Deep Learning in NLP

Homework P1

Solution to be sent (pdf, zip) to `waszczuk@phil.hhu.de` and `cwurm@phil.hhu.de` by 27.10.2019. You can do homework in pairs.

Exercise 1

Download the language recognition dataset from the course's website. We will use this dataset later in order to train a neural network to classify person names according to their language (EN, DE, FR, . . .). Before we do that, however, we shall recall the good practice of first leaving out some parts of the dataset to perform development and evaluation:

- The development part (`dev` for short) serves to select the best architecture and/or hyperparameters for the particular task.
- The `test` part is used at the very end to evaluate the final model.

The goal of this exercise is to:

- Read the dataset with person names to an appropriate data structure.
- Divide it randomly into three parts: `train`, `dev`, and `test`. The relative sizes (e.g., 80%/10%/10%) of the respective parts should be arguments of the splitting function.
- Store the resulting `train`, `dev`, and `test` parts into three separate files.

A preliminary, partial solution to this exercise is available for download on the course's webpage. More hints can be found in the code. The missing parts are marked with `TODO`s.

Update 15/10:

- Hint: the `random` module can be useful to solve the exercise. One of the functions mentioned in the Randomness section of the Python refresher should do the job.
- Clarification: you should strive to make all possible `train`/`dev`/`test` splits of the dataset equally likely (formally: assume uniform distribution over possible splits).

General note: You do not have to stick strictly to the structures suggested in the code. In any case, make sure that anyone can understand your code, for example, by using comments and assertions.