Semestralklausur zur Vorlesung

Web Mining

Prof. J. Fürnkranz
Technische Universität Darmstadt — Sommersemester 2006
Termin: 26. 7. 2006

Name: **Vorname:** **Matrikelnummer:**

Fachrichtung:

Punkte: (1) .... (2) .... (3) .... (4) .... (5) .... **Summe:**

Wichtig!

- **Aufgaben:** Diese Klausur enthält auf den folgenden Seiten 5 Aufgaben zu insgesamt 100 Punkten. Jede Aufgabe steht auf einem eigenen Blatt. Kontrollieren Sie *sofort*, ob Sie alle sechs Blätter erhalten haben!
- **Zeiteinteilung:** Die Zeit ist knapp bemessen. Wir empfehlen Ihnen, daß Sie sich zuerst einen kurzen Überblick über die Aufgabenstellungen verschaffen, und dann mit den Aufgaben beginnen, die Ihnen am meisten liegen.
- **Papier:** Verwenden Sie nur Papier, das Sie von uns ausgeteilt bekommen. Sie können Ihre Lösungen beliebig auf die sechs Blätter verteilen, solange klar ersichtlich ist, welche Lösung zu welcher Aufgabe gehört. Sollten sich allerdings mehrere Lösungen zu derselben Aufgabe finden, suchen wir uns eine aus. Insbesondere können Sie auch auf den Rückseiten schreiben!
  Brauchen Sie zusätzlich Papier (auch Schmierpaper), bitte melden.
- **Hilfsmittel:** Zur Lösung der Aufgaben sind *keine* Hilfsmittel zulässig, ausgenommen gedruckte Wörterbücher für ausländische Studenten. Elektronische Wörterbücher sind verboten!
- **Fragen:** Sollten Sie Teile der Aufgabenstellung nicht verstehen, bitte fragen Sie!
- **Abschreiben:** Sollte es sich (wie in den letzten Jahren leider immer wieder) herausstellen, daß Ihre Lösung und die eines Kommilitonen über das zu erwartende Maß hinaus übereinstimmen, werden beide Arbeiten negativ beurteilt (ganz egal wer von wem und egal in welchem Umfang abgeschrieben hat).
- **Ausweis:** Legen Sie Ihren *Studentenausweis* sichtbar auf Ihren Platz.
- **Aufräumen:** Sonst darf außer Schreibgerät, Essbarem, von uns ausgeteiltem Papier und eventuell Wörterbüchern nichts auf Ihrem Platz liegen. Taschen bitte unter den Tisch! Wer bei diesen Temperaturen einen Mantel mithat, wird gebeten, ihn anzubehalten.

Gutes Gelingen!

Aufgabe 1 22 Punkte (5/3/4/3/4/3)

1–a Welcher Lernalgorithmus ist bezüglich Gesamt-Laufzeit (Trainieren und Testen von je n Dokumenten) am effizientesten? Welcher beim Trainieren? Geben Sie eine kurze Begründung für beide Antworten.

- Naive Bayes
- k-Nearest Neighbor classifier
- Support Vector Machine mit linearem Kernel

1–b Erklären Sie kurz den Unterschied zwischen Model-Based und Memory-Based Collaborative Filtering.

1–c Sie möchten den Authority-Score Ihrer Home-Page erhöhen, und setzen daher an vielen, mehr oder weniger zufällig ausgewählten Stellen im Web Links auf Ihre Home-Page (analog zum Link-Spamming bei PageRank). Würde das funktionieren? Begründung?

1–d Eine der Methoden zur Hypertext-Klassifikation, die wir in der Vorlesung kennen gelernt haben, schlägt vor, alle Vorgänger-Dokumente eines Dokuments zu einem Dokument zusammenzusetzen. Funktioniert dieser Ansatz gut oder schlecht? Warum?

1–e Die Verwendung von linguistischen Phrasen und von n-Grammen verfolgt ähnliche Ziele, da man in beiden Fällen versucht, längere Wortsequenzen als Features zu definieren. Welches Verfahren würden Sie eher für eine Klassifikationsaufgabe im Web einsetzen? Warum?

1–f Skizzieren Sie kurz eine Situation, in der ein HLRT-Wrapper versagt, aber SoftMealy erfolgreich angewendet werden könnte.

Aufgabe 2 24 Punkte (4/4/4/4/4/4)

Sie erhalten die Aufgabe, ein Nachrichten-Portal zu implementieren.

2–a Sie erhalten eine Liste von URLs, die Homepages von großen Nachrichtenagenturen darstellen (z.B. `http://www.cnn.com`, `http://www.spiegel.de`, usw.). Von diesen muß eine lokale Kopie täglich auf dem neuesten Stand gehalten werden. Skizzieren Sie kurz die Probleme, die Sie dabei lösen müssen.

2–b Die archivierten Seiten sollen in ein einheitliches Format gebracht werden, d.h. Sie müssen in jeder Quelle die Nachrichtentexte identifizieren können, und zumindest den zugehörigen Titel erkennen können. Nennen Sie eine konkrete Methode, die Sie hier einsetzen können und begründen Sie Ihre Wahl.

2–c Ein wichtiger Schritt ist die Realisierung eines Moduls zur automatischen Sortierung der Nachrichtentexte in verschiedende Ressorts (Politik, Sport, Wissenschaft, ...). Eine derartige Beschlagwortung ergibt sich bei einigen Sites automatisch (z.B. durch die Struktur des URLs), bei anderen muß sie das System selbst vornehmen, da eine manuelle Zuordnung zu teuer wäre. Wie würden Sie vorgehen?

2–d Sie stellen fest, daß einige Sites Ihre Nachrichtentexte zweisprachig anbieten, d.h. Sie können denselben Text gleichzeitig auf Englisch und auf Deutsch erhalten. Wie könnten Sie diese Information nützen, um eventuell eine bessere Performanz in dem in Punkt c. skizzierten Modul zu erreichen?

2–e Innerhalb der Ressorts sollen Sie erkennen, welche Nachrichtentexte sich auf das gleiche Ereignis beziehen, und unter diesen einen repräsentativen Text auswählen. Wie würden Sie vorgehen?

2–f Die Anforderung des Auftraggebers ist, daß zumindest 95% aller Nachrichtentexte dem korrekten Ressort zugeordnet werden müssen. Wie überprüfen Sie, ob Ihr System diese Anforderung erfüllt?

Aufgabe 3 18 Punkte (4/3/4/3/4)

Sie evaluieren Ihre Suchmaschine, indem Sie die Anzahl der zurückgegebenen Dokumente für eine Query variieren. Dabei erhalten Sie folgende Paare von Recall- und Precision-Werten:

Precision	76%	80%	80%	85%	88%	90%	70%	45%
Recall	5%	15%	25%	35%	55%	75%	85%	100%

3–a Wie beurteilen Sie die Qualität der Reihung der Suchergebnisse?

3–b Geben Sie den Breakeven Point an.

3–c Geben Sie die interpolierten Precision Werte für die Recall-Werte 0, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100 an.

3–d Berechnen Sie die 11-point Average Precision.

3–e Den dritten Punkt in der Tabelle (25% Recall und 80% Precision) erhielten Sie für 1000 zurückgegebene Dokumente. Wie viele relevante Dokumente gab es insgesamt?

Aufgabe 4 20 Punkte (7/5/4/4)

Gegeben seien die folgenden kurzen Dokumente (jede Zeile ist ein Dokument) mit ihrer entsprechenden Klassenzugehörigkeit.

braun schwarz gelb	+
blau braun schwarz orange	+
braun schwarz gelb	+
gelb braun	+
blau grün gelb	−
grün schwarz orange	−
blau grün gelb	−
grün gelb orange	−

4–a Schätzen Sie mit Laplace-Korrektur alle Wahrscheinlichkeiten, die Sie brauchen, um auf diesen Daten einen Naive Bayes Klassifizierer zu definieren. Sie können als Lösung selbstverständlich Bruchzahlen angeben.

4–b Welcher Klasse würde der Naive Bayes Klassifizierer folgendes Dokument zuordnen? Der Rechengang muß aus Ihrer Antwort ersichtlich sein.

grün braun orange blau	?

4–c Welches Problem wäre bei der Klassifikation des Beispiels aus Aufgabe b aufgetreten, wenn Sie die Wahrscheinlichkeiten mit relativen Häufigkeiten (ohne Laplace-Korrektur) geschätzt hätten? Zeigen Sie das Problem speziell an diesem Beispiel.

4–d Welche Wörter würden von einem supervised Filtering-Ansatz zur Feature Subset Selection als wichtig angesehen werden und welche nicht? Nennen Sie zumindest je ein Beispiel für ein wichtiges und ein unwichtiges Wort. Woran können Sie das erkennen?

Aufgabe 5 16 Punkte (3/8/5)

Gegeben sei folgende (symmetrische) Matrix von paarweisen Ähnlichkeiten zwischen Dokumenten.

	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8	d_9
d_1	—	0.90	0.81	0.50	0.43	0.67	0.35	0.38	0.27
d_2		—	0.85	0.49	0.52	0.43	0.23	0.22	0.33
d_3			—	0.65	0.42	0.50	0.34	0.24	0.24
d_4				—	0.95	0.88	0.40	0.34	0.39
d_5					—	0.87	0.24	0.45	0.33
d_6						—	0.33	0.37	0.31
d_7							—	0.86	0.75
d_8								—	0.80
d_9									—

Hinweis: Sie können sich vorstellen, daß es sich dabei um Werte der *cos*-Ähnlichkeit handelt, also ein Wert von 1.0 bedeutet maximale Ähnlichkeit/minimale Distanz und ein Wert von 0.0 bedeutet minimale Ähnlichkeit/maximale Distanz.

5–a Gegeben seien die Cluster $\{d_1, d_2\}$ und $\{d_8, d_9\}$, berechnen Sie die Distanz zwischen diesen beiden Clustern nach der *Average Link* Methode.

5–b Vollziehen Sie ein hierarchisches agglomeratives Clustering mit der *Single Link* Methode. Zeichnen Sie die hierarchische Struktur und geben Sie auch die Ähnlichkeiten zwischen den Clustern in der Struktur an.

5–c Auf diesen Daten wird ein Klassifikationsproblem definiert, wobei die Klassenzuordnungen wie folgt definiert werden:

d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8	d_9
+	−	+	−	+	−	+	−	+

- Würden Sie hier die Verwendung eines k-NN-Algorithmus empfehlen? Begründung?
- Können Sie sich eine Anwendungsaufgabe vorstellen, in der so eine Situation realistisch ist? Welche bzw. Warum nicht?