

# Algorithmen und Programmierung I

WS 2004 / 2005

## Übung 8 (20 + 5 Punkte)

Ausgabe: Mo. 13.12.2004

Abgabe: Mi. 5.1.2005, 10.00

### Verspätet abgegebene Lösungen werden nicht akzeptiert.

Die Abgabe erfolgt in Zweiergruppen. Die Gruppe darf nur mit ausdrücklicher Genehmigung des Tutors gewechselt werden.

Abgegeben werden müssen

- alle Lösungen einmal ausgedruckt. **Handschriftliche Lösungen werden nicht akzeptiert.**
- Programme, Testdaten, Testergebnisse in elektronischer Form in Absprache mit den Tutorinnen und Tutoren.

**Aufgabe 8.1 = 6.4 (5 Punkte)** War zurück gestellt worden.

### Aufgabe 8.2

In dieser Aufgabe geht es um Textverarbeitung und Sortieren. Es soll ein KWIC-Index (*keyword-in-context*) aus einer Liste von Buchtiteln o.ä. erstellt werden. Die Wörter jedes Titels werden so rotiert, dass jedes im Titel vorkommende Schlüsselwort einmal vorn steht. Für eine Titelzeile mit  $n$  Schlüsselwörtern werden also  $n$  Zeilen, von denen jede einer Rotation entspricht, erzeugt.

Beispiel:

"Spielberg: Schindlers Liste" führt zu drei Rotationen: dem ursprünglichen Eintrag sowie "Liste. Spielberg: Schindlers" und "Schindlers Liste. Spielberg:"

Am Ende jeder Originalzeile wird ein Punkt eingefügt, damit man den Originaltitel leicht wieder erkennen kann. Das erste Zeichen einer Zeile soll ein Großbuchstabe sein.

Alle aus der ursprünglichen Liste erzeugten Zeilen werden alphabetisch sortiert und liefern den KWIC-Index. Beispiel:

```
"Spielberg: Schindlers Liste\nKubrick: 2001 Odysse im Weltraum"
```

wird zu:

```
2001 Odysse im Weltraum. Kubrick:  
Kubrick: 2001 Odysse im Weltraum.  
Liste. Spielberg: Schindlers  
Odysse im Weltraum. Kubrick: 2001  
Schindlers Liste. Spielberg:  
Spielberg: Schindlers Liste.  
Weltraum. Kubrick: 2001 Odysse im
```

Schlüsselwörter sind alle Wörter, die nicht in einer Stoppwortliste vorkommen. Sie ist alphabetisch sortiert; alle Wörter bestehen nur aus Kleinbuchstaben. Stoppwörter sind z.B. Artikel (der, die, das) oder Konjunktionen wie und.

Die Stoppwortliste (englische und deutsche Stoppwörter, ohne Garantie auf Vollständigkeit) ist Teil einer Datei mit dem Programmrahmen `kwicProgFragm.lhs`.

Die Datei `titel.txt` enthält die Titel, für die ein KWIC-Index erzeugt werden soll. Ein Rahmenprogramm für das Einlesen der Dateien und das Schreiben des KWIC-Index befindet sich unter `Main.lhs`. Die Ausgabedatei soll `kwicTitel.txt` heißen.

**Main.lhs** erwartet die Eingabedatei **titel.txt** in einem Verzeichnis, dessen Pfad dort angegeben ist. Den müssen Sie ersetzen. Alle Dateien sollten sich in einem Verzeichnis **Kwic** befinden. Das ist nützlich, wenn Sie die Modularisierung von Haskell nutzen wollen (Gilt besonders für GHC / GHCi)

Planen Sie sorgfältig alle Schritte (Text in Zeilen zerlegen, Zeilen in Wörter, Rotieren der Zeilen, Sortieren, usw.) und schreiben Sie dafür jeweils handliche Funktionen, die dann geeignet zusammengesetzt werden müssen. Beachten Sie, dass Groß- und Kleinschreibung in der Stoppwortliste unterschieden wird, jedoch nicht bei der Sortierung. Das erste Wort einer Zeile im KWIC-Index beginnt immer mit einem Großbuchstaben, unabhängig von der Schreibweise in der Originalzeile.

Sie können Ihr Programm ohne das Rahmenprogramm mit Ein- und Ausgabe testen. Das empfiehlt sich natürlich. Der Ausdruck

```
putStr (kwicIndex "Spielberg: Schindlers Liste\nKubrick: 2001 Odysse im Weltraum")
```

liefert zum Beispiel die obige Ausgabe, wenn ihr **kwicIndex**-Programm korrekt ist.