Korpusbasierte Programmierprojekte

Bitte bilden sie Gruppen von 2-4 Personen und wählen sie ein Projekt.

Vorverarbeitung der Texte

- Wählen sie Texte aus dem Gutenbergprojekt (http://www.gutenberg.org/) und speichern sie sie in dem Ordner (C:/Texte/)
- Schreiben sie ein Programm,
  - das fragt, welcher Text eingelesen werden soll
  - das den Header und Footer der Gutenbergtexte ignoriert
  - das den Text tokenisiert
  - und eine Datei erzeugt, in der pro Zeile ein Token steht
  - und in der eine Leerzeile das Ende eines Satzes markiert.
  - Normalisieren sie den tokenisierten Text (z.B. indem sie Anführungszeichen durch Standardanführungszeichen ersetzen)

  Beispieldateien: minitext.txt, minitext_token.txt

Ausführende: Markus, Dustin, Roman, Martin

Textstatistik (I)

Schreiben sie ein Programm für folgende Aufgaben (lagern sie die Aufgaben sinnvoll in Funktionen aus):

- Ermittlung der Satzanzahl und der durschnittlichen Satzlänge
  Beispieloutput: Der Text enthält 15 Sätze. Jeder Satz besteht durchschnittlich aus 8.6 Wörtern
- Ermittlung der Wortzahl und der durschnittlichen Wortlänge
  Beispieloutput: Der Text enthält 6 Wörter. Jedes Wort besteht durchschnittlich aus 4.3 Buchstaben
- Ermittlung der Anzahl der Typen (vorkommende unterschiedliche Wortformen) und des Verhältnisses von Token (vorkommende Wortformen ~ Wortzahl) zu Typen
  Beispieloutput: Der Text besteht aus 6 Token (vorkommende Wortformen) und 4 Typen (vorkommende unterschiedliche Wortformen). Das Verhältnis von Token zu Types ist 1.25. Jede Wortform kommt also durchschnittlich 1.25 mal vor.

Ausführende: Stephanie, Victor, Juliana

Textstatistik (II)

Schreiben sie ein Programm für folgende Aufgaben (lagern sie die Aufgaben sinnvoll in Funktionen aus):

- Ermittlung der n längsten / kürzesten Wörter in einem Text
- Ermittlung der n Wörter, die am häufigsten / seltensten vorkommen
- Anteil eines Worttypens an einem Gesamttext
- Anteil der Worttoken mit einer Mindest-/Höchstlänge n am Gesamttext.
- Überprüfe, ob das Zipfsche Gesetz für einen Text Gültigkeit hat (http://de.wikipedia.org/wiki/Zipfsches_Gesetz)

Ausführende: Anna, Pascal, Sonia, Kim,

Konkordanz

Schreiben sie ein Programm, das zu einem Wort oder einer Phrase die Konkordanz (Vorkommensliste des Wortes mit Kontext) erzeugt (http://de.wikipedia.org/wiki/Konkordanz).

- Bemühen sie sich um eine gut lesbare Ausgabe.
- Ermöglichen sie es, die Kontextgröße zu variieren (ganzer Satz oder v Wörter vorher und n Wörter nachher)

Ausführende: Tobias, Anna, Alida

Automatisches Dichten

Schreiben sie ein Programm, das automatisch Gedichte erzeugt (4-Zeiler). Die Gedichte müssen inhaltlich keinen Sinn ergeben, sie sollten aber

- ein vorgegebenes wählbares Versmaß einhalten (http://wortwuchs.net/versmass/)
- ein vorgegebenes wählbares Reimschema einhalten (http://wortwuchs.net/reimschema/)

Verwenden sie das CMU-Wörterbuch und dichten sie in Englisch.

Ausführende: Marc, Veronika, Marianna, Konstantin

Automatisches Texten (basierend auf Wörter- oder Zeichen-N-Grammen)

Schreiben sie ein Programm, das automatisch Texte generiert. Ziel ist, dass die Texte möglichst so aussehen als seien es echte Texte der gewählten Sprache.

Nutzen sie häufige Teilstringe in der Sprache, die sie automatisch aus einem oder mehreren Texten ermitteln.

Beispiel: deutschen Text generieren aus Zeichen-3-Grammen.

Ein String „kr“ wird im Deutschen häufig mit einem Vokal fortgesetzt, fast nie mit einem Konsonanten. „kri“ kann also zu etwas fortgesetzt werden, das einem deutschen Wort ähnelt „krt“ nicht.

Ist es möglich, aus den automatisch generierten Texten die Sprache der Texte zu erkennen, aus denen die N-Gramme gewonnen wurde?

Wie groß sollte N gewählt werden, damit die Sprache erkennbar wird?

Automatisches Texten (basierend auf Wörter- oder Zeichen-N-Grammen)

Gehen sie wie folgt vor:

- Zerlegen sie einen Text in N-Gramme auf Basis der Zeichen (Buchstaben) oder der Wörter
- Fragen sie ein (N-1)-Gramm ab, mit dem gestartet werden soll.
- Generieren sie mögliche Fortsetzungen aufgrund der Häufigkeit der N-Gramme, die mit dem (N-1)-Gramm beginnen
- Fahren rekursiv fort.

Ausführende: Anna, Anna