

SCRAPING WEB APPLICATION

présenté par

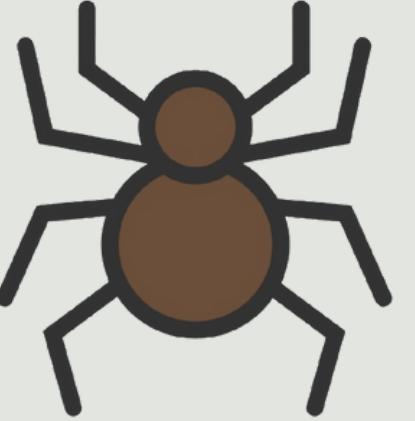
NAJIB ISMAIL

NAJIMDDIN MOHAMED

YOUSSEF BOUDOUAR

Supervise par MR :

Mohamed-Amine Chadi



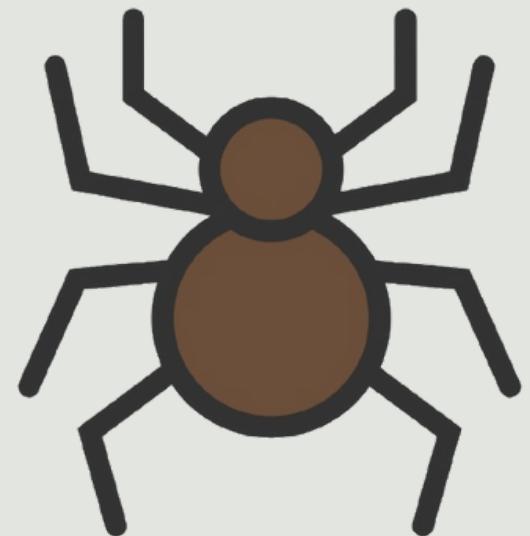
PLAN

- Introduction
- Objectif de projet
- La Conception
- Les outils
- Méthodologie
- L'interface
- Conclusion
- References

01

Introduction

INTRODUCTION



Qu'est-ce que le Web Scraping ?

Le web scraping est le processus automatisé d'extraction de données à partir de sites web. Il permet de collecter rapidement de grandes quantités d'informations en accédant à des pages web et en extrayant des contenus spécifiques.

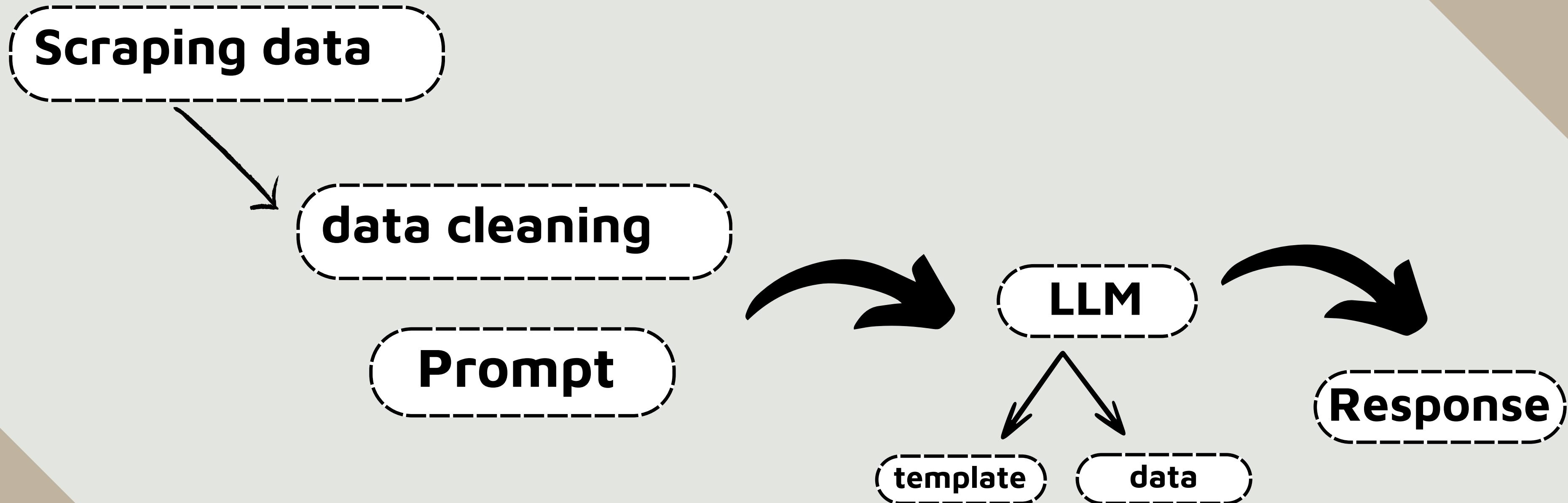
02

OBJECTIF

Créer une application de web scraping qui permet l'extraction de données à partir de sites web ciblés, transformant les données brutes en informations exploitables.

La Conception

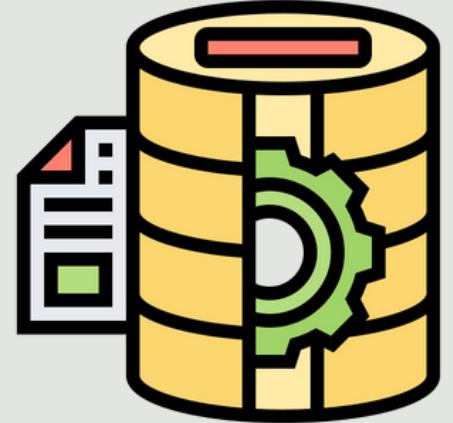
LA CONCEPTION



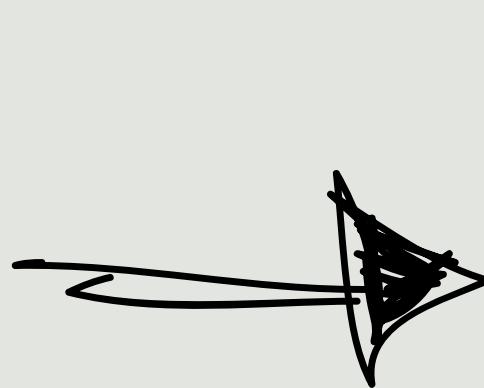
LES APPROCHES

Utilisation d'un LLM Local :

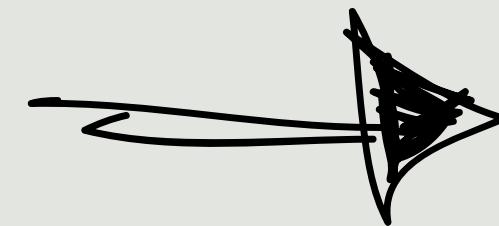
La première approche consiste à utiliser un modèle de langage (LLM) local pour analyser les données des pages web.



données extraites



Ollama



données traitées

LES APPROCHES

Inconvénients de l'Utilisation d'un LLM Local :

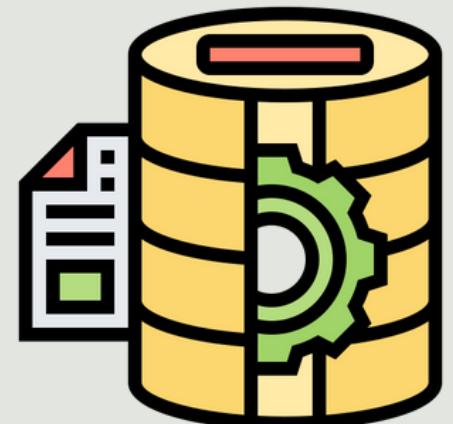
- **Consommation de ressources** : Nécessite une infrastructure puissante (processeur, mémoire vive), ce qui peut augmenter les coûts.
- **Temps de Réponse Très Long** : L'utilisation d'un LLM local peut entraîner des temps de réponse longs, surtout lorsqu'il s'agit de traiter des pages web volumineuses ou complexes.



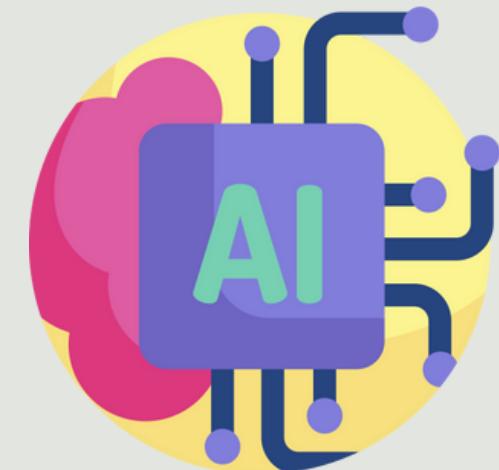
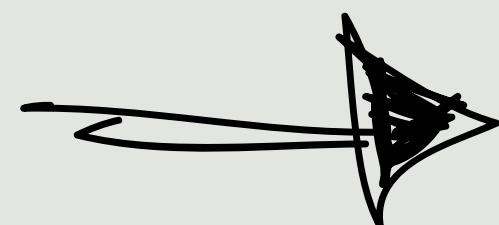
LES APPROCHES

Utilisation de BERT :

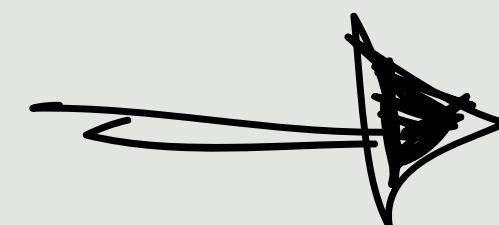
La deuxième approche consiste à utiliser BERT, un modèle de traitement du langage naturel (NLP) pré-entraîné, pour améliorer l'organisation et l'analyse des données extraites lors du web scraping.



données extraites



BERT



données traitées

LES APPROCHES

Inconvénients de l'Utilisation de BERT :

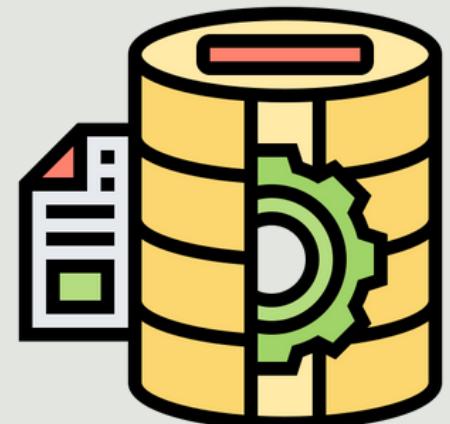
- **Consommation de ressources** : Nécessite une infrastructure puissante (processeur, mémoire vive), ce qui peut augmenter les coûts.
- **Temps de Réponse Très Long** : L'utilisation d'un LLM local peut entraîner des temps de réponse longs, surtout lorsqu'il s'agit de traiter des pages web volumineuses ou complexes.
- **Support linguistique limité**: ne prend pas en charge toutes les langues de manière égale.



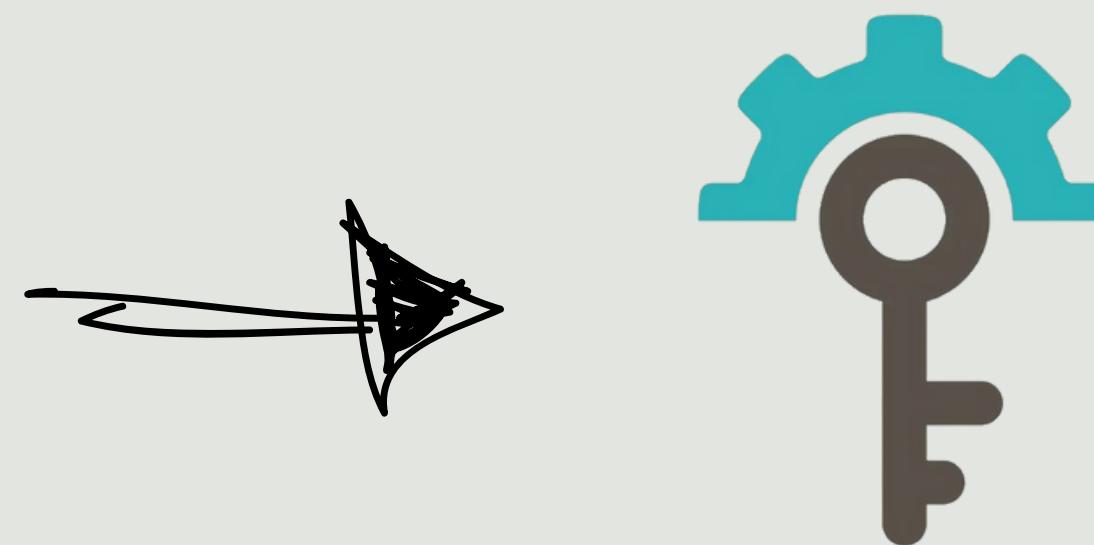
LES APPROCHES

Utilisation d'une API LLM:

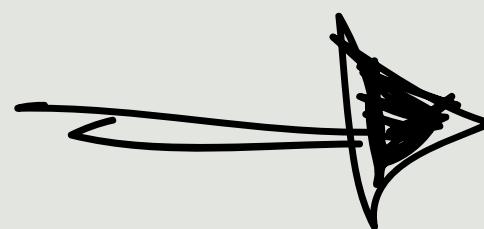
L'utilisation d'une API LLM permet d'intégrer les capacités d'un modèle de langage pré-entraîné sans la nécessité de gérer l'infrastructure pour l'exécution locale.



données extraites



API LLM



données traitées

LES APPROCHES

Avantages d'une API LLM:

- **Simplicité d'Intégration** : L'utilisation d'une API réduit la complexité technique en évitant d'avoir à gérer l'infrastructure ou l'entraînement des modèles.
- **Accès aux Modèles de Pointe** : Accès à des modèles de traitement du langage les plus avancés .
- **Mises à Jour Automatiques** : Les modèles sont régulièrement améliorés et mis à jour par les fournisseurs d'API .
- **Optimisation des Ressources** : API LLM permet une optimisation considérable des ressources pour le traitement du langage naturel.



10

LES APPROCHES

Inconvénients d'une API LLM:

- **Dépendance au Service Externe** : Cette approche dépend de la disponibilité et de la fiabilité du service API externe, ce qui peut poser problème en cas de panne ou de restrictions.
- **Coûts Récurrents** : L'utilisation d'API LLM peut entraîner des coûts récurrents basés sur le nombre de requêtes, ce qui peut devenir coûteux sur le long terme, surtout à grande échelle.



11

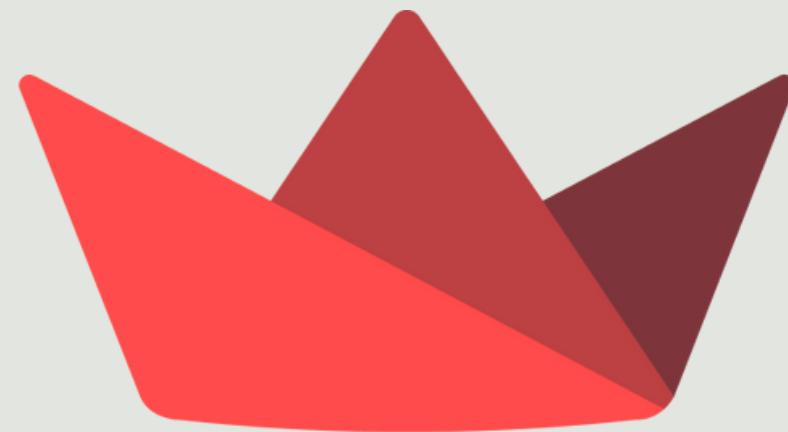
Les outils

LES OUTILS

Dans ce projet de web scraping, plusieurs outils et bibliothèques ont été utilisés pour faciliter l'extraction, le traitement et la présentation des données. Ces outils, principalement basés sur le langage Python



Python



Streamlit



Selenium



Chromedriver

Méthodologie

MÉTHODOLOGIE

01

L'extraction
de données

02

Déterminer
des Entités
Nommées :

03

Modélisation
des Données
avec NER et
API Groq

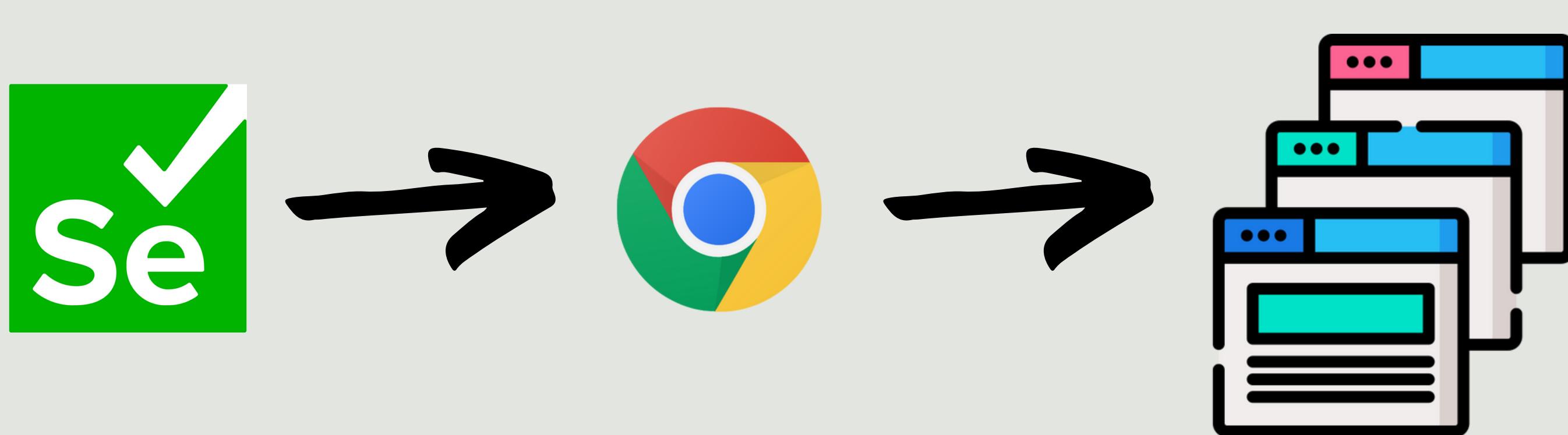
04

Visualisation
des Résultats

MÉTHODOLOGIE

L'extraction de données :

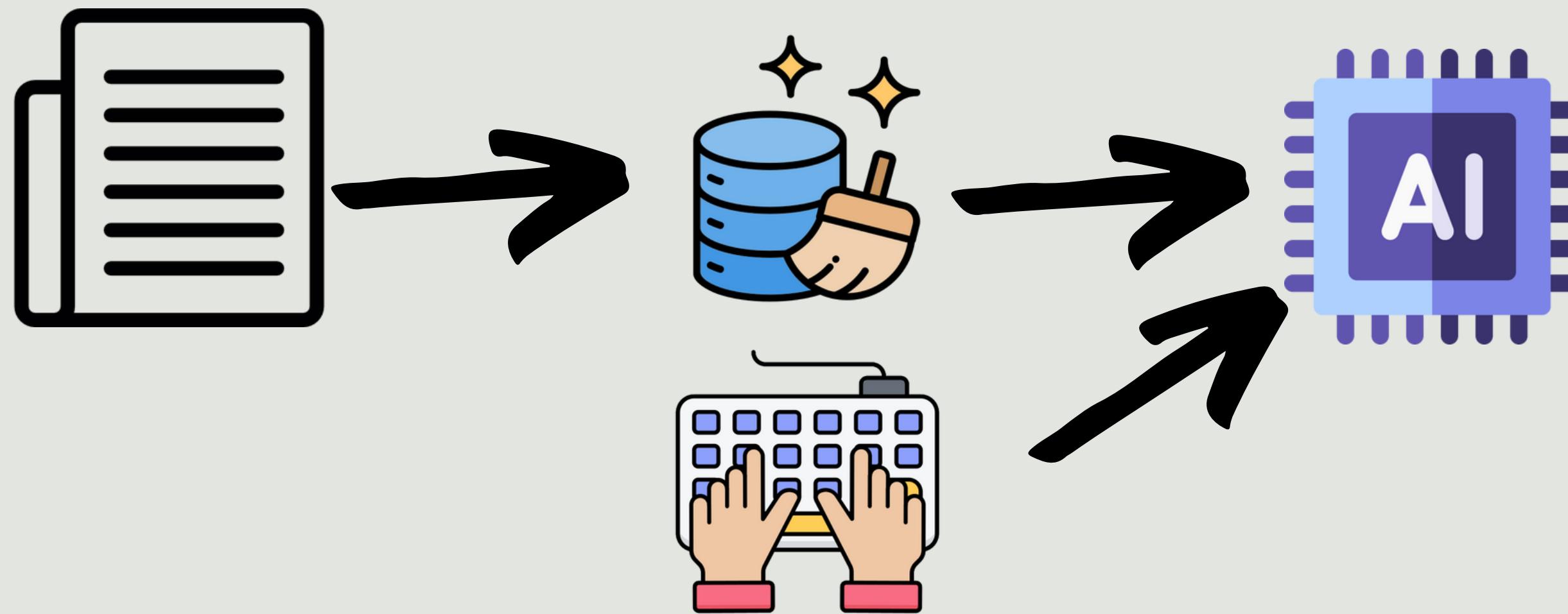
Pour l'extraction des données, nous avons utilisé Selenium et ChromeDriver afin d'accéder à la page web via une URL. Une fois la page chargée, nous avons extrait le code source. Ensuite, nous avons appliqué BeautifulSoup pour nettoyer les données en supprimant les balises HTML et autres éléments inutiles, afin de conserver uniquement le texte brut du contenu de la page.



MÉTHODOLOGIE

Déterminer des Entités Nommées (NER) :

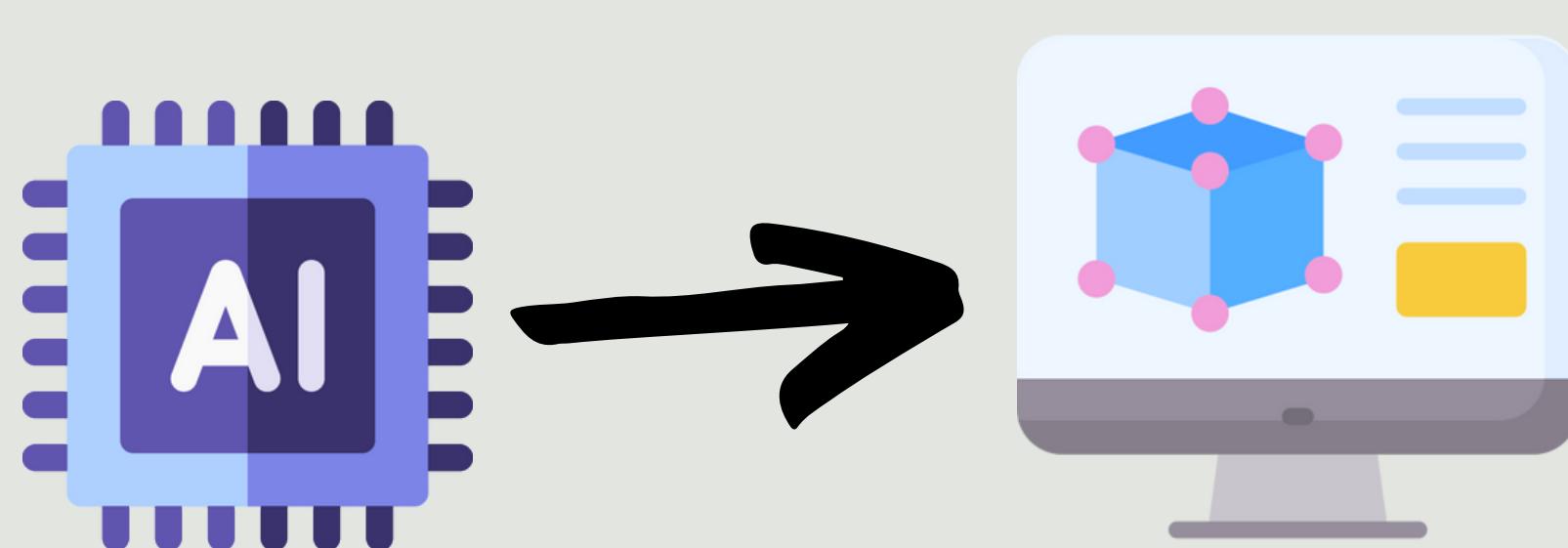
Après avoir nettoyé les données, nous avons utilisé une clé API Groq pour exploiter LLAMA3 en tant que modèle de langage (LLM) afin de traiter les données extraites. Le modèle organise les informations en fonction du prompt spécifié par l'utilisateur NER.



MÉTHODOLOGIE

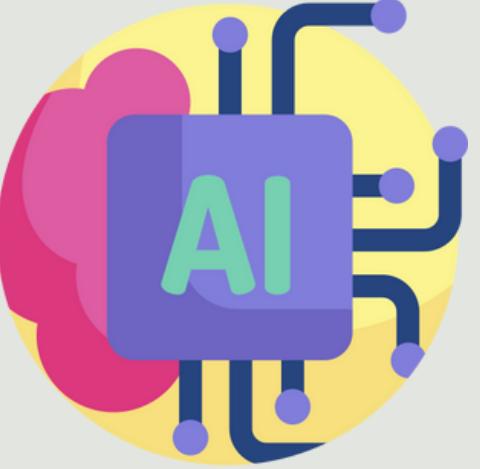
Modélisation des Données avec NER et API :

Pour utiliser le LLM, nous avons tenté de créer un template composé des données extraites, du prompt et de quelques instructions destinées à aider le modèle à comprendre le contexte et les tâches à réaliser. Une fonction de type "calling function" est utilisée pour renvoyer le résultat final.

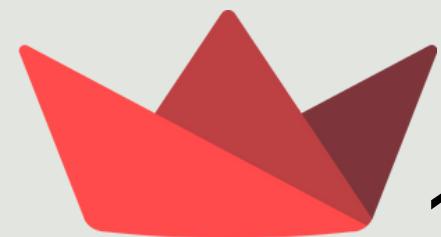


L'interface

L'INTERFACE



- L'interface utilisateur a été conçue avec Streamlit pour permettre une visualisation interactive des résultats du web scraping.
- L'utilisateur peut soumettre une URL via un champ de saisie, et les données de la page web cible sont extraites en utilisant Selenium avec ChromeDriver.
- Un modèle de Named Entity Recognition (NER) est utilisé pour extraire et afficher des informations pertinentes (comme des entités nommées) de manière structurée.



L'INTERFACE V1

The screenshot shows a Streamlit application titled "Web App Scraping with LLM". The title is displayed prominently at the top center of the main content area. Below the title, there is a descriptive text block: "By bridging web scraping and LLM capabilities, this app serves as a powerful tool for data-driven applications in today's information-rich environment." To the left of the main content area, there is a sidebar with a light beige background. It contains a navigation bar with three items: "app" (which is highlighted with a light brown background), "about", and "contact". At the bottom of the sidebar, there is a "Scrape" button. The main content area has a light blue gradient background. At the top of this area, there is a dark purple header bar with various icons and the URL "web-app-scraping.streamlit.app". In the bottom right corner of the main content area, there is a small black button with the text "Manage app" and a back arrow icon.

← → ⌂ web-app-scraping.streamlit.app

app

about

contact

Web App Scraping with LLM

By bridging web scraping and LLM capabilities, this app serves as a powerful tool for data-driven applications in today's information-rich environment.

Enter a website URL :

Scrape

Manage app

RESULTS

News | **Opinion** | **Sport** | **Culture** | **Lifestyle** | More ▾

Football ▶ Live scores Tables Fixtures Results Competitions Clubs

Football tables

Choose league:

Premier League

P	Team	GP	W	D	L	F	A	GD	Pts	Form
1	Liverpool	19	14	4	1	47	19	28	46	■■■-
2	Arsenal	20	11	7	2	39	18	21	40	■■■-
3	Notm Forest	19	11	4	4	26	19	7	37	■■■■
4	Chelsea	20	10	6	4	39	24	15	36	■■■-
5	Newcastle	20	10	5	5	34	22	12	35	■■■■
6	Man City	20	10	4	6	36	27	9	34	■■■-
7	AFC Bournemouth	20	9	6	5	30	23	7	33	■■■-
8	Aston Villa	20	9	5	6	30	32	-2	32	■■■-
9	Fulham	20	7	9	4	30	27	3	30	■■■-
10	Brighton	20	6	10	4	30	29	1	28	■■■■
11	Brentford	20	8	3	9	38	35	3	27	■■■■
12	Spurs	20	7	3	10	42	30	12	24	■■■■
13	Man Utd	20	6	5	9	23	28	-5	23	■■■-
14	West Ham	20	6	5	9	24	39	-15	23	■■■-
15	C Palace	20	4	9	7	21	28	-7	21	■■■-
16	Everton	19	3	8	8	15	25	-10	17	■■■■
17	Wolves	19	4	4	11	31	42	-11	16	■■■-
18	Ipswich	20	3	7	10	20	35	-15	16	■■■■
19	Leicester	20	3	5	12	23	44	-21	14	■■■■
20	Southampton	20	1	3	16	12	44	-32	6	■■■■



Web App Scraping with LLM

By bridging web scraping and LLM capabilities, this app serves as a powerful tool

Enter a website URL:

Describe what you want to parse ?

Parsing the content ...

Here is the extracted information:

Rank	Team	GP	W	D	L	F	A	GD	Pts	Form
1	Liverpool	19	14	4	1	47	19	28	46	Drew 2-2 with Fulham
2	Arsenal	20	11	7	2	39	18	21	40	Drew 0-0 with Everton
3	Notm Forest	19	11	4	4	26	19	7	37	Won 3-2 against Man Utd
4	Chelsea	20	10	6	4	39	24	15	36	Won 2-1 against Brentford
5	Newcastle	20	10	5	5	34	22	12	35	Won 4-0 against Leicester
6	Man City	20	10	4	6	36	27	9	34	Lost 1-2 to Man Utd
7	AFC Bournemouth	20	9	6	5	30	23	7	33	Drew 1-1 with West Ham
8	Aston Villa	20	9	5	6	30	32	-2	32	Lost 1-2 to Notm Forest
9	Fulham	20	7	9	4	30	27	3	30	Drew 2-2 with Liverpool
10	Brighton	20	6	10	4	30	29	1	28	Lost 1-3 to C Palace
11	Brentford	20	8	3	9	38	35	3	27	Lost 1-2 to Chelsea
12	Spurs	20	7	3	10	42	30	12	24	Won 5-0 against Southampton
13	Man Utd	20	6	5	9	23	28	-5	23	Won 2-1 against Man City
14	West Ham	20	6	5	9	24	39	-15	23	Drew 1-1 with AFC Bournemouth
15	C Palace	20	4	9	7	21	28	-7	21	Won 3-1 against Brighton
16	Everton	19	3	8	8	15	25	-10	17	Drew 0-0 with Arsenal
17	Wolves	19	4	4	11	31	42	-11	16	Lost 1-2 to West Ham
18	Ipswich	20	3	7	10	20	35	-15	16	Won 2-1 against Wolves
19	Leicester	20	3	5	12	23	44	-21	14	Lost 0-4 to Newcastle
20	Southampton	20	1	3	16	12	44	-32	6	Lost 0-5 to Spurs

Les limites

- Le nombre de tokens de l'API est limité

20

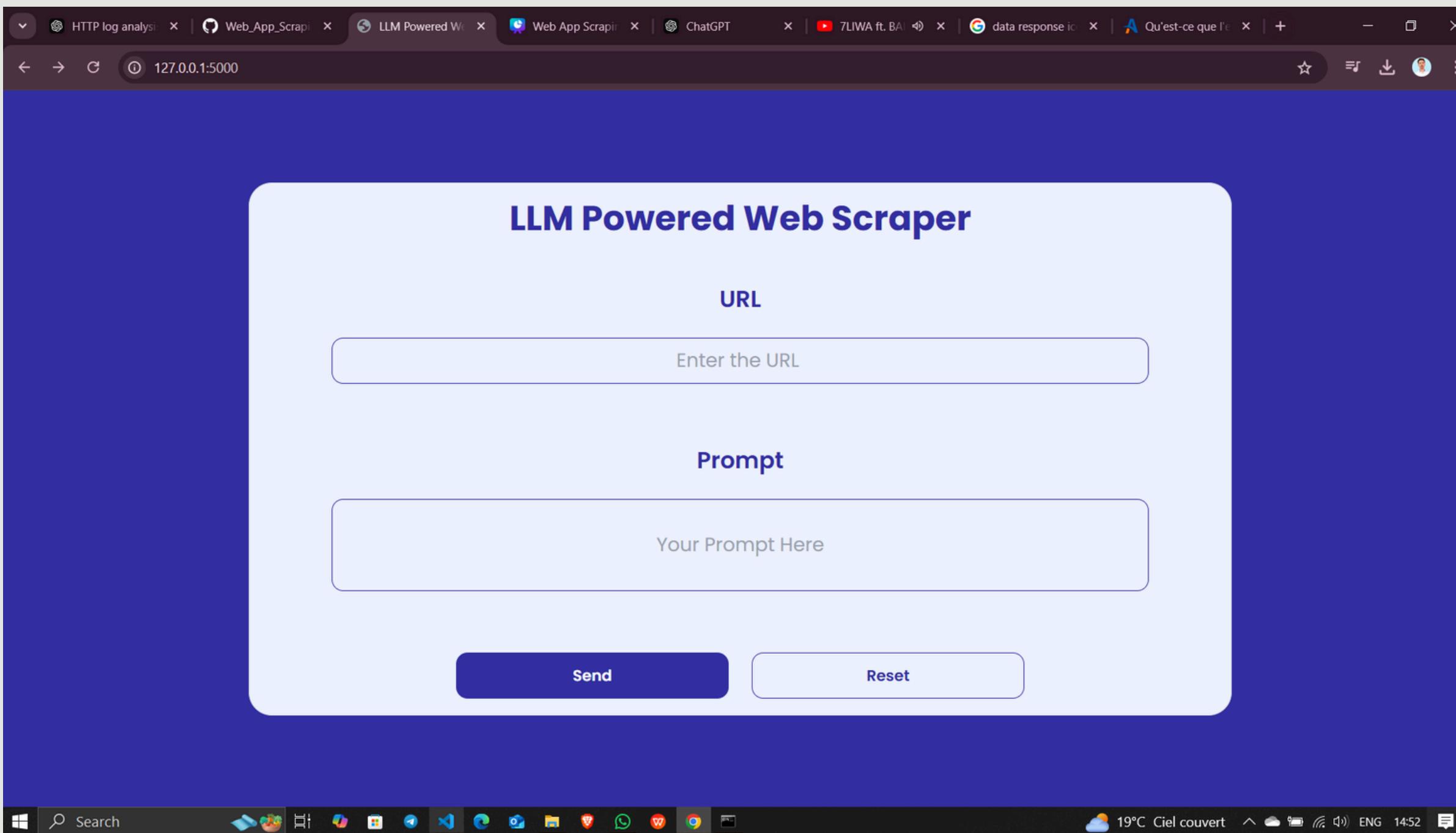


Les limites

- Le nombre de tokens de l'API est limité
- La plateforme est capable d'extraire uniquement des données textuelles
- Parfois, le modèle LLM génère des résultats désordonnés pour l'interface.

L'INTERFACE V2

Nous avons opté pour le framework Flask afin de concevoir des interfaces à la fois professionnelles et intuitives, dans le but d'améliorer l'expérience utilisateur (UI/UX).



Conclusion

En conclusion, ce projet a non seulement permis de démontrer l'intégration efficace de différentes technologies de scraping, traitement du langage naturel et interfaces interactives, mais a aussi mis en lumière les aspects pratiques et les défis techniques liés à l'automatisation de ces processus. Avec des optimisations futures, ce projet pourrait être étendu pour gérer des volumes de données encore plus importants, tout en offrant une performance et une fiabilité accrues.

REFERENCE

Documentations des outils: [Selenium](#), [Streamlit](#), [Flask](#)

Youtuble Toturiel: [Elzero](#), [Tech With Tim](#),

Thank You