Homework 1 Natural Language Processing

Irene Tallini: 1608460

April 16, 2018

1 W2V

1.1 SkipGram

I implemented SkipGram. This is a two-layer neural network. Let n be the vocabulary size, k the embedding size and d the window size. The first layer is a $n \times k$ lookup table (W) which embeds the input word (v) into a vector $w \in R^k$. The second layer computes $s = softmax(w \times M)$ where M is a $k \times n$ matrix. The model then adjusts W and M components via back-propagation in order to minimize the log likelihood of the model. The ground truth is given by a word in the context of v, i.e. a word at most d positions far from v in the training text data. To reduce training time I implemented negative sampling and kept a word (v) with frequency $f(v)$ with probability $f(v)^{\frac{1}{4}}$. Moreover I worked only with characters [a,...,z] and deleted some stop-words.

1.2 Optimal parameters and general findings

Due to technical problems I couldn't run serious tests. Optimal parameters: [WS=5, ES=150, VS=3000, NS=5, BS=60].
General findings

- WINDOW SIZE. Small WS is faster but has good performance just on words which tend to occur in standard sentences (like months or nationality). This is due to the fact that, for instance, after training the model ($WS = 1$) with the sentence "*author born in June of year...*", "*born*" becomes similar just to other words contained between "*author*" and "*year*". Large WS has better performance but takes much longer to train.
- EMBEDDING SIZE. Small (with respect of vocabulary size) ES doesn't allow for complex configurations. For instance, with $WS = n$ it's impossible to represent $n + 2$ pairwise equidistant words. On the other hand the increase in accuracy given by very large ES is not worth the expense in computational time.
- VOCABULARY SIZE. I trained on 1000, 3000 and 10000 words. $VS = 3000$ gave the best percentage accuracy on analogy reasoning, while 10000 gave the best absolute accuracy. This was due to the fact that, with the same number of steps, the 1000 model got trained mostly on UNK context, the 3000 one was evaluated on fewer questions than the 10000 one but, being a simpler model, at the end of training it was closer to convergence than the last one and so more precise. I chose $VS = 3000$ for final model, but, having more time, a better choice would have been to run a $VS \geq 10000$ model longer.

2 Using Vectors

2.1 Visualization

Since the vocabulary and the number of iterations are so small, accuracy on analogy reasoning is very low (4.5 %). The model behaves decently just for words which are common, not very ambiguous and usually found in standard sentences. The principal examples are months and nationality (see Figure 3.2), which are clustered and the two cluster are close since they largely occur together in Wikipedia sentences like "*Leonardo da Vinci was born in Italy*".

2.2 5 most similar words

german = [french, known, american, october, new], **general** = [national, american, october, november, european], **food** = [raw, treatment, sold, house, covered], **teach** = [station, oxford, meeting, breed, enroll]. **cat** and **eat** are not present.

3 Domain Classification

3.1 K-Nearest-Neighbors

I represented each file as the centroid of his words' embeddings, i.e. each file is an n-ple $c_1, \ldots, c_n$, where $c_i = \frac{1}{n} \sum_{j_1}^{n} w_i^j$ and w_i^j is the i-th component of the j-th word in the file. I then applied 5-nearest-neighbors algorithm.

3.2 Performance

Since the vocabulary is so small, every document is defined by the mean of his most common words, which are the less informative. Thus performance is awful: the average precision is 0.06%, which is very close to random guessing (since $\frac{1}{34} \cong 0.03$).

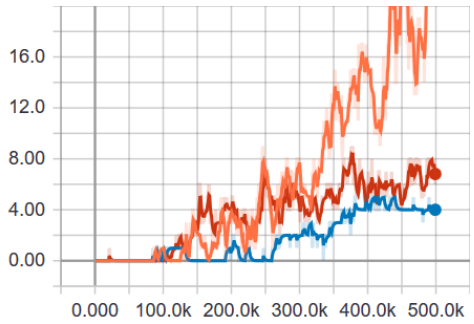

Figure 1: Analogical reasoning accuracy for EMBEDDING SIZE ∈ [50 (blue), 100 (orange), 300 (red)] with other parameters fixed

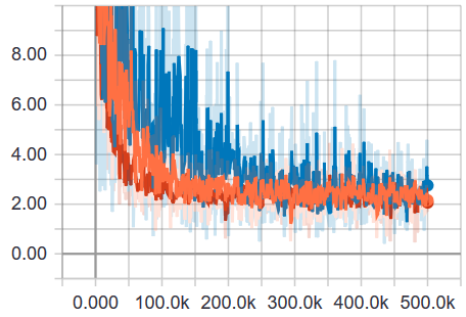

Figure 2: Loss for EMBEDDING SIZE ∈ [50 (blue), 100 (orange), 300 (red)] with other parameters fixed

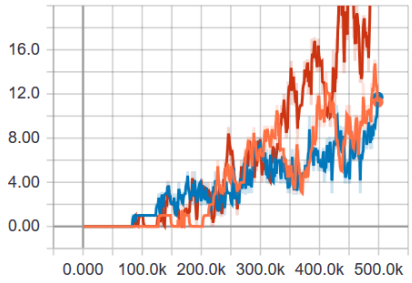

Figure 3: Analogical reasoning accuracy for WS ∈ [1 (blue), 5 (red), 10 (orange)] with other parameters fixed

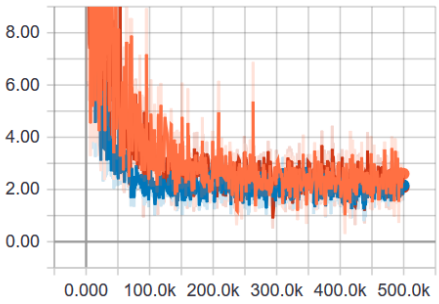

Figure 4: Loss for WS ∈ [1 (blue), 5 (red), 10 (orange)] with other parameters fixed

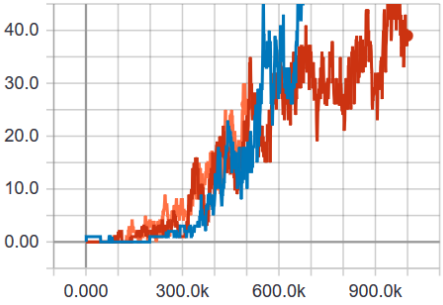

Figure 5: Absolute analogical reasoning accuracy for VOCABULARY SIZE ∈ [1000 (orange), 3000 (red), 10000 (orange)] with other parameters fixed

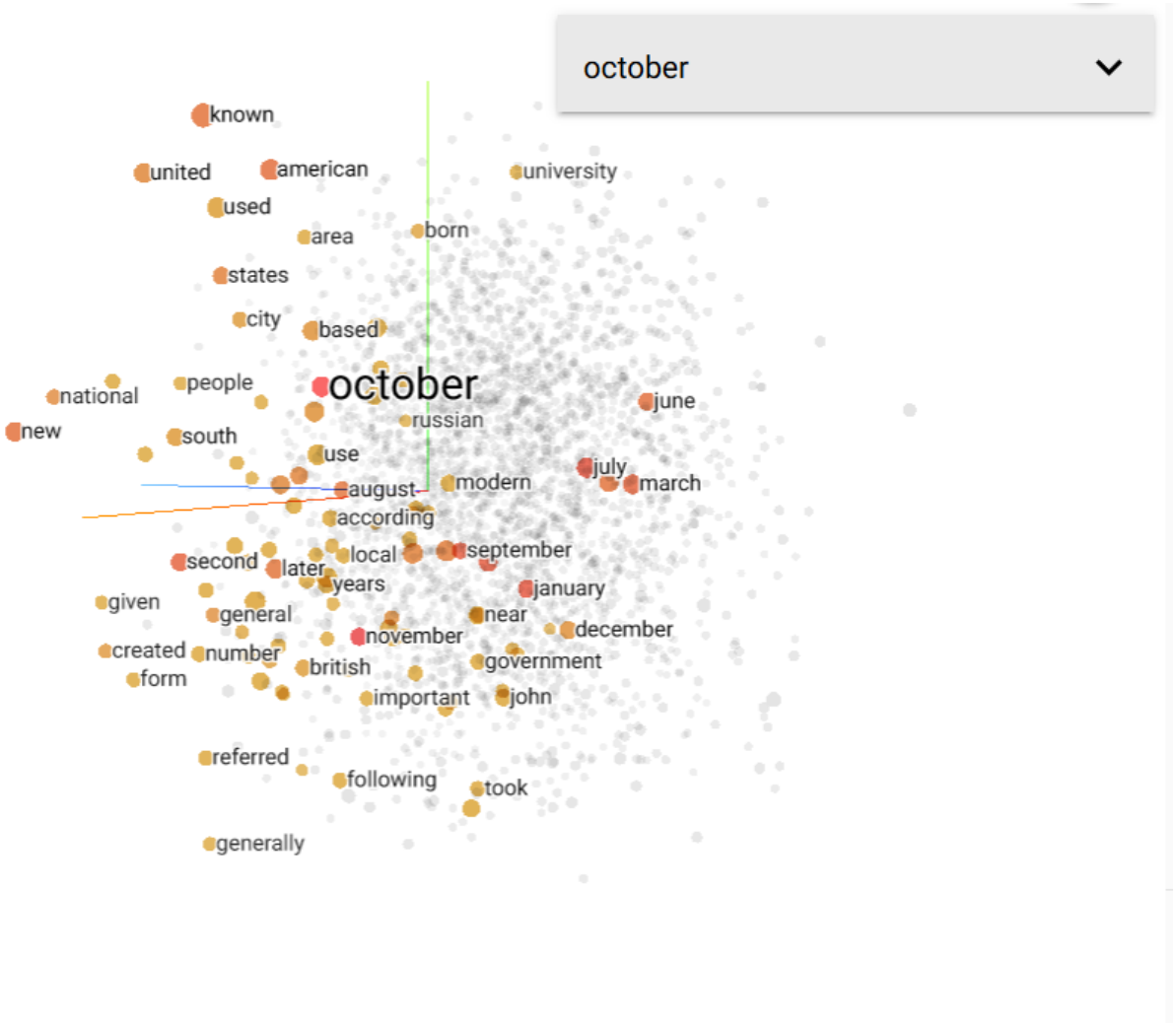

Figure 6: Nearest words to October (the more intense the red the less distant)

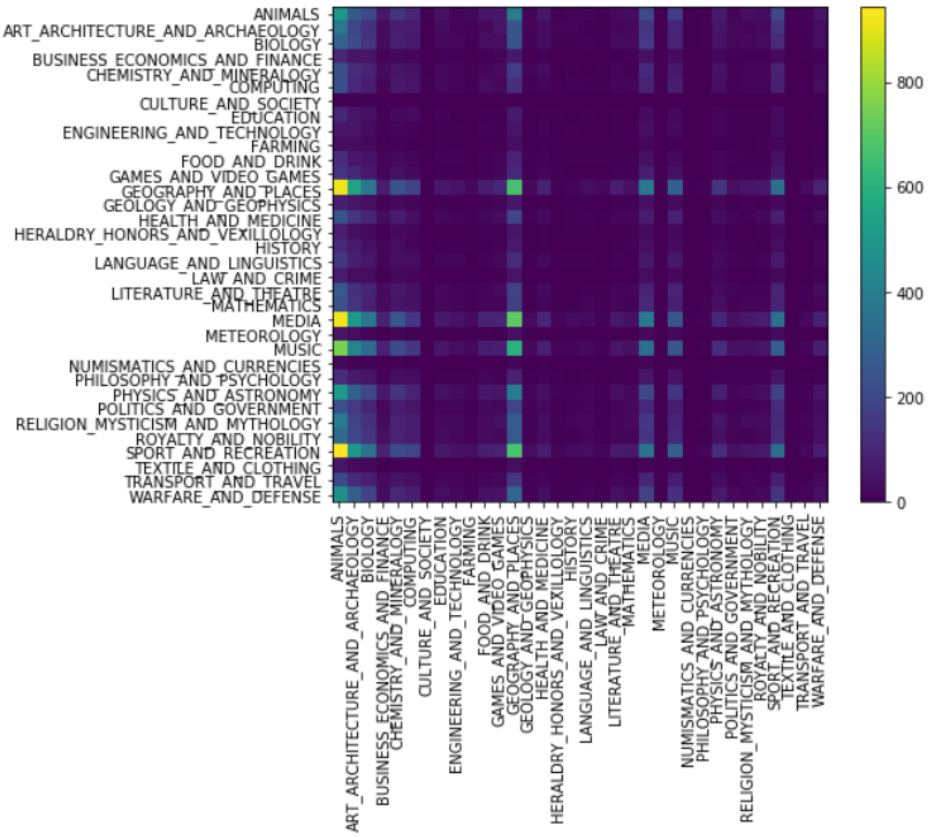

Figure 7: Confusion matrix

	precision	recall	f1-score	support
ANIMALS	0.05	0.18	0.08	2855
ART_ARCHITECTURE_AND_ARCHAEOLOGY	0.04	0.10	0.05	1937
BIOLOGY	0.04	0.08	0.05	1831
BUSINESS_ECONOMICS_AND_FINANCE	0.01	0.02	0.02	518
CHEMISTRY_AND_MINERALOGY	0.02	0.04	0.03	1279
COMPUTING	0.03	0.04	0.03	1205
CULTURE_AND_SOCIETY	0.00	0.00	0.00	39
EDUCATION	0.01	0.02	0.02	493
ENGINEERING_AND_TECHNOLOGY	0.01	0.01	0.01	401
FARMING	0.01	0.01	0.01	231
FOOD_AND_DRINK	0.01	0.01	0.01	600
GAMES_AND_VIDEO_GAMES	0.02	0.02	0.02	827
GEOGRAPHY_AND_PLACES	0.10	0.13	0.11	5051
GEOLOGY_AND_GEOPHYSICS	0.01	0.01	0.01	460
HEALTH_AND_MEDICINE	0.03	0.02	0.02	1365
HERALDRY_HONORS_AND_VEXILLOLOGY	0.01	0.00	0.00	378
HISTORY	0.00	0.00	0.00	545
LANGUAGE_AND_LINGUISTICS	0.02	0.01	0.01	879
LAW_AND_CRIME	0.01	0.00	0.00	360
LITERATURE_AND_THEATRE	0.03	0.01	0.02	1334
MATHEMATICS	0.02	0.01	0.01	1297
MEDIA	0.10	0.08	0.09	5090
METEOROLOGY	0.00	0.00	0.00	280
MUSIC	0.09	0.06	0.07	4350
NUMISMATICS_AND_CURRENCIES	0.08	0.01	0.02	119
PHILOSOPHY_AND_PSYCHOLOGY	0.00	0.00	0.00	688
PHYSICS_AND_ASTRONOMY	0.06	0.03	0.04	2823
POLITICS_AND_GOVERNMENT	0.02	0.01	0.01	1371
RELIGION_MYSTICISM_AND_MYTHOLOGY	0.05	0.01	0.02	1965
ROYALTY_AND_NOBILITY	0.03	0.01	0.02	1872
SPORT_AND_RECREATION	0.10	0.07	0.08	5006
TEXTILE_AND_CLOTHING	0.00	0.00	0.00	252
TRANSPORT_AND_TRAVEL	0.02	0.00	0.01	1148
WARFARE_AND_DEFENSE	0.04	0.02	0.02	2429
avg / total	0.06	0.06	0.05	51278

Figure 8: Precision, recall and f1 measure table