IN4080, 2020, Exercise set 2: 1. 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
- 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/. (By the way, for they who follow Norwegian politics, it is mystery that this was only one year ago.) You can find it on the IFI-cluster at /projects/nlp/in4080/crisis.txt or download it from the semester web page.

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. 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.

c) How many tokens are there in each group? How many different words does the text contain?

Exercise 2

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

b) Inspect the result of the splitting. Do you see any mistakes? Can you propose steps that could systematically avoid some of these mistakes? Are there mistakes that would be harder to detect? Correct the mistakes.

c) Tokenize the sentences.

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

Working with tagged text

Exercise 3

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?

Exercise 4

a) Consider the NLTK Brown corpus with the full tagset. How many different tags are there? It is many more than the original 87 tags.

b) Inspect some of the tags. (It can also be useful to inspect the frequencies of the various tags.) We will recognize the original tags, like VBD, with a decent frequency (26167). In addition, we find many longer tags, like VBD-HL, and most of them have a low frequency (8).

Upon closer inspection, we see that each of these more complex tags contain an original tag and in addition some suffixes or prefixes. The prefixes or suffixes are not relevant to all tasks, and it may be useful to strip them. For example, VBD-HL, indicates a VBD occurring in a headline. We will change this to VBD. Similar suffixes, we will strip are NC (the token occurs in cited passage), NP (complex title or name), TL (title) (see http://clu.uni.no/icame/manuals/BROWN/).

There is one prefix, FW, meaning foreign word. For the following, we will ignore whether a word is foreign or not and strip off the prefix FW.

In addition to original tags, we are then left with pairs of original tags, like VB+TO and starred tags, like HV*. The plus sign indicates that two words have been contracted into one token, like *wanna* and *it's*. The star means contraction with negation as in *doesn't*. Since, we are interested in the frequencies of different tags we will use a simple form of normalization where we replace, e.g. ('gonna', 'VBG+TO') with two tokens ('gonna', 'VBG') ('gonna', 'TO').

Perform the transformations.

c) What is the frequencies for the remaining tags in the Brown corpus?

d) 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