Take Home Assignment

Intermediate Methods and Programming in Digital Linguistics

Noun Chunk Counting

Using spaCy's built-in functionality for noun chunk identification, write a program `nc_counter.py` that lists the most frequent noun chunks found in a given text file.

Example program invocation and output (on the command line):

```
$ python nc_counter.py corpus.txt

55 I
22 we
19 it
19 you
17 Tajikistan
7 order
7 me
7 us
6 the Community
5 the European Union
```

The command-line interface (CLI) of your program should support an optional flag `--min-words` that takes a number. If present, only noun chunks with at least `--min-words` should be considered:

```
$ python nc_counter.py --min-words 3 corpus.txt

5 the European Union
4 the Member States
2 a greater success
2 the New Year
2 a technical correction
2 a procedural motion
1 their own countries
1 the Berlin Summit
1 a limited budget
1 little political interest
```

Make sure that your program is able to handle large text files, and that it doesn't break if there are no noun chunks in a text. You can use the enclosed `corpus.txt` file for testing, but your program must be able to handle any – and substantially larger – text files.

Documentation

https://spacy.io/usage/linguistic-features#dependency-parse

Grading Criteria

Your grade will be composed as follows:

- 75% Functional correctness
- 25% Style (naming conventions, code formatting, docstrings, etc.)

Type hints and unit tests will not be strictly required, but reflected positively in your grade if implemented properly.

~~Working in Pairs~~

This assignment is to be completed individually. Working in pairs or groups is not permitted.

Submission

Submission is due by Monday, 7 March, 18.00h CET as a single .zip or .tar.gz archive via OLAT. Please only include Python source files. Any other files (such as test texts or written instructions) will be ignored – your code, including inline comments and docstrings, should be self-explanatory.