Machine Learning Challenge

Build a Part-Of-Speech tagger.

Deliverables:

- Source code to train the POS tagger available on Github. Add `@jameslafa` and `@abustany` as a collaborator.
- Three scripts, e.g. `python train.py [train_file] [dev_file]`, `python eval.py [test_file]` and `python generate.py [text_file]` that allow us to train and test your model (accuracy), as well as to generate POS tags for unlabelled text (one sentence per line, tokenized).
- `README.md` containing the following information:
  - what assumptions did you make?
  - what is your testing strategy and what is your degree of confidence in the model?
  - list of trade-off you made
  - how much time did you spend on the challenge (please be honest)

What do we evaluate:

- The quality, clarity, and maintainability of your source code.
- Efficiency and effectiveness of the chosen algorithm, also with respect to batching and normalization
- Ease with which to use the code.

What we do not evaluate:

- This code is not going on production, security is not a concern
- Performance is not a strong criterion but it's worth adding a comment when you know you could do something more performant but decide to make a compromise for a lack of time.

Annex

Corpus

Please train the model on the Georgetown University Multilayer Corpus (https://github.com/UniversalDependencies/UD_English-GUM/tree/master).

POS Tag Sets

We will use the part-of-speech tags used in the Universal Dependencies Project:

ADJ	adjective
ADP	adposition
ADV	adverb
AUX	auxiliary
CCONJ	coordinating conjunction
DET	determiner
INTJ	interjection
NOUN	noun
NUM	numeral
PART	particle
PRON	pronoun
PROPN	proper noun
PUNCT	punctuation
SCONJ	subordinating conjunction
SYM	symbol
VERB	verb
X	other