**DSP181 – Final Project: Semantic Relations Analysis**

Ron Yehuda, Tamir Eyal

Program Flow:

Step1: Aggregating 5Grams (MapReduce)

Step1.5: Filtering 5Grams that contain blessed words

Step2: Word count on 1gram file (MapReduce)

Step2.5: Aggregating 2Grams (for PMI calculation, MapReduce)

Step3: Marking HFW, Hook, CW according to the parameters (MapReduce)

Step4: Creating two lists. First, contains a list of targets and patterns by hookword and Second, a list of hookswords by pattern. (Map Reduce)

Step4.5: Sorting by PMI score (optional) (MapReduce)  
 High PMI score means that the probability of two words co-occurrence   
 is similar to the probability of each word individually.  
 For example, Hong Kong.  
 Low PMI score means that the wourd couple is less likely to have a  
 co-occurrence even though they appear in great number individually.  
  
All the above steps are implemented using MapReduce and require no memory space assumptions except for the PMI calculation we had to consider the size of 1gram and 2gram files which were loaded into memory. The list of patterns and target by hook words was also kept in memory.

Step5:   
1. Creating clusters built of Hook word and its patterns per target.  
2. For each cluster (hookword), merging pattern lists that share more than 2/3  
  
Step6 (Section 3.4 in the article):  
1.Removing pattern that appear in one corpus (hook word) only  
2. Marking all clusters as unconfirmed.  
3. Find minimal cluster (from unconfirmed) and merge or delete it.

Step7 – Calculate Hits: (MapReduce)  
For every 5gram, if the second and the fourth words are blessed words, get hits vector. The output of this step is a list of hits vector for every blessed words couple. The hits calculation is based on the formula provided in the article.

Step8:  
Creating an ARFF outputfile compatible with the Weka application. This file contains all the attribute declerations and vector data for classification.

Parameters used in the process:

totalNumberOfWords = 301187285413  
Fc = ((totalNumberOfWords / 1000000) \* 100) \* 50  
Fb = ((totalNumberOfWords / 1000000))  
Fh = ((totalNumberOfWords / 1000000) \* 5)  
N = 1000 (Num of Hooks)  
S = 2 / 3 (minimum shared percent before merge)  
L = 1 / 3 (for PMI, dropped percent for target)  
ALPHA = 1;

Statistics:

Step1:

Map-Reduce Framework

Map input records=247491358

Map output records=4526

Map output bytes=142941

Map output materialized bytes=122223

Input split bytes=13452

Combine input records=0

Combine output records=0

Reduce input groups=4035

Reduce shuffle bytes=122223

Reduce input records=4526

Reduce output records=4035

Step2:

Map-Reduce Framework

Map input records=472764897

Map output records=396499851

Map output bytes=6646705702

Map output materialized bytes=1287150417

Input split bytes=21235

Combine input records=0

Combine output records=0

Reduce input groups=4423745

Reduce shuffle bytes=1287150417

Reduce input records=396499851

Reduce output records=4423745

Step3:

Map-Reduce Framework

Map input records=4423745

Map output records=5400

Map output bytes=84722

Map output materialized bytes=61682

Input split bytes=2204

Combine input records=0

Combine output records=0

Reduce input groups=5400

Reduce shuffle bytes=61682

Reduce input records=5400

Reduce output records=0

Step4:

Map-Reduce Framework

Map input records=4035

Map output records=2776

Map output bytes=69123

Map output materialized bytes=62587

Input split bytes=2318

Combine input records=0

Combine output records=0

Reduce input groups=1105

Reduce shuffle bytes=62587

Reduce input records=2776

Reduce output records=0

Step4 (For PMI):

Map-Reduce Framework

Map input records=24492478978

Map output records=8994424127

Map output bytes=307870340962

Map output materialized bytes=32416515682

Input split bytes=971056

Combine input records=0

Combine output records=0

Reduce input groups=206610607

Reduce shuffle bytes=32416515682

Reduce input records=8994424127

Reduce output records=41322113

Step 7:

Map-Reduce Framework

Map input records=247491358

Map output records=2583

Map output bytes=78792274

Map output materialized bytes=3795808

Input split bytes=13224

Combine input records=0

Combine output records=0

Reduce input groups=249

Reduce shuffle bytes=3795808

Reduce input records=2583

Reduce output records=249