**Knowledge-base for Word Prediction**

Ohad Abramovitz: 307880112

Ori Suchy: 204051114

The task:

Map-Reduce system for calculating the conditional probability of each trigram found in a given corpus.

This system input corpus is Hebrew 3-Gram dataset of Google Books Ngrams.

The output of the system is a list of word trigrams and their conditional probabilities.

The list is ordered:

1. By ascending
2. By the probability , descending.

The probabilities are calculated as suggested by [Thede & Harper](https://dl.acm.org/doi/10.3115/1034678.1034712):

Where:

* is the number of times occurs.
* is the number of times sequence ) occurs.
* is the number of times sequence occurs.
* is the total number of word instances in the corpus.
* is the number of times occurs.
* is the number of times sequence occurs.

Output file: [link](https://oo-dspsp-ass2.s3.amazonaws.com/output4/part-r-00000)

Description

The system is divided to 4 different Map-Reduce steps, the first 3 steps calculate the sums of the probability equation, the last step is for sorting the results:

1. Step1 – First sum:

in this step we arrange the input for the reducer such that every w1w2 pair appear right before all of the triplets that start with w1w2 so we can count the pair's occurrences one time and use it with all of the related triplets.

Input: 3-Gram dataset

* + Map:
    - Parse each of the input's records
    - Write to context:
  + Combiner:

Local aggregation of:

* + Reducer:
    - Count
    - Count s
    - For each Count s related to it
    - Calculate for each triplet
    - Write to context:
      * (all the pairs)

Output – "output1"

1. Step2 – Second sum:

In this step we arrange the input for the reducer such that triplets are sorted by their 2 last words and this way we can make <w2, counts> and <w2w3, counts> appear just before them and use the counts to calculate the second addend.

Input: "output1"

* + Map:
    - Parse each of the input's records
    - Write to context:
  + Combiner:

Local aggregation of:

* + Reducer:
    - Get
    - Count s
    - For each Count s that start with it to it
    - Calculate for each pair
    - Write to context:
      * (all singles)

Output – "output2"

1. Step3 – Third sum:

In this step we arrange the input for the reducer such that triplets are sorted by their last word and this way we can make <w3, counts> appear just before them and use the counts to calculate the third addend.

Input: "output2"

* + Map:
    - Parse each of the input's records
    - Write to context:
  + Combiner:
    - No need
  + Reducer:
    - Get
    - Get for each
    - Get for each pair
    - Get for each triplet
    - Write to context:

Output – "output3"

1. Step4 – Sort:

Input: "output3"

* + Map:
    - Parse each of the input's records
    - Write to context (*value2key*):
      * <w1\_w2\_w3\_
  + Combiner:
    - No need
  + Sort:

Sort function- CompareProbs():

* + - Sort the elements to be ordered by ascending and by probability for descending.
  + Reducer:
    - Parse each of the input's keys
    - Write to context:

Output – "output4"

Statistics:

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Local**  **Aggregation** | **#Records** | **Size** |
| Step 1 | With | 8,046,830 | 0.16 GB |
| Without | 277,762,412 | 5.64 GB |
| Step 2 | With | 2,676,065 | 0.07 GB |
| Without | 7,470,120 | 0.21 GB |
| Step 3 | No Combiner | 1,747,596 | 0.15 GB |
| Step 4 | No Combiner | 1,650,268 | 0.07 GB |

Analysis:

אבא היה אומר 0.08390877619029964

אבא היה אדם 0.043354840408451945

אבא היה איש 0.04188389455682147

אבא היה אז 0.025505551572616345

אבא היה בעל 0.024172061219457572

אמא שלי היתה 0.06299522116278222

אמא שלי לא 0.0583776198914247

אמא שלי אמרה 0.0351185598575306

אמא שלי אומרת 0.021796749948546277

אמא שלי סיפרה 0.016765299268192323

כמו שאני רואה 0.05519024699243442

כמו שאני אוהב 0.0530664370246349

כמו שאני עתיד 0.052370806982896895

כמו שאני רוצה 0.051247772975093885

כמו שאני עושה 0.04554411685421052

לאחר הקמת המדינה 0.4988759630881891

לאחר הקמת מדינת 0.09101880734260945

לאחר הקמת בית 0.01735615019190387

לאחר הקמת המשכן 0.012040858895309073

לאחר הקמת הממשלה 0.01148229377462036

בא לי בירושה 0.030442840836153277

בא לי להקיא 0.028712101445171924

בא לי לבכות 0.028424830109145798

בא לי כל 0.027796929294451878

בא לי רעיון 0.027613385642938543

אין עוד מלבדו 0.09012695843471483

אין עוד מקום 0.05887202426876536

אין עוד צורך 0.03948289838575427

אין עוד כל 0.03935233041172316

אין עוד שום 0.0223238975698425

היום יום ראשון 0.06126977270826241

היום יום של 0.05967458222835546

היום יום שני 0.033294472279010535

היום יום אחד 0.0288950056631746

היום יום טוב 0.027126088440835826

זה מה שאני 0.05533620592576643

זה מה שהוא 0.03319555335174059

זה מה שיש 0.03039975964427078

זה מה שהיה 0.027057470569441848

זה מה שאתה 0.025793019415178815

גם אני לא 0.07809195775074985

גם אני הייתי 0.05088855601492911

גם אני רוצה 0.025771958835374593

גם אני את 0.018953113252328672

גם אני יודע 0.009692755564576914

היו זמנים שבהם 0.2197534642028255

היו זמנים בהם 0.12343414837223808

היו זמנים של 0.09958415680414125

היו זמנים אחרים 0.08115146275446616

היו זמנים קשים 0.07709960353658679