

## Vorlesung Information Retrieval, WS 2006/07 Übungsblatt 1, Abgabe am 30.10.06

### Aufgabe 1 (Vektorraummodell)

Gegeben seien vier Textabschnitte aus der Rede zum Schuljahresabschluß am 29. September 1809 von Georg Wilhelm Friedrich Hegel. Um Zeit zu sparen können Sie auch Nullwerte weglassen und Logarithmen symbolisch schreiben, d.h.  $\log 5 = 0.69$ .

**Teilaufgabe a:** Bestimmen Sie zu folgendem Vokabular: {Bildung, Boden, Erziehung, Kinder, Menschen, Nützlich, Wesentlich} die Beschreibung der Texte im Vektorraummodell. Zählen Sie bei dem Auftreten von Termen auch Zusammensetzungen (z.B. Erziehung  $\leq$  Erziehungsanstalten) und Worte mit gleichem Wortstamm mit. Gegeben Sie alle Zwischenergebnisse mit an, d.h. Häufigkeiten, normalisierte Häufigkeiten, IDF und Gewichte. Benutzen Sie den Logarithmus zur Basis 10.

**Teilaufgabe b:** Bestimmen Sie für folgende Anfragen die Gewichtsvektoren und die Ähnlichkeitswerte zu allen Texten bezüglich des Vektorraummodells:

- $q_1 = \{ \text{Bildung, Kinder, Wesentlich} \}$
- $q_2 = \{ \text{Erziehung, Menschen, Nützlich, Wesentlich} \}$
- $q_3 = \{ \text{Bildung, Boden, Menschen} \}$

#### Text 1

Es sind zwei Zweige der Staatsverwaltung, für deren gute Einrichtung die Völker am erkenntlichsten zu sein pflegen, gute Gerechtigkeitspflege und gute Erziehungsanstalten; denn von keinem übersieht und fühlt der Mensch die Vorteile und Wirkungen so unmittelbar, nah und einzeln als von jenen Zweigen, deren der eine sein Privateigentum überhaupt, der andere aber sein liebstes Eigentum, seine Kinder, betrifft.

#### Text 2

Der Geist und Zweck unserer Erziehungsanstalt ist die Vorbereitung zum gelehrten Studium, und zwar eine Vorbereitung, welche auf den Grund der Griechen und Römer erbaut ist. Seit einigen Jahrtausenden ist dies der Boden, auf dem alle Kultur gestanden hat, aus dem sie hervorgesprosst und mit dem sie in beständigem Zusammenhange gewesen ist. Wie die natürlichen Organisationen, Pflanzen und Tiere, sich der Schwere entwinden, aber dieses Element ihres Wesens nicht verlassen können, so ist alle Kunst und Wissenschaft jenem Boden entwachsen; und obgleich auch in sich selbstständig geworden, hat sie sich von der Erinnerung jener älteren Bildung nicht befreit. Wie Anteus seine Kräfte durch die Berührung der mütterlichen Erde erneuerte, so hat jeder neue Aufschwung und Bekräftigung der Wissenschaft und Bildung sich aus der Rückkehr zum Altertum ans Licht gehoben.

#### Text 3

So wichtig aber die Erhaltung dieses Bodens ist, so wesentlich ist die Abänderung des Verhältnisses, in welchem er ehemals gestanden hat. Wenn die Einsicht in das Ungenügende, Nachteilige alter Grundsätze und Einrichtungen überhaupt und damit der mit ihnen verbundenen vorigen Bildungszwecke und Bildungsmittel eintritt, so ist der Gedanke, der sich zunächst auf der Oberfläche darbietet, die gänzliche Beseitigung und Abschaffung derselben. Aber die Weisheit der Regierung, erhaben über diese leicht scheinende Hilfe, erfüllt auf die wahrhafteste Art das Bedürfnis der

Zeit dadurch, daß sie das Alte in ein neues Verhältnis zu dem Ganzen setzt und dadurch das Wesentliche derselben ebenso sehr erhält, als sie es verändert und erneuert.

#### Text 4

Erstlich hat allerhöchste Regierung durch die Vervollkommnung der deutschen Volksschulen die allgemeine Bürgerbildung erweitert, es werden dadurch allen die Mittel verschafft, das ihnen als Menschen Wesentliche und für ihren Stand Nützliche zu erlernen; denen, die das Bessere bisher entbehrten, wird dasselbe hierdurch gewährt; denen aber, die, um etwas Besseres als den ungenügenden allgemeinen Unterricht zu erhalten, nur zu dem genannten Bildungsmittel greifen konnten, wird dasselbe entbehrlicher gemacht und durch zweckmäßigere Kenntnisse und Fertigkeiten ersetzt. Auch die hiesige Stadt sieht der vollständigen Organisation dieser dem größten Teil des übrigen Königreichs bereits erwiesenen Wohltat, erwartungsvoll entgegen -- einer Wohltat, deren wichtige Folgen für das Ganze kaum zu berechnen sind.

#### **Aufgabe 2 (Recall & Precision)**

Gegeben sind die Ergebnisse der folgenden Anfragen durch eine sortierte Reihenfolge der ersten 15 gefundenen Dokument-IDs. Ebenfalls sind die durch Experten ermittelten relevanten Dokumente zu jeder Anfrage gegeben. Um Zeit zu sparen können Sie Brüchen schreiben. Bestimmen Sie zu jeder Anfrage

**Teilaufgabe a** die nicht-standardisierten Werte für Recall und Precision, **Teilaufgabe b** konvertieren diese dann zu Standard-Recall-Precision Werten

- Ergebnis 1: 11, 2, 7, 8, 4, 9, 18, 207, 40, 16, 14, 6, 5, 10  
Expertenauswertung 1: 2, 4, 5, 7, 16
- Ergebnis 2: 55, 24, 62, 63, 53, 59, 52, 93, 20, 66, 18, 94, 71, 58, 81  
Expertenauswertung 2: 18, 20, 52, 53, 55, 58

#### **Aufgabe 3 (Soundex)**

Ermitteln Sie zu den gegebenen Namen den Soundex Kode:

{ Bayer, Grimmett, Strizaker, Borwein, Biggs, Flajolet, Agarwal }

#### **Aufgabe 4 (Lemur System)**

Installieren Sie sich das Lemursystem von <http://www.lemurproject.org/> . Lesen Sie die Dokumentation um das Programmpacket kennenzulernen. Für praktische Übungen können Sie die MED-Daten von der Vorlesungsseite herunterladen. Versuchen Sie eine invertierte Liste zu erzeugen und das System zu evaluieren. Auf der Vorlesungsseite sind auch Anfragen mit entsprechenden relevanten Antwortmengen bereitgestellt.