Guidelines for ML4NLP Competition

USTH Master 2 - ICT Department

This competition is organized within the teaching unit *Machine Learning and Deep Learning* and will be part of evaluation

Introduction

In this competition, you will have to develop a system/model/framework to perform Named-Entity Recognition (introduced in lecture and in notebook tutorial).

The participation to the competition is mandatory and will be evaluated for the final mark of the Teaching Unit.

You will have to submit :

- the implementation of your models. You can submit several version if necessary (for example to compare several methods or models)
- an official submission to the platform where the competition is hosted : https://codalab.lisn.upsaclay.fr/competitions/17508
- a report presenting your understanding of the problem, presentation of your solution (explanation of the models), discussion on results and conclusion

Due date : **Wednesday, February 7th 2024**

Resources

To help you in the development for your system you will have access to several resources :

- A dataset available on Moodle made of training set (train data `train_data.csv` and train groundtruth `train_gt.csv`, validation data `valid_data.csv` and validation groundtruth `valid_gt`). These files should be used to perform teh development of your system (training model if you are using a mchine learning model)
- an evaluation script in python `eval_stud.py` that is the same as the one used on the platform. This should help you to verify and validate the output format of your model.

Submission of results

First, You must subscribe to codalab platform and register to the competition (linked above). The platform provides an automatic evaluation system of an output file generated by your system.

The submitted file must follow these requirements :

- the extension of the text file is csv, but this text file must be **ZIPPED** (Mac Users must be careful because macos system include a `_MACOSX_` file by default, you are enclined to use `zip` command line in terminal to generate your zip file)
- CSV File is formatted like the `*_gt.csv` files that are given in train dataset. To be clear, your csv file should contain the same number of lines as the `test_data.csv` file provided, with the identification of the label for each token of `test_data.csv` (*i.e.* The Tokens are labeled under one of the following tags :
  - O
  - B-ORG
  - I-ORG
  - B-PER
  - I-PER
  - B-MISC
  - B-MISC
  - B-LOC
  - I-LOC

Assistance

To help you in this competion, you are enclined to use knowledge acquire during the class but you are also encouraged to have a look to the platform https://huggingface.co gathering a lot of resources for nlp problems.

For any help you can contact me on : nicolas.sidere