

Protokoll 25.06.

Organisatorisches

Grundsätzlich ist es sinnvoll und wünschenswert, wenn alle Gruppenmitglieder zu den Präsentationen kommen, egal ob sie selbst präsentieren oder nicht. BA-Seminar mal ausgenommen, seid bitte in Zukunft alle da (und pünktlich ...).

Zweite Präsentation zum Zwischenstand

Hr. Potthast war nicht anwesend, daher Antworten auf unsere Fragen unter einem gewissen Vorbehalt zu betrachten.

Hinweise zu unseren Fragen

- Was sollen wir als Baseline betrachten?
Im Grunde das, was die SimpleSearch schon macht. Wenn zu bestimmten Topics gar nichts gefunden wird, ist das eben so und fließt entsprechend in die Evaluation ein.
- Wortlisten einbauen:
Fällt unter lexikonbasierte Erweiterung, ist also schon Optimierung und kein Teil der Baseline.
Tipp: dafür vector space models anschauen.
- Known-Item-Search:
kein Muss für unsere Suchmaschine, eher nice-to-have und definitiv auch Teil möglicher Optimierungen.
- Ergebnisebenen:
Ausprobieren. Episoden/Szenen/Quotes anbieten sei an sich gut. Bei den quotes von der Darstellung her so machen, dass im Ergebnis trotzdem die ganze scene angezeigt wird und die Trefferquotes darin dann gehighlighted werden.
- Nächster Schritt: Evaluationsmodus einbauen.
Dazu „Relevance judgement“ bei den Topics (gepinnt in unserem Slack-Channel).

Wir haben danach noch zu dritt diskutiert und folgende Vorgehensweise für sinnvoll befunden:

1. Zuerst die Topics ausformulieren. Das wird Grundlage unserer Evaluation sein und daraufhin wird optimiert → Issue im git für alle.
2. Auf den Topics aufbauend schauen, was für Querytypen, Suchmodi etc. Lucene anbietet und passendes raussuchen.

V.a. zwei Aspekte von Lucene brauchen wir dafür:
Indexing und Query Syntax / Möglichkeiten zur Query Expansion.

Muss recherchiert werden, Material ist genug vorhanden. Wir teilen die Recherche auf zwei „Expertenteams“ auf:

Indexing: Sonja, Carlos.
Query Syntax: Paula, Esther.
(Robin hat den Code ;))

„Experte“ meint, dass man versteht, wie das jeweils im Detail funktioniert, so dass man es auch den anderen erklären kann.

Dazu gibt in Kürze 2 Issues im git, wo jeweils ein Überblick zum Nachlesen gemacht werden soll.

Sonja hatte dazu schon ein paar Links ins Slack gestellt:

<https://www.ionos.de/digitalguide/server/konfiguration/apache-lucene/>
(besonders informativ)

<https://linuxhint.com/introduction-to-lucene/>

<http://www.lucenetutorial.com/lucene-in-5-minutes.html>

Außerdem sehr anzuraten dieses Video, am Schluss mit Beispiel in Lucene:

<https://www.youtube.com/watch?v=yst6VQ7Lwpo>

- Interessanter Fund: Tool „Luke“ (auch im Video) ermöglicht, den Index direkt anzuschauen. Nicht verfügbar für unsere .net-Version.

- Gibt es weitere Tools / Funktionen, die wir uns mit einer anderen Version zunutze machen könnten? Arbeitsaufwand?