



---

# 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.

1 #begin document (nw/xinhua/02/chtb\_0223); part 000 ← name of the document  
2 nw/xinhua/02/chtb\_0223 0 0 Xinhua NNP (TOP (FRAG (NP\* - - - - (ORG\* -  
3 nw/xinhua/02/chtb\_0223 0 1 News NNP \* - - - - \*) -  
4 nw/xinhua/02/chtb\_0223 0 2 Agency NNP \*) - - - - \*) -  
5 nw/xinhua/02/chtb\_0223 0 3 , , \* - - - - \* -  
6 nw/xinhua/02/chtb\_0223 0 4 Hangzhou NNP (NP\* - - - - (GPE) -  
7 nw/xinhua/02/chtb\_0223 0 5 , , \* - - - - \* -  
8 nw/xinhua/02/chtb\_0223 0 6 September NNP (NP\* - - - - (DATE\* -  
9 nw/xinhua/02/chtb\_0223 0 7 2nd NN \*) - - - - \*) -  
10 nw/xinhua/02/chtb\_0223 0 8 , , \* - - - - \* -  
11 nw/xinhua/02/chtb\_0223 0 9 by IN (PP\* - - - - \* -  
12 nw/xinhua/02/chtb\_0223 0 10 reporters NNS (NP (NP\* - - - - \* -  
13 nw/xinhua/02/chtb\_0223 0 11 Haixiong NNP (NP (NP\* - - - - (PERSON\* -  
14 nw/xinhua/02/chtb\_0223 0 12 Shen NNP \*) - - - - \*) -  
15 nw/xinhua/02/chtb\_0223 0 13 and CC \* - - - - \* -  
16 nw/xinhua/02/chtb\_0223 0 14 Xiquan NNP (NP\* - - - - (PERSON\* -  
17 nw/xinhua/02/chtb\_0223 0 15 Shen NNP \*) ) ) ) ) ) - - - - \*) -  
18

blank line separates two sequences

position word POS

position	word	POS
1 0	Xinhua	NNP
2 1	News	NNP
3 2	Agency	NNP
4 3	,	,
5 4	Hangzhou	NNP
6 5	,	,
7 6	September	NNP
8 7	2nd	NN
9 8	,	,
10 9	by	IN
11 10	reporters	NNS
12 11	Haixiong	NNP
13 12	Shen	NNP
14 13	and	CC
15 14	Xiquan	NNP
16 15	Shen	NNP
17 *		

\* separates two sequences

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