Aufgabe: Textanalyse

Im vorangegangenen Semester hatten Sie eine einfache Aufgabe zur Textanalyse bearbeitet:

Eine Textdatei wird eingelesen und die Anzahl von Wörtern und Zeichen im Text bestimmt. Whitespace wird dabei ignoriert. Die durchschnittliche Anzahl von Zeichen pro Wort wird berechnet. Zudem wird die Frequenz des Auftretens von Vokalen bestimmt. Der häufigste Vokal wird ausgegeben.

Gehen Sie von der Lösung dieser Aufgabe aus. Sie können die in bereitgestellte Version oder (noch besser) Ihre eigene Lösung aus dem letzten Semester verwenden.

Umstellung und Erweiterung

1. Verwenden Sie in der neuen Version der Lösung ein Dictionary, um die Vokale zu zählen und den bzw. die häufigsten zu ermitteln.
2. Ermitteln Sie das häufigste Wort bzw. die häufigsten Wörter im Text. (siehe `most_common.txt`).
3. Erweiterung: Es soll ermittelt werden, welche 10 Wörter am häufigsten im Text verwendet werden (Top-Ten-Liste). Wenn ein Rang mehrfach besetzt ist, werden solange Wörter aufgegeben, bis insgesamt 10 genannt sind. Wenn mehrere Wörter auf dem letzten Rang landen, werden alle ausgegeben (siehe `top_ten.txt`)

Wenn Sie zunächst noch keine Lösungsidee für 3. haben, implementieren Sie zunächst die Umstellung auf Dictionaries (1. und 2.).

Lösungshinweise (1)

Zählen Sie die Vokale bzw. Wörter jeweils in einem Dictionary, wobei jeder Vokal bzw. jedes Wort einen Schlüssel darstellt. Da erst zur Laufzeit des Programms bekannt ist, welche Wörter im Dictionary enthalten sind, tritt ein „Henne-Ei-Problem“ auf: Der neue Wert zum Schlüssel hängt vom alten Wert ab. Beim ersten Zugriff ist der Schlüssel allerdings noch nicht bekannt, daher schlägt der Zugriff mit einem `KeyError` fehl:

```
counter_dict = {}
counter_dict["a"] = counter_dict["a"] + 1
```

Dieses Problem kann durch Vorgabe eines Rückgabewerts für den Fall, dass der Schlüssel noch nicht vorhanden ist, gelöst werden. Die Methode `get` lässt eine solche Definition zu (hier den Vorgabewert `0`):

```
counter_dict = {}
counter_dict["a"] = counter_dict.get("a", 0) + 1
```

Lösungshinweise (2)

Um die häufigsten Vokale bzw. Wörter zu ermitteln, können Sie bspw. über das Dictionary iterieren und sich die Schlüssel für die Gewinner merken.

Um die Top-Ten-Liste der häufigsten Wörter zu ermitteln, könnte man bspw. zunächst die Werte aus dem Dictionary auslesen und sortieren und anschließend wieder über das Dictionary iterieren und sich die entsprechenden Schlüssel merken.

Wenn Sie nach Lösungen für dieses Problem suchen (Such-Prompt bspw.: `sort python dictionary by value`) werden Sie viele verschiedene (oft elegante) Lösungswege finden, die teilweise aber auch sehr spezielle Python-Sprachkonstrukte verwenden.

In dieser Aufgabe geht es darum, dass Sie sich eine Lösung erarbeiten, die Sie verstehen :-)