block6-funktionen-textpreprocessing

September 20, 2018

1 Block 6: eigene Funktionen schreiben - Textpreprocessing

1.1 Text-Datei einlesen

```
In [ ]: with open("DATEINAME.txt", encoding="utf-8") as infile:
            rezension = infile.read()
```

1.2 Kleinschreibung vereinheitlichen

```
In [ ]: def text_lower(text):
            return text.lower()

        rezension = text_lower(rezension)
        print(rezension[:250])
```

1.2.1 Punktation entfernen

```
In [ ]: import string

        def remove_punctuation(text):
            punctuation = string.punctuation
            for marker in punctuation:
                text = text.replace(marker, "")
            return text

        rezension = remove_punctuation(rezension)
        print(rezension[:250])
```

1.3 Liste erstellen

```
In [ ]: rezension_words = rezension.split()
```

```
In [ ]: print("Anzahl aller Worte des Textes: ", (len(rezension_words)))
        print("=======")
        print(rezension_words[:25])
```

1.4 Zählen eines bestimmten Worts

```
In [ ]: def count_item_in_text(item_to_count, list_to_search):
            number_of_hits = 0
            for item in list_to_search:
                if item == item_to_count:
                    number_of_hits += 1
            return number_of_hits

        print(count_item_in_text("information", rezension_words))
```

1.5 Alle Wörter zählen mit Hilfe eines Dictionarys

```
In [ ]: def counter_dict(list_to_search):
            counts = {}
            for word in list_to_search:
                if word in counts:
                    counts[word] = counts[word] + 1
                else:
                    counts[word] = 1
            return counts

        print(counter_dict(rezension_words))
```

1.6 Worthäufigkeiten sortieren

```
In [ ]: def freq_sort(list_to_search):
            counts = counter_dict(list_to_search)
            counts = [(counts[key], key) for key in counts]
            counts.sort()
            counts.reverse()
            return counts

        print(freq_sort(rezension_words)[:25])
```

1.7 Entfernen von Stoppwörtern

```
In [ ]: import requests

        def remove_stopwords(list_to_search):
            stopword_url = "http://members.unine.ch/jacques.savoy/clef/germanST.txt"
            response = requests.get(stopword_url)
            stopwords = response.text
            stopwords = stopwords.split()
            return [w for w in list_to_search if w not in stopwords]

        print(remove_stopwords(rezension_words))
```

1.8 Funktionsaufrufe

```
In [ ]: rezension = text_lower(rezension)
        rezension = remove_punctuation(rezension)
        rezension_words = rezension.split()
        rezension_words = remove_stopwords(rezension_words)
        print(freq_count(rezension_words)[:25])
```