Textklassifikation

Problembeschreibung und Aufgabenstellung

Bei der dritten und letzten Aufgabe geht es darum, natürlichsprachliche Texte zu klassifizieren. Wir verwenden dafür nur zwei Kategorien von Texten, Nachrichtentexte (aus dem Online-Archiv der "Zeit") und Filmbeschreibungen (aus Wikipedia). Die zu entwickelnde Software soll für einen gegebenen Text möglichst gut entscheiden können, in welche der beiden Kategorien er fällt. Als Klassifikationstechnik soll die "Naive Bayes-Klassifikation" verwendet werden (s. Skript, Teil 2, Folien 101 - 109), wobei Sie davon ausgehen dürfen, dass die Texte der beiden Kategorien mit derselben Wahrscheinlichkeit auftreten. Das zugehörige Modell soll mittels beaufsichtigtem Lernen trainiert werden (s. Skript, Teil 3, Folien 224 - 226).

Die Klassifikation muss auf allgemeinen Merkmalen von Texten basieren und darf keinen Bezug auf den Inhalt nehmen. Die Anwesenheit oder Häufigkeit konkreter, inhaltsbezogener Wörter (wie z.B. "Film" oder "Handlung") darf also nicht herangezogen werden.

Das ZIP-Archiv *Datensatz_1.zip* (s.u.) enthält Trainings- und Testdaten. Für das Training dürfen ausschließlich die dafür vorgesehenen Daten verwendet werden. Die Testdaten dienen dazu, die Wirksamkeit der Klassifikation zu ermitteln und geeignete Merkmale auszuwählen.

Am 11.01.2016 wird ein zweites ZIP-Archiv (*Datensatz_2.zip*) bereitgestellt, das weitere Texte aus den beiden Kategorien enthält. Anhand dieser Texte soll dann die Evaluation Ihres Klassifikationsmodells erfolgen. Ermitteln Sie dazu die Rate der korrekt klassifizierten Texte der Evaluationsmenge. Pro Textkategorie soll diese Rate mindestens 70% betragen. Insgesamt, d.h. über beide Kategorien hinweg, sollen mindestens 80% erreicht werden.

Abzugeben sind der Quellcode der Java-Klassen sowie eine kurze schriftliche Darstellung der Vorgehensweise und der Ergebnisse (gerne stichwortartig) als PDF-Dokument.

Weitere Hinweise

Denken Sie an eine software- und programmiertechnisch korrekte Form der Implementierung. Denken Sie auch an den angemessenen Einsatz von Datenstrukturen und Algorithmen. Zu einer tadelosen Lösung gehört auch eine aussagekräftige und vollständige Kommentierung des Quelltextes. Die Kommentierung soll javadoc-fähig sein.

Die Bearbeitung der Aufgabe soll in Zweierteams erfolgen. Es wird vorausgesetzt, dass sich beide Teampartner mit dem Lösungsweg und auch mit der konkreten Implementierung der Lösung bestens auskennen.

Packen Sie alle zu Ihrer Lösung gehörenden Klassen sowie das PDF-Dokument in ein Zip-Archiv, benennen Sie es nach dem Schema *vorname1.nachname1-vorname2.nachname2-ISys3.zip* (Beispiel: stefan.krause-michael.breuker-ISys3.zip) und laden Sie das Archiv rechtzeitig im Moodle-Kurs hoch. Verspätete Abgaben werden nicht berücksichtigt.