UNIVERSITÄT WIEN
CSLEARN - EDUCATIONAL TECHNOLOGIES

Natural Language Processing

Exercise Sheet 1

Language Processing and Python

Exercise 1

How many words are there in `text2` from `nltk.book`? How many distinct words are there? Calculate the lexical diversity.

Exercise 2

Produce a dispersion plot of the four main protagonists in Sense and Sensibility: Elinor, Marianne, Edward, and Willoughby. What can you observe about the different roles played by the males and females in this novel? Can you identify the couples?

Exercise 3

Find the collocations in `text6`.

Exercise 4

Use only the `index()` function to find all the indexes of the word "sunset" in `text9`.

Exercise 5

What is the difference between the following two lines? Calculate the two values:

```
len(sorted(set(w.lower() for w in text1)))
len(sorted(w.lower() for w in set(text1)))
```

Exercise 6

Write the slice expression that extracts the last two words of `text2`.

Exercise 7

Find all the four-letter words in `text6`. With the help of a frequency distribution (`FreqDist`), show these words in decreasing order of frequency.

Exercise 8

Create a set for the words in `text6`. Use a `for` and an `if` statement to loop over the words in the set and print all titlecased words with more than one character, one per line.

Exercise 9

Write expressions for finding all words in `text6` that meet the conditions listed below:

a) ending in "ing",

b) containing the letter "z",

c) containing the letter sequence "pt".

Exercise 10

Define `sent` to be the list of words `['she', 'sells', 'sea', 'shells', 'by', 'the', 'sea', 'shore']`. Now write code to perform the following tasks:

a) print all words beginning with "sh",

b) print all words longer than four characters.

Exercise 11

What does the following Python code do?

```
sum(len(w) for w in text1)
```

Can you use it to work out the average word length of `text1`?

Exercise 12

Define a function `freq(word, text)` that calculates how often a given word occurs in a text, not using `count()` but a `FreqDist`. Use the function to calculate how often "promise" appears in `text4`.