Praktikum 1

Hinweise zur Abgabe

- Bitte geben Sie Ihre Lösung als PDF-Datei über Moodle ab. Die Deadline für die Abgabe finden Sie ebenfalls dort.
- Benennen Sie die Datei wie folgt: gruppe_n_pm.pdf, wobei n Ihre Gruppennummer ist und m die Nummer des Praktikums.

Aufgabe 1 - Grundlagen

1. Sie schauen in das Verkaufsprospekt eines bekannten Discounters, darin finden Sie das folgende Angebot:

   1 Glas Nutella, 400 Gramm, nur 99 Cent

   Erklären Sie den Unterschied zwischen den Begriffen *Daten*, *Wissen* und *Information* möglichst exakt anhand dieses Beispiels.

2. Geben Sie drei Beispiele für real existierende IR-Systeme. Welche Arten von Dokumenten können in diesen Systemen nach welchen Kriterien durchsucht werden? Ordnen Sie die Systeme den kennengelernten *Aufgabenstellungen von IR-Systemen* zu.

Aufgabe 2 - Evaluierung

1. Beschreiben Sie kurz, inwiefern sich die Evaluierung von IR-System von der Evaluierung von Datenbank-Systemen unterscheidet!
2. Beschreiben Sie kurz, was man unter den Begriffen *Relevanz*, *Precision* und *Recall* versteht!
3. Wie sind Precision und Recall mathematisch definiert? Visualisieren Sie die nötigen Kategorien, in welchen sich die Dokumente befinden können (relevant/nicht relevant, im Ergebnis/nicht im Ergebnis) in einem Euler-Venn-Diagramm!
4. Beim Vergleich verschiedener IR-Systeme existieren *zwei Ansätze zur Mittelwertbildung* über mehrere Anfragen hinweg. Beschreiben Sie die zwei unterschiedlichen Ansätze kurz mit eigenen Worten!
5. Bestimmen Sie mit den zwei Ansätzen zur Mittelwertbildung jeweils *Recall* und *Precision* für die im Folgenden aufgeführten Anfragen (auf 2 Nachkommastellen genau). Geben Sie die Rechenwege mit den einzelnen Formeln und Zwischenergebnissen an!

   a Relevante Dokumente im Ergebnis (Hits)

b Nicht relevante Dokumente im Ergebnis (Noise)

c Nicht gefundene, relevante Dokumente (Misses)

(a)

Anfrage	Alle im Erg. (a+b)	a+c	a
1	27	12	4
2	31	13	6
3	58	28	13
4	70	23	15

(b)

Anfrage	Alle im Erg. (a+b)	a+c	a
1	80	5	1
2	80	21	7
3	80	26	11
4	80	31	19

6. Wie bestimmt man in der Praxis den Recall? Erläutern Sie die vier in der Vorlesung kennengelernten Methoden kurz!

Aufgabe 3 - Entwurf

1. Ziel der Praktikumsblätter 2–4 ist die Entwicklung eines einfachen Information-Retrieval-Systems. Erstellen Sie einen vorläufigen Entwurf des Systems mithilfe eines UML-Klassen-Diagramms und erläutern Sie ihn kurz schriftlich. Das System soll mindestens über folgende grundlegende Funktionalitäten verfügen:

   - Erstellung einer Dokumentenkollektion aus einer vorgegebenen Textdatei (durch Zerlegung der Datei in einzelne Dokumente).
   - Eliminierung von Stoppworten sowie Grund-/Stammformreduktion in Dokumenten.
   - Durchsuchen der Dokumentenkollektion anhand eines einzelnen Suchterms oder mehrerer miteinander verknüpfter Suchterme und Ausgabe der Suchergebnisse.
   - Benutzeroberfläche (grafisch oder textuell) zum Starten der einzelnen Funktionen (Zerlegung, Stoppworteliminierung, Stammformreduktion, Suche, ...).
   - Unterstützung unterschiedlicher Retrieval-Modelle für die Suche (boolesches Modell und Vektorraummodell).
   - Unterstützung unterschiedlicher Implementierungen eines Retrieval-Modells (z.B. lineare Suche, invertierte Liste und Signaturverfahren für boolesches Modell).
   - Möglichkeit zur textuellen Ausgabe eines Dokuments nach unterschiedlichen Verarbeitungsschritten (Originaldokument; nach Stoppworteliminierung; nach Grund-/Stammformreduktion; nur Titel/Dateiname).
   - Speicherung der Dokumente (inkl. der einzelnen Verarbeitungsschritte) auf der Festplatte und Laden dieser Daten von der Festplatte nach Neustart des Programms.

- Automatische Ermittlung der Kennzahlen Recall und Precision für vorgegebene Testanfragen.