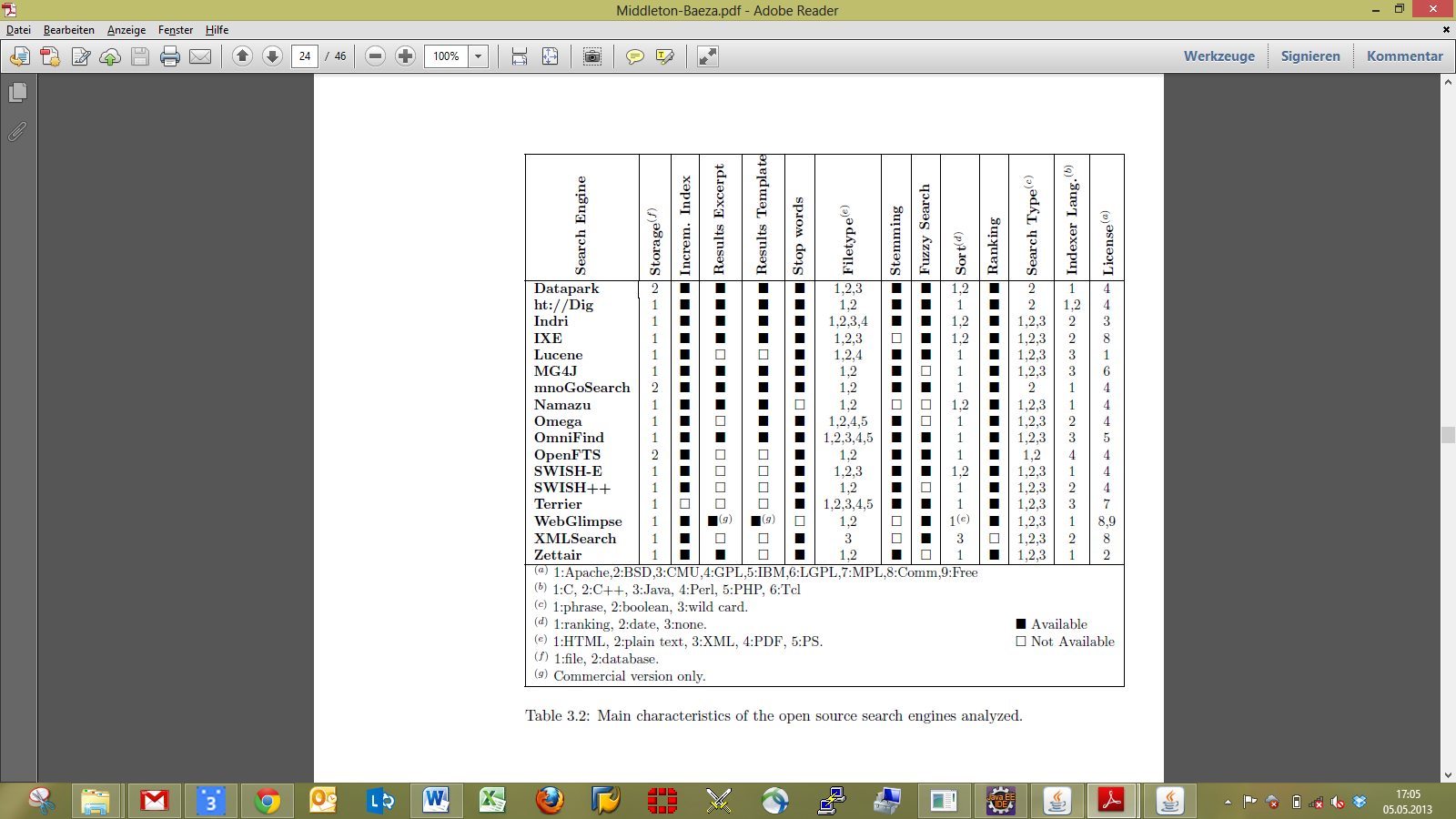
# IRG Semesterkurzbeitrag

## Autoren: Sandro Scalco, Nadri Mamuti

### Eingesetztes System:

Wir haben für die Realisierung unseres SKB Java als Programmiersprache gewählt, da wir schon etwas Java Erfahrung haben. Als Framework verwenden wir die Opensource Variante Lucene der Apache foundation, da dieses Framework sehr gut dokumentiert ist und auf unsere Bedürfnisse gut passen sollte.



### Motivation:

<http://members.unine.ch/jacques.savoy/clef/index.html>

Thanks to this international effort, many quality tools have now beenmade available to access information written in several language groups, includingthe Germanic (German, English, Dutch, and Swedish), Romance (Spanish, Portuguese, French, and Italian), Slavic (Russian, Bulgarian, and Czech) family groups, as well as members of the Uralic family (Finnish, Hungarian).

As described in this book, language diversity has generated an extensive variety of challenges to be solved by computer scientists. In German, for example, compound constructions are very frequent but the real concern is that the same concept might be expressed by two (or more) formulations, thus rendering it more difficult to find useful matches between search keywords and target items. The numerous grammatical cases found in Finnish grammar represent another example. Here the real problem involves irregularities imposed by vowel harmonies, which in turn generate complex morphological processing requirements. More research is required in order to promote better automatic processing of such linguistic constructions.

Wie wir im IRG Unterricht gesehen haben kann die Sprache auf diverse Aspekte mehrdeutig sein. So muss man bei der Suche auf Synonyme und Homonyme achten, wenn der Suchbegriff selbst nicht bekannt ist werden diese Suchbegriffe paraphrasiert. Es gibt Begriffe welche Sprach-, Region- und Traditionsabhängig sind, so werden in verschiedenen Sprachen Metaphern verwendet, welche nicht einfach Wort für Wort übersetzt werden können.

In erster Linie wollten wir untersuchen ob es einen Unterschied macht, wenn wir mit der selben Query und Collection in unterschiedlichen Sprachen suchen, wenn dem nicht so ist, wäre unsere Untersuchung nicht weiter spannend und wir hätten dann etwas neues suchen müssen.

Wir haben jedoch festgestellt, dass wir mit der Suche in unterschiedlichen Sprachen auch unterschiedliche Resultat erhalten.

Deshalb haben wir entschieden für unseren Semesterkurzbeitrag auf sechs verschieden Sprachen aus unterschiedlichen Sprachfamilien mit und ohne Stoppwörter zu suchen um dann entscheiden zu können, welche Suche die besten Resultat erzielt und was die Gründe dafür sind.

Wir möchten herausfinden ob der Durchschnitt aller Sprachen das beste Resultat erzielt.

### Aufbau Index:

Wie wurden die Anfragen an die Systeme gestellt (Manuell, automatisch, iterativ etc.)

Wir haben entschieden sechs Ranglisten der Sprachen Englisch, Deutsch, Finnisch, Französisch, Italienisch und Russisch mit- und sechs Ranglisten ohne Stoppwörter und eine Rangliste über alle Resultate zu erzeugen.

Um die Collection und Query, welche ursprünglich auf Englisch war, in die von uns gewählten Sprachen zu übersetzen mussten wir eine Lösung suchen, wie wir 40'000 Einträge effizient und möglichst schnell übersetzen können. Übersetz wurden die Querys und Collections dann mit Hilfe von Google Translator Toolkit.

Die Liste der Stoppwörter konnten wir von <http://members.unine.ch/jacques.savoy/clef/index.html> entnehmen.

### Interpretation der Resultate

Vergleicht man zwei Ranglisten der selben Sprache mit und ohne Stoppwörter und beobachtet dabei die 10 besten Ergebnisse, stellt man fest dass 90% der Ergebnisse bei beiden Ranglisten unter den besten 10 erscheinen. Beobachtet man nun die besten 15 Ergebnisse, stellt man fest dass nur noch 86.6% der Ergebnisse bei beiden Ranglisten unter den besten 15 enthalten sind. Analysiert man nun die Ergebnisse Stichprobenartig, stellt man fest dass die Rangliste mit Stoppwörtern die bessere ist, als ohne Stoppwörter und einen höheren Score erzielt.

Vergleicht man nun die Rangliste über alle Resultate mit den einzelnen Resultaten, stellt man fest, dass der Scor bei der Master Rangliste noch höher ist als bei den anderen.