Lab 7

Big Data Spring 2016

Today's Lab

- The Query Phrase Popularity script (script1-local.pig or script1-hadoop.pig) processes a search query log file from the Excite search engine and finds search phrases that occur with particular high frequency during certain times of the day.

- The Data: 3 columns: user, time, query

```
9593C58F7C1C5CE4    970916072134    levis
9593C58F7C1C5CE4    970916072311    levis strause & co
9593C58F7C1C5CE4    970916072339    levis 501 jeans
45531846E8E7C127    970916065859
45531846E8E7C127    970916065935
45531846E8E7C127    970916070105    "brazillian soccer teams"
45531846E8E7C127    970916070248    "brazillian soccer"
45531846E8E7C127    970916071154    "population of maldives"
082A665972806A62    970916123431    pegasus
```

- We will go through the script line by line to see what is happening
- For details on the PigLatin programming language, see http://pig.apache.org/docs/r0.14.0/basic.html

Pig Script

- Register the tutorial JAR file so that the included UDFs can be called in the script.

```
REGISTER ./tutorial.jar;
```

- Use the PigStorage function to load the excite log file into the raw bag as an array of records.
- Input: (user,time,query)

```
raw = LOAD 'excite-small.log' USING
PigStorage('\t') AS (user, time, query);
```

- Call the NonURLDetector UDF to remove records if the query field is empty or a URL.

```
clean1 = FILTER raw BY
org.apache.pig.tutorial.NonURLDetector(query);
```

Pig Script

- Call the ToLower UDF to change the query field to lowercase.

```
clean2 = FOREACH clean1 GENERATE user, time,
org.apache.pig.tutorial.ToLower(query) as query;
```

- Because the log file only contains queries for a single day, we are only interested in the hour.
- The excite query log timestamp format is YYMMDDHHMMSS.
- Call the ExtractHour UDF to extract the hour (HH) from the time field.

```
houred = FOREACH clean2 GENERATE user,
org.apache.pig.tutorial.ExtractHour(time) as
hour, query;
```

- Call the NGramGenerator UDF to compose the n-grams of the query.

```
ngramed1 = FOREACH houred GENERATE user, hour,
flatten(org.apache.pig.tutorial.NGramGenerator(qu
ery)) as ngram;
```

Terminology

- From Wikipedia:

In the fields of computational linguistics and probability, an **n-gram** is a contiguous sequence of **n** items from a given sequence of text or speech. The items can be phonemes, syllables, letters, words or base pairs according to the application. The **n-grams** typically are collected from a text or speech corpus.

Pig Script

- Use the DISTINCT command to get the unique n-grams for all records.

```
ngramed2 = DISTINCT ngramed1;
```

- Use the GROUP command to group records by n-gram and hour.

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

- Use the COUNT function to get the count (occurrences) of each n-gram.

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

Pig Script

- Use the GROUP command to group records by n-gram only.
- Each group now corresponds to a distinct n-gram and has the count for each hour.

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

- For each group, identify the hour in which this n-gram is used with a particularly high frequency.
- Call the ScoreGenerator UDF to calculate a "popularity" score for the n-gram.

```
uniq_frequency2 = FOREACH uniq_frequency1
GENERATE flatten($0),
flatten(org.apache.pig.tutorial.ScoreGenerato
r($1));
```

Pig Script

- Use the FOREACH-GENERATE command to assign names to the fields.

```
uniq_frequency3 = FOREACH uniq_frequency2
GENERATE $1 as hour, $0 as ngram, $2 as
score, $3 as count, $4 as mean;
```

- Use the FILTER command to move all records with a score less than or equal to 2.0.

```
filtered_uniq_frequency = FILTER
uniq_frequency3 BY score > 2.0;
```

Pig Script

- Use the ORDER command to sort the remaining records by hour and score.

```
ordered_uniq_frequency = ORDER
filtered_uniq_frequency BY hour, score;
```

- Use the PigStorage function to store the results.
- Output: (hour, n-gram, score, count, average_counts_among_all_hours)

```
STORE ordered_uniq_frequency INTO 'script1-
local-results.txt' USING PigStorage();
```