Sprachverarbeitung & Text Mining
Winter semester 2021/22

Prof. Dr. Andreas Hotho,
M.Sc. Janna Omeliyanenko
Lecture Chair X for Data Science, Universität Würzburg

10. Exercise for “Sprachverarbeitung und Text Mining”

28.01.2022

1 Knowledge Questions

1. What different types of vector models were discussed in the lecture and how do they differ in terms of vectors used?

   - Sparse-Vector models (e.g. Mutual-Information, word co-occurence)
   - Dense-Vector models (SVD, GloVe, Brown Cluster)

   In general, dense-vector models attempt to represent information with minimal dimensions and few zero entries.

2. What are possible applications for *Vector Semantics*?

   - Question Answering
   - Translation
   - Plagiarism detection
   - Summarisation

3. Describe the difference between first order co-occurrence and second order co-occurrence.

   First order (Syntagmatic association): two words often occur next to each other. Second order (Paradigmatic association): two words often occur next to similar context words.

4. List 3 possible reasons why only positive results of Pointwise Mutual Information are generally used.

    - Problems with very rare terms ($p(w_1), p(w_2) << 0$): If two words occur very rarely by themselves, then a very rare co-occurrence ($p(w_1, w_2) < (p(w_1) \cdot p(w_2))$) may only be caused by a too small corpus, and not by actual "unrelatedness"
    - Problems if a term occurs very often: $p(w_1, w_2) << p(w_1)$ for most w_2, thus $PMI(w_1, w_x)$ always negative for each $x \rightarrow$ no information content
    - No study availalbe that shows that humans have good intuition for "unrelatedness". "Unrelatedness" may not exist in human language.

5. Explain the process of a singular value decomposition in your own words.

    Approximate an N-dimensional dataset using fewer dimensions by first rotating the axes into a new space in which the highest order dimension captures the most variance in the original dataset and the next dimension captures the next most variance, etc.

    Many such (related) methods:

    - PCA – principle components analysis
    - Factor Analysis
    - SVD

6. When learning word embeddings through Glove, what information can be learned with small and large window size respectively?

    - Small window size: more syntactic information, since this information can mostly be taken from the immediate context
    - Large window size: semantic information is often not local, and more of this information can be discovered with larger window sizes.

2 Vector Semantics and Word Similarities

1. For the following Brown clustering C, calculate the clustering quality $Quality(C)$ as defined in the lecture.

c_1	c_2	c_1	c_2	c_3
January	and	February	its	cold

$$MutualInformation(C) = \frac{2}{4}\log\frac{\frac{2}{4}}{\frac{2}{5}\cdot\frac{2}{5}} + \frac{1}{4}\log\frac{\frac{1}{4}}{\frac{2}{5}\cdot\frac{2}{5}} + \frac{1}{4}\log\frac{\frac{1}{4}}{\frac{2}{5}\cdot\frac{1}{5}} = 0.4196$$

$$WordEntropy(C) = 0.4\log 0.4 + 3\cdot 0.2\log 0.2 = -0.5786$$

$$Quality(C) = MutualInformation(C) + WordEntropy(C) = -0.159$$

2. Given is the following word/context-word matrix:

word/context-word	**Zoo**	**Steak**	**Mammal**	**Cow**	**Farm**
Elephant	4	0	5	3	1
Snake	6	1	0	0	0
Tractor	1	0	0	2	4
Beef	2	5	3	4	4
Calf	2	5	2	2	4

Determine which word in the first column of the table is the most similar to the word **Beef** by using:

1. the word frequencies
2. the word probabilities
3. the associated PPMI values

. . . and applying the cosine similarity.

1

cos-sim	**Elephant**	**Snake**	**Tractor**	**Beef**	**Calf**
Beef	0.653	0.334	0.678	1.0	0.969

2 Probability matrix:

word/context-word	**Zoo**	**Steak**	**Mammal**	**Cow**	**Farm**
Elephant	0.067	0.0	0.083	0.05	0.017
Snake	0.1	0.017	0.0	0.0	0.0
Tractor	0.017	0.0	0.0	0.033	0.067
Beef	0.033	0.083	0.05	0.067	0.067
Calf	0.033	0.083	0.033	0.033	0.067

Cosine similarity:
Same as **1**

3 PPMI matrix:

word/context-word	**Zoo**	**Steak**	**Mammal**	**Cow**	**Farm**
Elephant	0.3	0.0	1.206	0.332	0.0
Snake	1.778	0.0	0.0	0.0	0.0
Tractor	0.0	0.0	0.0	0.64	1.399
Beef	0.0	0.599	0.0	0.278	0.037
Calf	0.0	0.862	0.0	0.0	0.3

Cosine similarity:

cos-sim	**Elephant**	**Snake**	**Tractor**	**Beef**	**Calf**
Beef	0.108	0.0	0.225	1.0	0.874

⇒ Calf is most similar to the word Beef.

3 GloVe Embeddings

Given is the following word/context-word matrix:

word/context-word	wolf	predator	cat	eats	meat	pot	grass	contains
wolf	0	2	0	3	2	1	0	0
predator	2	0	2	0	0	0	0	0
cat	0	2	0	3	2	0	1	0
eats	3	0	3	0	4	1	1	0
meat	2	0	2	4	0	4	0	4
pot	1	0	0	1	4	0	1	5
grass	0	0	1	1	0	1	0	1
contains	0	0	0	0	4	5	1	0

1. Calculate the Probability Ratios for the word pairs (cat, wolf) and (cat, pot) for all context words (use $x/0 = \infty$ and $0/0 = 1$).

(prob/ratio)/context-word	predator	eats	meat	pot	grass	contains
P(k | cat)	0.25	0.375	0.25	0	0.125	0
P(k | wolf)	0.25	0.375	0.25	0.125	0	0
P(k | cat) / P(k | wolf)	1	1	1	0	∞	1

(prob/ratio)/context-word	predator	eats	meat	wolf	grass	contains
P(k | cat)	0.25	0.375	0.25	0	0.125	0
P(k | pot)	0	0.084	0.34	0.084	0.084	0.417
P(k | cat) / P(k | pot)	∞	4.465	0.735	0	1.48	0

2. Interpret your results by discussing how probability ratios reflect the semantic similarity of words. Which word context-word co-occurrences are needed for high and low semantic similarites?

- For context words that often co-occur with the word cat (cat) and rarely co-occur with pot (wolf) in the corpus, the probability ratio $\frac{P_{k,cat}}{P_{k,pot}}(\frac{P_{k,cat}}{P_{k,wolf}})$ is large. Similarly, for context words that often co-occur with pot (wolf) but rarely co-occur with cat, the probability ratio is small. This can be interpreted as the two words being dissimilar in this context.
- For context words that often occur together with the word cat (cat) and the word pot (wolf), the ratio is close to 1. This can be interpreted as these words being similar in this context.
- For the context words that rarely occur together with both cat and wolf, the ration is also close to 1. This also indicates that these words are similar, as they rarely or never occur together with the specific contexts.
- For the context words meat and grass, the ratio for the words cat and pot is close to 1. This indicates a similarity between these two words. However, intuitively they are very different. This means that when learning vector semantics, the corpus must be large enough to provide enough contextual information about words to learn good word representations.