IN4080, 2021, Exercise set 2: 6. Sept.

The goal of this second lab session is to get practical training in some of the concepts from the lectures, in particular

- Text processing, first steps
- Alternative tokenizations
- Working with tagged text

Text processing, first steps

Exercise 1

We have extracted a text from the web from page https://www.newsinenglish.no/2019/08/24/solberg-wards-off-government-crisis/. It comes bundled with this exercise set. You can also find it on the IFI-cluster at /projects/nlp/in4080/crisis.txt.

Read it into an interactive ipython/notebook session as a string.

```
with open("crisis.txt", 'r') as f:
    raw = f.read()
```

a) Tokenize the raw-string. How many tokens are there in the text? How many different tokens?

b) The text may contain different sorts of tokens. Separate between words, punctuation marks, numbers and other tokens. How many tokens are there in each group? How many different words does the text contain?

c) Inspect each class. Are the tokens sorted correctly? In particular, are there tokens in the *other* class that should belong to one of the other classes? Revise the rules until you are satisfied with the result.

Exercise 2

a) Return to the raw-text and split it into sentences. How many sentences does the text contain?

b) Tokenize the sentences.

d) What is the average sentences length in terms of words? Punctuation marks should not count.

Alternatives for tokenization

Exercise 3

Work through the enclosed notebook called tokenization.ipynb

Working with tagged text

Exercise 4

a) Consider the Brown corpus with the universal tagset. Count how many occurrences there are of each tag and compare to the table from the lectures.

b) How many word forms occur with only one tag, how many with two tags etc.?
What is the largest number of different tags for a word form?
Which word forms are these?

THE END