**Project Summary**

**Team Name:**

**AiTrio**

**Group members:**

**Nardine Hanfi**

**Akram Sahli**

**Shayma Oueslati**

**Full project name:**

**Gov QA (passage retrieval then answer generation)**

**Jira Link:** [**Jira**](https://ai-trio.atlassian.net/jira/software/projects/TASK/boards/1/backlog?epics=visible&issueParent=10012&selectedIssue=TASK-17)

**Git Link:** [**Git**](https://github.com/HanfiNardine/esprit-nlu-2023-AiTrio)

**2022-2023**

1. **Data**

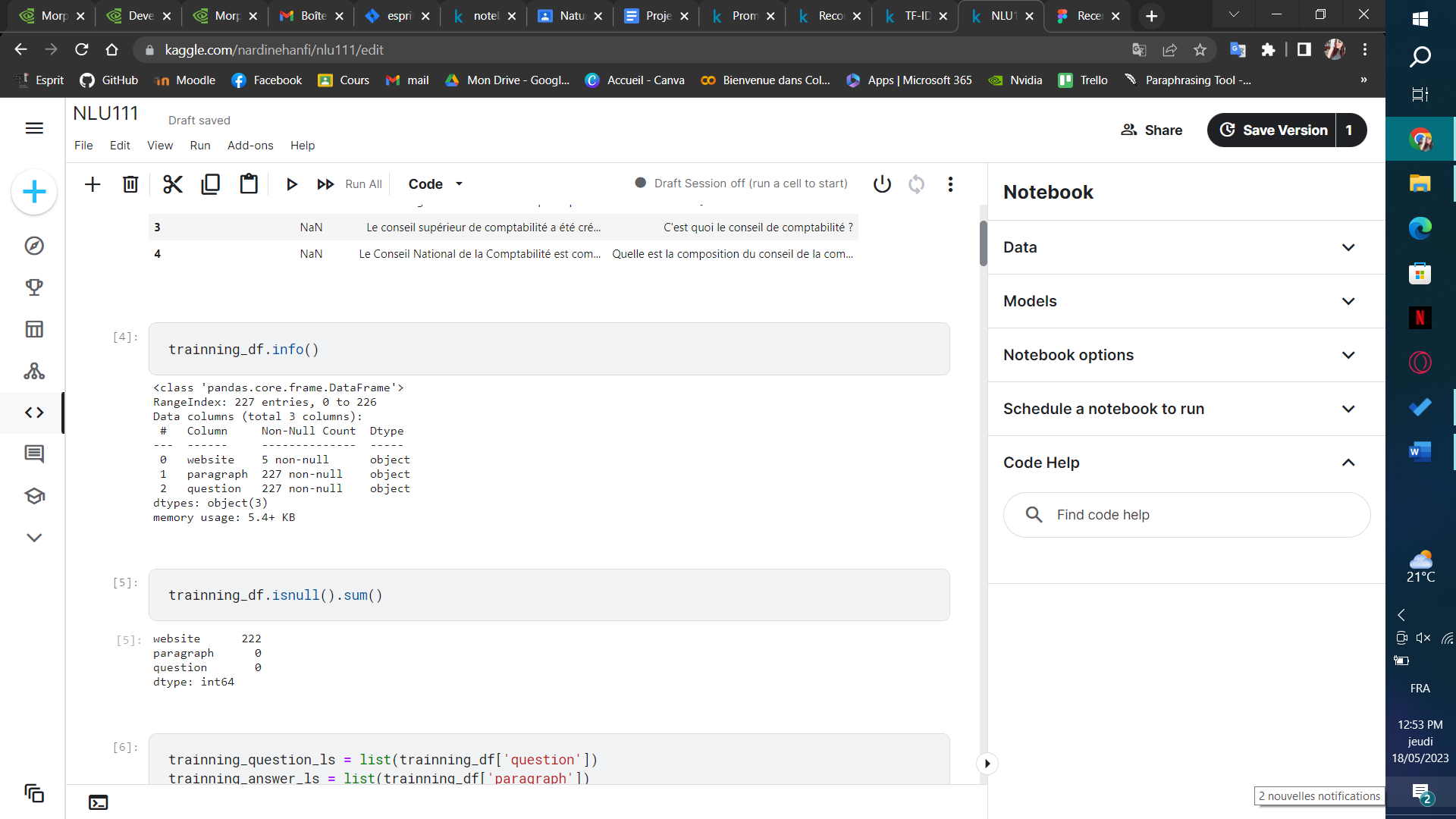
The first step in our project is scrapping.

We collected data by scraping gov.tn websites related to finance.

<http://www.cga.gov.tn/>

<http://www.finances.gov.tn/>

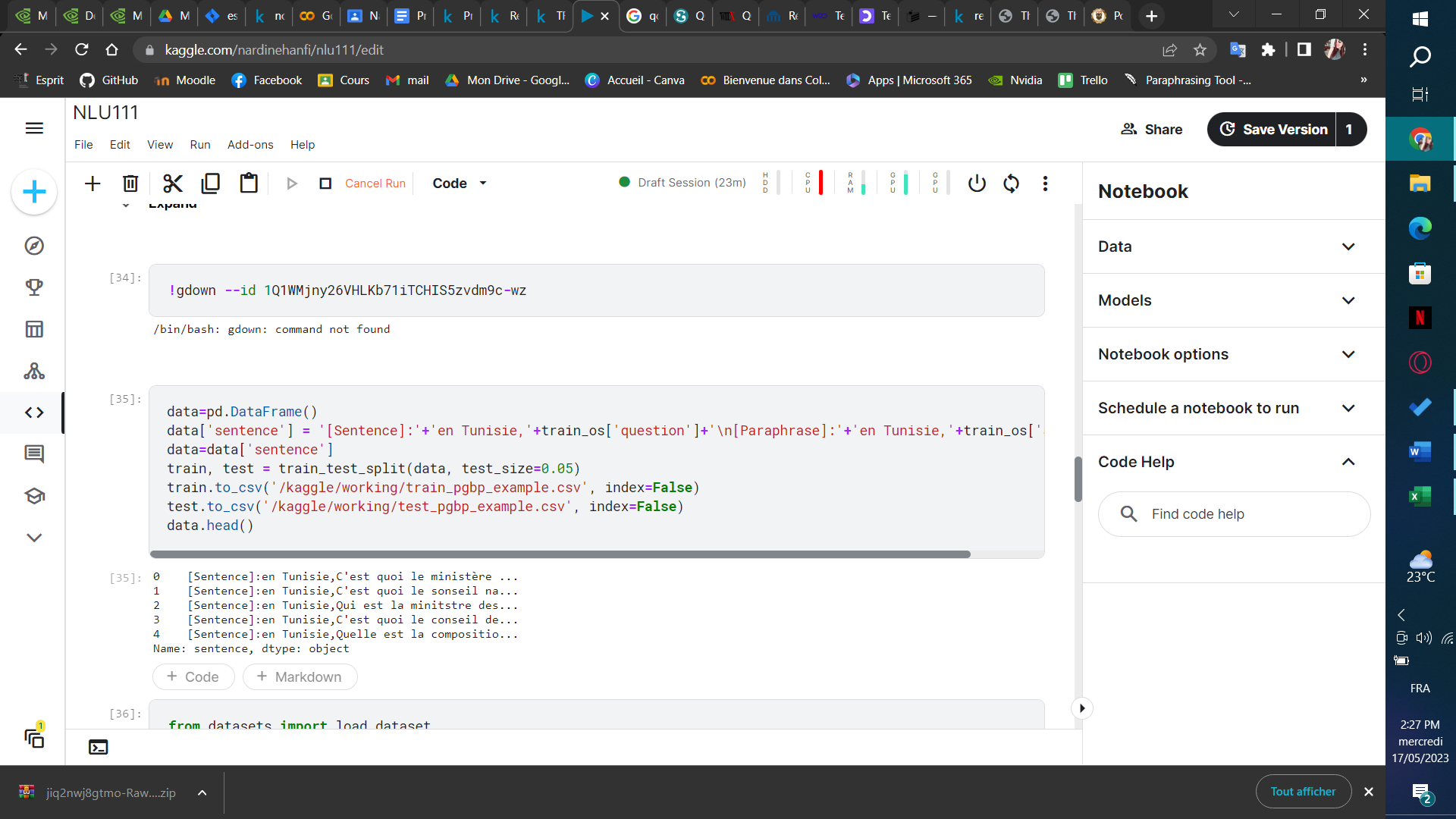
<https://www.douane.gov.tn/>



After getting our data, we cleaned it by removing any unwanted characters (e.g \n \r). We then removed any stop words, emojis or arabic characters.

Then we created a target column with the intended information for each listing.

Now we created prompts with our description and the intended target :



Now our data is ready to be passed to the next phase.

1. **GPT-J**

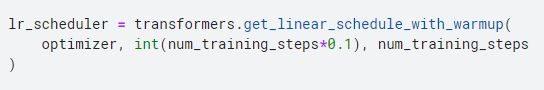
To extract the relevant information from our user input, we finetuned a gpt model (gustavecortal/fr-boris-8bit) which is based on the hivemind/gpt-j-6B-8bit . Boris was trained on around 78B tokens of French text from the C4 dataset.

This is a version that is modified so we can generate and fine-tune the model in Colab.

We tokenized our data using the model tokenizer.

For the model, we used Adam8bit as an optimizer

,we ran the model for 5 epochs and we set up a learning rate scheduler



Then we started the finetuning process, which ended with a loss of 2.3138 .

1. **Top-N paragraph retrieval**

For the this part, we used TF-IDF to get the top 5 paragraphs for a question

Une image contenant texte, capture d’écran, logiciel, Icône d’ordinateur

Description générée automatiquement

Une image contenant texte, capture d’écran, logiciel, Icône d’ordinateur

Description générée automatiquement

and we display the top 5 paragraphs

Une image contenant texte, capture d’écran, logiciel, Icône d’ordinateur

Description générée automatiquement

1. **Answer generation**

For this part we generated answers for our questions using gpt-j model

Une image contenant texte, capture d’écran, logiciel, Icône d’ordinateur

Description générée automatiquement