Collocation Extraction Using Elastic Map Reduce

Submitted by:

Tal Yitzhak (talitz), 204260533

Assaf Magrisso (assafmag), 201247020

In this assignment 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, with 15 instances of M1.Xlarge:

eng-us-all [38.3 GB] [3,923,370,881 rows] – 4 hours, 8 minutes

How did we run this?

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

Good Examples

bigram	npmi	decade
smoking tobacco	1.0	170
contagious Diseases	1.0	170
bra vest	1.0	170
babe sucking	1.0	170
MILITARY DUTIES	1.0	170
JESUS CHRIST	0.9447681140206563	170
HARVARD COLLEGE	0.9418573159865764	170

Bad Examples

bigram	npmi	decade
scandalous thing	0.5000927289746182	170
Altho '	0.5007369686313973	170
Affliction makes	0.5008472380534792	170
little provident	0.5021682442821844	170
ind <	0.5026132435321512	170
page 88	0.5032098954527111	170
Book II	0.5036524357700779	170

heb-all [2.4 GB] [252,069,581] – 27 minutes

How did we run this?

java -jar ElasticMapReduceRunner.jar 0.5 0.2 heb 1 s3://datasets.elasticmapreduce/ngrams/books/20090715/heb-all/2gram/data

Good Examples

bigram	npmi	decade
תרתי משמע	1.0	169
תכנון ופיתוח	1.0	169
תופעת לוואי	1.0	169
שנוי במחלוקת	1.0	169
רחב ידיים	1.0	169
בסיס צבאי	1.0	169
קרית מוצקין	0.9274667477478143	169
תשעה באב	1.0	167

Bad Examples

bigram	npmi	decade
שונים •	0.5016701027299401	169
מספר מטבעות	0.5043882649911482	169
בתחום צור	0.5034743433450243	169
נשאר בלתי	0.5081259252861062	169
' כיבוש	0.5100460408704407	169
169 10	0.5160272895369052	169
שנת 1952	0.5160272895369052	169

Why wrong collocations were extracted?

- Small dataset of stop words – we didn't filter enough.

- The value we ran the map reduces with wasn't high enough; therefore bigrams with value close to 0.5 were extracted.

- We didn't filter special characters like #, @, $, etc.