Information Retrieval - Exercise Sheet 1

Colm O' Riordan

September 15, 2022

Exercise sheet 1

1. Stemming involves reducing variations of words to a common root form. Can you identify, in any language, cases where this may cause problems in retrieval?

2. Stop word removal involves identifying words with little resolving power/meaning. Can you identify where this may have a negative effect on the performance of the IR system.

3. Consider the following illustrative document collection:

   - How would the sample documents be represented in the Boolean model
   - How would they be represented in the vector space model.

   ```
   D1:  java coffee shop sugar
   D2:  coffee shops in java
   D3:  java programming compilers
   ```

4. Given the conceptual models (Boolean, vector space), suggest an approach to physical design - what indexes would be appropriate?

5. The *tf-idf* scheme discussed in the notes takes two factors into account - *the term frequency* and *inverse document frequency.* Can you suggest any other features/heuristics that could be used/incorporated?

6. Check online for a stemmer and examine its performance on a passage of text. For example, you can find the code for Porter's original stemmer for English on line (coded in c). There are also several well known packages that include stemming rules for a wide range of languages.