Assignment 1

Deadline: Friday, 15th May (23:59)

May 1, 2020

Question 1 - Keras implementation (10pt):

Build word embeddings with a Keras implementation where the embedding vector is of length 50, 150 and 300. Use the Alice in Wonderland text book (`alice.txt`) for training . Use a window size of 2 to train the embeddings (`window_size` in the jupyter notebook).

1. Build word embeddings of length 50, 150 and 300 using the Skipgram model
2. Build word embeddings of length 50, 150 and 300 using CBOW model
3. Analyze the different word embeddings:
   - Implement your own function to perform the analogy task as explained in [1]. Use the same distance metric as in the paper. Do not use existing libraries for this task such as Gensim. Your function should be able to answer whether an analogy as in example 1 is true.

     $$\begin{aligned} &\text{A king is to a queen as a man is to a woman} \\ &e_{\text{king}} - e_{\text{queen}} + e_{\text{woman}} \approx e_{\text{man}} \end{aligned} \tag{1}$$

     , where e_x denotes the embedding e of word x. We want to find the word p in the vocabulary, where the embedding of p (e_p) is the closest to the predicted embedding (i.e. result of the formula). Then, we can check if p is the same word as the true word ("man" in example 1).
   - Give at least 5 different examples of analogies.
   - Compare the performance on the analogy tasks between the word embeddings and briefly discuss your results.
4. Discuss:
   - Given the same number of sentences as input, CBOW and Skipgram arrange the data into different number of training samples. Which one has more and why?

Question 2 - Peer review (0pt):

Finally, each group member must write a single paragraph outlining their opinion on the work distribution within the group. Did every group member contribute equally? Did you split up tasks in a fair manner, or jointly worked through the exercises? Do you think that some members of your group deserve a different grade from others?

[1] Mikolov, Tomas, et al. "Efficient estimation of word representations in vector space." arXiv preprint arXiv:1301.3781 (2013). (`https://arxiv.org/abs/1301.3781`)