Third Lab

Língua Natural – MEIC-A e MEIC-T

Luísa Coheur

luisa.coheur@inesc-id.pt

2019

Goal:

- Conduct a proper experiment;
- Test similarity measures.

1. In Table 1 you can find a golden collection, representing the "perfect" results in a Named Entity Recognition (NER) task, and, in addition, the results obtained by a system, called NER-X, in that task. Determine the precision and recall obtained by NER-X.

Reference	Jared, Drew, Albert, Kyoko, Archie, Paula, Dame Penelope, Jean Lovegood III, Kyle, Rita, Lois, Lois, Rita, Marion, Raylan, Pavel, Kathy, Stelu, Scarlet, Nicole, Marie
NER-X	Lisboa, Jared, Drew, Kyoko, Dame, Jean, Marion, Pavel, Que, Kathy, Argentina, Marie

Table 1: Reference vs. NER-X

2. In this lab you will also develop a program that classifies sentences based on a given taxonomy. A baseline is implemented in DIST.py (to run it: python3 DIST.py). It works as follows:

   - First, the development set is split in two sets: sentences and associated (correct) tags.
   - Then, the tag that should be given to each sentence of the development set, based on similarity measures, and taking the training set as a "knowledge base"[1] is (automatically) determined. The idea is the following: each sentence from the development set is compared with all the sentences in the training set; the tag of the most similar (or less different) sentence in the training set will give its tag to the sentence from the development set that is being analysed.
   - Finally, the obtained tag is compared with the correct one from the development set.

   Analyse the given corpora (***don't look at the test set***). Run DIST.py. Analyse the obtained results (the evaluation measure is *accuracy*). Try to improve the baseline (pre-procesing or new measures; some code is already implemented in DIST.py). You can find more instructions in the script.

3. At the end, test your work with the test set.

[1] Note that the training set, despite its name, it is not used as a traditional machine learning training set; it is just a "knowledge base".