Assignment 2: sequence labelling

Computational Linguistics course

This is an **individual hand-in assignment**. Your report should not be longer than 3 pages

Goals of this assignment

- You can pre-process existing annotated text data into the data structure that you need for classifier learning
- You can perform a sequence labelling task with annotated data in CRFsuite
- You understand the effect of using different features and context sizes
- You can evaluate sequence labelling with the suitable evaluation metrics

Preliminaries

- You have followed the CRFsuite tutorial https://sklearn-crfsuite.readthedocs.io/en/latest/tutorial.html
- You have all the required Python packages installed

Tasks

1. Go to https://www.kaggle.com/nltkdata/conll-corpora and download all files for conll2002. You can choose either the Dutch (ned*) or Spanish (esp*) files.
2. Add a function to the CRFsuite tutorial script that reads the train and test files and processes them in the same way as the benchmark data from the tutorial (an array of triples `(word,pos,biotag)`).
3. Run a baseline run with the features directly copied from the tutorial.
4. Extend the features: add a larger context (-2 .. +2) and engineer a few other features that might be relevant for this task
5. Experiment with the effect of different feature sets on the quality of the labelling.

Write a two-page report (3 pages is the hard maximum) in which you:

- describe the data (give a few statistics)
- describe the features;
- show a results table with both the baseline results and a number of experimental results with changes in the feature set;
- write your conclusions.