Preliminary Results

Cédric Milinaire

Summer period 2019

This document is used to summarize current results.

List of Figures

List of Tables

This document is used to summarize current results.

1 Sampling

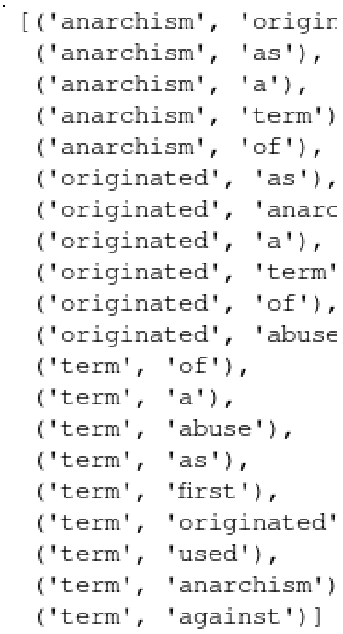
For Table 2 the text8 dataset was used, all words that occured less than 5 times were deleted from the dataset before each sampling technique.

Table 1: Statistics about dataset with different sampling techniques

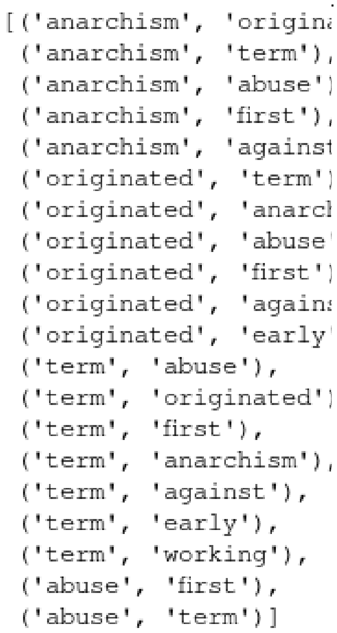
	w/o sampling	**Online Sampling**	**Mikolov**	**w/o outliers**
Min	34	34	34	7
Max	9 Mio.	0.36 Mio.	0.36 Mio	590
QTR1	79	79	79	30
Median	166	166	166	54
QTR3	533	533	534	123
QTR3 + 1.5IQR	1214	1214	1217	227
Mean	2227	1125	1125	100

Table 2: Statistics about dataset with different sampling techniques

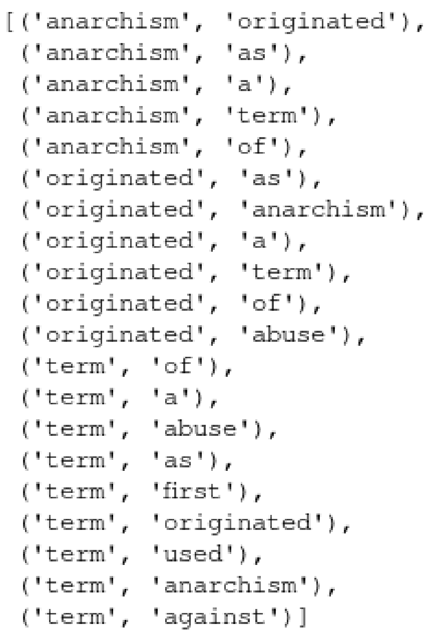
	w/o sampling	**Online Sampling**	**Mikolov**	**w/o outliers**
Size of Ds (in words)	17 Mio.	17 Mio.	8 Mio.	2.8 Mio.
Number of Pairs	141 Mio.	71 Mio.	71. Mio	6 Mio.
Number of Sentences	0.8 Mio	0.8 Mio	0.4 Mio	0.1 Mio.
Vocabulary Size	0.25 Mio	0.25 Mio	0.25 Mio	0.06 Mio.



```
[('anarchism', 'origir
 ('anarchism', 'as'),
 ('anarchism', 'a'),
 ('anarchism', 'term')
 ('anarchism', 'of'),
 ('originated', 'as'),
 ('originated', 'anarc
 ('originated', 'a'),
 ('originated', 'term'
 ('originated', 'of'),
 ('originated', 'abuse
 ('term', 'of'),
 ('term', 'a'),
 ('term', 'abuse'),
 ('term', 'as'),
 ('term', 'first'),
 ('term', 'originated'
 ('term', 'used'),
 ('term', 'anarchism')
 ('term', 'against')]
```



```
[('anarchism', 'origina
 ('anarchism', 'term'),
 ('anarchism', 'abuse')
 ('anarchism', 'first'),
 ('anarchism', 'against
 ('originated', 'term')
 ('originated', 'anarch
 ('originated', 'abuse'
 ('originated', 'first')
 ('originated', 'agains
 ('originated', 'early'
 ('term', 'abuse'),
 ('term', 'originated')
 ('term', 'first'),
 ('term', 'anarchism'),
 ('term', 'against'),
 ('term', 'early'),
 ('term', 'working'),
 ('abuse', 'first'),
 ('abuse', 'term')]
```



```
[('anarchism', 'originated'),
 ('anarchism', 'as'),
 ('anarchism', 'a'),
 ('anarchism', 'term'),
 ('anarchism', 'of'),
 ('originated', 'as'),
 ('originated', 'anarchism'),
 ('originated', 'a'),
 ('originated', 'term'),
 ('originated', 'of'),
 ('originated', 'abuse'),
 ('term', 'of'),
 ('term', 'a'),
 ('term', 'abuse'),
 ('term', 'as'),
 ('term', 'first'),
 ('term', 'originated'),
 ('term', 'used'),
 ('term', 'anarchism'),
 ('term', 'against')]
```

Figure 1: Online Sampling Figure 2: Preprocessing S. Figure 3: No Sampling

2 Boxplots

Figure 4: Distribution of the number of pairs per context word without sampling

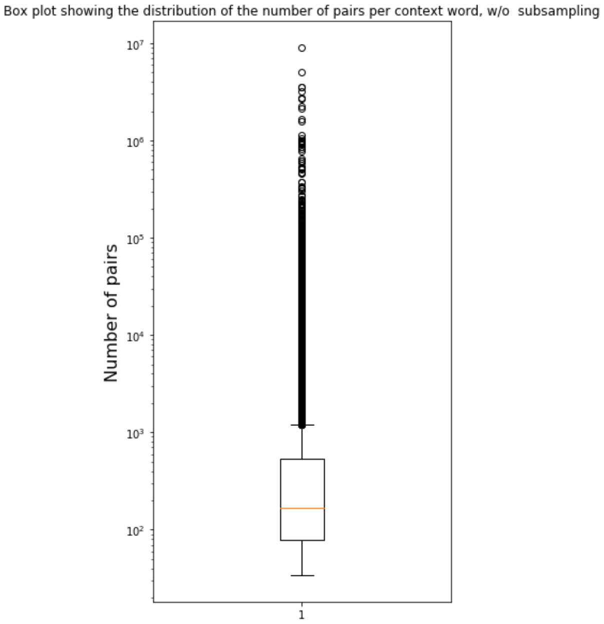


Figure 5: Distribution of the number of pairs per context word with online sampling

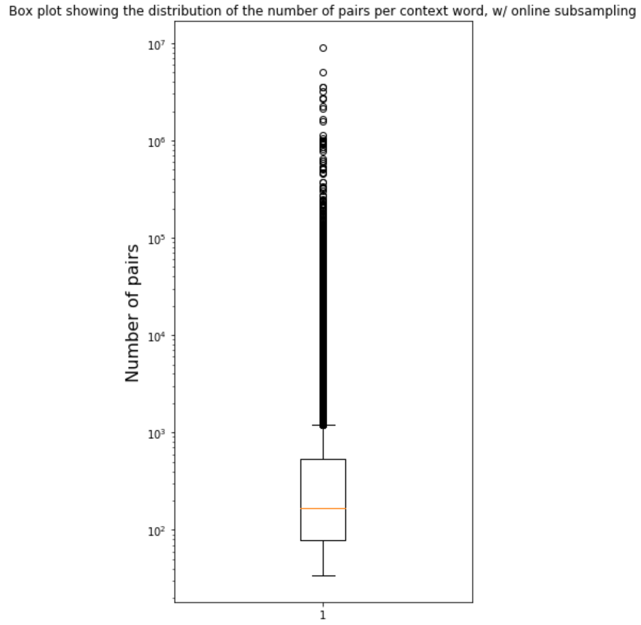
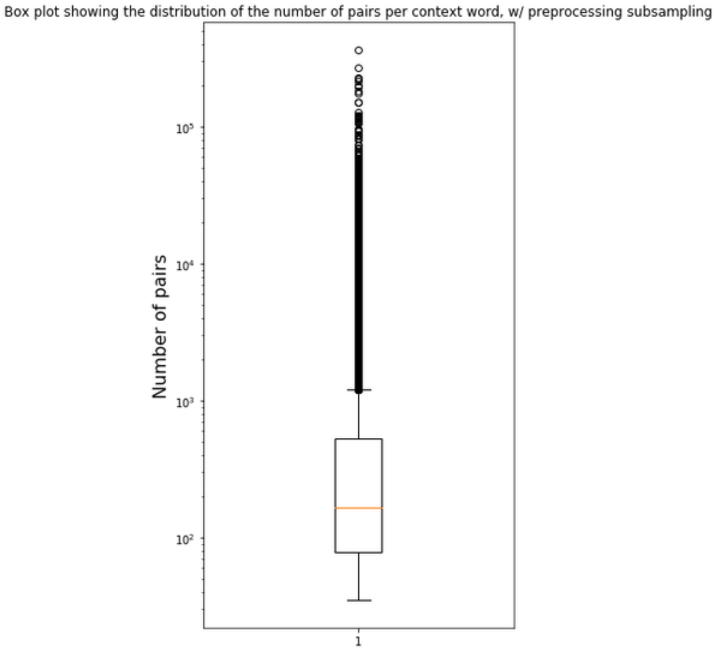


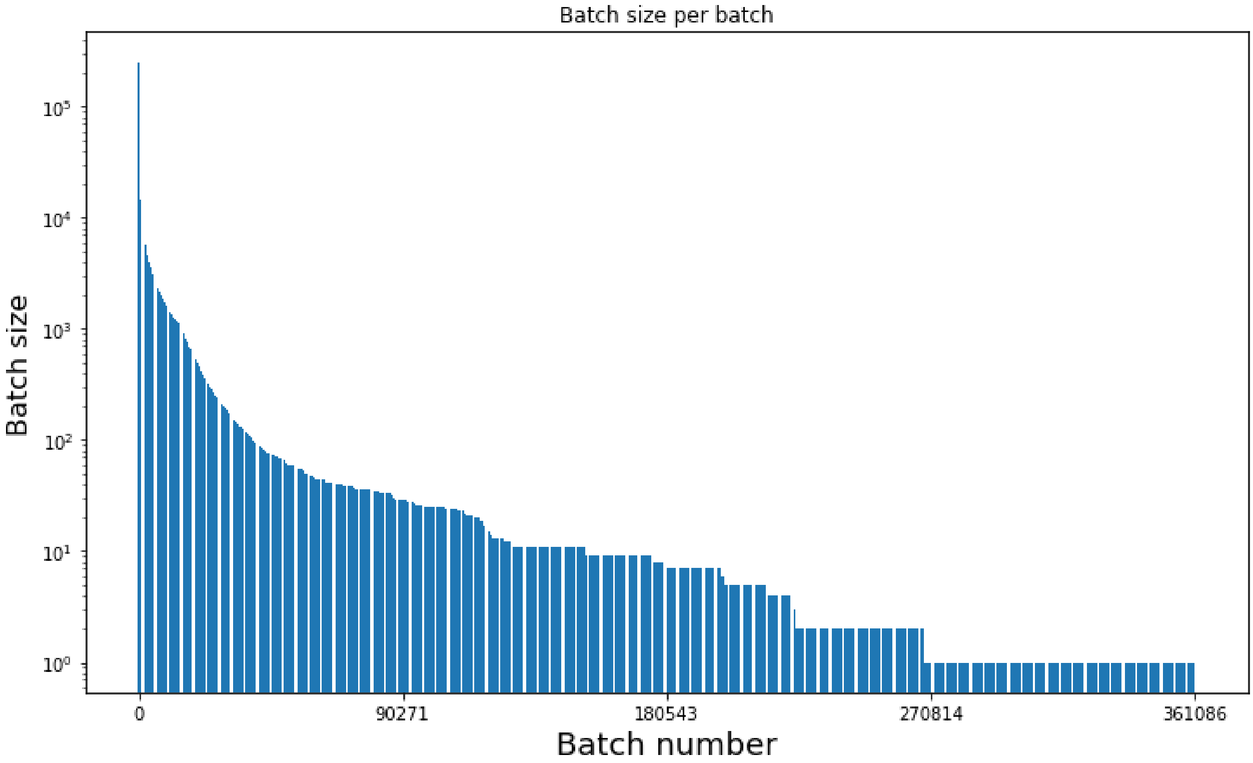
Figure 6: Distribution of the number of pairs per context word with preprocessing sampling



3 Batch_size

This section shows a plot of the batch size per batch.

Figure 7: Batch size per batch



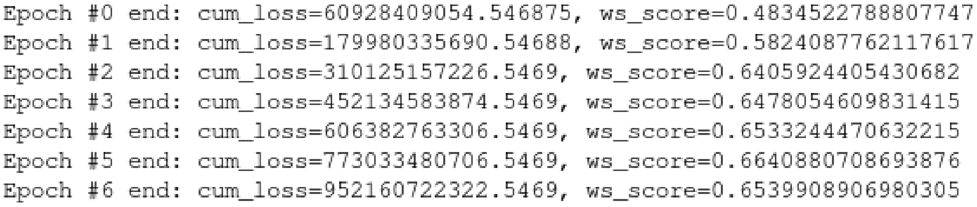
4 Results

Figure 9 shows the results of the training of the text8 dataset without outliers. Those results are to be taken very lightly for two reasons:

1. The data set is very small, i.e 60k vocabulary
2. To assess word similarity only 12 words from the data set were taken. As those were the only one that are in the data set without outliers and in the wordsim dataset.

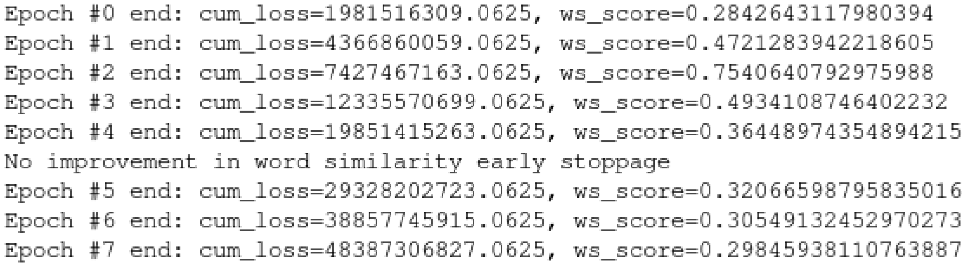
Figure shows that the deletion of the specific word “the“ does not hinder performance.

Figure 8: Results text8 without the word “the“



```
Epoch #0 end: cum_loss=60928409054.546875, ws_score=0.4834522788807747
Epoch #1 end: cum_loss=179980335690.54688, ws_score=0.5824087762117617
Epoch #2 end: cum_loss=310125157226.5469, ws_score=0.6405924405430682
Epoch #3 end: cum_loss=452134583874.5469, ws_score=0.6478054609831415
Epoch #4 end: cum_loss=606382763306.5469, ws_score=0.6533244470632215
Epoch #5 end: cum_loss=773033480706.5469, ws_score=0.6640880708693876
Epoch #6 end: cum_loss=952160722322.5469, ws_score=0.6539908906980305
```

Figure 9: Results text8 without outliers



```
Epoch #0 end: cum_loss=1981516309.0625, ws_score=0.2842643117980394
Epoch #1 end: cum_loss=4366860059.0625, ws_score=0.4721283942218605
Epoch #2 end: cum_loss=7427467163.0625, ws_score=0.7540640792975988
Epoch #3 end: cum_loss=12335570699.0625, ws_score=0.4934108746402232
Epoch #4 end: cum_loss=19851415263.0625, ws_score=0.36448974354894215
No improvement in word similarity early stoppage
Epoch #5 end: cum_loss=29328202723.0625, ws_score=0.32066598795835016
Epoch #6 end: cum_loss=38857745915.0625, ws_score=0.30549132452970273
Epoch #7 end: cum_loss=48387306827.0625, ws_score=0.29845938110763887
```