Praktikum 6b – IR Visualizer Tool (BETA)

Sie verwenden für die folgende Analyse ein neues Software-Tool, welches im Rahmen einer Projektarbeit hier an der ZHAW entstanden ist, und seither in Details abgeändert wurde. Betrachten Sie das Tool als Beta, melden Sie gerne Beobachtungen hinsichtlich Bugs/Eigentümlichkeiten, damit diese in einer späteren Version behoben werden können.

Starten Sie das Tool auf geeignete Weise (unter Windows bei korrekter Installation von Java per Doppelklick auf das JAR-File). Das Tool bietet den Zugriff auf Auswertungsdaten für ca. 300 Anfragen aus den Testkollektionen der CLEF-Kampagne. Bitte beachten Sie, dass einige Anfragen fehlen (z.B. Anfrage 54) – dies ist kein Fehler, sondern eine Eigenheit der verwendeten Testkollektionen.

Sie können mittels diesem Tool die Auswirkungen verschiedener IR-Massnahmen auf die Retrievaleffektivität nachvollziehen. Dazu wählen Sie auf dem ersten Screen beliebig viele Anfragen (Mehrfachselektion möglich), und selektieren dann auf dem zweiten Screen unterschiedliche Konfigurationen des IR-Systems. Diese sind kurz beschrieben (im Wesentlichen unterscheiden sich die Konfigurationen in der Stopwortelimination, dem Stemming und der Verwendung von Relevance Feedback). Sie können dann Balken oder Matrixgrafiken erstellen.

Grundsätzliche Überlegungen

Schauen Sie sich die Unterschiede zwischen den einzelnen Stemming-Algorithmen an. Wann funktionieren diese, wann versagen Sie? Interessant ist z.B. Der Unterschied zwischen "s-Stemmer" (entfernt nur plural-s), Porter (regelbasiert) und 5-gram (Buchstabensequenzen). Wann funktioniert das eine, wann das andere?

Schauen Sie sich die Unterschiede an, die sich bei Verwendung von Rocchio ergeben. Studieren Sie die zusätzlichen Begriffe. Können Sie sich das Verhalten erklären? Warum kamen die Begriffe wohl ins System (Tip: es hilft, die lange Version der Anfragen zu konsultieren).

Vergleichen Sie die Performance für lange und kurze Anfragen.

Analyse einzelner Anfragen

Ihre Aufgabe ist es, einzelne Anfragen und deren Retrievaleffektivität genauer unter die Lupe zu nehmen. Sie analysieren dabei, wie sich Massnahmen wie Stemming und Relevance Feedback auf EINZELNE Anfragen auswirken, ohne dass die Effekte durch Durchschnittsbildung "verwischt" werden.

Erstellen Sie für die gewählten Anfragen, und ggf. für weitere Anfragen zum Vergleich, geeignete Diagramme. Scheuen Sie sich auch nicht, weitere Anfragen mit speziellen Eigenschaften zu suchen (das Matrixdiagramm eignet sich sehr gut für diese Aufgabe).

Die einzelnen Anfragen wurden natürlich bewusst so gewählt, dass sie Auffälligkeiten zeigen. Beantworten Sie unter anderem die folgenden Fragen:

- was ist die Eigenheit der Anfrage hinsichtlich der Retrievaleffektivität?
- Wie verhalten sich die unterschiedlichen Stemmingalgorithmen? Warum?
- Was war ausschlaggebend für das auffällige Verhalten?

- Query 53: Tip: schauen Sie eine Abfolge Tokens-No Stop-Porter-Rocchio an.
- Query 55: Tip: schauen Sie die verschiedenen Stemming-Algorithmen an
- Query 90: Tip: schauen Sie die verschiedenen Stemming-Algorithmen an
- Query 104: Tip: schauen Sie alle unterschiedlichen Konfigurationen an. Die Stopwortelimination macht etwas anderes, als Sie vielleicht erwartet haben. Kann man dieses "Problem" kompensieren?
- Query 137: Tip: vergleichen Sie die Effektivität mit und ohne Relevance Feedback
- Query 180: Tip: schauen Sie eine Abfolge Tokens-No Stop-Porter-Rocchio an. Es lohnt sich, den "s-Stemmer" zusätzlich zu konsultieren
- Query 197: Tip: vergleichen Sie die Effektivität mit und ohne Relevance Feedback
- Query 217: Tip: schauen Sie eine Abfolge Tokens-No Stop-Porter-Rocchio an.
- Query 242: Top: vergleichen Sie mit/ohne Stemming
- Query 268: Tip: vergleichen Sie die Effektivität mit und ohne Relevance Feedback

Bonus: das Tool enthält auch Daten von Experimenten mit sprachübergreifender Suche. Hierzu wurde eine fremdsprachige Anfrage per Google Translate Service übersetzt. Schauen Sie sich die Ergebnisse an, und überlegen Sie, was für Probleme auftreten. Leider kann das Tool aufgrund fehlender Daten die übersetzten Anfragen noch nicht anzeigen; dies wird für folgende Durchführungen der Vorlesung behoben werden.