Number of key values without local aggregation-
Runtime- ***18 minutes, 25 seconds***

First Step Mapper	**First Step Reducer**
Map input records=252069581 Map output records=235217148 Map output bytes=5217250979	Reduce input groups=28211634 Reduce input records=235217148 Reduce output records=26608679
Second Step Mapper	**Second Step Reducer**
Map input records=26608679 Map output records=53217358 Map output bytes=1328603153	Reduce input groups=28378043 Reduce input records=53217358 Reduce output records=26608679
Third Step Mapper	**Third Step Reducer**
Map input records=26608679 Map output records=24861465 Map output bytes=1200427928	Reduce input groups=21503650 Reduce input records=24861465 Reduce output records=21503650

Number of key values with local aggregation-
Runtime- ***18 minutes, 9 seconds***

First Step Mapper	**First Step Combiner**	**First Step Reducer**
Map input records=252069581 Map output records=235217148 Map output bytes=5217250979	Combine input records=235217148 Combine output records=38254594	Reduce input groups=28211634 Reduce input records=38254594 Reduce output records=26608679
Second Step Mapper	**Second Step Combiner**	**Second Step Reducer**
Map input records=26608679 Map output records=53217358 Map output bytes=1328603153	Combine input records=53217358 Combine output records=28798168	Reduce input groups=28378043 Reduce input records=28798168 Reduce output records=26608679
Third Step Mapper	**Third Step Combiner**	**Third Step Reducer**
Map input records=26608679 Map output records=24861465 Map output bytes=1200427928	**No need for combiner as each key is unique**	Reduce input groups=21503650 Reduce shuffle bytes=717039323 Reduce input records=24861465 Reduce output records=21503650

Scalability report :

The time running for two different input sizes

- 100% of the 2-Grams Hebrew input - 18 minutes, 9 seconds
- ~50% of the 2-Grams Hebrew input[with local aggregation] – 15 minutes, 32 seconds, we ran on approximately half the input by using the Math.random() function and checking if it's bigger then 0.5[in the first step mapper.

The time running for two different number of input splits:

- 77 input splits on 2-Grams Hebrew input- 18 minutes, 25 seconds
- 289 input splits on 2-Grams Hebrew input- 20 minutes, 40 seconds

5 good collocations:

1890 כהנה וכהנה 0.9908603126152467

1990 פשוטו כמשמעו 0.958266903239367

1760 בית דין 0.7454060579520386

1820 הפלא ופלא 0.9903730572152806

2000 גוג ומגוג 0.9880370691028305

5 bad collocations:

1790 העין שולטת 0.9470336894327951

1850 טבעתי ביון 0.9878200441886431

1890 בעצב תלדי 0.9955003730980672

1980 סונים ושיעים 0.9946102368027737

1670 ידבר עמנו 0.9370750925625704