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

Technology
Arts Sciences
TH Köln

# Information Retrieval – Übung Index-Konstruktion

## I Stemming

Nehmen Sie zu den folgenden Fragen jeweils Stellung und begründen Sie Ihre Antwort. Bringen Sie ein Beispiel an, dass Ihre Aussage unterstützt!

1. Verringert Stemming in einem Booleschen Retrieval-System niemals die Precision?
2. Verringert Stemming in einem Booleschen Retrieval-System niemals den Recall?
3. Erhöht Stemming die Größe des Vokabulars / des Dictionaries?
4. Sollte Stemming zur Indexierungszeit angewendet werden, aber nicht bei der Verarbeitung einer Anfrage?

## II 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. Spärck Jones
2. Smith, Schmidt, Schmitz
3. Mr, Mayer, Meier
4. through, thru, trough
5. data, date, dito

Beantworten Sie die folgenden Fragen:

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

## III 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 das u.g. Beispieldokument.

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

## Beispieldokument

**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.

1. Analysieren sie das Dokument 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 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. **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 das Dokument mit dem Tool **dbpedia Spotlight**.
   Kopieren Sie dazu das **Abstracts des Dokumentes** 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 mit einem beliebigen anderen Dokument, jedoch diesmal ein deutschsprachiges Dokument.