tetis

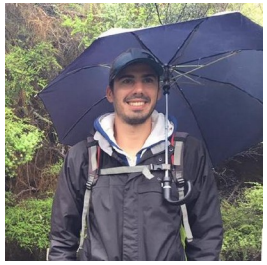
TERRITOIRE ENVIRONNEMENT TÉLÉDÉTECTION
INFORMATION SPATIALE

TETIS Text Mining

Rémy Decoupes
Nejat Arinik
Roberto Interdonato

ITU GeoAI LMR

Introduction



Rémy Decoupes

INRAE

Nejat Arinik

INRAE

Roberto Interdonato

cirad

Motivation

In the context of our research, we apply text mining:

- Epidemiology Event Based Surveillance
- Food Security

On tasks such as:

- Graph based analysis
- Text classification

… And we already use QCRI data (crisisNLP)

tetis

TERRITOIRE ENVIRONNEMENT TÉLÉDÉTECTION
INFORMATION SPATIALE

Methodology

Hypothesis

1. Fine-tune Bert like models accessible on HuggingFace

2. Use a Gazetteer (OSM) to improve results

3. Apply Data augmentation to enlarge the training dataset

Hypothesis

Reality

1. Fine-tune Bert like models accessible on HuggingFace
→ The best trained models are the best models
(see Where to start? Analyzing the potential value of intermediate models)

2. Use a Gazetteer (OSM) to improve results
→ The models were good enough for City / State / Country

3. Apply Data augmentation to enlarge the training dataset
→ It introduces to much noise

Architecture

Draw.io schema

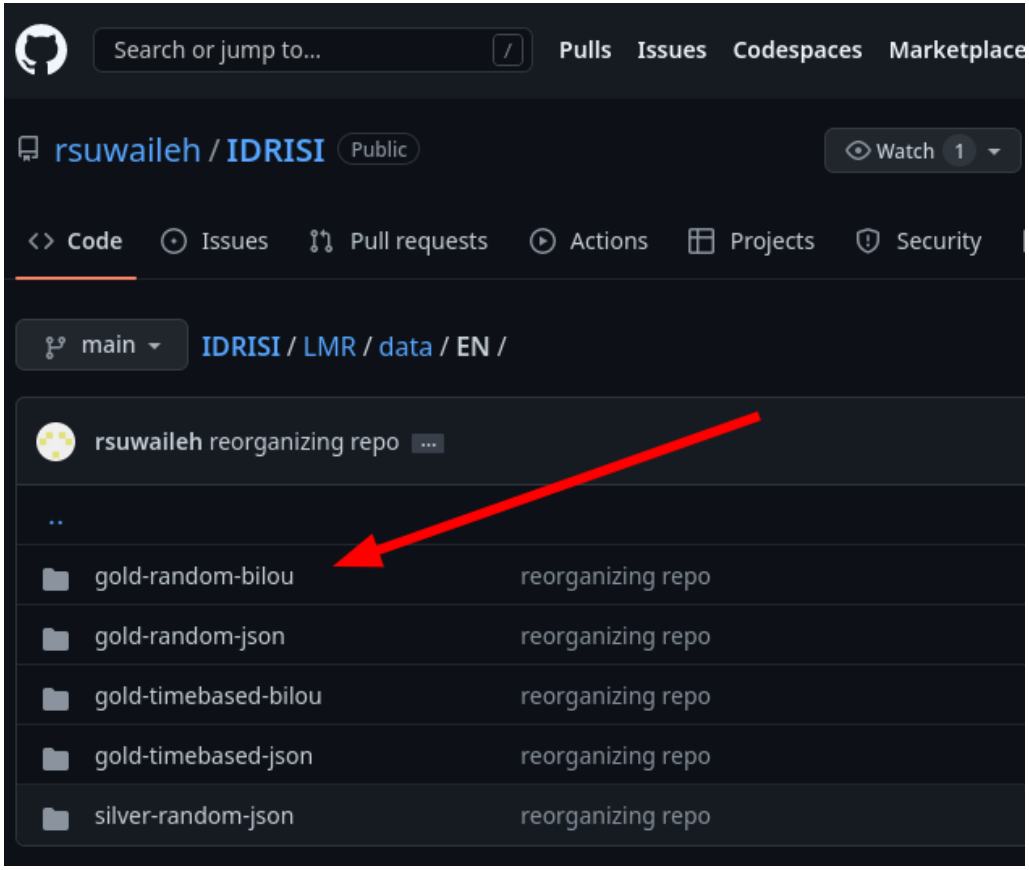
Input: BILOU file
Models: differents models from huggingFace
Engrenage: data augmentation + hyperparameters tuning

Then we packaged it into a dockers (with a post-processing steps)


tetis
TERRITOIRE ENVIRONNEMENT TÉLÉDÉTECTION
INFORMATION SPATIALE

Evaluation

Training data



Mettre taille des corpus

Training data

```
label_encoding_dict =
{"B-CONT" : 1, "B-CTRY" : 2, "B-STAT" : 3, "B-CNTY" : 4, "B-CITY" : 5, "B-DIST" : 6,
"B-NBHD" : 7, "B-ISL" : 8, "B-NPOI" : 9, "B-HPOI" : 10, "B-ST" : 11, "B-OTHR" : 12, "I-
CONT" : 13, "I-CTRY" : 14, "I-STAT" : 15, "I-CNTY" : 16, "I-CITY" : 17, "I-DIST" : 18,
"I-NBHD" : 19, "I-ISL" : 20, "I-NPOI" : 21, "I-HPOI" : 22, "I-ST" : 23, "I-OTHR" : 24,
"L-CONT" : 25, "L-CTRY" : 26, "L-STAT" : 27, "L-CNTY" : 28, "L-CITY" : 29, "L-DIST" :
30, "L-NBHD" : 31, "L-ISL" : 32, "L-NPOI" : 33, "L-HPOI" : 34, "L-ST" : 35, "L-OTHR" :
36, "U-CONT" : 37, "U-CTRY" : 38, "U-STAT" : 39, "U-CNTY" : 40, "U-CITY" : 41, "U-DIST"
: 42, "U-NBHD" : 43, "U-ISL" : 44, "U-NPOI" : 45, "U-HPOI" : 46, "U-ST" : 47, "U-OTHR"
: 48,
"O":0}
```

```
B: Begining of a NER
I: Inside the current NER
L: Last: the final token of a multi-token
U: Unit: a single-token entity
O: Out: a non-entity token
```

Tuning the model

Compare the results with baseline

Issues to address

Mettre une table latex comparant les types full : pour dire ce qu'on peut améliorer

tetis
TERRITOIRE ENVIRONNEMENT TÉLÉDÉTECTION
INFORMATION SPATIALE

Takeaway messages

Good to know

1. Use a tools to tracks experiments (such as mlflow)

2. As always analyse manually the data