Python exercise sheet A

This exercise sheet uses the dataset of Taylor Swift quotes provided on the Moodle page of the course.

Exercise A.1: Preprocessing

1. Lowercase the text.
2. Remove digits from the text.
3. Remove punctuation marks.
4. Remove superfluous blanks.

Hints:

- The Python package 're', for regular expressions, provides the 'sub' function to remove substrings with certain patterns.
- The Python package 'string' contains a collection of string constants, such as punctuation marks.

Exercise A.2: Tokenization

1. Tokenize the text.
2. Remove stopwords.

Hints:

- The Python package 'nltk.tokenize' includes the function 'tokenize'.
- The Python package 'spacy' includes collections of stopwords for many languages.

Exercise A.3: Stemming and lemmatization

1. Perform stemming.
2. Perform lemmatization.

Hints:

- The Python package 'nltk.stem' includes the function 'PorterStemmer'.
- The Python package 'nltk.stem' includes the function 'WordNetLemmatizer'.

Exercise A.4: Vocabulary

Obtain a vocabulary of words (a 'dictionary' objects that contains all the unique words in the quotes.)

Exercise A.5: One-hot encoding

Perform one-hot encoding with respect to the vocabulary.

Exercise A.6: Bag-of-words

Encode every quote in the dataset as a bag-of-words.

Exercise A.7: Bag-of-n-grams

1. Encode every quote as a bag-of-bigrams.
2. Encode every quote as a bag-of-trigrams.

Hint: The Python package 'sklearn.feature_extraction.text' includes the function 'CountVectorizer'.

Exercise A.8: TF-IDF

Encode every quote via the TF-IDF method.

Hint: The Python package 'sklearn.feature_extraction.text' includes the function 'TfidfVectorizer'.