Par DUBOIS Vincent (mis à jour le 30/03/2020)

JOURNAL DE BORD : 30/03/2020

- Installation du module Pytorch (autre module permettant une accélération CUDA AVEC prise en charge des chaînes de caractères et une possibilité de faire de la classification avec appui GPU (malheursement impossible avec mon gpu actuel...)).

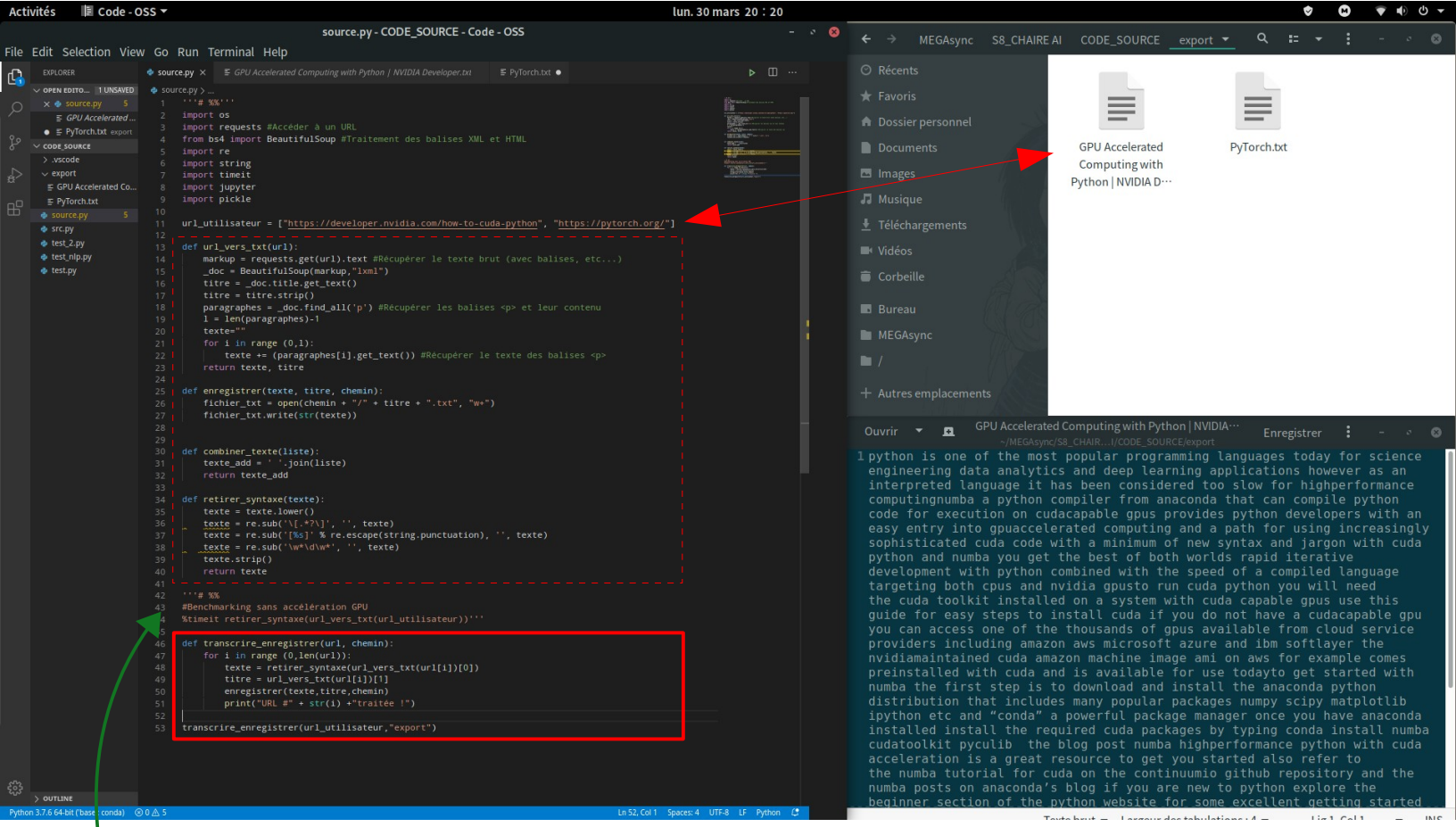
- Entrevue du Cloud Natural Language API :
  - Outil de parsing de texte émettant une requête au service dans le Cloud (avec une IA derrière)
  - /!\ https://cloud.google.com/natural-language/pricing
    Outil payant dès 5000 caractères
  - Peu d'intérêt pour exploiter un GPU étant donné que le calcul est effectué dans le cloud…

- Revue de NLTK https://www.nltk.org/api/nltk.html

  - Outil Open Source, avec un module à importer dans python
  - Un module contient les outils d'analyse de texte, et plusieurs packages additionnels contiennent des données de référence.

- Réalisation de la première partie du code source :

  - Parsing de texte depuis une liste d'URL au choix (HTML et XML pour l'instant)
  - Fonctions de requête, d'export au format txt et de nettoyage (enlever la ponctuation, etc.) programmées
  - Fonction maîtresse faisant appel aux fonctions imbriquées ci-dessus en préparation (Ici : URL → texte brut en .txt avec comme titre le contenu de la balise <title> contenue dans la tête de la page HTML)



OBJECTIFS SUIVANTS :

- Préparer le système de benchmark avec la fonction%timeit en Ipython présentée dans le JDB du 23/03 (pas urgent car pas de GPU Nvidia corrects sous la main)

- Commencer la classification des textes récoltés (avec NLTK, et si possible Pytorch en mode CPU pour l’instant).

- (Compléter davantage les fonctions pour la récolte de textes)