Text-Statistik

Es geht in dieser Aufgabe darum, verschiedene Statistiken zu einem Text, der als Array von Strings (ein String pro Zeile) gegeben ist, zu ermitteln. Später wird die Aufgabe so erweitert, dass der Text aus einer Datei gelesen wird.

Hier ein Beispiel für so ein String-Array und die zu erzeugende Ausgabe:

"Alan Mathison Turing (1912-1954) war ein britischer Logiker,",
"Mathematiker, Kryptoanalytiker und Informatiker.",
"Das von ihm entwickelte Berechenbarkeitsmodell der Turingmaschine ",
"bildet eines der Fundamente der Theoretischen Informatik.",
"Während des Zweiten Weltkrieges war er maßgeblich an der Entzifferung",
"der mit der deutschen Rotor-Chiffriermaschine Enigma verschlüsselten deutschen Funksprüche beteiligt.",
"Nach ihm benannt sind der Turing Award, die bedeutendste Auszeichnung in der Informatik,",
"sowie der Turing-Test zum Überprüfen des Vorhandenseins von künstlicher Intelligenz."

```
Der Text enthält:
- 8 Zeilen
- 70 Wörter
- 573 Zeichen
Längstes Wort: "Rotor-Chiffriermaschine" (Zeile 6, Position 5)

Häufigkeiten der Buchstaben:
a:    26; Häufigste Kombination: at (6x)
b:     9; Häufigste Kombination: be (5x)
c:    15; Häufigste Kombination: ch (14x)
d:    25; Häufigste Kombination: de (17x)
e:    74; Häufigste Kombination: er (21x)
f:    10; Häufigste Kombination: fo (3x)
g:    12; Häufigste Kombination: ge (3x)
h:    21; Häufigste Kombination: he (9x)
i:    45; Häufigste Kombination: in (15x)
j:     0;
k:    12; Häufigste Kombination: ke (6x)
l:    15; Häufigste Kombination: li (4x)
m:    16; Häufigste Kombination: ma (10x)
n:    48; Häufigste Kombination: nd (6x)
o:    14; Häufigste Kombination: or (6x)
p:     3; Häufigste Kombination: pr (2x)
q:     0;
r:    43; Häufigste Kombination: ri (7x)
s:    25; Häufigste Kombination: sc (7x)
t:    37; Häufigste Kombination: te (9x)
u:    14; Häufigste Kombination: un (5x)
v:     4; Häufigste Kombination: vo (3x)
w:     8; Häufigste Kombination: wa (3x)
x:     0;
y:     2; Häufigste Kombination: yp (1x)
z:     5; Häufigste Kombination: ze (1x)
```

Den Beispiel-Text finden Sie im Moodle als Code-Schnipsel zum Rauskopieren.

Für die Ermittlung müssen Sie den Text einmal komplett durchlaufen und die benötigten Informationen in passenden Variablen und Feldern sammeln.

Wie immer ist es am besten, die Aufgabe in kleinen Schritten anzugehen.

Teilaufgabe 1: Gesamtstatistik

a) Ermitteln Sie, wie viele Zeilen, wie viele Wörter und wie viele Zeichen der Text insgesamt hat und geben Sie diese Informationen entsprechend aus.

   Sie dürfen an dieser Stelle die Standard-Funktion String.Split() verwenden, um den Text einer Zeile in ein String-Array zu verwandeln. Trennen Sie bei Leerzeichen, aber zählen Sie sie mit.

b) Ermitteln Sie, das längste Wort im Text, die Zeile in der das Wort steht sowie die Position des Wortes in der Zeile und geben Sie diese Informationen aus.

Teilaufgabe 2: Buchstabenhäufigkeiten (zunächst noch ohne häufigste Kombination)

Die Idee ist, zum Mitzählen der Häufigkeiten ein Integer-Feld der Länge 26 zu verwenden, mit einem Speicherplatz für die Häufigkeit jedes Buchstabens.

Der Code-Schnipsel im Moodle enthält bereits eine Funktion **int Index(char c)**, die zu einem gegebenen Buchstaben b seine Position im Alphabet zurückliefert. Bei Umlauten, Sonderzeichen etc. wird -1 zurückgegeben.

Definieren Sie sich ein Int-Array der Länge 26 und erweitern Sie Ihr Programm so, dass beim Durchlaufen des Textes für jeden gelesenen Buchstaben der entsprechende Feldwert inkrementiert (=hochgezählt wird). Die richtige Position im Feld liefert die Index-Funktion.

Lassen Sie dann die Häufigkeiten pro Buchstabe ausgeben. Tipp hier: Zu einem Index i bekommt man den zugehörigen Buchstaben einfach mit der umgekehrten Rechnung wie in der Index-Funktion angegeben, es ist allerdings ein Type-Cast nötig.

Teilaufgabe 3: Häufigster Nachfolger pro Buchstabe

In dem Beispieltext zu Turing ist der häufigste Nachfolger für ein 'a' der Buchstabe 't', denn die Kombination 'at' kommt 6x vor, z.B. bei "M**at**hison" oder "Inform**at**ik".

Sie sollen für jeden Buchstaben ermitteln, welcher der häufigste Nachfolger ist und die entsprechende Buchstabenkombination mit ihrer Häufigkeit ausgeben.

Damit Sie das ermitteln können, benötigen Sie ein zweidimensionales Integer-Array der Größe 26x26, in dem Sie für jeden Buchstaben mitzählen, welcher andere Buchstabe auf ihn folgt.

Angenommen, der aktuelle Buchstabe ist ein 'a' und der darauffolgende ein 't', dann erhöhen Sie den Zähler in der Zeile, die dem 'a' entspricht (Index('a')) und der Spalte, die dem 't' entspricht (Index(t)).

Für Ausgabe des häufigsten Nachfolgers müssen Sie am Ende in dieser 2D-Matrix jeweils das Zeilen-Maximum bestimmen. Dabei brauchen Sie sowohl den Maximalwert der Häufigkeit (bei 'at' wäre das der Wert 6) als auch den Spalten-Index, wo der Wert aufgetreten ist (Spalte, die dem 't' entspricht).

Teilaufgabe 4: Lesen des Textes aus einer Datei

Hinweis: Das Thema Dateien wird in der ersten Dezember-Woche in der Vorlesung behandelt und in der Woche darauf in der Übung. Es ist sinnvoll, mit der Bearbeitung dieser letzten Aufgabe solange zu warten, bis der Stoff behandelt wurde.

Ändern Sie Ihr Programm so ab, dass Sie den Text aus der Datei mit einem StreamReader lesen und zeilenweise mit sr.ReadLine() den Text durchlaufen. Benutzen Sie **nicht** sr.ReadToEnd() oder File.ReadAllLines(), aber schauen Sie sich die Dokumentation zu diesen Funktionen ruhig einmal an.

Als Beispiel finden Sie im Moodle die Datei "kafka_verwandlung.txt", die den kompletten Text des Buches "Die Verwandlung" von Franz Kafka enthält.

In der Endversion Ihrer Lösung ist es ausreichend, wenn Sie eine Datei verarbeiten können. Das String-Array vom Anfang dürfen Sie gern rauswerfen oder auskommentieren.