TP3

This assignment focuses on the NLP task known as named entity recognition (NER), which consists in identifying the spans that refer to specific real-world entities, as opposed to common words, which denote classes (or categories) of phenomena. For instance, university is a common word which denotes all higher-education institutions where multiple disciplines are studied and taught, but the span University of Geneva refers to a single, particular entity that exists in the real world.

We will work with annotated data in several languages using two Python libraries.

Data

We use one data set per language, in particular:

- UNER_Portuguese-Bosque
- UNER_Chinese-GSDSIMP
- UNER_Swedish-Talbanken
- UNER_Serbian-SET
- UNER_Slovak-SNK
- UNER_Croatian-SET
- UNER_English-EWT
- UNER_Danish-DDT

Tools

- spaCy is a general NLP tool that can perform various tasks including NER
- python-crfsuite is a more specific tool for sequence classification, most commonly used for NER

What to do

Your task is to find the appropriate way to use the two libraries and make NER predictions for the test portion of each data set. You will compare the output of the models with the gold labels given in the data and print one full classification report with the F-score evaluation per language and per tool. The reprot should contain per-class and overall performance. Save all the spaCy reports in the text file spacy_rerports.txt and the CRF reports in crf_reports.txt.

For creating classification reports, you can use sklearn or your own functions. Submit to moodle as a ZIP archive:

- Script 1 named spacy_tp3.py
- Script 2 named crf_tp3.py
- spacy_rerports.txt
- crf_reports.txt
- README.txt containing the following (and in this order)
  1. instructions for running your scripts
  2. running time for both scripts
  3. options selected for running spaCy