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	
Max	9Mio	0.36 Mio.	0.36 Mio	
QTR1	79	79	79	
Median	166	166	166	
Mean	2227	1125	1125	
QTR3	533	533	534	
QTR3 + 1.5IQR	1214	1214	1217	

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.	
Number of Pairs	141 Mio.	71 Mio.	71. Mio	
Number of Sentences	0.8 Mio	0.8 Mio	0.4 Mio	

2 Boxplots

Figure 1: Distribution of the number of pairs per context word without sampling

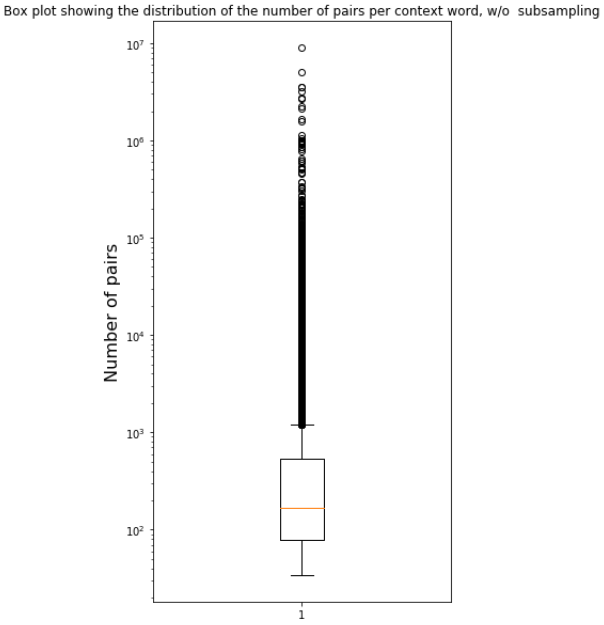


Figure 2: Distribution of the number of pairs per context word with online sampling

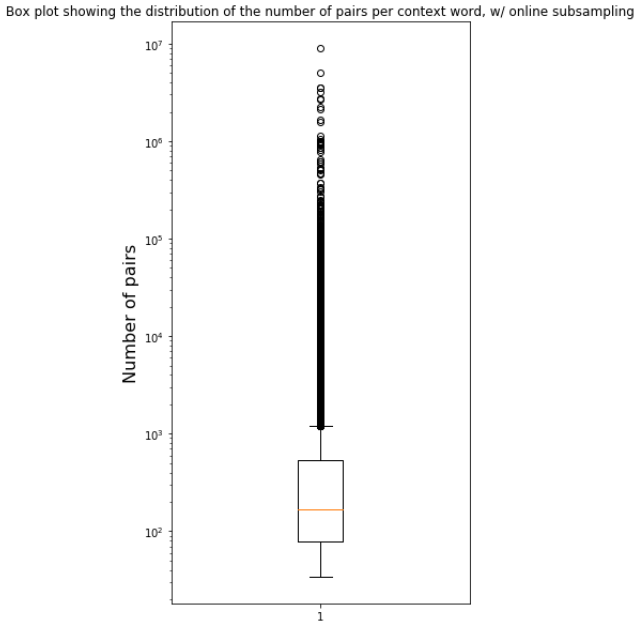
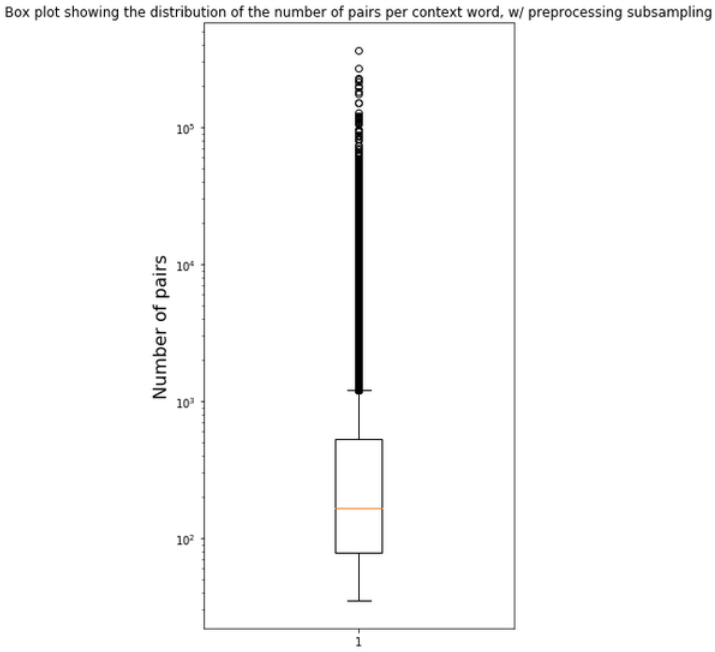


Figure 3: Distribution of the number of pairs per context word with preprocessing sampling



3 Batch_size

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

Figure 4: Batch size per batch