*, *hour*, *count*. Thus, each record of the result consists of a list containing information of a ngram used for search, when it was used, and the number of search attempts.