Deep Learning for NLP 2020
Exercise 04

May 12, 2020

1 Pingo

Try to find the right answer(s) to each question on your own or in a group with your colleagues. The interactive survey will be conducted near the end of the practice class.

- In neural networks, the choice between different pretrained embeddings can be considered like a hyperparameter.
  - ☐ True
  - ☐ False
- Which of the following approaches for word representations are extrinsic methods?
  - ☐ Testing as part of named entity recognition
  - ☐ Calculating vector distances for similar words
  - ☐ Evaluation by humans, for example by word analogy
  - ☐ Performance comparison of two word representations trained within the same model
- Which of the following statements on the skip-gram model are correct?
  - ☐ It it used to learn a lower-dimensional representation for words
  - ☐ It attempts to infer the missing word from its context
  - ☐ It attempts to infer the missing context from a word
  - ☐ It is implemented in word2vec
  - ☐ It learns exactly one matrix
- Which of the following statements apply on negative sampling?
  - ☐ Frequent words are omitted in the calculation
  - ☐ Random combinations for contexts are used as negative training examples
  - ☐ It is applied to reduce the number of trainable parameters of the skip-gram model
  - ☐ It is applied because a full computation of all softmax probabilities in skip-gram is infeasible

2 Word Embeddings

1. What is unsupervised pre-training and why is it desirable?
2. Vectorized representations can be trained for many applications and on all kinds of different data. State at least three different kinds of embeddings that one could train.

3 Tokenization

Tokenization is the NLP task of splitting a sentence (or an arbitrary character sequence) into useful parts, called tokens.

1. A fellow student of yours uses `str.split(" ")` to tokenize his/her input sentences, which come from children's novels. How would you convince him/her that this is a bad idea?
2. A neural network of yours should classify sentences. You decide to use pretrained embeddings downloaded from the internet (say, word2vec). How would you preprocess your input sentences in order to reach the best possible performance? Explain in up to three sentences.