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

Technology
Arts Sciences
TH Köln

DIS12: Information Retrieval – Online-Sitzung 15.01.2021
Übung 06: Query Expansion

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 Jaccard-Index berechnen

Berechnen Sie den Jaccard-Index für die vier folgenden Begriffskombinationen (wobei A jeweils aus dem Titel stammt und B aus den Schlagwörtern):

(1) A = Jugend / B = Jugendlicher

(2) A = Jugend / B = Arbeitsmarkt

(3) A = Jugend / B = Jugend

(4) A = Arbeitslosigkeit / B = Arbeit

Interpretieren Sie die jeweiligen Ergebnisse und erläutern Sie Ihre Schlussfolgerungen.

	Dokument 1	**Dokument 2**	**Dokument 3**
Titel	Jugend und Arbeit : ein Fazit zum Stand der Forschung	Jugend und Arbeit : empirische Bestandsaufnahme und Analysen	Erwartungen und Lebensorientierungen der polnischen und deutschen Jugend in Zeiten von Unsicherheit
Schlagwörter	Arbeit, Jugendlicher, Erwerbsarbeit, Jugend, Einstellung, Arbeitslosigkeit, Arbeitsmarkt, Berufsorientierung	Schule, Arbeit, Jugendlicher, Erwerbsarbeit, Jugend, Arbeitslosigkeit, Migrant, Bundesrepublik Deutschland, Benachteiligung, Berufseinmündung, Berufsvorbereitung, Europa	Zufriedenheit, postsozialistisches Land, Jugendlicher, Einstellung, Arbeitslosigkeit, Arbeitsmarkt, Bundesrepublik Deutschland, Erwartung, Bildungsziel, Polen

II Normalized Google Distance

Eine Erweiterung des Jaccard-Index ist die Normalized Google Distance[1] (NGD). Diese berechnet eine semantische Ähnlichkeit oder Nähe zwischen zwei Termen, basierend auf der Anzahl der Treffer, die Google (oder eine andere Suchmaschine) liefert. Details finden Sie u.a. im Wikipedia-Artikel oder in der Original-Quelle unter https://arxiv.org/abs/cs/0412098.

(1) Implementieren Sie eine einfache Berechnung der NGD z.B. mit einem Python-Skript oder Excel. Zur Extraktion der Treffer von Google, Bing oder DuckDuckGo können Sie bspw. BeautifulSoup aus dem letzten Semester verwenden.

(2) Vergleichen Sie die Jaccard-Index-Werte mit den NGD-Werten für fünf beliebige Wortpaare; jeweils auf Bing und Google. Fallen bestimmte Effekte auf?

[1] https://en.wikipedia.org/wiki/Normalized_Google_distance