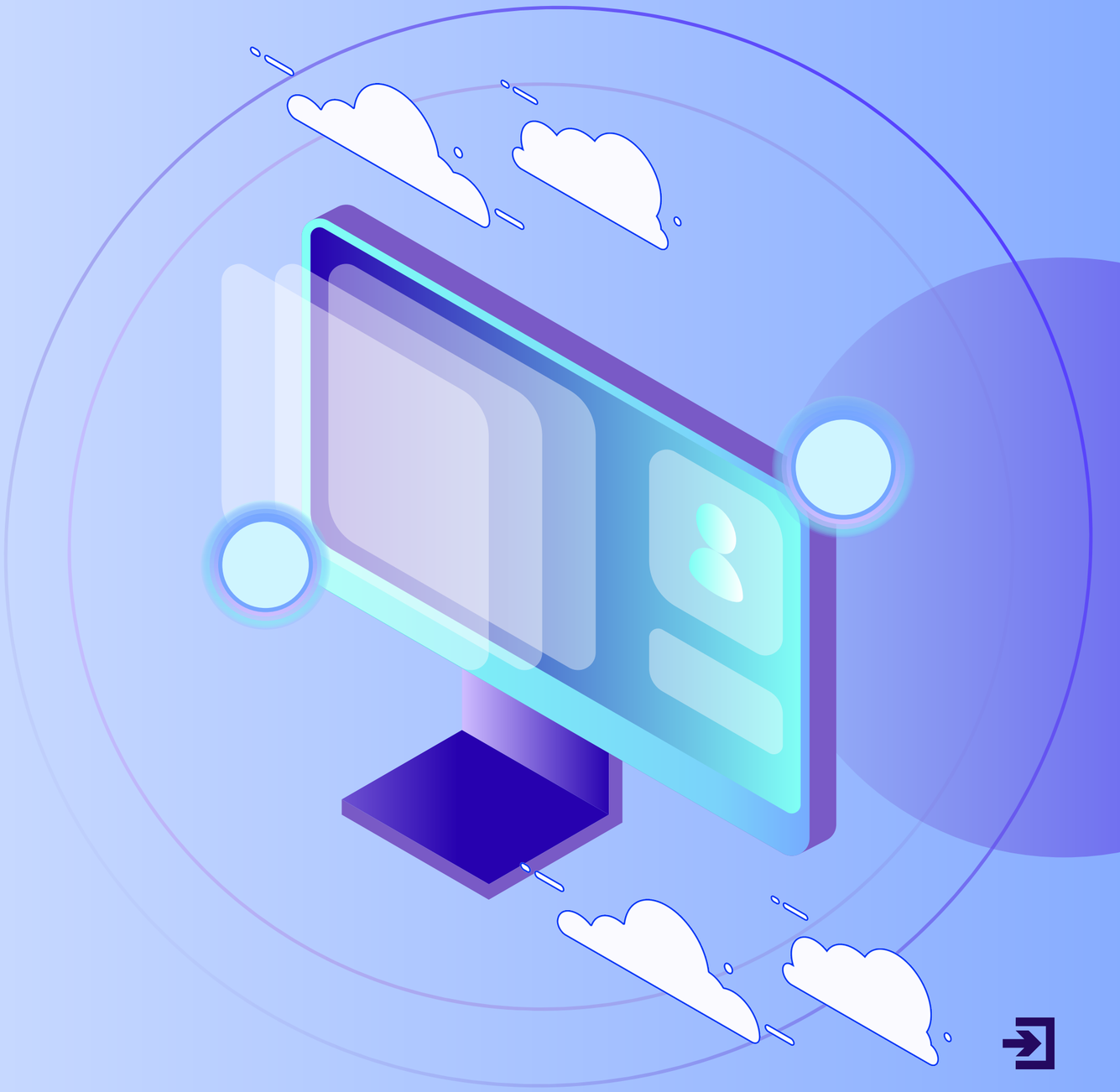
Francesca
Romain
Paul

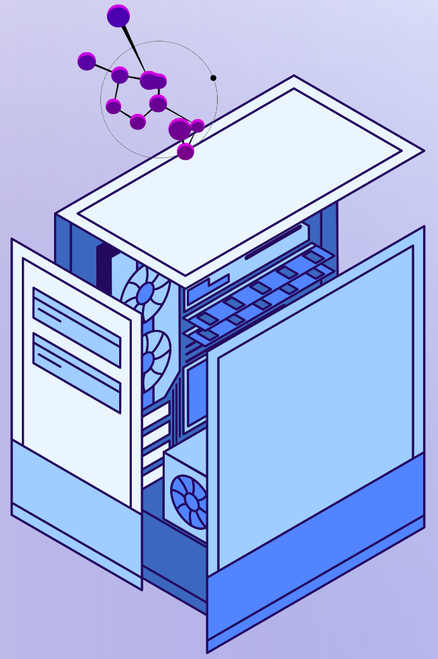
U1-ADVANCED PROGRAMMING

LINKEDIN PROJECT

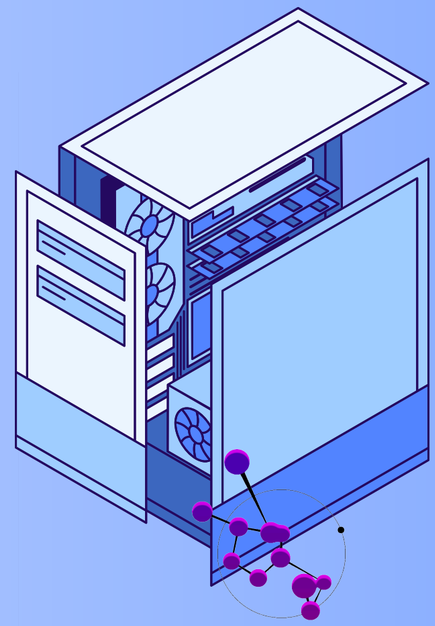


www.Linkedin/scraping.fr

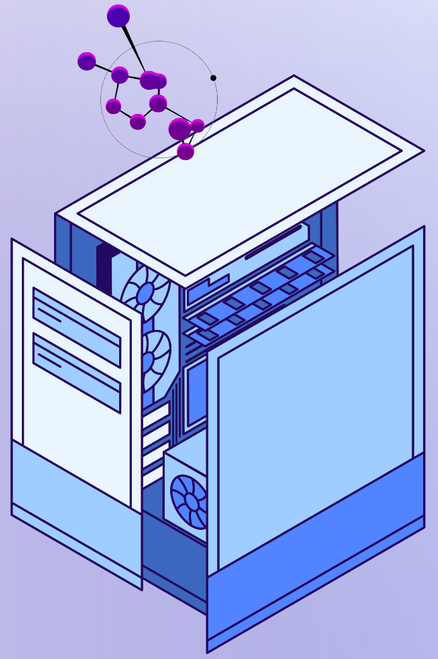
WHY THIS PROJECT ?



- Because it was challenging
- Useful for the students in M1
- We know how difficult it is to find an internship
- Searching job offers on the web can be overwhelming and require a lot of time

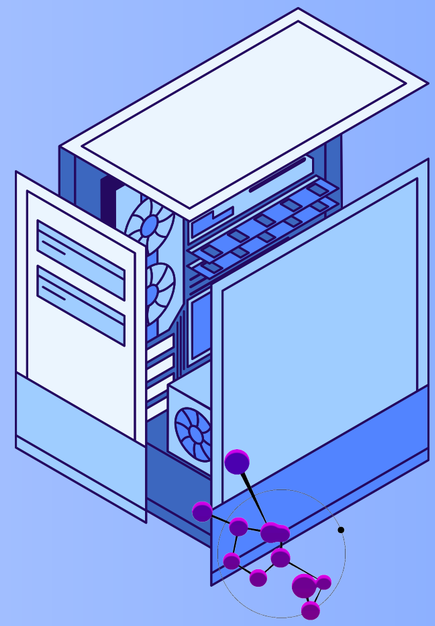


OUR GOAL



Create an algorithm that:

- Collects relevant information based on the conditions set by the user
- Gives the user an overview of the current employment trends in France using the data retrieved



WORK SET-UP

Our project will be presented in two parts, reflecting exactly how we organized our work:

Part 1: Focused on ***web scraping*** and data collection

Part 2: Focused on ***generative AI*** for quick analysis and text generation

PART 1
WEB - SCRAPING

- Since we used Linkedin for the project, we had to create a **special account for the master**.
- We used **Selenium** instead of Beautiful Soup because it was easier to replicate human-like behaviour and interact with the driver.

We defined in advance on the type of job offers and details we wanted to collect, specifically:

The most recent internship offers for the role of “Data Scientist”.

[Job title, Job Link, Company Name, Company Link, Location, Job Description]

FUNCTIONS THAT WE INTRODUCED

There were various dynamic elements to interact with and we defined the following functions to help us:

- **click_element_CSS**
- **scroll_list_to_bottom**
- **scrape_jobs_on_page** to obtain all the jobs on the page while scrolling
- **click_button_page** to click on subsequent pages and load more results
- **scrape_all_pages** which collects all the previous functions
- **get_job_details** which iterates on each job offer link to collect data

EXPECTED RESULTS

A complete dataset stored in a csv file

We have successfully created a table that stores all the information about the job offers scraped from LinkedIn. The table doesn't have a fixed size, as the number of rows depends on the available job offers that satiasfy the conditions at the moment the algorithm starts.

CHALLENGES

- No prior experience with HTML, which made the task more challenging
- **XPath couldn't be used reliably** due to the dynamic structure of the page and changing element identifiers. Instead, we relied on other indicators like CLASS or CSS_Selectors to locate the desired elements and extract the data.

PART 2
GENERATIVE AI

Before applying generative AI, we performed a **preliminary cleaning of Job Description** and **preliminary descriptive analysis**

To carry on AI tasks, we exploited ***HuggingFace*** open-source **transformers** library

- google/flan-T5-base
- BERT uncased
- Helsinki-NLP/opus-mt-fr-en

TRANSFORMER MODELS THAT WE USED

- **translator = pipeline("translation_fr_to_en", model="Helsinki-NLP/opus-mt-fr-en", device=-1)** to translate all the job descriptions into english.

- **qa_pipeline = pipeline("question-answering", model="distilbert-base-uncased-distilled-squad")** to derive different kinds of information from the Job Description: industry, duration of the internship and required technical skills.

- **tokenizer = T5Tokenizer.from_pretrained("google/flan-t5-base") and model = T5ForConditionalGeneration.from_pretrained("google/flan-t5-base")** were imported to rewrite the text based on a clear prompt.

EXPECTED RESULTS: JOB OFFER ANALYSIS

The goal was to use AI to derive job offers details from the Job Description and enrich the dataset.

We have derived the industry, internship contract duration and required skills.

Leveraing TD/IDF vectorization, we also obtained a word-cloud to visually summarize the most required skills

CHALLENGES

- The Job Descriptions were very dirty, with emojis and symbols. So we had to clean the text.
- Cleaning the strings was not enough. We also had to translate some of the Job Descriptions from French to English.
- We had **no test or validation set to check the accuracy** of the predictions

Industries appearing more frequently

```
automotive, industrial and tertiary, technology, engineering, finance
```

Common contract lenghts

```
     6 months
up to a year
   13 months
     6-month
up to a year
```

Some of the skills identified by the model

Not so satisfactory

```
deep technical expertise
you or up to a year and be notified about jobs and updates
experts' and engineers' deep technical expertise
teamwork
Pleasure, Transmission, Excellence, Empathy, Integrity and Commitment

financial modeling and valuation
safety, excellence, innovation, team spirit and transparency
Respect, transparency and performance
You will enrich your synthesis spirit and your technological expertise
real sales skills
```

The wordcloud performed better because it focused on the TD/IDF of each single word

```
Top 20 Keywords with their TF-IDF Scores:
team: 7.075
data: 5.999
company: 5.513
group: 5.111
development: 4.648
project: 4.643
business: 4.625
customer: 4.486
solution: 4.472
service: 4.199
technology: 3.912
engineering: 3.877
management: 3.804
work: 3.794
mission: 3.738
support: 3.645
digital: 3.643
learn: 3.589
experience: 3.584
product: 3.484
```

EXPECTED RESULTS: TEXT GENERATION

Give the user an overview of the current employment market in France

The goal was to obtain a detailed description of the current French employment market and expected trends, emphasizing the distribution of jobs offer across various regions, the top industries and the common contract lenght.

↓

the result was not satisfactory

CHALLENGES

- ✕ Difficult to find the right prompt and instructions
- ✕ Limited access to more advanced Text Generation tools (they required a subscription and an API)

Expected result

```
At present, there are a total of 118 job offers available across various regions of France.
Among these, 71 job opportunities are located in the northern part of the country, 16 in the south,
5 in the west, and 12 in the eastern regions.
The demand for flexible work arrangements has seen an uptick, with 3 positions offering hybrid work options, indicating a
growing preference for a balanced work environment that combines both remote and in-office elements.
The top industries actively seeking employees span various sectors, with automotive, industrial and tertiary, technology, engineering, finance leading
the charge in recruitment efforts.  These industries reflect the evolving needs of the workforce in response to economic shifts
and technological advancements.
In terms of contract duration, the majority of job offers are structured around up to a year contracts.
This pattern highlights the current employment landscape, which leans toward more secure, longer-term engagements.
This overview reflects the present trends in France’s labor market, where regional demand, work flexibility, industry growth, and contract
stability are key focal points for both employers and job seekers.
```

True result

```
France has a relatively high number of job offers.
```