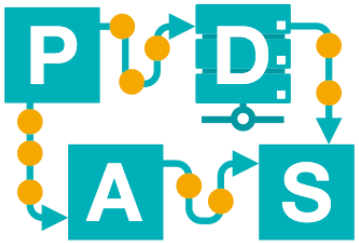
Text Mining - Instruction

Lecture 16

IDS-L16-I



Chair of Process and Data Science

RWTH AACHEN UNIVERSITY

Short summary of the lecture

- **Basics and definitions of Text Mining.**
- **Text preprocessing: tokenization, stopword removal, stemming and lemmatization.**
- **The BoW model and some applications.**
- **Tfidf weighting.**

Definitions - Recap

- **Corpus**: collection of pieces of text with a consistent nature (articles, forum posts, tweets, etc)
- **Documents**: the fragments of text in a corpus
- **Annotated corpus**: corpus in which the documents or fractions of documents have been enriched with metadata

Preprocessing - Recap

Necessary to preprocess text to get structured data.

1. **Dividing the text in discrete units (tokenization)**
2. **Removing irrelevant tokens (stopwords)**
3. **Normalize the tokens so that the same concept is always represented by the same token**
    1. **Represent concepts with stems**
    2. **Represent concepts with lemmas**

Bag of Words - Recap

You can use the Bag of Words model (BoW) to represent documents in a corpus.

It is simply the bag or multiset of the tokens contained in a document of the corpus.

Advantages:

- Simple
- Effective in some applications

Disadvantages:

- The order of the words in a document is lost
- The representation is very sparse (one feature per word in the dictionary)

Tfidf - Recap

Tfidf weighting is a simple way to represent the relevance of a word in a document.

$$tf(w,d) = \#of\ occurrences\ of\ word\ w\ in\ document\ d$$

$$idf(w) = log_2\left(\frac{N}{\#of\ documents\ that\ contain\ w\ at\ least\ once}\right)$$

$$tfidf(w,d) = tf(w,d) * idf(w)$$

Tfidf - Recap

Even if extremely simple, many querying systems rely on (variations of) tf-idf!

- Given a query and a corpus
- For each document in the corpus
  - Compute $score(query, d) = \sum_{w \in query} tfidf(w, d)$
- Rank documents by score
- Return first n documents

Querying systems - exercise

D1: ‘Cats are the only pet of the felines family, while dogs are canids.’

D2: ‘Cats are the third-most popular pet in the US.’

D3: ‘Dogs have been selected for millennia as pet animals.’

D4: ‘Normally, dogs are not aggressive towards other dogs outside their territory.’

This is the example corpus shown in the lecture.

Can you find the ranking of the four documents in the corpus for these three queries?

Q1: ‘dog’

Q2: ‘dog pet’

Q3: ‘dog cat’

Assume that plural normalizes on singular. Stopword removal not necessary.