Intelligente Datenanalyse in Matlab

Prüfung: Projektaufgabe 6

Michael Brückner
Thomas Vanck
Prof. Tobias Scheffer

Datum: 07.12.2010

Die Projektaufgabe ist Teil der Prüfung *Intelligente Datenanalyse in Matlab*. Jede Aufgabe soll durch einen Studenten selbstständig bearbeitet und die Lösung innerhalb der mündlichen Prüfung vorgestellt werden. Ein Ausdruck des MATLAB-Programmcodes und der Ergebnisse in Form eines Diagramms, Tabelle o.ä. werden vorausgesetzt; die Art der Präsentation der Ergebnisse ist dem Studenten freigestellt.

Problemstellung

Ein Anbieter von Dokumentenmanagementsystemen möchte ein Softwaremodul entwickeln welches neue Korrespondenz (Emails, gescannte Briefe/Faxe usw.) des Nutzers basierend auf deren Relevanz sortiert. Das System soll später lokal auf dem Computer des Nutzers arbeiten und selbständig neue Dokumente sortieren. Es wird angenommen, dass für einen Großteil älterer Dokumente die Bewertung durch die Nutzer bekannt ist und neue Dokumente nach dem Lesen ebenfalls (teilweise) bewertet werden. Zum Zweck der Entwicklung und Evaluation der Software wurden 11.269 Dokumente gesammelt und durch einen repräsentativen Nutzer auf einer Skala von 1 (irrelevant) bis 20 (sehr relevant) bewertet.

Sie wurden damit beauftragt ein Lernverfahren für das Softwaremodul zu entwickeln. Es soll basierend auf bereits bewerteten Dokumenten eines Nutzers Vorhersagen über die Relevanz neuer Dokumente treffen und insbesondere die 10 relevantesten Dokumente korrekt sortieren. Als Eingabe erhält das Modul die Bag-of-Word-Darstellung des zu bewertenden Dokuments. Als Ausgabe soll es einen Score liefern welcher die Relevanz des Dokumentes wiederspiegelt, d.h. ein hoher Score entspricht einer hohen Relevanz.

Aufgabe

Die Daten enthalten die Attribute *Dokument-ID*, *Wort-ID* und *Anzahl*, wobei *Anzahl* angibt wie häufig das Wort *Wort-ID* in Dokument *Dokument-ID* vorkommt. Lesen sie die Daten in MATLAB ein und führen sie eine Datenvorverarbeitung durch. Überlegen sie ob eine Feature-Normalisierung notwendig ist und erklären sie welche Art von Analyseproblem vorliegt. Identifizieren sie ein geeignetes Verfahren zum Lösen des Problems, implementieren sie dieses in MATLAB und trainieren sie ein Modell. Evaluieren sie das Modell indem sie die Werte für MAP@10 und NDCG auf einer zufällig gewählten Holdout-Menge bestimmen. Begründen und dokumentieren sie kurz alle durchgeführten Schritte.