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

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

Bad Examples

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

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

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

Bad Examples

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

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.