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

Technology
Arts Sciences
TH Köln

DIS12: Information Retrieval – Online-Sitzung 04.12.2020
Übung 01: Boolean Retrieval

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 Mengen und Sets

Gegeben ist das folgende Vokabular:

{Auto, Straße, Schild, Ampel, Verkehr, Lärm, Benzin, Diesel, Umwelt}.

Zusätzlich seien folgende Dokument gegeben:

Dok1 = {Auto, Schild, Verkehr, Benzin, Umwelt}

Dok2 = {Straße, Ampel, Lärm, Diesel}

Dok3 = {Schild, Verkehr}

Bestimmen Sie folgende Kombinationen der Bag-of-words:

1. Dok1 AND Dok2
2. Dok1 OR Dok2
3. Dok1 AND Dok3
4. Dok1 NOT Dok3
5. Dok2 OR Dok3
6. Dok2 XOR Dok3

II Venn-Diagramme

Eine Umfrage mit 100 Teilnehmern zur Nutzung von Suchmaschinen ergab folgende Ergebnisse:

- 25 Teilnehmer nutzen DuckDuckGo.
- 30 Teilnehmer nutzen Bing.
- 40 Teilnehmer nutzen Google.

- 6 Teilnehmer nutzen sowohl Google als auch DuckDuckGo, aber nicht Bing.
- 7 Teilnehmer nutzen alle drei Technologien.
- 8 Teilnehmer nutzen sowohl Bing als auch Google, aber nicht DuckDuckGo.
- 10 Teilnehmer nutzen ausschließlich Bing.

Benutzen Sie ein Venn-Diagramm zur Beantwortung folgender Fragen

1. Wie viele Teilnehmer nutzen keine der drei Technologien?
2. Wie viele Teilnehmer nutzen sowohl Bing als auch DuckDuckGo, aber nicht Google?
3. Wie viele Teilnehmer nutzen ausschließlich DuckDuckGo?
4. Wie viele Teilnehmer nutzen ausschließlich Google?
5. Wie viele Teilnehmer nutzen entweder Google oder Bing?
6. Wie viele Teilnehmer nutzen entweder Google oder DuckDuckGo?

III Term-Dokument-Matrix

Sie haben einen Dokumentkorpus, der aus vier Dokumenten besteht:

Dokument 1: population below the poverty line

Dokument 2: calculation of the poverty gap index

Dokument 3: population, poverty and economic growth

Dokument 4: the impact of economic growth on poverty

1. Erstellen Sie die Term-Dokument-Matrix (auf Papier, mit Excel, Python, egal).
2. Was sind die Ergebnisse für die folgenden beiden Anfragen:
   - poverty AND population
   - the AND NOT (economic OR poverty)

IV Suche in Shakespeares Gesamtwerk

Im Moodle finden Sie den Gesamtbestand aller Werke von Shakespeare. Diese Dateien können Sie mit den Ihnen bekannten Tools wie grep oder Python verarbeiten.

Versuchen Sie ein kleines Bash-Skript oder Python-Programm zu schreiben, dass alle Textdateien durchsucht und die aus der Vorlesung bekannte boolesche Anfrage nach

Brutus *UND* ***Caesar*** *UND NICHT* ***Calphurnia***

auswertet. Im Optimalfall kann ihre Lösung natürlich mit beliebigen Termen und booleschen Operatoren umgehen.

(achten Sie auf die besondere Schreibweise von Caesar im Korpus)