# Bericht

Verfassen Sie einen Bericht, welcher lesbar auf maximal zwei A4 Seiten Platz hat

Darin erklären Sie:

2. Führen Sie aus, welche Systeme (Beispiele: Eigenbau, Laborumgebung, Google Desktop Search, Microsoft Search Server, etc.) Sie eingesetzt haben und was die Motivation war.

3. Beschreiben Sie den Aufbau des Index, welchen Sie erzeugt haben. Zeigen Sie, wie die Anfragen an die Systeme gestellt wurden (manuell, automatisch, iterativ, etc.) und wie Sie die Resultate interpretieren.

# 1. Einleitung

Basierend auf der Aufgabenstellung, 50 Anfragen für eine Kollektion von über 20‘000 Dokumenten zu bearbeiten und Retrieval Systeme auf ihre Fähigkeit hin zu untersuchen, diese Informationsbedürfnisse innerhalb der Dokumentenkollektion zu befriedigen, wird in dieser Arbeit grundsätzlich der Frage nachgegangen, wie ein solches System seine Ausbeute sowie seine Präzision erhöhen kann. Das im folgenden erläuterte System nimmt keine Rücksicht auf Performance, kombiniert jedoch verschiedene Ansätze zur Verbesserung der Resultate. Der erste untersuchte Ansatz ist, dass durch Filterung derjenigen Terme1 mit relativ kleinem Informationsgehalt, die übrigen Terme1 eine höhere Gewichtung erhalten und die Suchanfragen somit präzisere Resultate liefern. Der zweite Ansatz versucht, die Ausbeute zu erhöhen indem die Terme1 auf einen Stamm2 reduziert werden. Im dritten Ansatz werden Terme aus den Anfragen, welche nicht in der Kollektion auftreten, durch ähnliche Terme1 ersetzt, so dass mögliche Schreibfehler in der Anfrage korrigiert werden. Mit dem vierten und aufwändigsten Ansatz der verfolgt wird soll erreicht werden, dass Anfragen durch Terme1 ergänzen werden, welche Gemeinsamkeiten bzw. Ähnlichkeiten mit den Termen aus den ursprünglichen Anfragen haben.

# 2. Material und Methoden

Das in dieser Arbeit benutzte System wurde vom Forschungsteam selber implementiert unter Zuhilfenahme der in den Referenzen angegebenen Quellen.

## 2.1 Grobaufbau

Zur Relevanz-Beurteilung von Termen in Dokumenten einer Dokumentenkollektion wird tf.idf verwendet. Auf dieses Beurteilungsverfahren wird nicht genauer eingegangen. Zur Erreichung der Zielsetzungen aus den einleitend genannten Ansätzen wird folgendes verwendet:

- Ansatz 1: Englisch-Stoppwort-Filter3

- Ansatz 2: Porter-Stemmer Algorithmus

- Ansatz 3: Levenshtein Algorithmus

- Ansatz 4: Erweiterung der Anfragen mit Hilfe eines Ähnlichkeit-Wörterbuchs

## 2.2 Ablauf

Schritt 1: Automatisches Einlesen der Kollektion aus vorgegebenem Verzeichnis

# 3. Rangliste

Erklärung 1. Ihre abgegebene Rangliste.

# 4. Diskussion

# 5. Referenzen

Qiu Y., Frei H.-P. (Jahr unbekannt). Improving the Retrieval Effectiveness by a Similarity Thesaurus.

http://e-collection.library.ethz.ch/eserv/eth:4438/eth-4438-01.pdf

# 6. Glossar

1 bezieht sich auf Terme aus allen Anfragen und der gesamten Kollektion

2 Stamm im Sinne eines Wortstammes (Abstammung)

3 Englisch-Stoppwort-Filter mit der Kollektion von Stoppwörter aus dem Programm IRLab