Web Mining im SoSe 2017 – Übung 1

Ingo Adrian und Steffen Pegenau
4. Mai 2017

Aufgabe 1

Überlegen Sie sich eine neuartige, originelle Web mIning Anwendung, die mit Text-Klassifikationsverfahren gelöst werden könnte. Skizzieren Sie eine mögliche Umsetzung (z.B. Sammlung der Trainingsdaten, Klassifikation der Trainingsdaten, Einsatz des gelernten Klassifikators in der Praxis, etc.) (2 Punkte)

Aufgabe 2

Schreiben Sie ein einfaches Programm, das eine sortierte Liste der in einem Text vorkommenden Worte (im weitesten Sinn alles was durch Leerzeichen begrenzt wird) mit den assoziierten Häufigkeiten (absolut und prozentual) erstellt und sortiert ausgibt. (2 Punkte)

Vergleichen Sie anhand der Ausgabe Ihres Programms die 30 am häufigsten vorkommenden Worte in zwei oder mehreren längeren Texten der gleichen Sprache (z. B. E-books, Projekt Gutenberg, etc.). Wählen Sie einge geeignete Darstellung für Ihren Vergleich. Sind diese Worte als Merkmale für Text-Klassifizierungs-Aufgaben geeignet? Warum? Modifizieren Sie Ihr Programm dahingehend, daß es eine Liste von Stoppwörtern erhalten kann, die ignoriert werden. Wiederholen Sie die vorherige Aufgabe, indem Sie jedoch diesmal die Stoppwörter der jeweiligen Sprache ignorieren (eine Auswahl finden Sie unter http://www.nltk.org/nltk$_{data/packages/corpora/stopwords.zip}$).$Wiew$ürden$SienundieEignungder30häufigstenW$örtereinschätzen$?

Aufgabe 3