02.01.2023

AYDIN ADNAN MENDERES UNIVERSITY

CSE 431 NATURAL LANGUAGE PROCESSING WITH MACHINE LEARNING

FOURTH TERM PROJECT

Due Date: 17.01.2023 23:59

Your third term project is related to named entity recognition (NER) using Conditional Random Fields (CRF) in Turkish.

Initially, you need a labeled dataset (MilliyetNER dataset) [1] involving the words with their NER tags in Turkish. Then apply the same steps in Figure 1 of [2]. You can benefit from [2] and [3] for the features used in CRF.

After appying the steps in [2], show the accuracy, precision, recall and F1 scores for sentence and word levels in dev/test sets. [4] involves the steps and the sample code in detail for an English dataset. You can make modifications by changing the features for Turkish and obtain the results similarly as given in [4].

Finally, compare the results (accuracy, precision, recall and F1 scores for sentence and word levels) of HMM and CRF based solutions in one table.

[1] https://data.tdd.ai/#/effafb5f-ebfc-4e5c-9a63-4f709ec1a135

[2] Şeker, G., & Eryiğit, G. (2018). State of the art in Turkish named entity recognition.

[3] ÇEKINEL, R. F., AGRIMAN, M., KARAGÖZ, P., & YILMAZ, B. Türkçe Haber Metinlerinde Sartlı Rastgele Alanlar ile Varlık Ismi Tanıma: Özniteliklerin Gözden Geçirilmesi Named Entity Recognition with Conditional Random Fields on Turkish News Dataset.

[4] https://github.com/TeamHG-Memex/sklearn-crfsuite/blob/master/docs/CoNLL2002.ipynb