SW03 - Gruppe 1

Einführung

M. Nebroj, S. Hauri, S. Ineichen, R. Schwarzentruber

2019-10-07

Testatübung SW03

Aufgabe 1 - Tokenizieren

Tokenizieren sie folgenden Text:

```
text = "When Alexander Graham Bell invented the telephone he had three missed
 ↪  calls from Chuck Norris."
```

Code

```
from nltk import tokenize
tokens = tokenize.word_tokenize(text)
print(tokens)
```

Output

```
['When', 'Alexander', 'Graham', 'Bell', 'invented', 'the', 'telephone', 'he',
 ↪  'had', 'three', 'missed', 'calls', 'from', 'Chuck', 'Norris', '.']
```

Aufgabe 2 - Stop Words

Entfernen sie die Stop Words aus dem tokenizierten Text.

Code

```
from nltk.corpus import stopwords
stop_words = set(stopwords.words('english'))
tokens_no_stopwords = [word for word in tokens if word not in stop_words]
print(tokens_no_stopwords)
```

Output

```
['When', 'Alexander', 'Graham', 'Bell', 'invented', 'telephone', 'three',
 ↪  'missed', 'calls', 'Chuck', 'Norris', '.']
```

Aufgabe 3 - Stemming / Lemmatization

Führen sie auf dem gleichen Text Stemming und Lemmatization aus. (Welche Problematik entdecken sie in den Resultaten?)

Code

```
from nltk.stem import PorterStemmer
stemmer = PorterStemmer()
stemmed_words = [stemmer.stem(word) for word in tokens]
print(stemmed_words)
```

Output

```
['when', 'alexand', 'graham', 'bell', 'invent', 'the', 'telephon', 'he', 'had',
 ↪  'three', 'miss', 'call', 'from', 'chuck', 'norri', '.']
```

Code

```
from nltk.stem import WordNetLemmatizer
lemmatizer = WordNetLemmatizer()
lemmatized_words = [lemmatizer.lemmatize(word) for word in tokens]
print(lemmatized_words)
```

Output

```
['When', 'Alexander', 'Graham', 'Bell', 'invented', 'the', 'telephone', 'he',
 ↪  'had', 'three', 'missed', 'call', 'from', 'Chuck', 'Norris', '.']
```

Problematik

Stemming schneidet oft unnötigerweise Endungen weg, sodass Wörter verfälscht werden; wie zum Beispiel bei den Namen “alexand” und “norri”.

Lemmatization dagegen ist zu konservativ; in dem Beispielsatz wird nur ein Wort reduziert. Der eigentlich erwünschte Nutzen bleibt dadurch eher gering.