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

Technology
Arts Sciences
TH Köln

DIS12: Information Retrieval – Laborsitzung 11.12.2020
Übung 04: Indexing, Terms & Tokens

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

Die Präsentation der Lösung kann beispielsweise über "Teilen des Bildschirms", Kurzreferate oder interaktive Kollaborationswerkzeuge wie Google Jamboards erfolgen.

I Stemming

Sind die folgenden Aussagen wahr oder falsch? Begründen Sie und bringen Sie ein Beispiel an!

1. In einem Booleschen Retrieval-System verringert Stemming niemals die Precision.
2. In einem Booleschen Retrieval-System verringert Stemming niemals den Recall.
3. Stemming erhöht die Größe des Vokabulars / des Dictionaries.
4. Stemming sollte zur Indexierungszeit aufgerufen werden, aber nicht bei der Verarbeitung einer Anfrage.

II Indexkonstruktion und Vorverarbeitung

Diese Übung dient dem ersten Verständnis für die Möglichkeiten und Grenzen der automatischen Indexerstellung und Entitätserkennung. Machen Sie sich zunächst mit **dbpedia Spotlight und der dbpedia** vertraut. Für die weitere Übung verwenden Sie die u.g. Beispieldokumente.

- dbpedia Spotlight https://www.dbpedia-spotlight.org/demo/
- Hintergrundinformationen https://wiki.dbpedia.org/about

Beispieldokumente

Dokument 1

Titel: Management und Soziale Arbeit : Beiträge zu Freire, Lévinas und Luhmann zur Frage nach dem Menschenbild

Abstract: Ziel des Verfassers ist es, Gedanken zum Umgang mit dem Anderen zu formulieren, die sich sowohl auf die Basisarbeit als auch auf die Leitungsebene der sozialpädagogischen Arbeit beziehen und eine 'Fundierung und Sensibilisierung des Kontaktes' bewirken. Im Mittelpunkt steht die Rolle des Managements in der Sozialarbeit. Den Bezugsrahmen hierfür bilden Levinas, Luhmann, Freire und Boal als radikale Kritiker des klassischen Bildungsideals. Die Schrift schließt mit einer 'Ethik des Lebens' und formuliert das Ergebnis der Untersuchung so: 'Es geht um den Einzelnen'.

Dokument 2

Titel: Is education the cause for Iberian economic growth? : a study in econometric history

Abstract: Recent models of growth, such as Romer (1986, 1990) and Lucas (1988), following Arrow (1962) and Uzawa (1965), emphasise human capital investment as an important factor contributing to long-run growth. In the literature, human capital investment takes several forms (educational attainment, learning by doing, etc.). The focus in this paper is on human capital accumulation through the formal schooling. It is the author's thesis that education is more an accompanying investment than a 'driving force' behind growth. They test this argument with the concept of the causal relationship formulated by Granger. All the tests are performed on the basis of the aggregate series of public expenditures on education (EXPEDU), total public expenditures (EXPTOT), population (Population) and Gross domestic product (GDP) in Portugal and Spain before World War II.

Übungsaufgaben (für mindestens 2-3 Personen)

1. Analysieren sie die zwei Dokumente zunächst **manuell**.
   Führen Sie folgende Analysen durch:
   a. **Stoppworte erkennen**: Streichen Sie **alle** Stoppworte im Abstract!
      Beispiel: „Ziel ~~des~~ Verfassers ist ~~es~~, Gedanken ~~zum~~ Umgang ~~mit dem~~ Anderen ~~zu~~ formulieren …“. Nehmen Sie dazu eine deutsche bzw. engl. Stoppworteliste zur Hilfe, z.B. diese hier: http://members.unine.ch/jacques.savoy/clef/index.html.
   b. **Vereinheitlichen Sie** einzelne Worte (mind. 10): führen Sie eine typische Wortformreduzierung (Lemmatization, Stemming) durch!
      Beispiel: Gedanken → Gedanke, Umgang → umgehen, formulieren → formulier. Nehmen Sie dazu die entsprechenden Versionen des Snowball-Stemmers zur Hilfe, siehe http://text-processing.com/demo/stem/
   c. **Eigennamenerkennung**: Kennzeichnen Sie Eigennamen!
      Beispiel: „Den Bezugsrahmen hierfür bilden Levinas, Luhmann, Freire und Boal als radikale Kritiker des klassischen Bildungsideals.“
   d. **Kompositazerlegung**: Zerlegen Sie Komposita (mind. 10)! Achten Sie dabei darauf, dass die Semantik der Komposita durch die Zerlegung nicht verloren geht.
      Beispiel: Bildungsideal → Bildung + Ideal, Bezugsrahmen → besser nicht zerlegen!
   e. **Phrasenerkennung**: Kennzeichnen Sie semantisch zusammengehörige Phrasen, die nicht getrennt werden sollten!
      Beispiel: "democratic citizenship community", "dem Anderen", "sozialpädagogischen Arbeit", "Fundierung und Sensibilisierung des Kontaktes", "Es geht um den Einzelnen", "educational attainment", "learning by doing"

2. Analysieren Sie die beiden Dokumente mit dem Tool **dbpedia Spotlight**.
   Kopieren Sie dazu die **Abstracts der Dokumente** in die Oberflächen und lassen Sie eine Analyse mit den Standardeinstellungen laufen. Speichern Sie das Analyseergebnis z.B. als Screenshot.
   a. **Kommentieren Sie** das Analyseergebnis.
      Beispiel dbpedia: Beurteilen Sie die Erkennungsleistung indem Sie die dbpedia-Entitäten aufrufen und die Kontextinformationen mit dem Dokument abgleichen.
   b. **Listen Sie die Entitäten auf**, die aus Ihrer Sicht korrekt erkannt wurden.
   c. **Vergleichen Sie die Erkennungsleitung** der Tools bei deutschsprachigen und englischsprachigen Dokumenten.

III Soundex

Codieren Sie folgende Worte über das Verfahren Soundex. Die Buchstabencodes finden Sie im Web z.B. unter http://de.wikipedia.org/wiki/Soundex oder in den Folien der heutigen Vorlesung.

1. Weber
2. Spärck Jones
3. Straßenbahn
4. Musician
5. Metaphysik

Beantworten Sie die folgenden Fragen:

- Wozu könnte die Soundex-Behandlung von Wörtern hilfreich sein?
- Welche Repräsentationen von Wörtern könnten sonst noch nützlich sein.

IV Vektorraummodell in Shakespeares Gesamtwerk

Stück für Stück wächst unsere Suchmaschine.

Greifen Sie ruhig auf die Lösungen Ihrer Kommilitonen zurück, die eine Lösung in das gemeinsame GitHub-Repository einstellen. Wenn Sie dort auch etwas hochladen wollen, schicken Sie mir kommentarlos Ihre Github-Kennung.

- Erweitern Sie Ihre Lösung der letzten Wochen (Erstellung einer Term-Dokumentmatrix mit tf-idf-Werten), sodass für alle Terme tf-idf-Werte berechnet werden und diese in einer Term-Dokumentmatrix gespeichert sind. Nutzen Sie z.B. Pandas, um eine solche Matrix/Tabelle zu erstellen.
- Implementieren Sie eine einfache Suche nach dem Vektorraummodell, anhand der Verfahren, die Sie in der Vorlesung kennengelernt (Kosinusähnlichkeit, Vektornormalisierung, Skalarprodukt) haben. Ein Tipp: Sie müssen nicht alles von Hand programmieren. Viele Dinge gibt es schon fertig, z.B. im Paket NumPy: https://www.python-kurs.eu/matrix_arithmetik.php
- Es ist vollkommen okay, wenn Sie im Netz nach Lösungsansätzen suchen und diese entsprechend anpassen. ABER: Geben Sie bitte Ihre Quellen an und stellen Sie den Code auf GitHub dem Rest des Kurses zur Verfügung. Stellen Sie bitte auch sicher, dass der Code lesbar ist und ein paar Kommentare enthält. Eine README.md hat auch noch nie geschadet.