Skriptsprachen
Naumann
SS 2014

Aufgabenblatt (4)

Es gibt verschiedene Möglichkeiten, um die Sprache, in der ein Dokument verfasst wurde, zu identifizieren. Ein sehr einfaches Verfahren besteht darin, für jede Sprache eine kleine Liste hochfrequenter Ausdrücke zu speichern und zur Sprachidentifikation zu verwenden.

Aufgabe (1) [2,5 Punkte]

Generieren Sie für 5 Sprachen ihrer Wahl eine mindestens 20 Elemente umfassende Liste von hochfrequenten Wortformen.

Aufgabe (2) [2,5 Punkte]

Definieren Sie die Klasse **Identify** inklusive einer ***initialize***- und einer ***to_s***-Methode. Jede Instanz dieser Klasse verfügt über folgende (Instanzen)

- Variable @sprachen(Hash):
  Speichert für jede Sprache die Liste hochfrequenter Wortformen.
- @text(String): Der letzte Text, dessen Sprache identifiziert wurde.
- @bericht(Array):
  Ein Array der Form [[Sprache, Treffer]*].

Aufgabe (3) [2 Punkte]

Definieren Sie in der Klasse **Identify** die Methode **add_language**, die einen Sprachnamen (*String*) und eine Folge von hochfrequenten Wörtern der Sprache (*String*) als Argument nimmt und diese Daten in der Instanzenvariable @sprachen speichert.

Beispiel

```
test_id = Identify.new
test_id.add_language('Deutsch', 'der die das ...')
test_id.add_language('Englisch', 'the a an ...')
```

Aufgabe (4) [4 Punkte]

Definieren Sie in der Klasse **Identify** die Methode **identify_language**, die einen Text (String) als Argument nimmt und für jede in @sprachen gespeicherte Sprache feststellt, wieviele Wörten des Texten in der für diese Sprache gespeicherte Liste mit hochfrequenten Wörtern vorkommt und diese Wert in @bericht speichert. Als Ergebnis liefert die Methode den Namen der Sprache, fü die dieser Wert am höchsten ist.

Beispiel

p test_id.identify_language 'Die Zentrale der CSU in München ist in...'
Deutsch

Aufgabe (5) [2 Punkte]

Definieren Sie die außerdem die Methode **bericht**, die für den zuletzt identifizierten Text für jede Sprache die Trefferzahl ausgibt.

Beispiel

test_id.bericht
Für den Text:
'Die Zentrale der CSU in München ist in die Jahre gekommen, aber...'
gab es folgende Ergebnisse
Englisch : 8
Deutsch : 53