Prof. Dr. Philipp Schaer, Timo Breuer, Björn Engelmann
Institut für Informationswissenschaft

Technology
Arts Sciences
TH Köln

Information Retrieval – Übung Evaluation von IR-Systemen

Bitte bearbeiten Sie die Aufgaben **vor den Übungsterminen**. In den Online-Sitzungen arbeiten wir dann mit den Ergebnissen Ihrer Vorarbeiten. Es stellt jemand aus Ihren Reihen die vorbereitete Lösung vor. Die Lösung kann wird anschließend im Plenum diskutiert, erweitert und in Kontext gesetzt. Auch Peer-Reviews sind möglich.

Die Präsentation der Lösung kann beispielsweise über zuvor eingereichte Videos oder Bildschirmaufnahmen, Kurzreferate mit vorbereiteten Folien oder interaktiven Kollaborationswerkzeuge wie Google Jamboards erfolgen.

I Precision und Recall

Ein IR-System liefert 23 relevante Dokumente und 20 nicht-relevante zurück. Es gibt insgesamt 50 relevante Dokumente in der Testkollektion.

a. Wie hoch ist die Precision?
b. Wie hoch ist der Recall?
c. Wie ist der Wert für das F-Measure?

II Precisionbestimmung in gerankten Systemen

Stellen Sie sich eine Anfrage vor, zu der es insgesamt fünf relevante Dokumente gibt. Zwei verschiedene Systeme liefern unterschiedlich gerankte Ergebnislisten. Die ersten 10 Ergebnisse wurden wie folgt auf ihre Relevanz hin bewertet (R = relevant; N = nicht-relevant):

	1.	**2.**	**3.**	**4.**	**5.**	**6.**	**7.**	**8.**	**9.**	**10.**
System 1	R	R	N	R	N	N	R	N	N	N
System 2	N	R	N	R	N	R	R	R	N	N

a. Bestimmen Sie Precision-at-k für k=1, 3, 5 und 10, jeweils für System 1 und 2.
b. Bestimmen Sie die R-Precision für beide Systeme. Welches hat die höhere R-Precision?
c. Wie lauten die AP-Werte für jedes System? Welches System lieferte bessere Werte?
d. Vergleichen Sie die jeweiligen Werte? Welches ist das bessere System? Gibt es Unterschiede im Ranking? Wie können Sie sich diese erklären?

III Topic-Erstellung und Pooling

Relevanzbewertungen von Testkollektionen werden üblicherweise anhand von sogenannten Topics vorgenommen. Diese enthalten neben der kurzen Angabe zum eigentlichen Thema (Title), eine kurze Beschreibung (Description) sowie eine ausführliche Erzählung zum Kontext des Informationsbedürfnisses (Narrative).

Ein kleines Beispiel aus der Testkollektion GIRT (German Information Retrieval Testdatabase):

```
<top>

<num> 100 </num>

      <title> Kneipenkultur </title>

      <description> Finde Dokumente, die die Gewohnheiten von Kneipengän-
      gern thematisieren. </description>

      <narrative> Relevante Dokumente analysieren die Gepflogenheiten von
      Kneipengängern und die Kultur der Räumlichkeiten und des Ambientes
      von Szenekneipen. </narrative>

</top>
```

Topics sind ein Weg Kriterien für die Relevanzbewertung aufzustellen, um so die rein subjektive Relevanzbewertung zu objektivieren.

In der folgenden Übung sollen Sie sich wie folgt mit Topics und dem Thema Relevanzbewertung auseinandersetzen:

a. Erstellen Sie zu einem beliebigen Thema ihrer Wahl ein solches Topic (inkl. Title, Description und Narrative).
b. Durchsuchen Sie auf Grundlagen Ihres gewählten Topics eine passende Onlinebibliothek (bspw. Livivo ZB MED oder SSOAR von GESIS). Führen Sie hierbei die Suche auf zwei unterschiedliche Arten durch (z.B. mit Hilfe von manueller Anfrageerweiterung, nur Termen aus dem Titel vs. zusätzlich auch Termen aus der Description, ...).
c. Bewerten Sie die ersten 10 Suchergebnisse beider Anfragearten nach Relevanz (binär). Entscheiden Sie, inwiefern das einzelne Suchergebnis Ihrer Vorstellung des von Ihnen gewählten Topic entspricht.
d. Führen Sie anhand der von Ihnen bewerten Ergebnisse ein Pooling durch wie es in der Vorlesung vorgestellt wurde. Dokumentieren Sie anschließend die resultierende Dokumentsammlung.