Temporal Query Phrase Popularity

The Temporal Query Phrase Popularity script (script2-local.pig or script2-hadoop.pig) processes a search query log file from the Excite search engine and compares the occurrence of frequency of search phrases across two time periods separated by twelve hours.

The script is shown here:

Register the tutorial JAR file so that the user-defined functions (UDFs) can be called in the script.

REGISTER ./tutorial.jar;

Use the PigStorage function to load the excite log file (excite.log or excite-small.log) into the “raw” bag as an array of records with the fields user, time, and query.

raw = LOAD 'excite.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);

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 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(query)) as ngram;

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

ngramed2 = DISTINCT ngramed1;

Use the GROUP operator to group the 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;

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

hour\_frequency3 = FOREACH hour\_frequency2 GENERATE $0 as ngram, $1 as hour, $2 as count;

Use the FILTERoperator to get the n-grams for hour ‘00’

hour00 = FILTER hour\_frequency2 BY hour eq '00';

Uses the FILTER operators to get the n-grams for hour ‘12’

hour12 = FILTER hour\_frequency3 BY hour eq '12';

Use the JOIN operator to get the n-grams that appear in both hours.

same = JOIN hour00 BY $0, hour12 BY $0;

Use the FOREACH-GENERATE operator to record their frequency.

same1 = FOREACH same GENERATE hour\_frequency2::hour00::group::ngram as ngram, $2 as count00, $5 as count12;

Use the PigStorage function to store the results. The output file contains a list of n-grams with the following fields: hour, count00, count12.

STORE same1 INTO '/tmp/tutorial-join-results' USING PigStorage();