D7041E - Lab2

Sergio Serrano Hérnandez (serser-1), Nicolas Scheidler (nicsch-3)

November 2023

1 Introduction

For this lab, I worked with the different python scripts and adapted them to be able to run several time. The two methods studied in this lab are Word2Vec and RandomIndexing.

2 Word2Vec

I chose to run the code with the following $vector_{size}$: 10, 50, 100, 500, 1000.

I also decided to train a different number of epochs ranging from 1 to 6. To be able to analyse and compare the models, I save every model on my computer and after having trained all of them, I analyzed their score on the TOEFL-test.

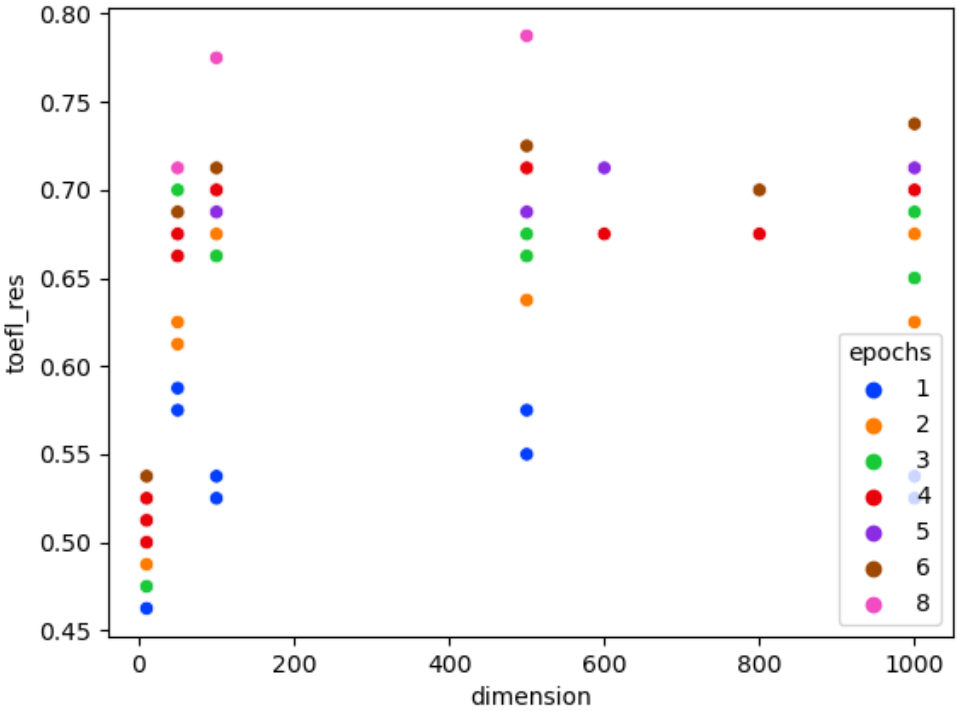


Figure 1: Results on TOEFL-test depending on dimension

What we can see here is that the dimension affects the result on the TOEFL test, as well as the epochs. With We can also see that the results seems to converge when reaching a $vector_{size}$ of hundred and gets slowly better.

What is interesting is also to analyze the graph of the TOEFL result depending on the training time. On figure 2, we notice the tendency of a model to be better the more time it has trained. This is logical because it means that the model has been several times through the data and has therefore learned better.

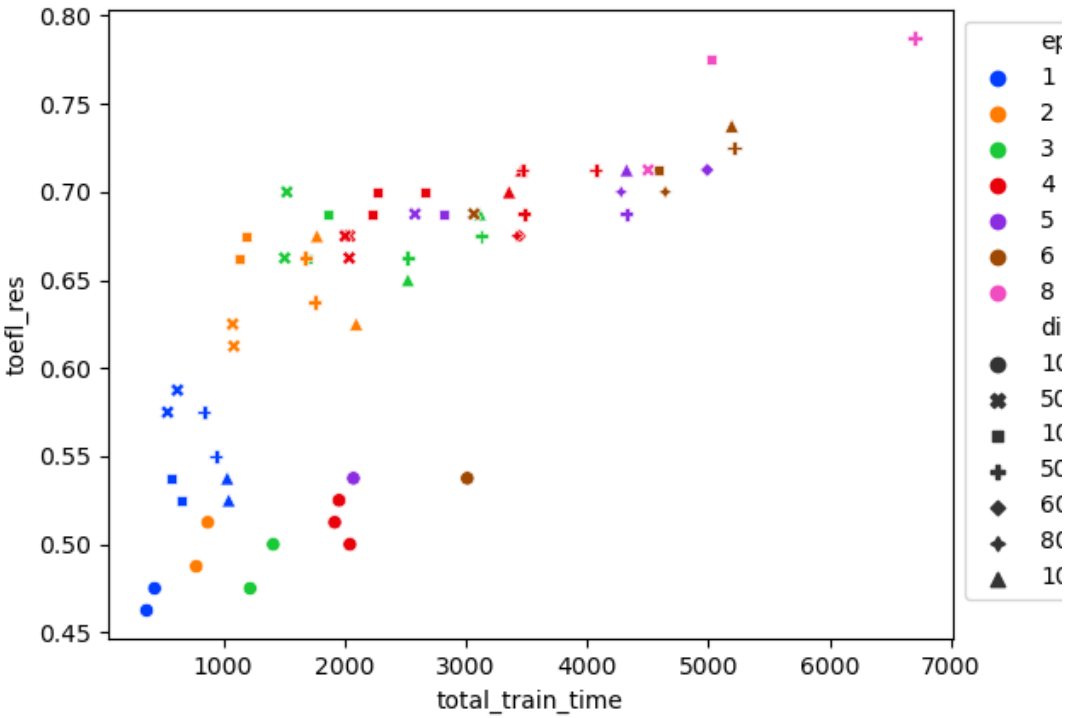


Figure 2: Results on TOEFL-test depending on training time

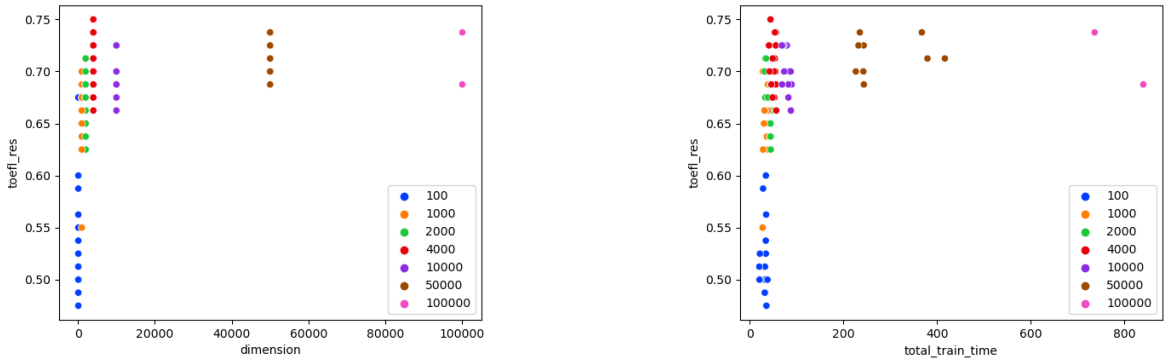
The complete results in table format can be found in the appendix. The best result is obtained with an dimension of 500 and 8 epochs. Unfortunately, I couldn't train all models with 8 epochs but what we can clearly see is that the accuracy grows with the epochs, it would be interesting to know at what point the model accuracy would begin to decrease.

3 RandomIndexing

In order to change the $window_{size}$, it is also necessary to adapt the code to other possibilities: when the word is at the beginning or end of the sentence, the neighboring words have to be found in other sentences.

By changing the $window_{size}$, it is also necessary to consider more sentences, since some sentences may not contain enough words.

After running the algorithm several times, we get the following graphs:



(a) TOEFL result depending on dimension

(b) TOEFL result depending on training time

Figure 3: TOEFL results using the RI models

Here we can see that a too low dimension is bad for the results on TOEFL-test, but augmenting dimensions over 4000 isn't really useful to have a better result, it just takes more time to train the model.

From this experience, we can conclude that the best dimension is around 4000. A summary containing the mean of the values can be found below, while the whole data will be made available in the appendix.

dimension	toefl_res	total_train_time
100	0.525000	31.342840
1000	0.657500	36.881264
2000	0.672917	41.226114
4000	0.704688	51.179313
10000	0.699107	80.235851
50000	0.715278	287.312040
100000	0.712500	787.950657

Table 1: Mean of the results grouped by dimension

4 Conclusion

To conclude, accuracy increases with the dimension but after a while, the accuracy stays stable. Also, the bigger the dimension, the longer the models need to be trained.

5 Appendix

Table 2: Results of the different trainings of Word2Vec models.

	dimension	epochs	total train time	toefl res	skipped tests
0	1000	1	1032.577860	0.537500	0
1	1000	1	1044.769646	0.525000	0
2	1000	2	1772.395053	0.675000	0
3	1000	2	2096.109997	0.625000	0
4	1000	3	2522.249426	0.650000	0
5	1000	3	3115.959609	0.687500	0
6	1000	4	3356.906780	0.700000	0
7	1000	4	3459.071747	0.712500	0
8	1000	5	4326.146627	0.712500	0
9	1000	6	5193.009208	0.737500	0
10	100	1	576.046493	0.537500	0
11	100	1	652.010240	0.525000	0
12	100	2	1136.760010	0.662500	0
13	100	2	1186.479172	0.675000	0
14	100	3	1698.054969	0.662500	0
15	100	3	1861.998322	0.687500	0
16	100	4	2271.002787	0.700000	0
17	100	4	2658.985630	0.700000	0
18	100	4	2230.254519	0.687500	0
19	100	5	2812.668609	0.687500	0
20	100	6	4585.861817	0.712500	0
21	100	8	5023.187048	0.775000	0
22	10	1	433.771569	0.475000	0
23	10	1	368.043657	0.462500	0
24	10	2	872.277663	0.512500	0
25	10	2	777.471842	0.487500	0
26	10	3	1411.912715	0.500000	0
27	10	3	1221.364882	0.475000	0
28	10	4	1919.099913	0.512500	0
29	10	4	1954.635626	0.525000	0
30	10	4	2043.834599	0.500000	0
31	10	5	2073.458292	0.537500	0
32	10	6	3011.976808	0.537500	0
33	500	1	840.998090	0.575000	0
34	500	1	938.567411	0.550000	0
35	500	2	1678.704688	0.662500	0
36	500	2	1755.316113	0.637500	0
37	500	3	2520.157741	0.662500	0
38	500	3	3126.127027	0.675000	0
39	500	4	3467.547649	0.712500	0
40	500	4	4071.236768	0.712500	0
41	500	4	3480.921776	0.687500	0
42	500	5	4322.864746	0.687500	0
43	500	6	5214.892884	0.725000	0
44	500	8	6700.719322	0.787500	0
45	50	1	540.288649	0.575000	0
46	50	1	621.576953	0.587500	0
47	50	2	1077.561187	0.625000	0
48	50	2	1087.289227	0.612500	0
49	50	3	1524.317160	0.700000	0
50	50	3	1506.328984	0.662500	0
51	50	4	2038.023572	0.662500	0
52	50	4	2040.829581	0.675000	0
53	50	4	2005.078184	0.675000	0

	dimension	epochs	total train time	toefl res	skipped tests
54	50	5	2580.607838	0.687500	0
55	50	6	3066.873204	0.687500	0
56	50	8	4503.627877	0.712500	0
57	600	4	3446.345368	0.675000	0
58	600	5	4990.757913	0.712500	0
59	800	4	3422.670128	0.675000	0
60	800	5	4281.618341	0.700000	0
61	800	6	4645.357568	0.700000	0

Table 3: Results of the different trainings of RandomIndexing models.

	dimension	toefl res	total traintime
0	100	0.500000	34.916443
1	100	0.512500	32.276720
2	100	0.475000	35.671387
3	100	0.500000	34.400108
4	100	0.600000	34.513933
5	100	0.500000	34.931172
6	100	0.500000	34.360076
7	100	0.675000	34.891156
8	1000	0.675000	40.218717
9	1000	0.700000	40.510515
10	1000	0.662500	40.465456
11	1000	0.625000	39.986143
12	1000	0.650000	39.680056
13	1000	0.625000	39.173488
14	1000	0.687500	38.747573
15	2000	0.687500	45.540065
16	2000	0.650000	44.613513
17	2000	0.687500	45.091192
18	2000	0.637500	45.304519
19	2000	0.662500	45.337558
20	2000	0.625000	44.930496
21	4000	0.687500	55.280128
22	4000	0.700000	55.199958
23	4000	0.687500	55.826441
24	4000	0.662500	56.643591
25	4000	0.737500	55.908565
26	10000	0.687500	88.224925
27	10000	0.687500	89.464084
28	10000	0.662500	87.812744
29	10000	0.700000	87.748875
30	100	0.487500	33.113844
31	100	0.500000	32.217875
32	100	0.537500	34.335949
33	100	0.500000	37.960791
34	100	0.562500	35.096313
35	100	0.487500	32.043427
36	100	0.512500	32.475282
37	100	0.525000	34.454965
38	1000	0.637500	36.656542
39	1000	0.637500	43.239447
40	1000	0.675000	41.299093
41	1000	0.700000	39.565771
42	1000	0.700000	39.822233
43	1000	0.662500	39.735162
44	1000	0.662500	39.009960

	dimension	toefl res	total train time
45	2000	0.675000	43.785615
46	2000	0.675000	44.613003
47	2000	0.687500	44.850488
48	2000	0.687500	44.352297
49	2000	0.637500	44.397798
50	2000	0.650000	44.362802
51	4000	0.737500	53.829792
52	4000	0.725000	55.603248
53	4000	0.712500	53.623150
54	4000	0.675000	52.671753
55	4000	0.700000	52.303224
56	10000	0.687500	82.504359
57	10000	0.675000	82.555211
58	10000	0.700000	83.522112
59	10000	0.700000	87.286630
60	50000	0.712500	379.319796
61	50000	0.712500	416.417895
62	50000	0.737500	367.270416
63	100000	0.687500	839.917757
64	100000	0.737500	735.983558
65	100	0.500000	22.370330
66	100	0.512500	20.516488
67	100	0.500000	20.979350
68	100	0.525000	21.739649
69	100	0.550000	27.703306
70	100	0.587500	28.573918
71	1000	0.662500	33.368012
72	1000	0.662500	31.450046
73	1000	0.550000	27.547650
74	1000	0.650000	30.449791
75	1000	0.625000	28.567760
76	1000	0.700000	28.131871
77	2000	0.712500	32.895768
78	2000	0.675000	32.656843
79	2000	0.675000	33.451817
80	2000	0.700000	32.000717
81	2000	0.675000	38.739717
82	2000	0.712500	35.145840
83	4000	0.712500	49.026782
84	4000	0.687500	46.290347
85	4000	0.750000	44.410486
86	4000	0.700000	42.416716
87	4000	0.725000	41.093316
88	4000	0.675000	48.741514
89	10000	0.725000	79.162662
90	10000	0.725000	72.576900
91	10000	0.687500	69.088249
92	10000	0.700000	73.546879
93	10000	0.725000	71.081867
94	10000	0.725000	68.726417
95	50000	0.737500	234.990070
96	50000	0.700000	242.529083
97	50000	0.687500	243.565132
98	50000	0.700000	226.439641
99	50000	0.725000	243.145508
100	50000	0.725000	232.130821