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

Technology
Arts Sciences
TH Köln

Information Retrieval – Übung Boolean Retrieval

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

Die Präsentation der Lösung kann beispielsweise über zuvor eingereichte Videos oder Bildschirmaufnahmen, Kurzreferate mit vorbereiteten Folien oder interaktiven 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:

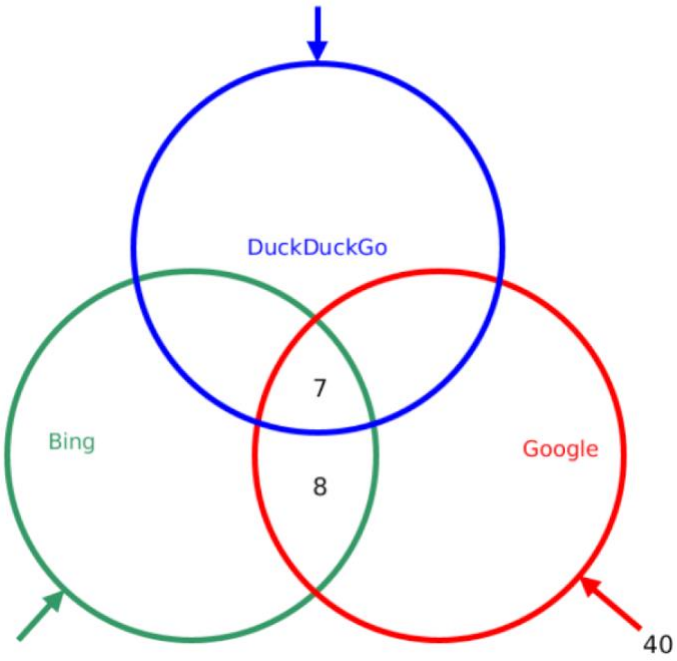
1. Dok1 AND Dok2
2. Dok1 OR Dok2
3. Dok1 AND Dok3
4. Dok1 NOT Dok3
5. Dok2 OR Dok3
6. Dok2 XOR Dok3

II Venn-Diagramme

Im Rahmen einer Studie zur Nutzung von Websuchmaschinen möchten Sie zunächst ermitteln, wie häufig ausgewählte Suchmaschinen genutzt werden. Sie beschränken Ihre Untersuchungen auf die drei Suchmaschinen Google, Bing und DuckDuckGo. Für Ihre Befragungen konnten Sie insgesamt 100 TeilnehmerInnen gewinnen. Außerdem können Sie Ihren Befragungen folgende Aussagen entnehmen:

- 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.

Sie sind an weiteren Ergebnissen interessiert und entscheiden sich deshalb mit Hilfe eines Venn-Diagramms, die Resultate zu untersuchen und zu visualisieren:



Vervollständigen Sie das Venn-Diagramm und beantworten folgende Fragen:

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

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? Zeigen Sie jeweils die einzelnen Rechenschritte und nicht nur das Endergebnis!
   - poverty **AND** population
   - the **AND NOT** (economic **OR** poverty)