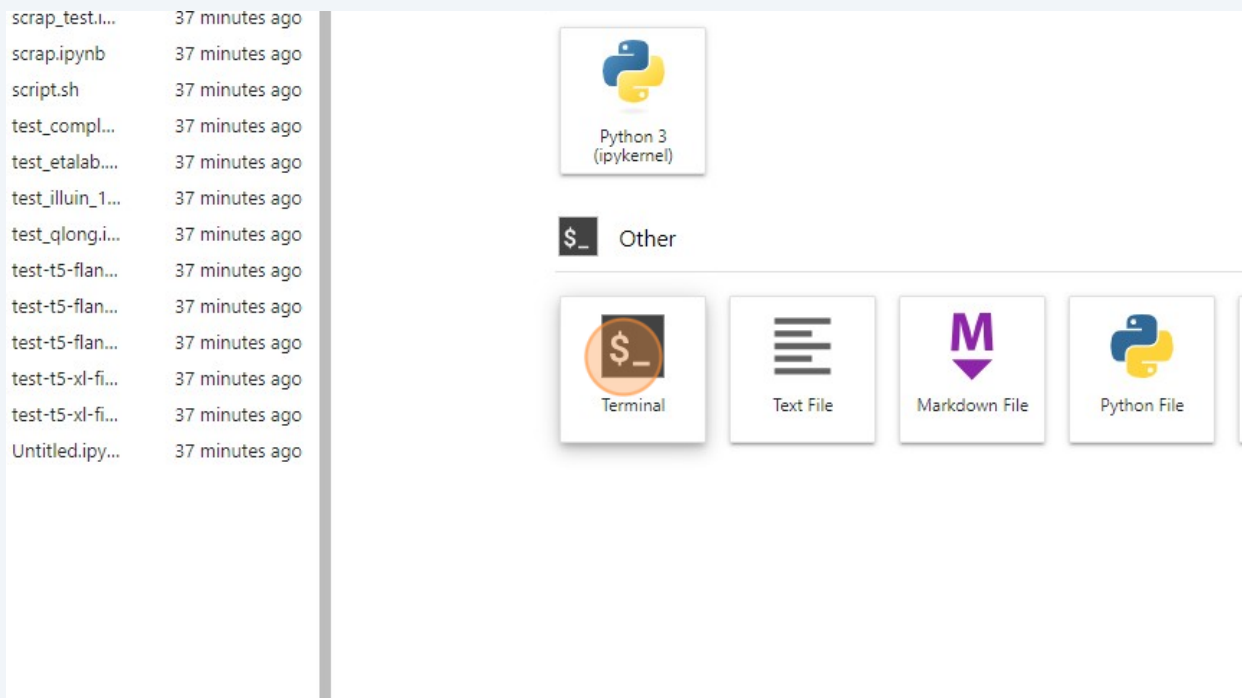


Exécuter le code de collecte de PDF en ligne Scribe[®]

1

ouvrir une fenêtre de terminal dans une instance de jupyter-pytorch-gpu sur le ssp cloud.



2

Type "cd/sispea/NLP" puis "pip install -r requirements.txt" pour installer les librairies nécessaires

3

Modifier le numéro "start", qui représente le premier lien à récupérer en ligne. La commande search va par défaut retourner 100 liens.

```
[55]: from googlesearch import search
      requête='RPQS'
      links=[]

      for url in search(requête,start=100,pause=2):
          print("URL: ",url)
          try:

              read = requests.get(url)

          # full html content
          html_content = read.content

          # Parse the html content
          soup = BeautifulSoup(html_content, "html.parser")
          for link in soup.find_all('a'):
              current_link = link.get('href')
              #print(current_link)
              if current_link!=None:
                  if current_link.endswith('.pdf'):
                      if ((re.findall("rpqs",current_link)!=[]) or (re.findall("RPQS",current_lin
```

4

Exécuter toutes les cellules du notebooks

Requirements

On importe toutes les librairies nécessaires

```
[1]: import numpy as np
      import re
      import pandas as pd
      #!pip install unicode
      from unicode import unicode
      import torch
      import ast
```

ModuleNotFoundError Traceback (most recent call)

```
Cell In[1], line 3
      1 import numpy as np
      2 import re
----> 3 import pandas as pd
      4 #!pip install unicode
      5 from unicode import unicode
```

5

Le fichier csv final avec les liens des PDF sera enregistré sur le dépôt s3