Hausaufgabe

Abgabe bis 17.6.

Wir haben gesehen, daß in Texten zumeist wenige Wörter häufig und viele Wörter selten vorkommen. Wählen Sie einen Text von `http://www.gutenberg.org/catalog/` und speichern Sie ihn mit der Endung `.txt`. Laden Sie das Skript `textstat.pl` von der geschützen Bereich der Kursseite herunter. Starten Sie das Skript und lesen Sie ihren Text ein. Betrachten Sie die Dateien `stat_$name.txt` und `frequ_$name.txt`. Welche für den Text zentralen Wörter finden sich unter den 20 häufigsten? Wieviel Prozent aller Typen kommen nur einmal vor? Welchen Anteil haben die drei häufigsten Typen am Gesamttext? Erläutern Sie, welche Auswirkung diese Verteilung für computerlinguistische Aufgaben hat. (Insgesamt nicht mehr als 300 Wörter)

Vorsicht, das Skript `textstat.pl` ist nicht ausgiebig getestet worden, es könnte noch Fehler beinhalten! Sie sind eingeladen, das Skript zu verbessern.