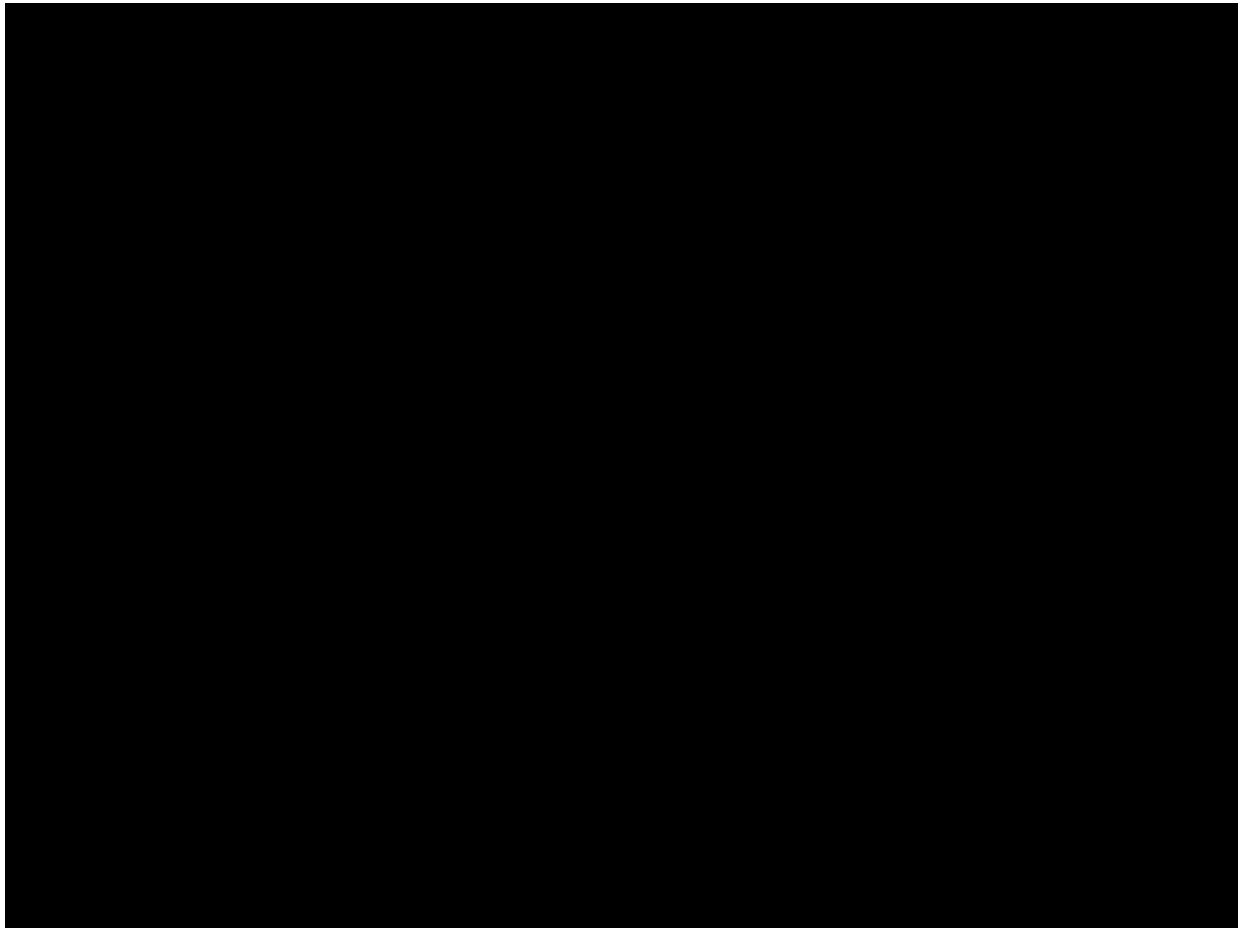
EFREI / LMG HACKATON

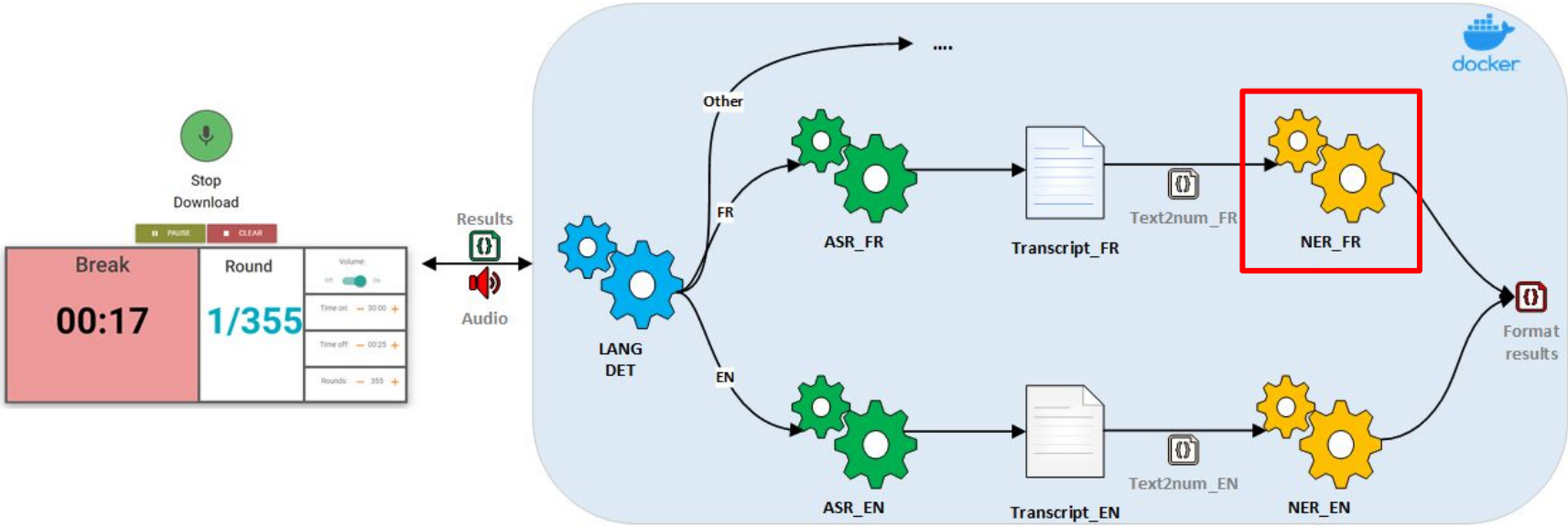
BUILD A NLP MODEL FROM SCRATCH

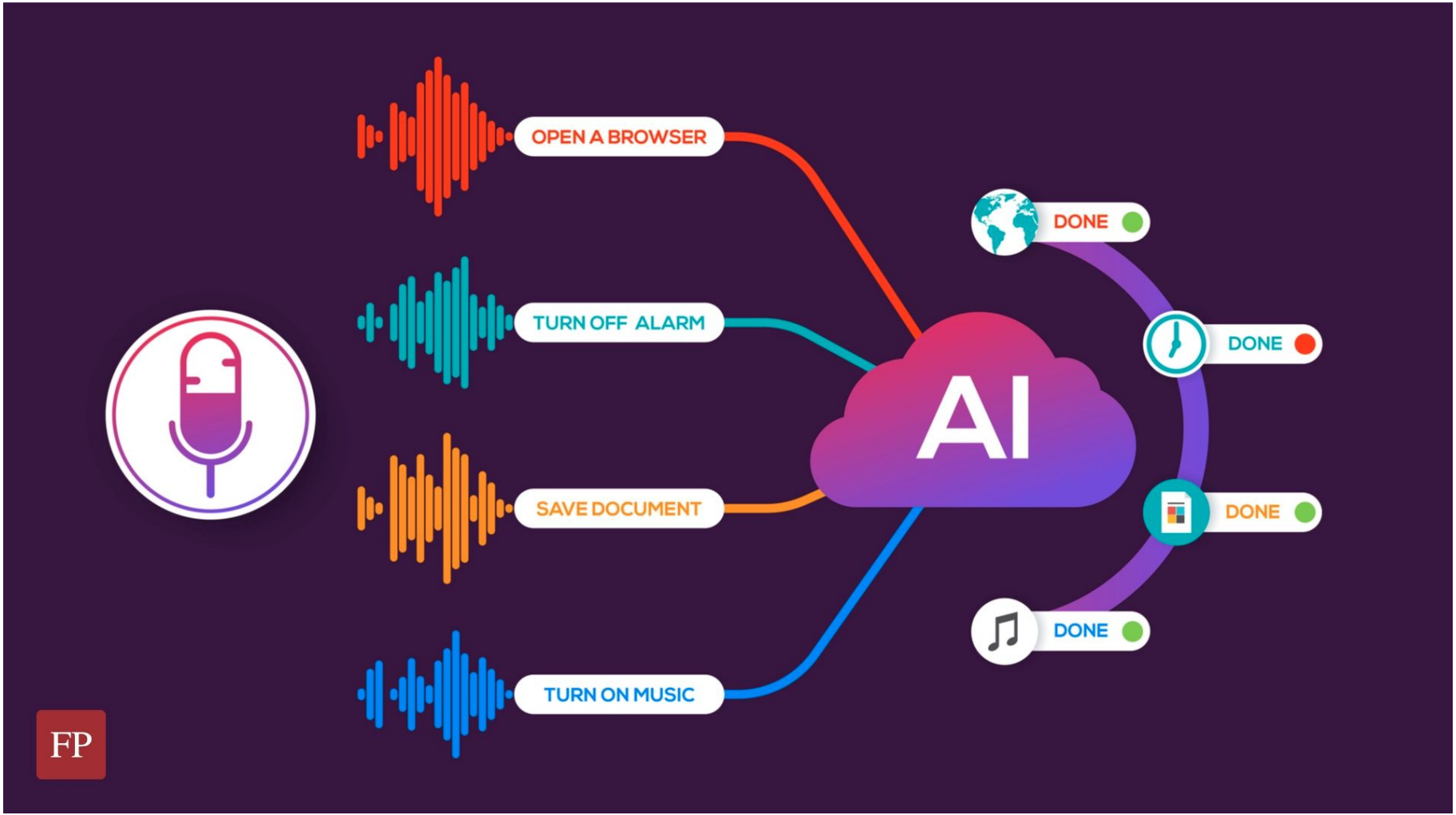
DEMO



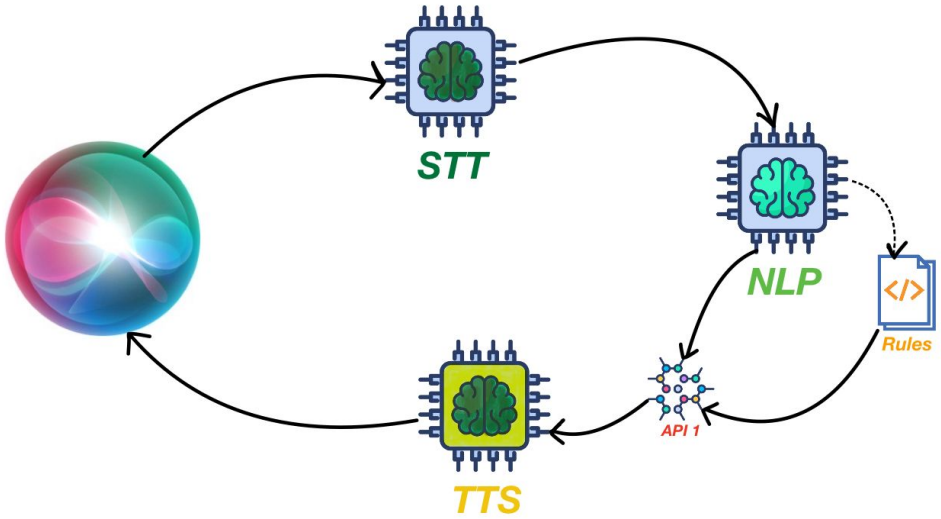
SpeechTimer : Use your voice to setup a sport interval timer

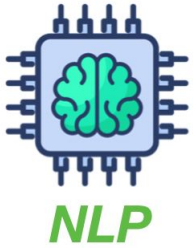
- The camp consists on building from scratch NLP/NER model in French
- Goal: a more complex dataset and a better model




OPEN A BROWSER
TURN OFF ALARM
SAVE DOCUMENT
TURN ON MUSIC
AI
DONE
DONE
DONE
DONE
FP


JUL
17


STT
NLP
Rules
API 1
TTS



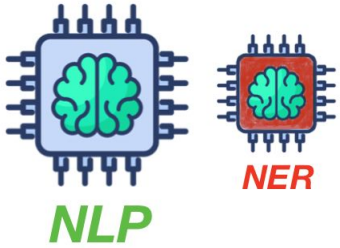
NLP : Natural Language Processing
Set of techniques for understanding, analyzing and extracting text

Technological advancements timeline :

- ML (BOW / TF-IDF) > RNN > LSTM > BERT (2019)

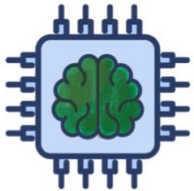
Some types on NLP models :

- Classification
- Named Entity Recognition
- Translation
- Sentiment Analysis
- Q/A
- …



Place Time

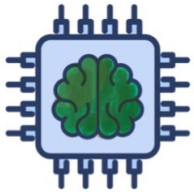
I hear Berlin is wonderful in the winter

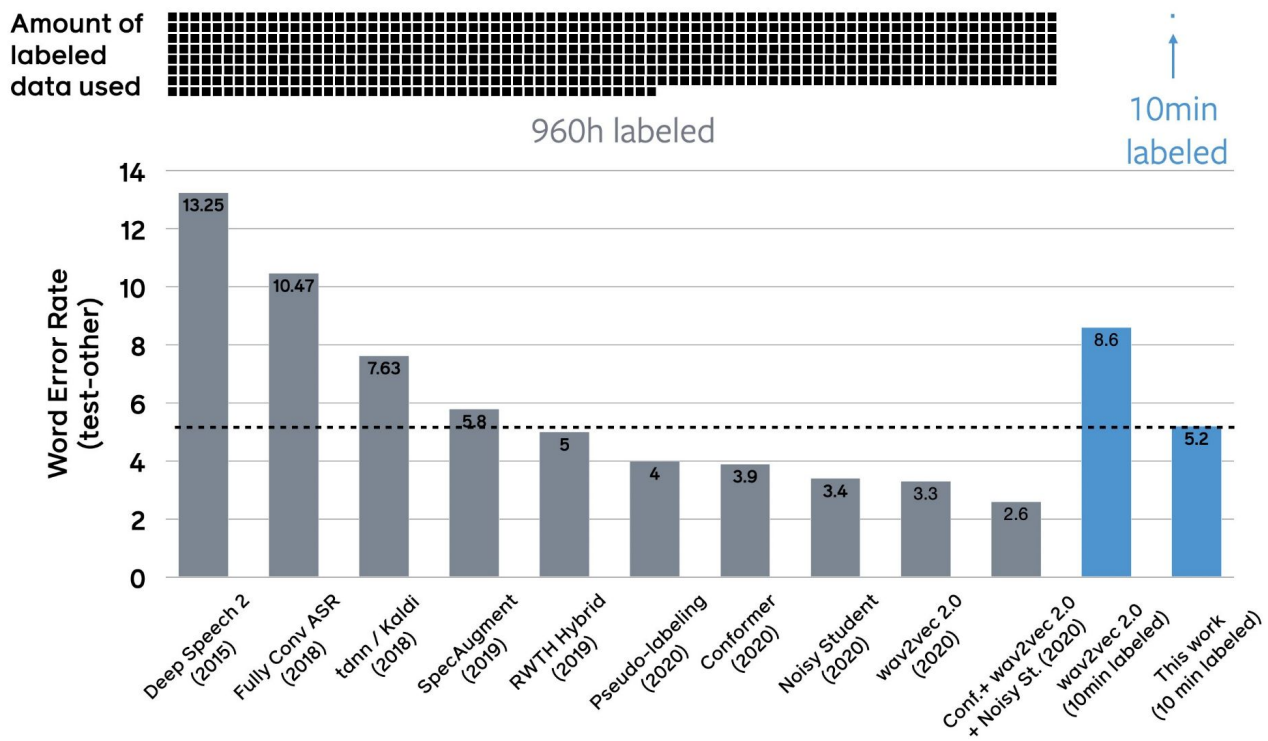


ASR : Automatic Speech Recognition

Technological advancements timeline :

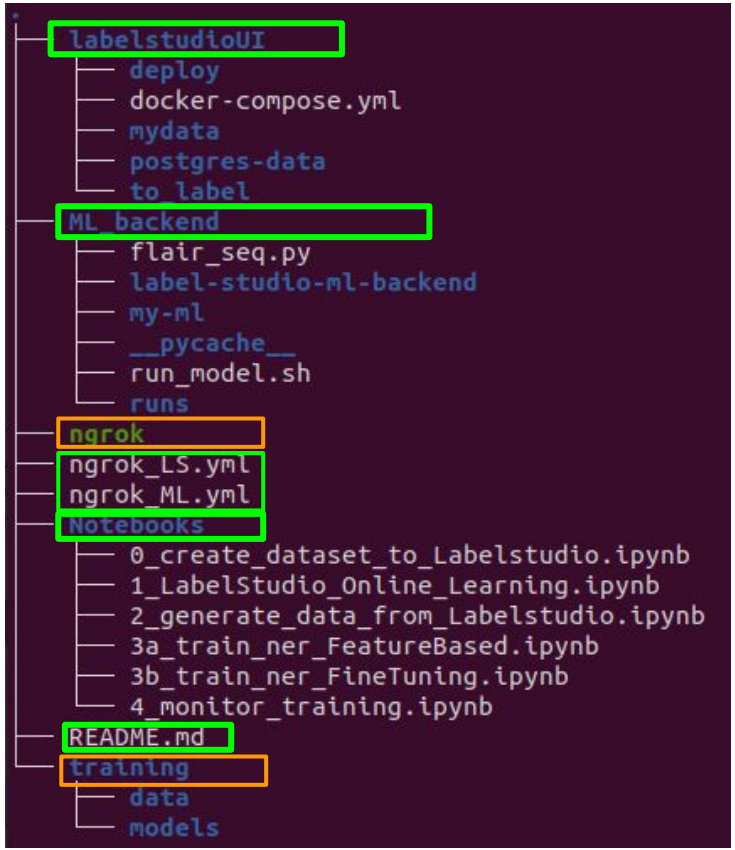
- Standard: Kaldi (acoustic model to train phonemes - langage model to convert phonemes to readable text)
- End-to-End : Deepspeech (2014)
- Self-supervised : Wav2vec2.0 (2020)



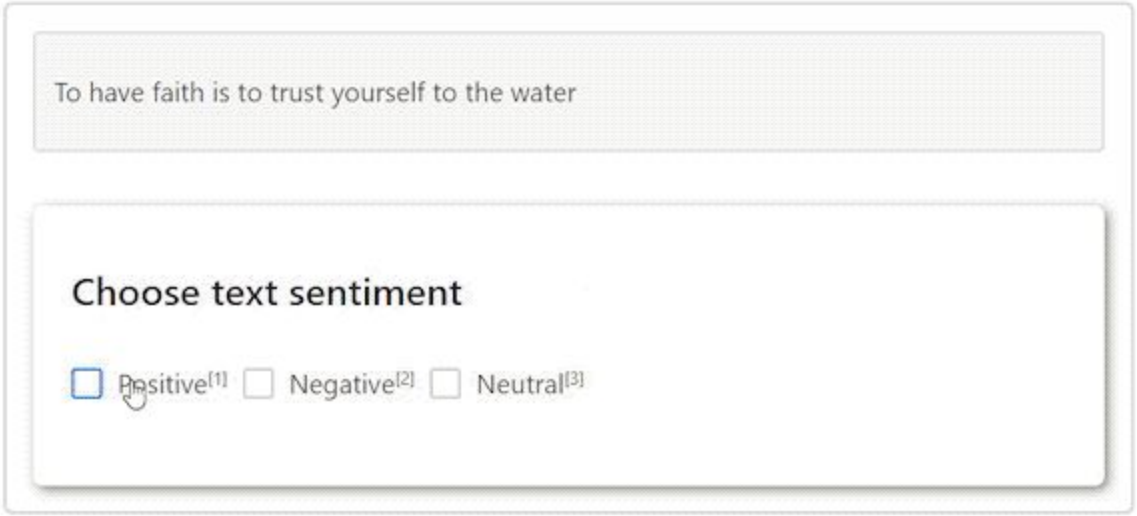


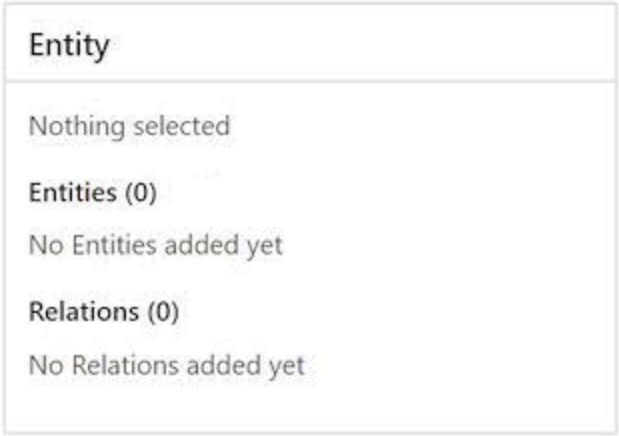
Librispeech benchmark, WER on test-other

Project folders

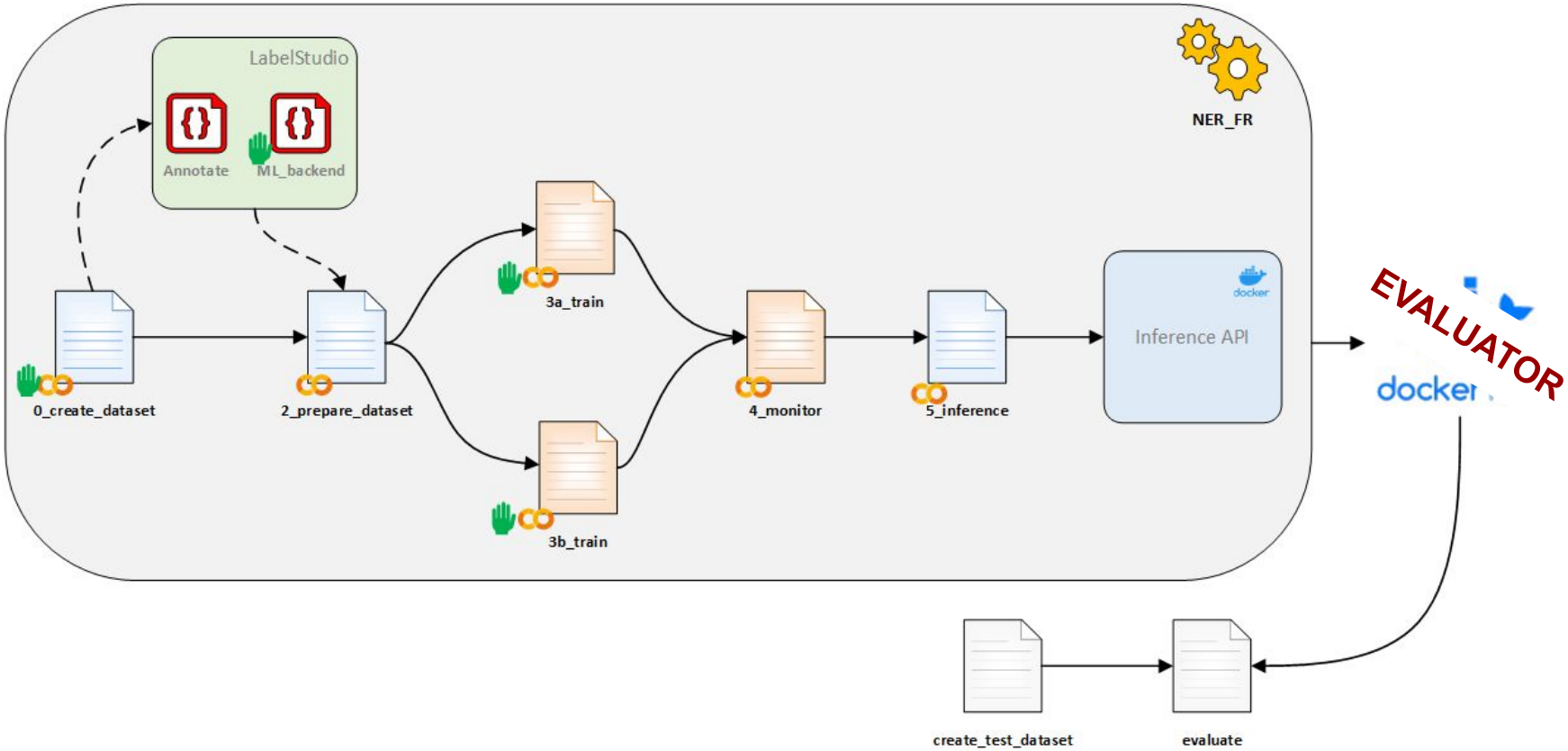


Text Classification





Application / NER French



Day main stages

STAFF DEADLINE

- Send project link 9h45
- Upload project to Google Drive
- Setup Labelstudio in PC
- Build Dataset & extract labeled data - Notebook 0 & 2
- Start labelling
- Send Evaluator link 11h30
- Setup Labelstudio Model backend in PC
- Submit a basic model to Evaluator before 13h30
  - API
  - Model
- (Pre-)Labelling + correction and monitoring
- Training and optimization (model architecture, hyperparams) - Notebook 3a & 3b
- Test final model before submission - Notebook 5
- Submit final model & Api (dernier 17h15)
- Display winners by 18h30

NOTES:

- Pay attention to the download time, the Model can be large, up to 2Gb (it can take up to 20 min to upload)

Sourcing / setup labelling tool

Dataset is built with code (Notebook 0)

- You can edit code parts:
  - Dataset volume
  - Get representative dataset

- The dataset template contains already all the type of sentences needed for this use case

NOTES:

- This step should be done once, choose carefully the dataset size you will be using during all the day

Dataset / Labelling

Goal is to add more complexity to the Dataset:

1 er ajout

- 19 séries de 3 minutes 30duration_wt_sd 21 initie 25 NB_ROUNDS seminutes 25duration_br_sd entre chaque série série

2ème ajout

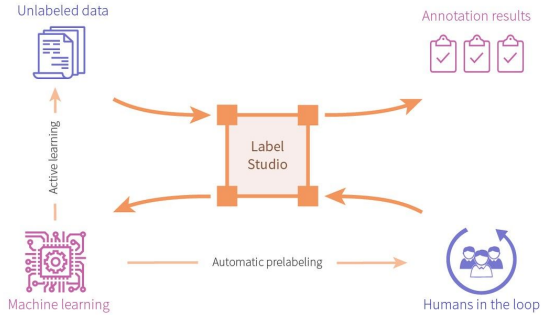
- pr[illegible]
- 3 sets de 20 heures **et quart**duration_wt_min
- 3 sets de 20 heures **moins quart**duration_wt_min débute 3 NB_ROUNDS séq[illegible] DURATION_WT_MIN minutes 51 DURATION_BR_MIN minute 20 DURATION_BR_SD entre series
- 3 sets de 3 minutes, 3 heures **et demi**duration_br_min entre chaque série

tag	meaning
nb_rounds	Number of rounds
duration_br_sd	Duration btwn rounds in seconds
duration_br_min	Duration btwn rounds in minutes
duration_br_hr	Duration btwn rounds in hours
duration_wt_sd	workout duration in seconds
duration_wt_min	workout duration in minutes
duration_wt_hr	workout duration in hours

NOTES:

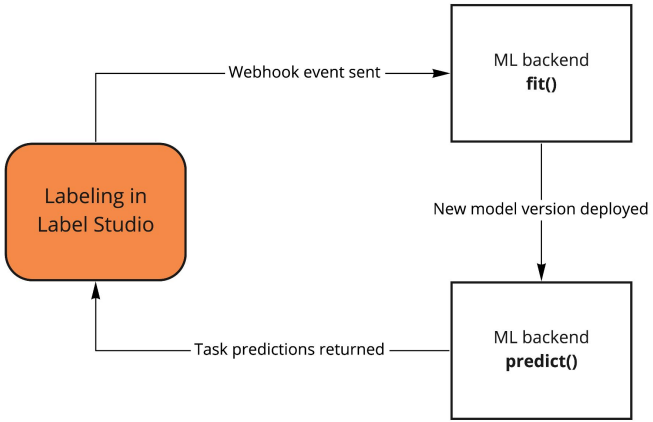
- Highlight only numeric characters (except “et quart” “moins quart”...)
- Avoid making mistakes, 10 sentences labelled correctly better than 50 with 20% errors
- Check instructions carefully before starting to label
- You have all admin rights in the UI, be careful what you’re doing (Ex. remove labeled data !)

Model pre-labelling setup



Model pre-labelling is a manually intensive Task !
Automatic pre-labelling / Active learning can help!

- Model templates:
  - Flair seq2seq (LSTM + CRF)
  - Flair Transformers (fine-tuning Transformer pretrained models)



NOTES:

- At least 50 annotated sentences before running the first training
- Take into account the training time before re-launching a model
- The more data there is, the longer the model takes to train
- One person per Team is charged to Train the model - Communicate with your teammates

Training / optimization

- 2 Notebooks templates for training use 2 different approaches
  - Lstm + CRF
  - Transformers
- Included Hyper-parameter optimization / real time monitoring on Tensorboard
- Check links and tags in training notebooks for more detail
  - #FINETUNE #TIPS #PAPER

NOTES:

- Start this step only when you decide to stop the annotation (except for the basic model)
- Model evaluation is not based only on performance ! (check Final Evaluation slide)

Model uploading to Evaluator

- Notebook 5 to test and evaluate your model and Api before uploading

- To submit your model:
  - if <= 500 Mo:
    - submit to EVALUATOR App
  - if > 500 Mo:
    - Share downloadable link by email (respond to TOKEN EMAIL)
      - Google drive
      - Wetransfer
- To submit your API:
  - Submit to EVALUATOR App

NOTES:

- Start this step only when you decide to stop the annotation (except for the quick & dirty model)

Final evaluation

- Deliverables
  - Api (Https link) to submit to Evaluator
  - a model named model.pt

- Metrics used for evaluation:
  - Performance (F1-micro)
  - Model size (Mb)
  - Inference speed (seconds)

Penalties :

- Teams submit similar model (may infringe School regulations)
- Api URL not submitted
- Late submission

GOOD LUCK !