Collocation Extraction Using Elastic Map Reduce

Submitted by:

*Tal Yitzhak (talitz), 204260533*

*Assaf Magrisso (assafmag), 201247020*

In this assignment you we automatically extracted collocations from the Google 2-grams dataset using Amazon Elastic Map Reduce.

We used chaining of 5 map reduces:

First map reduce – filtering stop words if requested, and calculating c(w1,w2).

Second map reduce – calculating c(w1).

Third map reduce – calculating c(w2).

Fourth map reduce – calculating N and npmi for each bigram.

Fifth map reduce – calculating relative npmi for each bigram, and extract all collocations above each of the two minimum inputs (minimal pmi & relative minimal pmi).

We ran out project on the following large corpuses:

eng-us-all [38.3 GB] [3,923,370,881 rows]

How did we run this?

0.5 0.2 eng 1 s3://datasets.elasticmapreduce/ngrams/books/20090715/eng-us-all/2gram/data

Good Examples

Bad Examples

Why wrong collocations were extracted?

heb-all [2.4 GB] [252,069,581]

How did we run this?

Good Examples

Bad Examples

Why wrong collocations were extracted?