Universität des Saarlandes
Prof. Dr. Dietrich Klakow
Lehrstuhl für Signalverarbeitung
NNIA Winter Term 2019/2020

Project Instructions (part 1)

Preprocessing Data

Deadline: 23.02.2021, 23:59

Step 1.1 - Download data

- Where to download: folder project/data/sample in Teams contains the following files:
  chtb_0223.gold_conll
  phoenix_0001.gold_conll
  pri_0016.gold_conll
  wsj_1681.gold_conll
- How to combine: see here how to concatenate files from the terminal. Concatenate the 4 files into one file named 'sample.conll'

Step 1.2 - Extract POS tags

In Figure 1 you can see a part of the 'sample.conll' file and the file containing only relevant information.

- Inspect the file 'sample.conll'. The data is in the conll format.
- The original file contains a lot of information that we do not need, so we will create a file 'sample.tsv' that contains only:

  a) position of a word (first column)

  b) the word itself (second column)

  c) POS tag (third column)

  We do not care about the ID's of the documents and sequences, so we ignore the lines starting with # in the 'sample.conll'.
  The sequences should be separated by *.
  The new file 'sample.tsv' should be a file with **tab separated values**, i.e. the columns are separated with tabs.

Step 1.3 - Get information about the data

Create a file 'sample.info' containing the following information about our data:

- Maximum sequence length;

- Minimum sequence length;
- Mean sequence length;
- Number of sequences;
- List of tags and the percentage of the words that have these tags.

An example of such a file you can find in the Figure 2.

Submission instructions

The following instructions are mandatory. If you are not following them, tutors can decide to not correct your exercise.

- For this part you do NOT have to submit anything. The submission instructions will follow later.
- If you have any problems or questions, **use the channel 'Project Questions'** in Teams. Feel free to answer the questions of your fellow students.

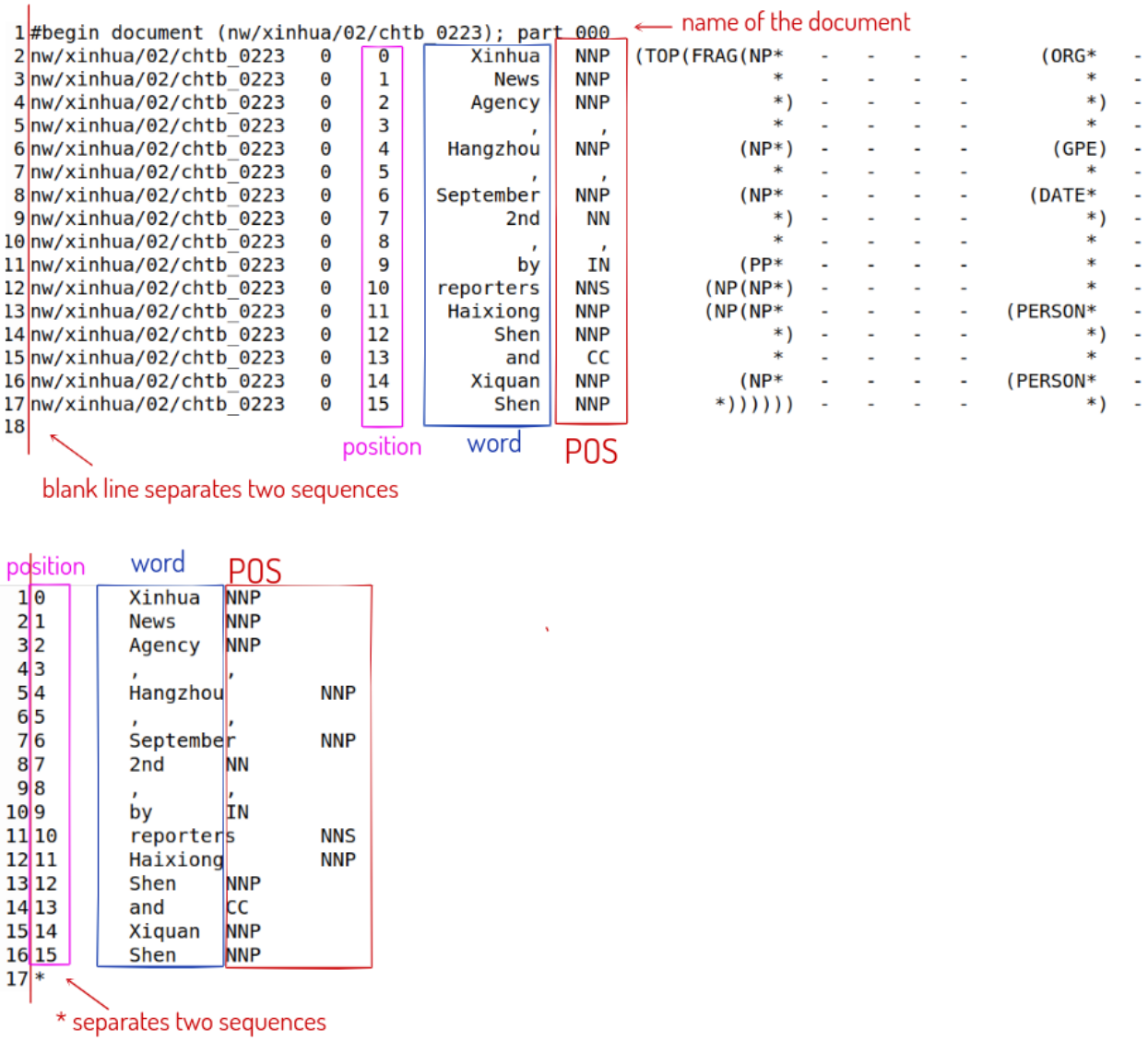

Figure 1: sample.conll (top) and sample.tsv (bottom)

```
Max sequence length: 73
Min sequence length: 2
Mean sequence length: 18.74757281553398
Number of sequences: 309

Tags:
''      0.00%
,       0.06%
-LRB-   0.00%
-RRB-   0.00%
.       0.05%
:       0.00%
CC      0.03%
CD      0.02%
DT      0.10%
```

Figure 2: sample.info file