PROJET ETL

PRESENTATION

Pipeline complet de traitement de données textuelles
Extract Transform Load

wail brimesse

SOMMAIRE

1 PRÉSENTATION DU PROJET

2 GESTION DE PROJET

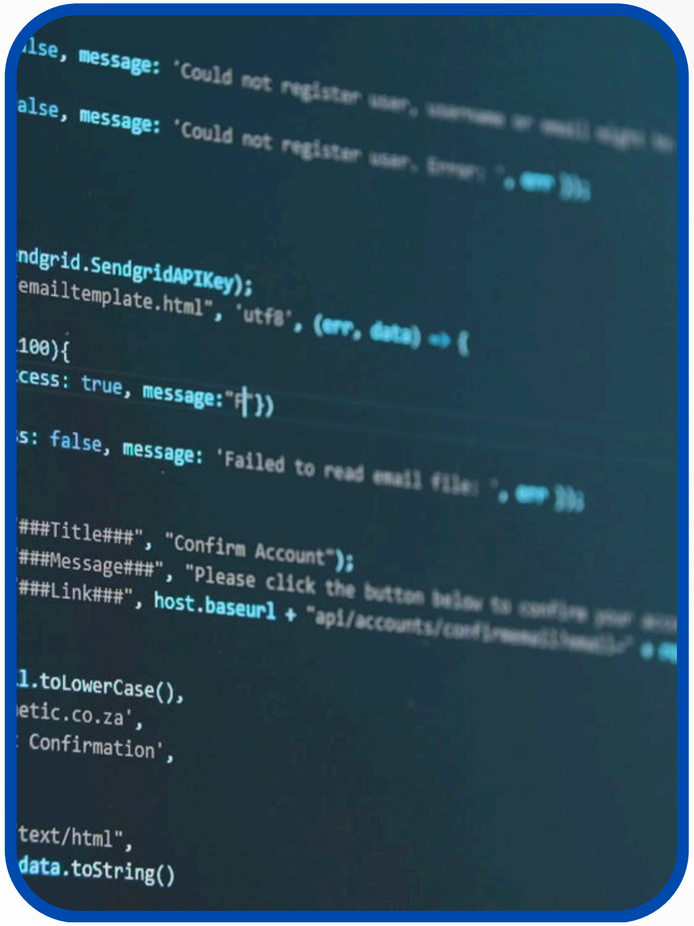
3 DÉTAILS TECHNIQUES

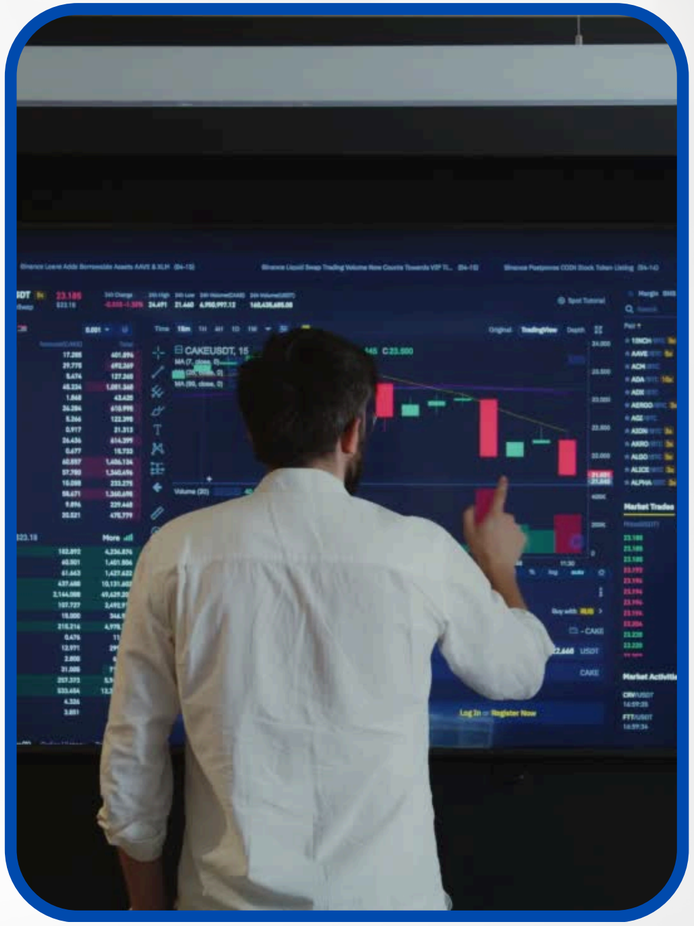
2 DIFFICULTÉS RENCONTRÉES

3 AMÉLIORATION FUTURE

PRÉSENTATION DU PROJET

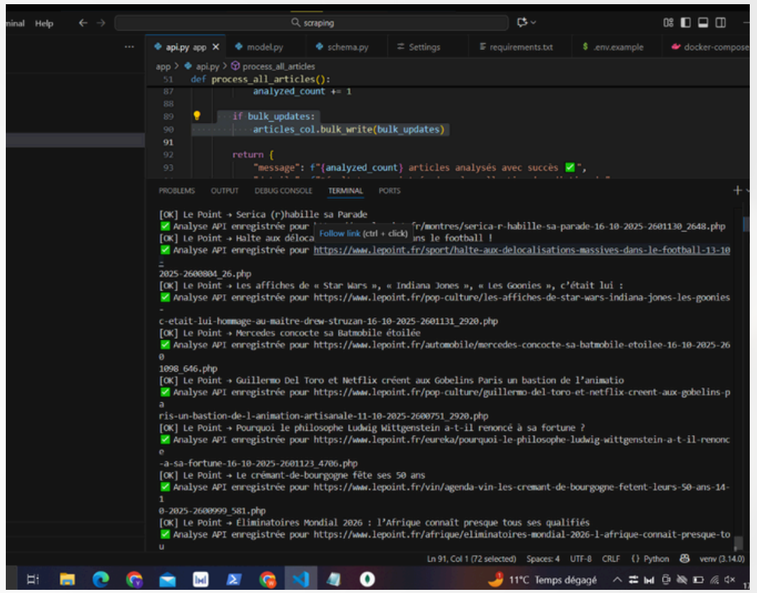
Objectif : créer un pipeline complet ETL de traitement de texte.
- Extraction : scraping d’articles sur 8 sites.
- Transformation : analyse NLP (toxicité).
- Chargement : MongoDB + visualisations + Docker.

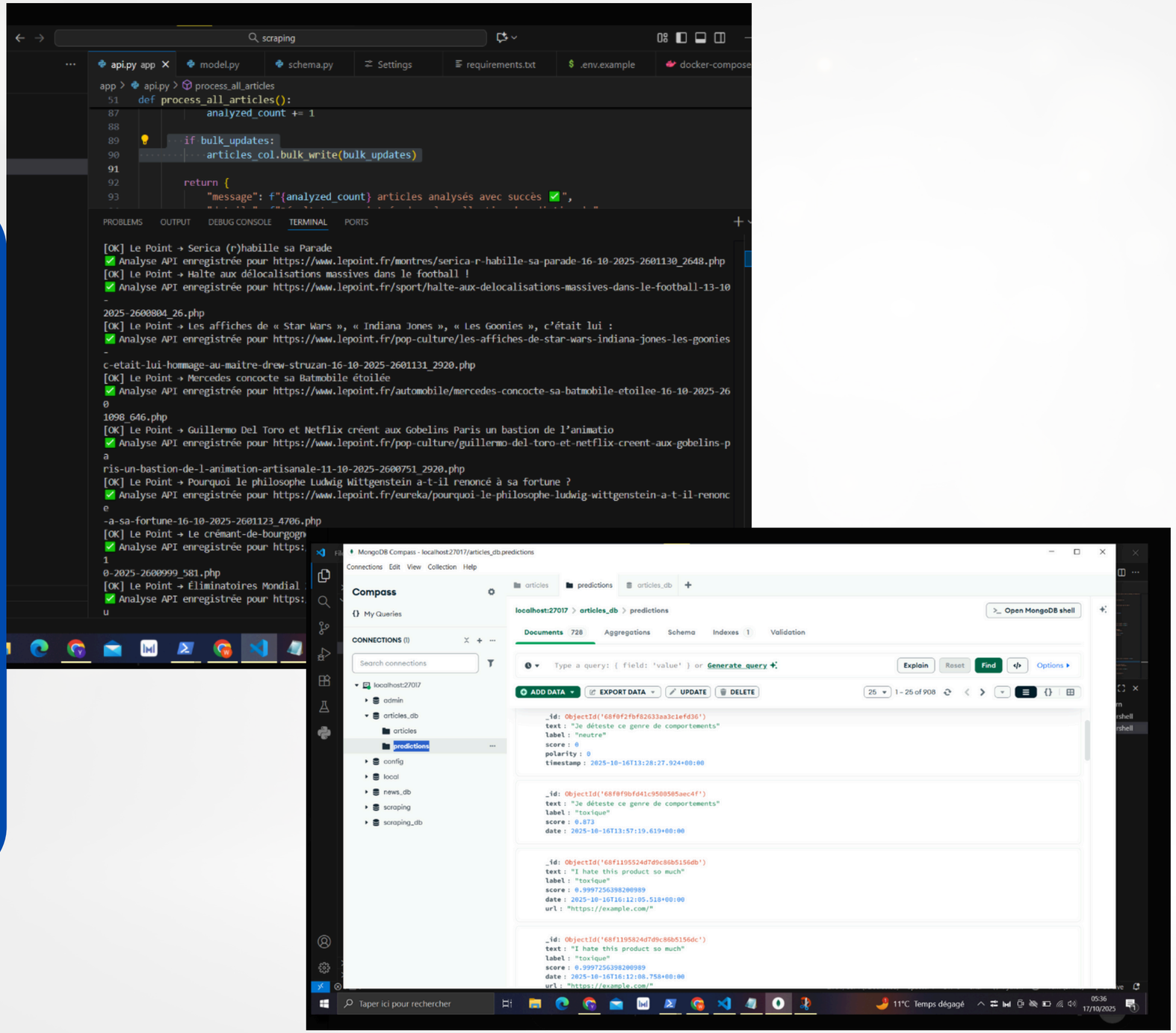




PARTIE 1 — COLLECTE ET STOCKAGE

Scraping via BeautifulSoup et Selenium.
Chaque site a sa propre fonction dédiée (balises différentes).
Nettoyage des doublons et insertion dans MongoDB.
7 sites sur 8 scrappés avec succès.





PARTIE 2

API NLP ET PRÉDICTION

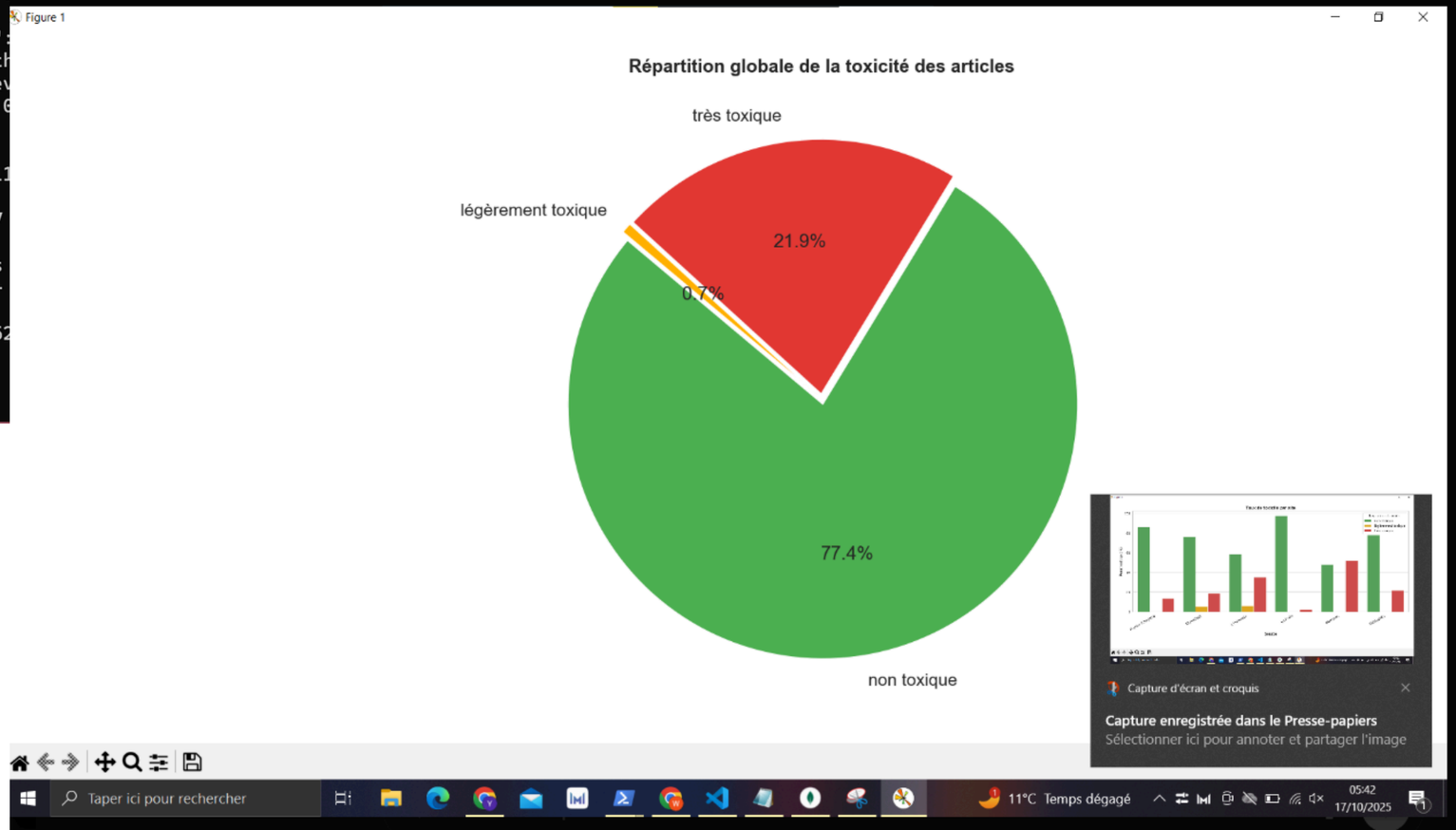
FastAPI + modèle sentiment-analysis (Hugging Face).
Endpoint /predict : texte → label (toxique / non toxique).
Stockage dans MongoDB avec score et date.
Difficulté : modèle en anglais

PARTIE 3

ANALYSE ET VISUALISATION

Objectif : analyser la toxicité par site.
Outils : Pandas, Matplotlib, Seaborn.
Graphiques : camembert + barres comparatives.
Résultats enregistrés dans MongoDB (toxicity_stats).

```
Windows PowerShell                    Windows PowerShell
mongo_express  | basicAuth credentials are "admin:pass", it is recommended you change this in you
toxicity_api   | /usr/local/lib/python3.11/site-packages/transformers/utils/hub.py:127: FutureWar
deprecated and will be removed in v5 of Transformers. Use `HF_HOME` instead.
toxicity_api   |   warnings.warn(
toxicity_api   | Chargement du modèle anglais (sentiment-analysis)...
toxicity_api   | No model was supplied, defaulted to distilbert/distilbert-base-uncased-finetuned
(https://huggingface.co/distilbert/distilbert-base-uncased-finetuned-sst-2-english).
toxicity_api   | Using a pipeline without specifying a model name and revision in production is n
toxicity_api   | INFO:     Started server process [7]
toxicity_api   | INFO:     Waiting for application startup.
toxicity_api   | INFO:     Application startup complete.
toxicity_api   | INFO:     Uvicorn running on http://0.0.0.0:8000 (Press CTRL+C to quit)
mongodb        | {"t":{"$date":"2025-10-17T00:09:52.812+00:00"},"s":"I",  "c":"NETWORK",  "id":22
ction accepted","attr":{"remote":"172.18.0.3:34262","uuid":"3e504bdb-0479-400d-a875-566e89f08a96"
5}}
mongodb        | {"t":{"$date":"2025-10-17T00:09:52.814+00:00"},"s":"I",  "c":"NETWORK",  "id":51
etadata","attr":{"remote":"172.18.0.3:34262","client":"conn6","negotiatedCompressors":[],"doc":{"
4.13.0"},"os":{"type":"Linux","name":"linux","architecture":"x64","version":"5
 v18.20.3, LE (unified)|Node.js v18.20.3, LE (unified)"}}}
mongodb        | {"t":{"$date":"2025-10-17T00:10:40.658+00:00"},"s":"I",  "c":
iredTiger message","attr":{"message":{"ts_sec":1760659840,"ts_usec":657972,"th
ckpoint","category":"WT_VERB_CHECKPOINT_PROGRESS","category_id":6,"verbose_lev
nt snapshot min: 3, snapshot max: 3 snapshot count: 0, oldest timestamp: (0, 0
9"}}}
mongo_express  | GET / 200 51.724 ms - 9260
mongo_express  | GET /public/css/bootstrap-theme.min.css 200 14.609 ms - 23411
mongo_express  | GET /public/css/style.css 200 11.405 ms - 1883
mongo_express  | GET /public/img/mongo-express-logo.png 200 10.882 ms - 17847
mongo_express  | GET /public/css/bootstrap.min.css 200 13.624 ms - 121457
mongo_express  | GET /public/vendor-93f5fc3ae20e0dfd68cb.min.js 200 11.998 ms
mongo_express  | GET /public/index-56afe067afbbbde795be.min.js 200 8.394 ms -
mongo_express  | GET /public/img/gears.gif 200 4.477 ms - 50281
mongo_express  | GET /public/fonts/glyphicons-halflings-regular.woff2 200 1.52
mongo_express  | GET /db/admin/ 200 33.373 ms - 8586
mongo_express  | GET /public/css/bootstrap.min.css 304 2.300 ms - -
```



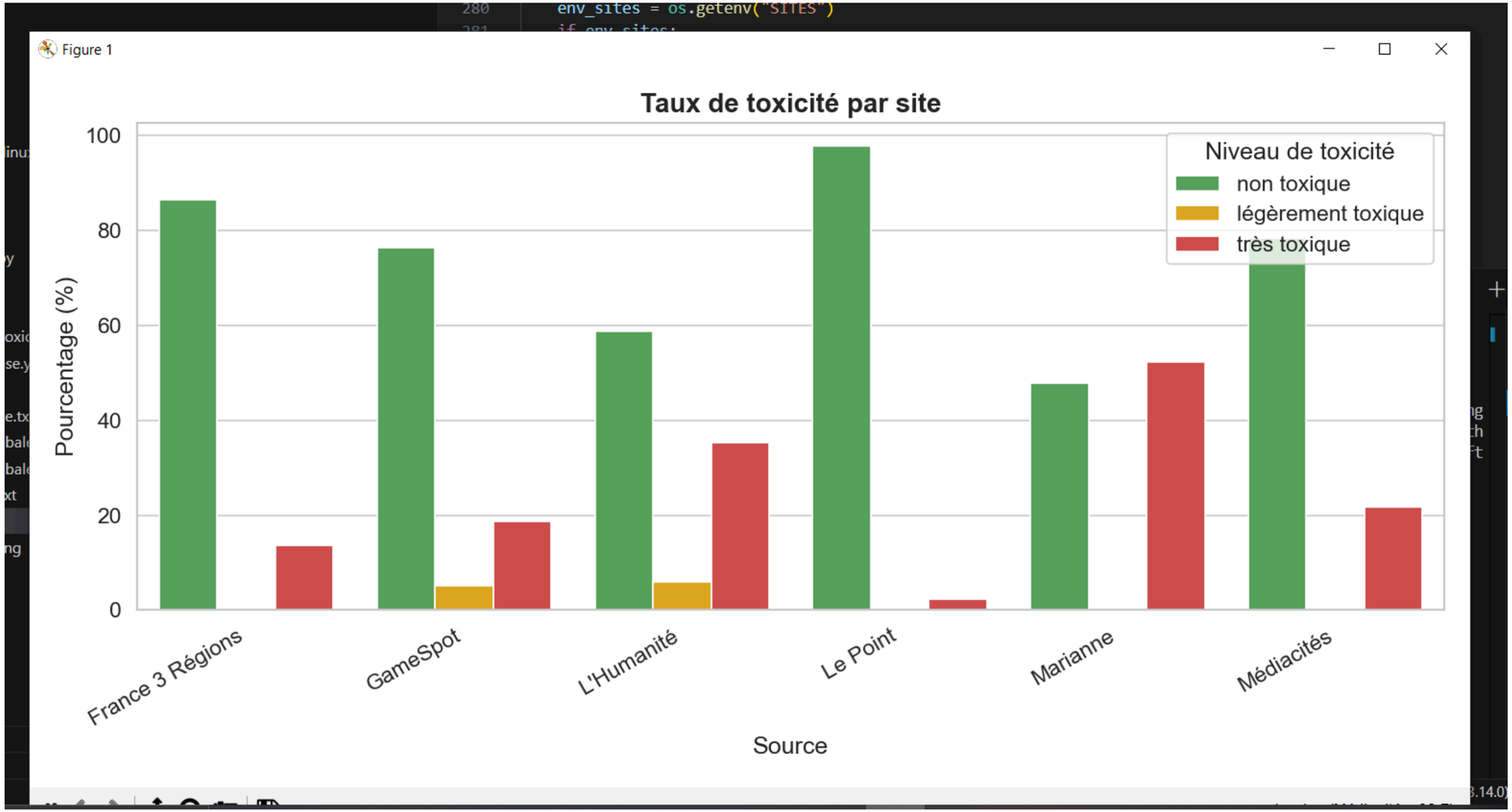
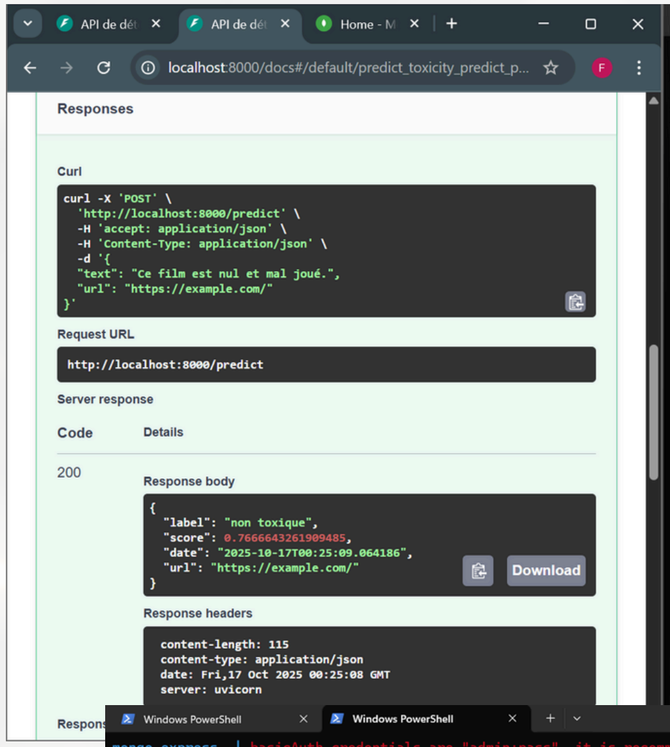
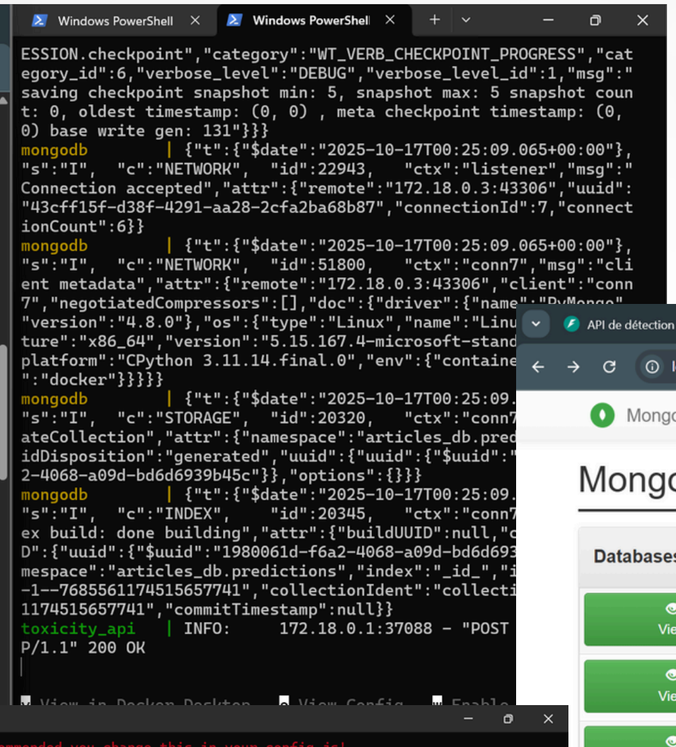


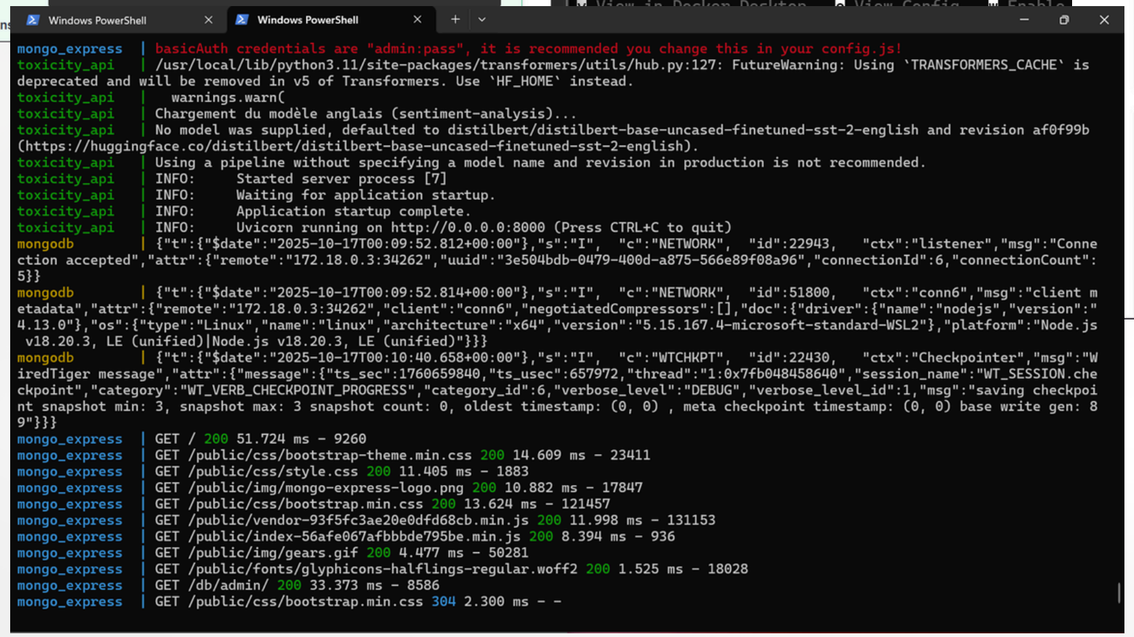
Figure 1
Taux de toxicité par site
Niveau de toxicité
non toxique
légèrement toxique
très toxique
Pourcentage (%)
0
20
40
60
80
100
France 3 Régions
GameSpot
L'Humanité
Le Point
Marianne
Médiacités
Source

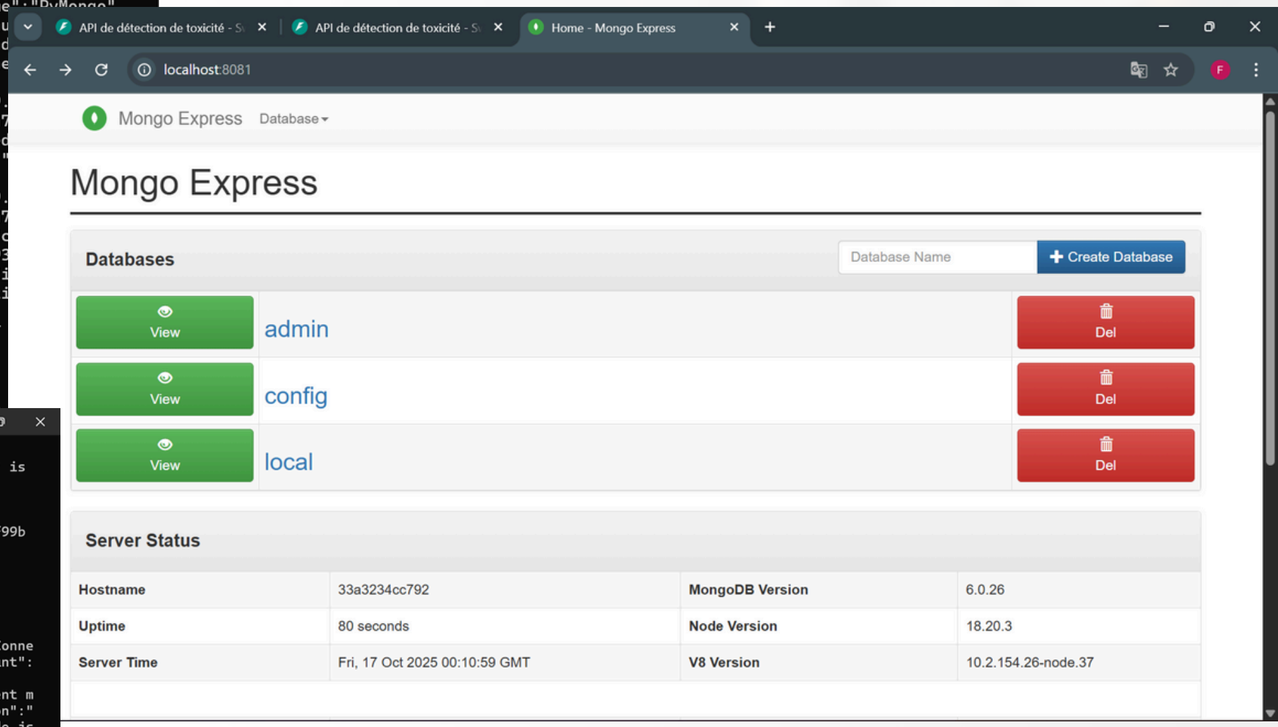
PARTIE 4
DÉPLOIEMENT AVEC DOCKER

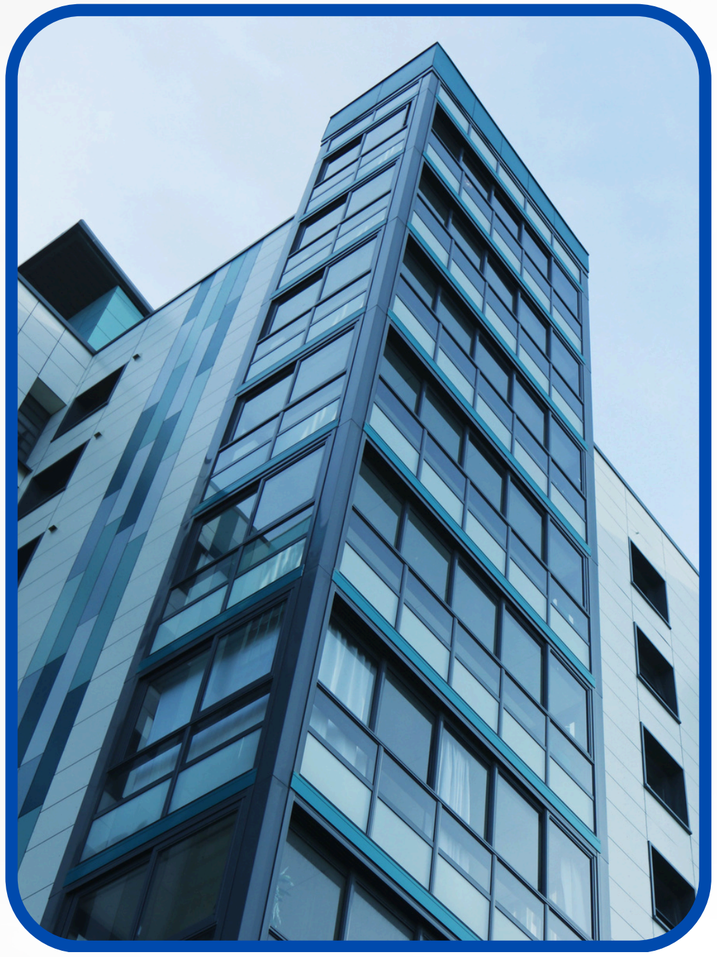
Conteneurisation de FastAPI, MongoDB et Mongo Express.
Fichiers : Dockerfile + docker-compose.yml.
Résolution des erreurs WSL et d’espace disque.
Résultat : pipeline complet exécutable en une commande.


API de dét
API de dét
Home - M
localhost:8000/docs#/default/predict_toxicity_predict_p...
Responses
Curl
curl -X 'POST' \
'http://localhost:8000/predict' \
-H 'accept: application/json' \
-H 'Content-Type: application/json' \
-d '{
"text": "Ce film est nul et mal joué.",
"url": "https://example.com/"
}'
Request URL
http://localhost:8000/predict
Server response
Code
Details
200
Response body
{
"label": "non toxique",
"score": 0.7666643261909485,
"date": "2025-10-17T00:25:09.064186",
"url": "https://example.com/"
}
Download
Response headers
content-length: 115
content-type: application/json
date: Fri,17 Oct 2025 00:25:08 GMT
server: uvicorn


Windows PowerShell
Windows PowerShell
ESSION.checkpoint","category":"WT_VERB_CHECKPOINT_PROGRESS","category_id":6,"verbose_level":"DEBUG","verbose_level_id":1,"msg":"saving checkpoint snapshot min: 5, snapshot max: 5 snapshot count: 0, oldest timestamp: (0, 0) , meta checkpoint timestamp: (0, 0) base write gen: 131"}}}
mongodb | {"t":{"$date":"2025-10-17T00:25:09.065+00:00"},"s":"I", "c":"NETWORK", "id":22943, "ctx":"listener","msg":"Connection accepted","attr":{"remote":"172.18.0.3:43306","uuid":"43cff15f-d38f-4291-aa28-2cfa2ba68b87","connectionId":7,"connectionCount":6}}
mongodb | {"t":{"$date":"2025-10-17T00:25:09.065+00:00"},"s":"I", "c":"NETWORK", "id":51800, "ctx":"conn7","msg":"client metadata","attr":{"remote":"172.18.0.3:43306","client":"conn7","negotiatedCompressors":[],"doc":{"driver":{"name":"PyMongo","version":"4.8.0"},"os":{"type":"Linux","name":"Linux","architecture":"x86_64","version":"5.15.167.4-microsoft-standard-WSL2","platform":"CPython 3.11.14.final.0","env":{"container":"docker"}}}}}
mongodb | {"t":{"$date":"2025-10-17T00:25:09.
"s":"I", "c":"STORAGE", "id":20320, "ctx":"conn7
ateCollection","attr":{"namespace":"articles_db.pred
idDisposition":"generated","uuid":{"uuid":{"$uuid":"
2-4068-a09d-bd6d6939b45c"}},"options":{}}}
mongodb | {"t":{"$date":"2025-10-17T00:25:09.
"s":"I", "c":"INDEX", "id":20345, "ctx":"conn7
ex build: done building","attr":{"buildUUID":null,"c
D":{"uuid":{"$uuid":"1980061d-f6a2-4068-a09d-bd6d693
mespace":"articles_db.predictions","index":"_id_","i
-1--7685561174515657741","collectionIdent":"collecti
11745156577741","commitTimestamp":null}}
toxicity_api | INFO: 172.18.0.1:37088 - "POST
P/1.1" 200 OK


Windows PowerShell
Windows PowerShell
mongo_express | basicAuth credentials are "admin:pass", it is recommended you change this in your config.js!
toxicity_api | /usr/local/lib/python3.11/site-packages/transformers/utils/hub.py:127: FutureWarning: Using `TRANSFORMERS_CACHE` is deprecated and will be removed in v5 of Transformers. Use `HF_HOME` instead.
toxicity_api | warnings.warn(
toxicity_api | Chargement du modèle anglais (sentiment-analysis)...
toxicity_api | No model was supplied, defaulted to distilbert/distilbert-base-uncased-finetuned-sst-2-english and revision af0f99b (https://huggingface.co/distilbert/distilbert-base-uncased-finetuned-sst-2-english).
toxicity_api | Using a pipeline without specifying a model name and revision in production is not recommended.
toxicity_api | INFO: Started server process [7]
toxicity_api | INFO: Waiting for application startup.
toxicity_api | INFO: Application startup complete.
toxicity_api | INFO: Uvicorn running on http://0.0.0.0:8000 (Press CTRL+C to quit)
mongodb | {"t":{"$date":"2025-10-17T00:09:52.812+00:00"},"s":"I", "c":"NETWORK", "id":22943, "ctx":"listener","msg":"Connection accepted","attr":{"remote":"172.18.0.3:34262","uuid":"3e504bdb-0479-400d-a875-566e89f08a96","connectionId":6,"connectionCount":5}}
mongodb | {"t":{"$date":"2025-10-17T00:09:52.814+00:00"},"s":"I", "c":"NETWORK", "id":51800, "ctx":"conn6","msg":"client metadata","attr":{"remote":"172.18.0.3:34262","client":"conn6","negotiatedCompressors":[],"doc":{"driver":{"name":"nodejs","version":"4.13.0"},"os":{"type":"Linux","name":"linux","architecture":"x64","version":"5.15.167.4-microsoft-standard-WSL2"},"platform":"Node.js v18.20.3, LE (unified)|Node.js v18.20.3, LE (unified)"}}}
mongodb | {"t":{"$date":"2025-10-17T00:10:40.658+00:00"},"s":"I", "c":"WTCHKPT", "id":22430, "ctx":"Checkpointer","msg":"WiredTiger message","attr":{"message":{"ts_sec":1760659840,"ts_usec":657972,"thread":"1:0x7fb0048486 40","session_name":"WT_SESSION.checkpoint","category":"WT_VERB_CHECKPOINT_PROGRESS","category_id":6,"verbose_level":"DEBUG","verbose_level_id":1,"msg":"saving checkpoint snapshot min: 3, snapshot max: 3 snapshot count: 0, oldest timestamp: (0, 0) , meta checkpoint timestamp: (0, 0) base write gen: 89"}}}
mongo_express | GET / 200 51.724 ms - 9260
mongo_express | GET /public/css/bootstrap-theme.min.css 200 14.609 ms - 23411
mongo_express | GET /public/css/style.css 200 11.405 ms - 1883
mongo_express | GET /public/img/mongo-express-logo.png 200 10.882 ms - 17847
mongo_express | GET /public/css/bootstrap.min.css 200 13.624 ms - 121457
mongo_express | GET /public/vendor-93f5fc3ae20e0dfd68cb.min.js 200 11.998 ms - 131153
mongo_express | GET /public/index-56afe067afbbbde795be.min.js 200 8.394 ms - 936
mongo_express | GET /public/img/gears.gif 200 4.477 ms - 50281
mongo_express | GET /public/fonts/glyphicons-halflings-regular.woff2 200 1.525 ms - 18028
mongo_express | GET /db/admin/ 200 33.373 ms - 8586
mongo_express | GET /public/css/bootstrap.min.css 304 2.300 ms - -


API de détection de toxicité - S
API de détection de toxicité - S
Home - Mongo Express
localhost:8081
Mongo Express
Database
Mongo Express
Databases
Database Name
Create Database
View
admin
Del
View
config
Del
View
local
Del
Server Status
Hostname
33a3234cc792
MongoDB Version
6.0.26
Uptime
80 seconds
Node Version
18.20.3
Server Time
Fri, 17 Oct 2025 00:10:59 GMT
V8 Version
10.2.154.26-node.37



CONCLUSION

Projet complet : scraping → NLP → analyse → Docker.
Approche ETL maîtrisée et documentée.
Apprentissage approfondi sur Docker et FastAPI.
« De la donnée brute à l’information utile. »

REMERCIEMENTS

Merci pour l’attention portée à ce projet.
Ce test technique a été une véritable opportunité d’apprentissage

WAIL.BR

+33 638010145

wbrimesse@gmail.com