A row of matrix $W^{(c)}$ represents

- 1. How relevant each word is for a dimension
- 2. How often a context word c co-occurs with all words
- 3. A representation of word c in concept space

©2023, Karl Aberer, EPFL-IC, Laboratoire de systèmes d'informations répartis

Embedding Models - 1

Answer 1

The rows of the matrix correspond to the dimensions of the embedding. Each entry in the row corresponds to a word from the vocabulary. Therefore the row represents the importance of each word for a given dimension.

Which of the following functions is not equal to the three others?

- 1. f(w, c)
- 2. $f_{\theta}(w,c)$ 3. f(w,c)4. $\sigma(c \cdot w)$

©2023, Karl Aberer, EPFL-IC, Laboratoire de systèmes d'informations répartis

Embedding Models - 2

Answer 1

f(w,c) is the function to be approximated. The three other functions are the same, with a more detailed specification for answers 2, 3 and 4.

From which data samples the embeddings are learnt?

- 1. Known embeddings for (w,c) pairs
- 2. Frequency of occurrences of (w,c) pairs in the document collection
- Approximate probabilities of occurrences of (w,c) pairs
- 4. Presence or absence of (w,c) pairs in the document collection

©2023, Karl Aberer, EPFL-IC, Laboratoire de systèmes d'informations répartis

Embedding Models - 3

Answer 4

In the skipgram model the sample data consists of word-context pairs that are present or absent in the document collection.

With negative sampling a set of negative samples is created for

- 1. For each word of the vocabulary
- 2. For each word-context pair
- 3. For each occurrence of a word in the text
- 4. For each occurrence of a word-context pair in the text

©2023, Karl Aberer, EPFL-IC, Laboratoire de systèmes d'informations répartis

Embedding Models - 4

Answer 4

For each occurrence of a word-context pair, a set of negative samples is produced. Note that this is also different from creating a set of negative samples for each word-context pair, since the same word-context pair can occur multiple times in the document collection.

The loss function is minimized

- 1. By modifying the word embedding vectors
- 2. By changing the sampling strategy for negative samples
- 3. By carefully choosing the positive samples
- 4. By sampling non-frequent word-context pairs more frequently

©2023, Karl Aberer, EPFL-IC, Laboratoire de systèmes d'informations répartis

Embedding Models - 5

Answer 1

The loss function is minimized by incrementally modifying the word embedding vectors. Answer 4 refers to an approach to improve the quality of the word embeddings achieved, but not to minimize the loss function.

A word embedding for given corpus ...

- 1. depends only on the dimension d
- 2. depends on the dimension d and number of iterations in gradient descent
- 3. depends on the dimension d, number of iterations and chosen negative samples
- 4. there are further factors on which it depends

©2023, Karl Aberer, EPFL-IC, Laboratoire de systèmes d'informations répartis

Embedding Models - 6

Answer 4

Other factors that can influence the outcome of the optimization are

- The order in which the documents are processed
- The initialization of the word embedding vectors

Fasttext speeds up learning by

- 1. Considering subwords of words
- 2. By selecting the most frequent phrases in the text as tokens
- 3. By selecting the most frequent subwords in the text as tokens
- 4. By pre-computing frequencies of n-grams

©2023, Karl Aberer, EPFL-IC, Laboratoire de systèmes d'informations répartis

Embedding Models - 7

Answer 3

Answer 1 improves the quality of embeddings, but may slow down learning. Answer 4 speeds up the selection of frequent phrases, but not learning.

The most important difference between Glove and skipgram is

- That Glove considers the complete context of a word
- 2. That Glove computes a global frequency for wordcontext pair occurrences
- 3. That Glove uses a squared error loss function
- 4. That Glove does not differentiate words and context words

©2023, Karl Aberer, EPFL-IC, Laboratoire de systèmes d'informations répartis

Embedding Models - 8

Answer 2

The main difference between Glove and earlier methods, including skipgram and CBOW, is the computation of global co-occurrence counts. This is an additional processing step, but provides additional information on the global statistics. Answer 3 refers to a difference that is rather a consequence of using a global statistics. As for answer 4, in a sense also skipgram does not really distinguish between words and context words.