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

Technology
Arts Sciences
TH Köln

Information Retrieval – Übung Query Expansion

I Manuelle Query-Expansion

Sie wollen einen Artikel zum Thema „Aktienkauf zur Altersvorsorge“ schreiben. Sie möchten sich zunächst einen Überblick über das Fachgebiet verschaffen und suchen dazu im System **EconBiz** nach einschlägigen Dokumenten.

- Entwerfen Sie eine **Blockstrategie**, indem Sie in die folgende Tabelle Begriffe eintragen, die Sie im STW – Standard Thesaurus Wirtschaft finden und die Ihnen logisch erscheinen.
- Notieren Sie eine **Begründung** für Ihre Auswahl an Begriffen.
- Wir würde eine Anfrage zu diesen Suchbegriffen in einer **booleschen Anfrage** aussehen?
- Setzen Sie Ihre Suche in EconBiz entsprechend der Anleitung unter https://www.econbiz.de/eb/de/hilfe-erweiterte-suche/ um.

Link zum Thesaurus: https://zbw.eu/stw/version/latest/about.de.html

Link zum Suchformular in EconBiz: https://www.econbiz.de/Search/Advanced

	Begriff 1	**Begriff 2**
	Aktienkauf	Altersvorsorge
Synonyme		
Verwandte Begriffe		

II Evaluation der Query Expansion

Die Ergebnisse aus Aufgabe I sollen sie nun mit Hilfe den Ihnen bekannten IR-Evaluationsmaßen mit einer einfachen Baseline vergleichen werden. Die Baseline entspricht hierbei der Suche in EconBiz nach dem Thema „Aktienkauf zur Altersvorsorge“, ohne besondere Verfeinerungen oder Erweiterung: https://www.econbiz.de/Search/Results?lookfor=Aktienkauf+zur+Altersvorsorge

Bewerten Sie die Ergebnisse bis zur Tiefe k=10 nach Relevanz. Nutzen Sie hierfür eine binäre Bewertung und lassen Sie alle Einträge von mindestens zwei Personen bewerten. Im Zweifel entscheidet die Mehrheit durch Hinzuziehen einer dritten Person.
Nutzen Sie die Relevanzbewertungen, um die folgenden Kennzahlen zu ermitteln:

- Average Precision (AP)
- Precision at 5 (P@5)
- Precision at 10 (P@10)

Diskutieren Sie kurz, welches der beiden Suchsysteme (mit oder ohne Query Expansion) bessere Ergebnisse liefert.

III Zusatzaufgabe für DIS: Normalized Google Distance

Die Normalized Google Distance[1] (NGD) 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.

Im Folgenden soll die Ähnlichkeit von Begriffen anhand der Normalized Google Distance bewertet werden. Zur Auswertung steht Ihnen dazu ein Google Colab Notebook zur Verfügung.

Link zum Colab Notebook:
https://colab.research.google.com/drive/1A7Lc7L7rTS9W6TkmKFewojQd7D068hkz?usp=sharing

In den letzten beiden Zellen können Sie Begriffe spezifizieren, für welche dann die Ähnlichkeit ausgewertet wird. Für diese Aufgabe stehen Ihnen die Suchmaschinen Google und Wikipedia bereit.

(1) Werten Sie die Ähnlichkeit der Begriffspaare aus, die Sie in die obige Tabelle eingetragen haben.

(2) Welche Begriffe ähneln sich am stärksten? Wie wirkt sich die semantische Ähnlichkeit auf die Ergebnisse einer Suche aus?

(3) Wie unterscheidet sich die Normalized Google Distance zwischen den verwendeten Suchmaschinen? Wodurch kommen die Unterschiede zwischen den Suchmaschinen zustande?

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