Praktikum 2

Hinweise zur Abgabe

- Die benötigten Materialien zum Lösen der Aufgaben finden Sie im Moodle-System.
- Sie können die verwendete Programmiersprache frei wählen.
- Bitte geben Sie Ihren Sourcecode inklusive (evtl. aktualisierten) Entwurfsdokumenten als ZIP-Archiv über Moodle ab. Die Deadline für die Abgabe finden Sie dort. Denken Sie außerdem an die rechtzeitige Vereinbarung eines Abnahmetermins.
- Benennen Sie die Datei wie folgt: gruppe_n_pm.zip, wobei n Ihre Gruppennummer ist und m die Nummer des Praktikumsblattes.
- Dokumentieren Sie Ihren Quellcode sowie Installation und Start des Programms! Das heißt, dass Sie aussagekräftige Kommentare im Code einfügen *und* eine grundlegende Dokumentation der wesentlichen Methoden schreiben.
- Vermeiden Sie die Verwendung von absoluten Pfaden oder plattformspezifische Pfadsyntax im Quelltext. Das Programm soll auch nach dem Wechsel des Rechners oder Betriebssystems problemlos und vollständig ausführbar sein!

Aufgabe 1 – Vorbereitung der Dokumentensammlung

1. Die Datei *aesopa10.txt* enthält einige Fabeln des griechischen Dichters Äsop. Zerlegen Sie das Textdokument, so dass jede Fabel in einer eigenen Textdatei abgelegt wird. Die einleitenden Nutzungshinweise und das Inhaltsverzeichnis können Sie ignorieren. Um Ihnen die Arbeit zu vereinfachen, sind die Fabeln wie folgt abgesetzt:

   - 3 Leerzeilen
   - Titel der Fabel
   - 2 Leerzeilen
   - Text der Fabel

   Speichern Sie die einzelnen Fabeln als separate Textdateien, die sogenannten *Originaldokumente*. Benennen Sie die Dateien nach folgendem Muster:

   - Nehmen Sie den Titel der Fabel als Dateinamen.
   - Schreiben Sie den Dateinamen *klein*.
   - Ersetzen Sie Leerzeichen durch einen Unterstrich (_).
   - Wählen Sie als Dateiendung *.txt*.

   Die Fabel „The Sick Lion“ würde beispielsweise in der Datei *the_sick_lion.txt* abgespeichert werden.

Aufgabe 2 – Stoppworteliminierung

1. Implementieren Sie eine Funktion zur Stoppworteliminierung für Ihre *Originaldokumente*. Eine Liste englischer Stoppworte finden Sie in der Datei *englishST.txt*[1]. Achten Sie darauf, dass die interne Verarbeitung der Dokumente unabhängig von Groß- und Kleinschreibung sein soll. Entfernen sie außerdem Satzzeichen und Zeilenumbrüche, achten Sie dabei jedoch auf die spezielle Bedeutung des Apostrophs im Englischen.

Aufgabe 3 – Lineare Suche

1. Implementieren Sie eine lineare Suche in der Dokumentenkollektion anhand eines einzelnen Suchterms. Lineare Suche bedeutet hierbei, dass Sie sequentiell für jedes einzelne Dokument der Kollektion prüfen, ob der Suchterm enthalten ist (boolesches Retrieval-Modell).

2. Ermöglichen Sie es festzulegen, ob entweder die Originaldokumente oder die um Stoppworte bereinigten Dokumente zur Suche herangezogen werden.

[1] Quelle: `http://members.unine.ch/jacques.savoy/clef/englishST.txt`