Midterm Probeklausur von 16.12.2010

Hilfsmittel: Taschenrechner

Bearbeitungszeit: 25 Minuten

Viel Erfolg!

I. Web Information Retrieval (20 Punkte)

1. Gegeben seien zwei Text-Dokumente ($N = 2$). [5+5=10 P.]

Dokument 1:
Morgen, morgen, nur nicht heute, sagen alle faulen Leute.
Dokument 2:
Herr A. sagt: „Was du heute kannst besorgen, das verschiebe nicht auf morgen“.

a. Erzeugen Sie aus den Dokumenten einen invertierten Index. Tokenisierung-Regeln: Wort-weise (Satzzeichen ignorieren), alles klein geschrieben. Geben Sie dabei jeweils an passender Stelle für jeden Term die entsprechenden Werte für TF und DF an.

b. Die Funktion für eine Suchanfrage-Dokument-Relevanz sei

$$w_{Q,d} = \sum_{q \in Q,d} \left(1 + \log\left(TF_{q,d}\right)\right) * log \frac{N}{DF_q}$$

Berechnen Sie die jeweilige Relevanz für die beiden Dokumente im Bezug auf die beiden folgenden Suchanfragen:

Q_1= *Klausur heute*
Q_2= *morgen sagen*

Erläutern Sie die Ergebnisse!

2. Von einer Dokumentensammlung mit insgesamt 50 Dokumenten seien 25% Klausurrelevant. Für die Suchanfrage „Klausur heute“ liefert die Suchmaschine folgende Reihe von relevanten (T) sowie nicht relevanten (F) Suchergebnissen:

TFFFF TFFTT

Berechnen Sie die Top-10 Precision, Recall sowie das F_1-Maß des Suchsystems. Zur Erinnerung: $F_\beta = \frac{(\beta^2+1)PR}{\beta^2 P+R}$ [5 P.]

3. Gegeben sei die Relevanz-Feedback-Formel: [5 P.]

$$\overrightarrow{q_m} = \propto \overrightarrow{q_0} + \beta \frac{1}{|D_r|} \sum_{\overrightarrow{d_j} \in D_r} \overrightarrow{d_j} - \gamma \frac{1}{|D_{nr}|} \sum_{\overrightarrow{d_j} \in D_{nr}} \overrightarrow{d_j}$$

Der Benutzer hat das Dokument
D_1=„*Morgen, morgen, nur nicht heute!*“
als relevant für die Suchanfrage „*Klausur heute*“ eingestuft.

Nennen Sie die Suchanfrage, die von dem System im Rahmen des Relevanz-Feedback Algorithmus mit α=1, β=0,5, γ=0 als nächstes ausgeführt wird. Geben Sie für jedes Wort der Suchanfrage die Gewichtung an.

II. Klassifikation (5 Punkte)

1. Gegeben seien die folgenden Dokumente: [5 P.]

Dokument 1:
Herr A. hat gesagt: „Morgenstund hat Gold im Mund“.

Dokument 2:
Es ist nicht alles Gold, was glänzt.

Gegeben sei ein naives Bayes-Klassifikator mit zwei Klassen c_1 und c_2.
$P(c_1)$=0,45; $P(t_x|c_1)$ = 0,01

$P(c_2)$=0,55; $P(\text{„alles“}|c_2)$=0,7; $P(t_y \neq \text{“alles“}|c_2)$=0,01

Ordnen Sie die beiden Dokumente den entsprechenden Klassen zu.
Erläutern Sie die Ergebnisse!