Master Thesis: Useful Words

Leonard Paál
Hochschule Karlsruhe
Betreuender Professor: Jannik Strötgen

Dezember 2024

Zusammenfassung

Useful words are nice to have.

Inhaltsverzeichnis

1 Main useful words

2 Evaluation of Approaches

2.1 Comparison of list similarities

In order to compare the results provided by the various approaches, metrics are needed that can be consistently calculated across the different approaches. While a human may be able to qualitatively analyze lists and gain a rough idea of their similarity, computed metrics provide an instantaneous (if simplified) outlook on similarities. A metric shall be defined as a function which takes as parameters two word lists of equal length which are word lists ordered descendingly by supposed relevance, and outputs a real number giving either a distance or similarity between the lists.

Considerations of the choice of metric

Metrics must be able to handle elements which occur in only one of the two lists. Thus, a metric which solely compares ranks of elements is not viable.

Changes at the beginning of the lists have a bigger impact on the metric output than changes further down. The rationale is that two lists which have large differences at the start should be regarded as more different than ones where words

Changes in rank should be counted as more important if they occur further up the list. A word that is rank 1 in list A but rank 101 in list B says more about the similarity than if a word is rank 2000 in list A but rank 2100 in list B.

2.1.1 Metrics used

2.1.2 Metrics rejected

Kendall's tau: This metric is bounded between 0 and 1 and compares the ranks of the elements of two lists. However, it cannot handle elements that only occur in one of the two lists, and thus is not suitable for our purposes.

3 Bibliography

Literatur