Unix-basierte Implementierung kleiner Datenbanken (SS 2005)

Abgabetermin 18.05.2005

Ähnlichkeit von Texten (10 Punkte)

In diesem Übungsblatt sollen Texte analysiert und verglichen werden. Ziel ist es festzustellen, ob verschiedene Texte in der gleichen Sprache verfaßt sind.

Statistik

Zunächst wird jeder Text für sich analysiert. Dabei wird der Text zunächst in Worte (Zeichenfolgen ohne Leerzeichen) zerlegt und alle Großbuchstaben werden in Kleinbuchstaben umgewandelt. Worte, die mindestens ein Zeichen enthalten, das kein Buchstabe ist, werden ignoriert. Schließlich wird gezählt wie oft jede Folge von 1, 2 oder 3 Buchstaben in den verbliebenen Worten vorkommt, d.h. es wird z.B. gezählt wir oft die Buchstabenfolge "abc" in dem Text vorkommt. Dabei zählen Vorkommen in ignorierten Worte und Vorkommen, die über eine Wortgrenze hinweggehen nicht mit.

Vergleich

Zwei auf diese Art erstellte Statistiken können verglichen werden. Dazu soll auf folgende Weise eine Kennzahl berechnet werden:

- Es wird für beide Statistiken die relative Häufigkeit jedes Buchstabens, jeder zweibuchstabigen Zeichenkombination und jeder dreibuchstabigen Zeichenkombination berechnet.
- Für jede Buchstabenkombination aus einem, zwei oder drei Buchstaben wird der Absolutbetrag der Differenz der beiden relativen Häufigkeiten gebildet.
- Alle diese Ergebnisse werden aufsummiert. Diese Summe ist das Ergebnis. Je kleiner die Summe, desto ähnlicher sind sich die beiden Texte.

Ein Hauptprogramm

Schreibt ein **Perl**-Programm, das zwei Dateinamen als Kommandozeilenparameter erhält. Diese beiden Dateien sollen geöffnet und wie oben beschrieben verglichen werden. Das berechnete Ergebnis soll ausgegeben werden.
Testet Euer Programm mit mehreren Texten in mindestens zwei Sprachen und vergleicht die berechneten Kennzahlen.

Hinweise

- Eine Übersicht über die Manualseiten von perl gibt es mit "`man perl`". Die in perl direkt verfügbaren Funktionen sind unter "`man perlfunc`" zusammengestellt.
- Interessant für dieses Blatt könnten unter anderem die Funktionen `split`, `chr`, `ord` und `lc` sein.
- Die Dokumentation zu perl-Modulen kann mit Hilfe von `perldoc` angezeigt werden, etwa "perldoc IO::File".