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

12. Exercise for "Sprachverarbeitung und Text Mining"

11.02.2022

1 Knowledge Questions

1. Describe the features and structure of WordNet in your own words.

   Describe and give an example for WordNets synsets.

2. What are potential problems of standard similarity metrics that are based on path length?

   Name one possible way to fix these problems.

3. Define the term *Information Content* in your own words.

4. Explain why the similarity metrics from the lecture that use Information Content use the Information Content of the lowest common subsumer of the words being compared in the concept hierarchy.

2 Word Similarity

2.1 Similarity measures

1. For the following ontology, compute the similarities of the following concepts:

   - sim(Pizza,Pasta)
   - sim(Pizza,Beer)
   - sim(Cola,Coffee)
   - sim(Cookies,Beer)

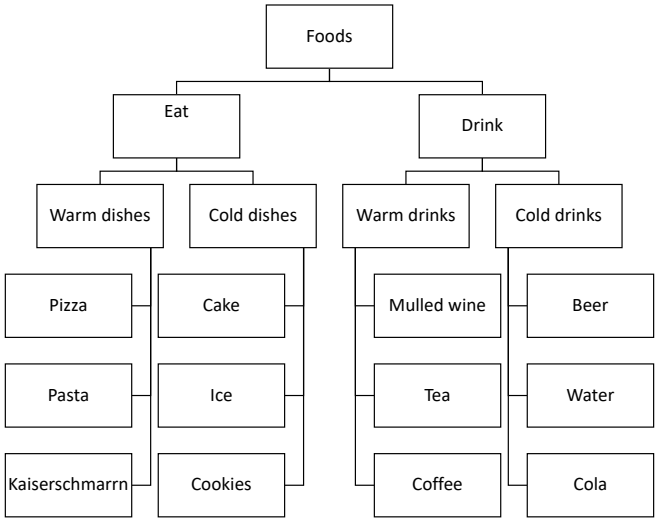
   Use the measures introduced in the lecture for this purpose:

   a) $sim_{path}(c_1, c_2)$

   b) $sim_{jiang-conrath}(c_1, c_2)$

   Use the following text for your calculations, if needed:

   ```
   My favorite foods are pizza and pasta. I have either
   beer or tea with it. Warm dishes are part of every meal
   for me. With cake I like to have coffee and cookies. The
   water is already frozen to ice when I have mulled wine
   with Kaiserschmarrn at the Christmas market. I would
   rather have pizza and beer, though.
   ```



We define the probability of a word from the taxonomy as:

$$P(c) = \frac{\sum_{w \in \text{words}(c)} \text{count}(w)}{N}$$

where words(c) contains all sub-concepts of the taxonomy from concept c including concept c itself. N is the number of words in the text that occur as a concept in the taxonomy.

2. Another similarity metric is the weighted path length (wpath) measure, which is defined as follows:

$sim_{wpath}(c_1, c_2) = \frac{1}{1+pathlen'(c_1,c_2)*k^{IC(LCS(c_1,c_2))}}$

$pathlen'(c_1, c_2) = Number\,of\,edges\,in\,graph\,between\,nodes\,c_1\,and\,c_2$

   - Describe the intuition behind wpath in your own words. What types of information that we introduced previously are used in wpath? What is the role of wpath's parameter k?
   - Calculate the wpath similarity for the concept pairs from subtask 1.

3 String Kernel

Calculate the String-Kernel of the words *Text* and *Test*. A completely simplified term depending on λ is sufficient as a result.

4 Minimum Edit Distance

Determine the minimum edit distance between the following words. We define the costs as triples $w = (w_1, w_2, w_3)$, where w_1 is the cost for Insertion, w_2 the cost for Deletion, and w_3 the cost for Replace.

- Initial-state: `datamining` Goal-state: `textmining` with $w = (1, 1, 2)$
- Initial-state: `gehend` Goal-state: `verstehen` with $w = (3, 7, 9)$