Twitter:

Para rodar o projeto:

1. Acessar a pasta ‘/crf_to_ner’
2. No terminal executar: ‘python3 -m venv .env’ (para setar um ambiente virtual local)
3. No terminal executar: ‘source .env/bin/activate’ (nesta pasta o python default sera 3.x)
4. No terminal executar: ‘pip install -U pandas scikit-learn sklearn-crfsuite spacy’ (instalar dependências)
5. No terminal executar: ‘python -m spacy download pt_core_news_sm’
6. No terminal executar: ‘python crftoner.py’ (rodar o código em si)

Implentação:
Após a leitura dos tweets, foi feito um POS tagging utilizando a biblioteca SPACy.

After tokenization, spaCy can parse and tag a given Doc. This is where the statistical model comes in, which enables spaCy to make a prediction of which tag or label most likely applies in this context. A model consists of binary data and is produced by showing a system enough examples for it to make predictions that generalize across the language – for example, a word following “the” in English is most likely a noun.

A modificação das features não resultou em nenhuma melhora significante no resultado final.

Leis:

Implementação

O foco na implementação foi na qualidade do dataset. Ele foi normalizado de uma forma consiza, em fique fácil a extração de características pelo CRF. Com esse dataset, vários testes foram realizados, onde o melhor foi sem POS Tag, assim, apenas as colunas ‘Word’ e ‘BIO’.

Para rodar o código é necessário estar no Google Colab, ou se rodar localmente é ter os seguintes módulos:

1) Python3
2) sklearn_crfsuite
3) pandas
4) nltk