Protokoll vom 21.06. nach Themen

2 Suchverfahren

derzeit als 2 Basisklassen mit jeweils eigenem Indexing implementiert:

A simpleSearch: Typ Datenbankabfrage
B advancedSearch: Typ interpretative Suche (→ Klasse AdvancedSearchBase noch nicht ausformuliert)

Soll dem User diese Unterscheidung in Form von 2 Buttons explizit vorgegeben werden, oder sollen die beiden Suchtypen "im Hintergrund" voneinander unterschieden werden, ohne Vorauswahl bei der Eingabe der Query?

Indexing für beide zusammenführen j/n?

AdvancedSearch

Wie bauen wir die Wortlisten ein? Gewichtung der Begriffe?

Modellierung der Wortlisten zwischen Query und Index - wie können wir die konkret einbauen? Ein abstrakter Suchbegriff führt auf die passende Wortliste, und dann müsste man nach jedem einzelnen Begriff dadrin extra suchen → extrem ineffizient?!

Definieren wir eine Schwelle, „ab x passenden Begriffen ist Dok. ein Treffer“, oder fließt das eben ohne Schwelle direkt ins Ranking ein?
Alle Wortkombinationen vorher schon (irgendwie) im Indizierungsverfahren den Dokumenten zuweisen? Ist das schummeln?

Ergebnisstruktur

Definition Suchergebnis/'Dokument': Derzeit entspricht eine komplette Zeile in den csv-Dateien einem Dokument.
→ in Datei *script_lines* nur Spalte *normalized_text* für Indizierung und als Ergebnis verwenden (die Infos zu character, location etc. brauchen wir natürlich trotzdem):

raw_text:
Bart Simpson: Well, you got that right. (TO TERRI AND SHERRI) Thanks for your vote, girls.

spoken_words:
Well, you got that right. Thanks for your vote, girls.

normalized_text:
well you got that right thanks for your vote girls

Reicht das als zu durchsuchende Textmenge, von der Wörterzahl her?

→ Verschiedene Ergebnisebenen ermöglichen: episode, scene, quote. Eine scene ist begrenzt durch die location. Bei scenes / quotes als Ergebnis trotzdem auch Angabe der episode.

Indexing

Wie passt die Vorlesung zum Indexing von Lucene? Welche key-value-Paare, Vektoren, ... stehen im Index, wie sieht das konkret aus? Gibt es Visualisierungsmöglichkeiten?
Wie wird gewichtet und wie können wir das anpassen?

Technologien

Welche Technologie eignet sich für welchen Typ Datensatz?

Abgrenzung Lucene - Solr:
"Solr runs as a standalone full-text search server. It uses the Lucene Java search library at its core for full-text indexing and search [...]" (kurz aus Wikipedia zitiert).

Mit Lucene mehr Kontrolle und Customizing, dafür auch mehr Arbeit für uns.
Generell: Ist es möglich, alle benötigten Funktionalitäten mit Lucene abzudecken?
→ Möglichkeiten von Lucene ausloten...

Workflow

Wortlisten / Topics überarbeiten erstmal aufschieben

Dann aber:
Wie groß sollen die Wortlisten sein?
Nur Substantive?
Überlappungen zwischen Topics?
...

Wunsch nach Gruppenleiter, Struktur, klarer Aufgabenverteilung

Weitere To Do's

Topics ausformulieren

Aufgabenverteilung für den Code?!

Fahrplan Evaluation

Weitere Fragen für die Präsentation am 25.06.

- Wir hätten gerne eine Checkliste: Must-haves und nice-to-haves

- In welchem Zustand muss die Suchmaschine sein (=welche Funktionalitäten muss sie abdecken), um sie als Baseline für die Evaluation zu benutzen?
Gibt es irgendwelche Vorgaben für die Evaluation, z.B. bestimmte Statistiken, die auf jeden Fall drin sein müssen?

- known-item-search (Anmerkung von Hr. Potthast):

Falsche Zitate sind noch nicht angedacht ... → nochmal nachfragen

"**Known-item search** is a specialization of information exploration which represents the activities carried out by searchers who have a particular item in mind.[1] In the context of library catalogs, known-item search means a search for an item for which the author or title is known.[2] Although the concept of known-item search originated in library science, it is now applied in the context of web search and other online search activities.[3]"

(https://en.m.wikipedia.org/wiki/Known-item_search)

User weiß, dass es das erinnerte Zitat so oder so ähnlich definitiv gibt. Richtig erinnert-Query wäre vermutlich reine DB-Anfrage. Aber was passiert bei teilweise falsch erinnert, z.B. durch Austausch eines Wortes?

Noch nicht näher angeschaut, evtl. interessant zum Thema known-item search:

https://www.phil-fak.uni-duesseldorf.de/fileadmin/Redaktion/Institute/Informationswissenschaft/stock/1078739217password_1.pdf

http://eprints.rclis.org/8748/1/Lee_Known-Item.pdf