Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

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?
2. What are possible applications for *Vector Semantics*?
3. Describe the difference between first order co-occurrence and second order co-occurrence.
4. List 3 possible reasons why only positive results of Pointwise Mutual Information are generally used.
5. Explain the process of a singular value decomposition in your own words.
6. When learning word embeddings through Glove, what information can be learned with small and large window size respectively?

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

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.

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 $0/0 = 1$).
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?