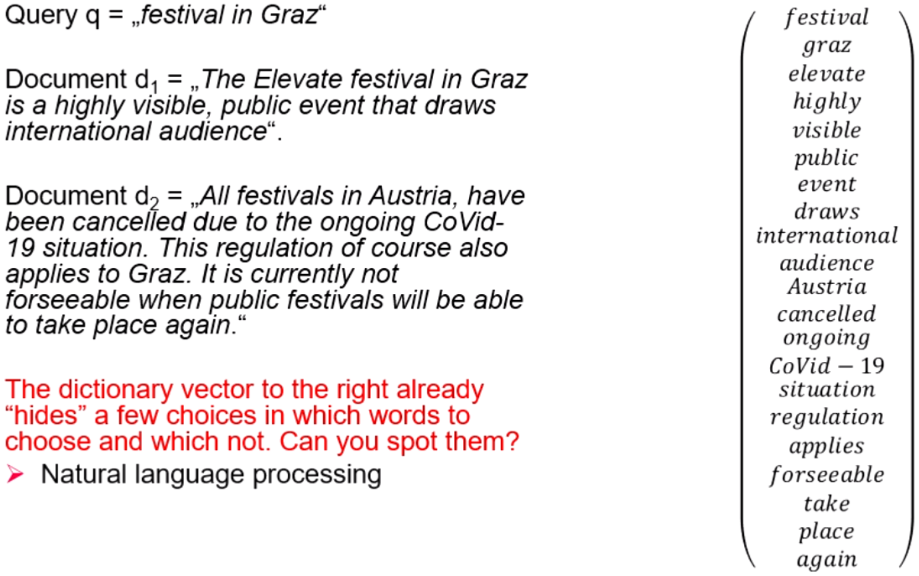
Overview

- search information retrieval
    - represent documents as [[Vectors as KR]]
    - assess its relevance (similarity)
    - ranking

Representing Documents as Vector

- vector represents content
    - disregards metadata (mostly)
- vector elements correspond to content elements
    - words
    - word stems
    - concepts
    - phrases
- dictionary vector
    - contains all selected content elements from documents and query
    - each element has index
    - decides which elements are relevant
        - each vector element represents a different topic
        - each document has value corresponding to each topic
            - different ways to generate this value
            - e.g. [[TFIDF]]
    - hides some words with [[Natural Language Processing]]
    - weird example



Query q = „*festival in Graz*“

Document d_1 = „*The Elevate festival in Graz is a highly visible, public event that draws international audience*“.

Document d_2 = „*All festivals in Austria, have been cancelled due to the ongoing CoVid-19 situation. This regulation of course also applies to Graz. It is currently not forseeable when public festivals will be able to take place again.*“

The dictionary vector to the right already “hides” a few choices in which words to choose and which not. Can you spot them?

➢ Natural language processing

$$\begin{pmatrix} festival \\ graz \\ elevate \\ highly \\ visible \\ public \\ event \\ draws \\ international \\ audience \\ Austria \\ cancelled \\ ongoing \\ CoVid-19 \\ situation \\ regulation \\ applies \\ forseeable \\ take \\ place \\ again \end{pmatrix}$$

        -

[[Information Retrieval]]