Gruppe 1: Vincent Schuster, Maximilian Raspe

Praktikum 1

Aufgabe 1 – Grundlagen

1.)
Beispiel – Die rote Ampel
Daten:

- syntaktisch: Farbe Rot, leuchtender Kreis
- definierte Verfahren der Datenverarbeitung

Wissen:

- semantisch: Stopp! Halt!
- begründete Verfahren der Wissensrepräsentation

Informationen:

- pragmatisch: Empfänger des Zeichens soll anhalten!
- kontrollierte Informationsverarbeitung zur informationellen Handlungsabsicherung

Die Daten sind in diesem Beispiel das gesamte gegebene Angebot, also „1 Glas Nutella“, „400 Gramm“, „nur 99 Cent“. Anhand dieser Daten können wir Wissen interpretieren, nämlich, dass ein Glas 400 Gramm Nutella nur 99 Cent kosten. Die Informationen daraus wiederum sind, dass 1 Glas Nutella preislich reduziert ist, man also die Nutella zu einem niedrigeren Preis als normalerweise kaufen kann.

2.)
Beispiele für real existierende IR-Systeme:

- Suche nach Textinhalten im WWW (Google, Yahoo!)
    - sowohl Suche nach einzelnem Begriff, als auch nach Phrase / Sätzen in textuellen Dokumenten
    - Aufgabe: Anfragebearbeitung, Browsing, Information Filtering
- Suche nach Bildern im WWW
    - Aufgabe: Anfragebearbeitung
- elektronische Kataloge (Yahoo)
    - Dokumente in Klassifikationsschema eingeordnet
    - Aufgabe: Klassifikation

Aufgabe 2 – Evaluierung

1.)

- Evaluierung von Datenbank-Systemen
  - Funktionsumfang des Systems
    - unterstützte SQL-Standard
    - Vorhandensein erweiterter Funktionalitäten
  - Performance (Laufzeit- und Speicherplatzeffizienz)
    - Ergebnis muss bei allen Systemen gleich sein
    - Systeme mit falschen Ergebnissen werden vom Vergleich ausgeschlossen
- Evaluierung von IR-Systemen
  - Funktionsumfang und Performance ebenfalls von Interesse
  - aber: dominierendes Kriterium ist Qualität der Ergebnisse
    - es kann nicht davon ausgegangen werden, dass sie auf Anfrage gleichwertige oder gar gleiche Antworten liefern
    - jedes System kann für gleiche Anfrage unterschiedliche Dokumente relevant halten

    -> möglichst objektive Beurteilung der Ergebnisse

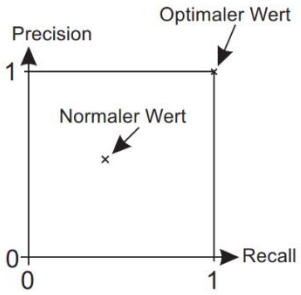
2.)

- Relevanz
  - Relevanz eines Dokuments auf eine Anfrage
  - lineare Skala: 0 bis 10
  - binär: relevant oder nicht relevant
  - R_q bezeichnet die Menge der zur Anfrage q relevanten Dokuimente in der Dokumentenkollektion
- Recall
  - drückt Fähigkeit eines IR-Systemns aus, relevante Dokumente in Ergebnismenge zu liefern
  - Wie hoch ist der Anteil der relevanten Dokumenten im Ergebnis im Verhältnis zu allen relevanten Dokumenten in der Dokumentenkollektion?
  - 80 % Recall = unter gefundenen Dokumenten befinden sich 80 % aller verfügbaren relevanten Dokumente (100 % wäre maximal)
- Precision
  - Quote der gefundenen relevanten Dokumente unter allen gefundenen Dokumenten
  - Wie gut ist IR-System in der Lage, irrelevante Dokumente in Ergebnismenge zu vermeiden?
  - 40 % Precision = 60 % der Dokumente im Ergebnis sind nicht relevant

3.)

$$\text{Recall} = \frac{\text{gefundene relevante Dokumente}}{\text{relevante Dokumente insgesamt}}$$

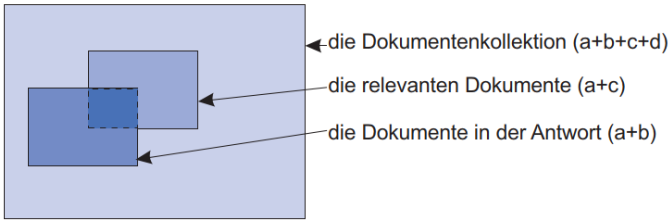
$$\text{Precision} = \frac{\text{gefundene relevante Dokumente}}{\text{gefundene Dokumente insgesamt}}$$



	relevant	nicht relevant	Σ
im Ergebnis	a (Hits)	b (Noise)	$a+b$
nicht im Ergebnis	c (Misses)	d (Rejected)	$c+d$
Σ	$a+c$	$b+d$	$a+b+c+d$

$$Recall = \frac{a}{a+c} = \frac{\text{Anzahl der relevanten Dokumente im Ergebnis}}{\text{Gesamtzahl der relevanten Dokumente}}$$

$$Precision = \frac{a}{a+b} = \frac{\text{Anzahl der relevanten Dokumente im Ergebnis}}{\text{Gesamtzahl der Dokumente im Ergebnis}}$$



a (hits) relevant und auch im Ergebnis [korrekt]

b (noise) nicht relevant und trotzdem im Ergebnis [falsch]

c (misses) relevant und trotzdem nicht im Ergebnis [falsch]

d (rejected) nicht relevant und auch nicht im Ergebnis [korrekt]

4.)

Ansätze zur Mittelwertbildung:

- Makrobewertung
  - nutzerorientierter Ansatz
  - Ermittlung von wirklichen Durchschnittswerten
    (Summe der Einzelwerte wird durch Anzahl Experimente geteilt)
  - alle Experimente gehen unabhängig von Ergebnisgröße mit gleichem Gewicht in Durchschnittsberechnung ein
  - Problem: wenn einzelne Anfragen leere Ergebnisse liefern, kann es zur Division durch 0 kommen

$$Recall_{\varnothing u} = \frac{1}{m} \cdot \sum_{i=1}^{m} \frac{a_i}{a_i + c_i} \qquad Precision_{\varnothing u} = \frac{1}{m} \cdot \sum_{i=1}^{m} \frac{a_i}{a_i + b_i}$$

- Mikrobewertung
  - systemorientierte Sichtweise
  - Summe der einzelnen Suchanfragen werden als große Anfrage betrachtet
  - Anfragen mit größerer Ergebnismenge werden im Gesamtergebnis stärker gewichtet

$$Recall_{\varnothing 2} = \frac{\sum_{i=1}^{m} a_i}{\sum_{i=1}^{m} (a_i + c_i)} \qquad Precision_{\varnothing 2} = \frac{\sum_{i=1}^{m} a_i}{\sum_{i=1}^{m} (a_i + b_i)}$$

5.)

a – relevante Dokumente im Ergebnis (Hits)
b – nicht relevante Dokumente im Ergebnis (Noise)
c – nicht gefundene, relevante Dokumente (Misses)

a)

Anfrage	a+b	a+c	a	Recall	Precision
1	27	12	4	0,33	0,15
2	31	13	6	0,46	0,19
3	58	28	13	0,46	0,22
4	70	23	15	0,65	0,21
Summe	**186**	**76**	**38**	**1,90**	**0,77**

Makrobewertung:

Recall = 0,25 * 1,90 = 0,48
Precision = 0,25 * 0,77 = 0,19

Mikrobewertung

Recall = 38/76 = 0,50
Precision = 38/186 0,20

b)

Anfrage	a+b	a+c	a	Recall	Precision
1	80	5	1	0,20	0,01
2	80	21	7	0,33	0,09
3	80	26	11	0,42	0,14
4	80	31	19	0,61	0,24
Summe	**320**	**83**	**38**	**1,56**	**0,48**

Makrobewertung:

Recall = 0,25 * 1,56 = 0,39
Precision = 0,25 * 0,48 = 0,12

Mikrobewertung

Recall = 38/83 = 0,46
Precision = 38/320 = 0,12

6.)

Bestimmung des Recalls:

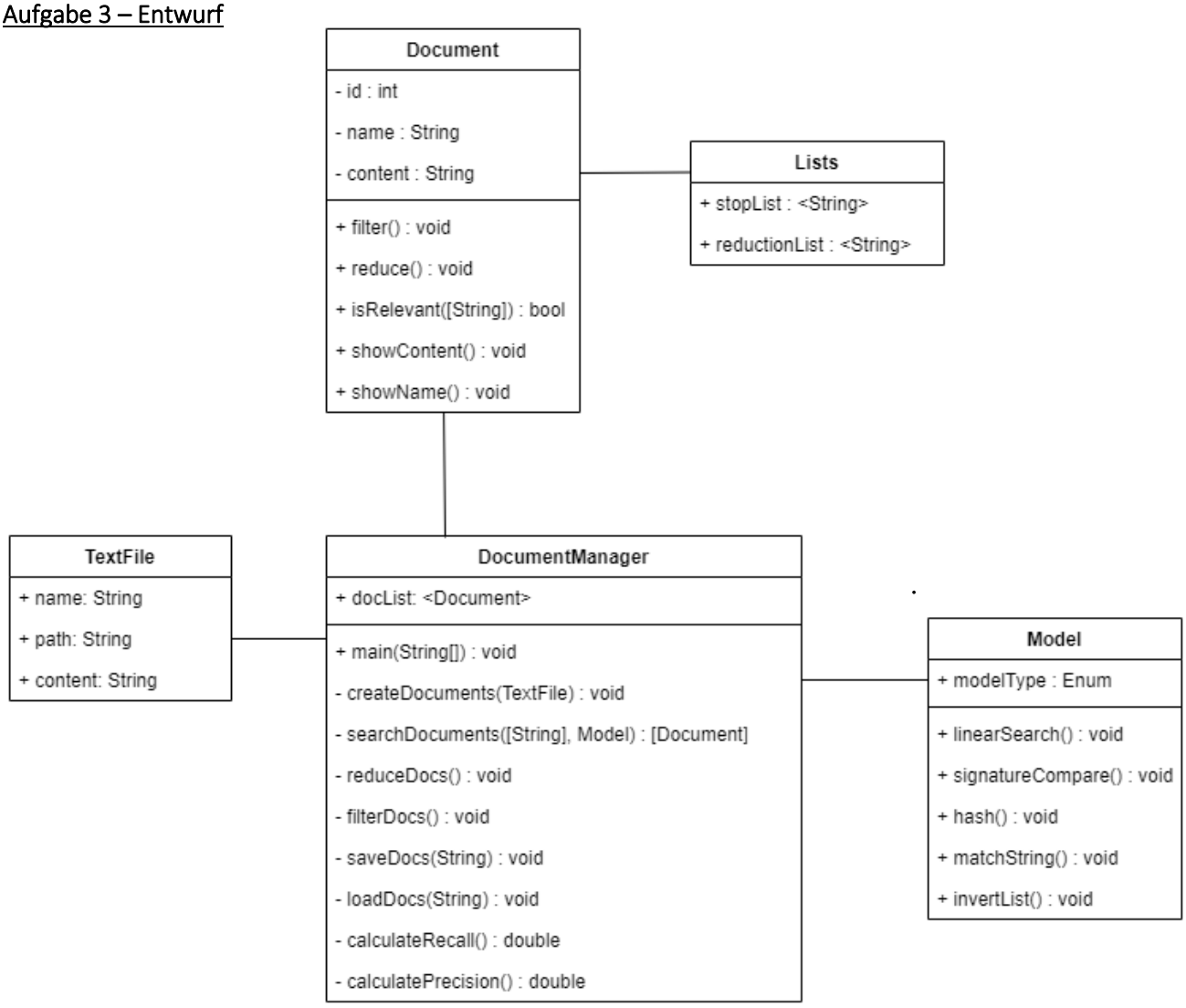
- vollständige Relevanzbeurteilung einer repräsentativen Stichprobe
    - Problem: Anteil der relevanten Dokumente im Allgemeinen sehr gering

  -> Stichprobe muss sehr groß sein, was zu großem Aufwand führt

$$Recall = \frac{a}{a+c} = \frac{\text{Anzahl der relevanten Dokumente im Ergebnis}}{\text{Gesamtzahl der relevanten Dokumente}}$$

- Dokument-Source-Methode
    - wählt zufälliges Dokument aus Datenbasis
    - formuliert Anfrage, zu der Dokument tatsächlich relevant ist
    - prüft, ob Dokument in Antwort ist
    - über mehrere Dokumente und Anfragen gemittelt ergibt sich Wert für Recall
    - Problem: es handelt sich nicht um „echte“ Benutzerfragen
- Anfrageerweiterung
    - erweitert Anfrage so, dass Obermenge der ursprünglichen Antwortmenge gefunden wird
        - z.B. mehrere Frageformulierungen von verschiedenen Bearbeitern
    - Problem: man erhält i.A. nur eine Teilmenge der relevanten Dokumente
    - Folge: Recall-Schätzungen werden i.A. zu hoch sein
- Abgleich mit externen Quellen
    - parallel mit unabhängigen Methoden relevante Dokumente bestimmen
    - Fragenden oder andere Fachleute befragen, welche relevanten Dokumente sie kennen

Aufgabe 3 – Entwurf



Erläuterungen:

Der DocumentManager erstellt durch die Funktion createDocuments() aus einer vorgegebenen Textdatei eine Dokumentenkollektion, nämlich die docList.
Durch die Funktionen filter() und reduce() in Document können einzelne Dokumente Stoppworte elemieren und eine Grund-/Stammformreduktion durchführen, indem sie die Listen aus Lists benutzen.
Mit searchDocuments() kann die Dokumentenkollektion durchsucht werden. Dabei kann ein einzelner oder mehrere Suchterme angegeben werden, mitsamt des zu verwendenden Retrieval-Modells, welches in der Klasse Model implementiert ist.
Die textuelle Benutzeroberfläche ist in main() realisiert, in der man dann die einzelnen Funktionen des DocumentManagers aufrufen kann.
Die Dokumente können durch showContent() und showName() nach beliebigen Bearbeitungsschritten ausgegeben werden. Ebenso können die Dokumente durch loadDocs() und saveDocs() gespeichert, bzw. geladen werden.
Recall und Precision können durch calculateRecall() und calculatePrecision() berechnet werden.