How to Teach the Concepts of „*Information Retrieval*“ using Shell Commands

Andreas Schmidt

Institute for Automation and Applied Informatics
Karlsruhe Institute of Technology
Karlsruhe, Germany
andreas.schmidt@kit.edu

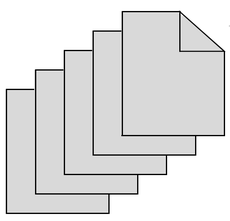
Department of Computer Science and Business Information Systems
Hochschule Karlsruhe
University of Applied Sciences
Karlsruhe, Germany
andreas.schmidt@h-ka.de

General Approach

- **Using shell commands to get informations about the vocabulary (size of vacabularity for different languages, frequency of words, Zipfs Law)**
- **Building a stopwordlist from a corpus**
- **Construction of an Inverted Index based on ASCII files using shell commands**
  - Boolean Retrieval
  - Ranking (using tf, idf)
  - Phrase match
- **Implementation of character-filter, tokenizer, token-filter using shell commands**
- **Similarity of documents**
- **Simple Entity Regognition**
- **Used tools:**
  - grep
  - sed
  - cut
  - tr
  - uniq
  - sort
  - awk
  - cat
  - comm
  - join
  - less
  - wc
  - merge
  - split/csplit
  - head/tail
  - find
  - xargs

Principle Idea of Index Construction

BBCsport Document Collection [1]



Preprocessing (tokenization, at least 2 characters long, lowercase)

```
grep -Eo '[A-Za-z]{2,}' bbcsport/*/*.txt | \
  tr 'A-Z' 'a-z' | sort | uniq | \
  tr ':' '\t' > tmp/doc-term.tsv
```

```
football/001.txt
football/001.txt
football/001.txt
football/001.txt    alert
football/001.txt    alex
football/001.txt    allowed
football/001.txt    almost
football/001.txt    always
football/001.txt    an
football/001.txt    and
```

Building Inverted Index

```
awk '{print $1 >> "InvIndex/"$2}' tmp/doc-term.tsv
```

```
$ ls InvIndex/a* | head
InvIndex/aaron
InvIndex/aberdeen
InvIndex/abeyie
InvIndex/ability
InvIndex/able
InvIndex/about
InvIndex/above
InvIndex/abramovich
InvIndex/abroad
...
```

```
football/00
football/00
football/01
football/017.txt
football/018.txt
football/021.txt
...
football/063.txt
tennis/113.txt
tennis/121.txt
...
```

Query: Looking for documents containing „*ronaldo*“ and „*barcelona*“:

```
$ cat InvIndex/ronaldo InvIndex/barcelona | sort | uniq -d
bbcsport/football/057.txt
bbcsport/football/101.txt
...
```

documents containing both terms

[1] D. Greene and P. Cunningham. "Practical Solutions to the Problem of Diagonal Dominance in Kernel Document Clustering", Proc. ICML 2006.