**TALB330A Algorithmique et structuration des données 2**

**Projet 1 S6, 30 janvier 2023**

**à rendre pour le 6 février 2023 15h30**

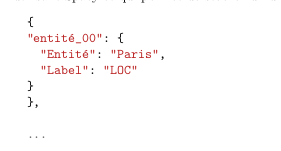
Vous pouvez travailler soit à partir des corpus que vous avez créés au semestre précédent, soit à partir d’un nouveau corpus.

Le but de l’exercice est de :

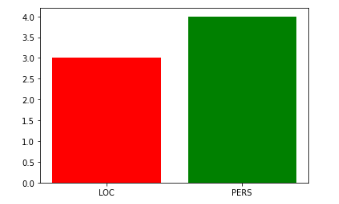
* Utiliser l’outil de Reconnaissance d’Entités Nommées (NER), Spacy : LOC, PERS, ORG et en déterminer les limites
* Utiliser la library [Matplotlib](https://matplotlib.org/stable/plot_types/basic/bar.html#sphx-glr-plot-types-basic-bar-py) pour la représentation graphique des données.

Votre programme devra présenter les étapes suivantes :

1. Lire les textes
2. Pratiquer la Reconnaissance d’Entités Nommées avec [Spacy](https://spacy.io/usage/linguistic-features#named-entities) et obtenir en sortie des documents au format json :



1. Observer les résultats et les commenter (20 à 30 lignes)
2. Représenter les résultats avec un diagramme comme ci-dessous :



Vous pouvez utiliser Matplotlib : <https://matplotlib.org/stable/plot_types/basic/bar.html#sphx-glr-plot-types-basic-bar-py>

1. À partir d’une sortie au format json :
   1. Préparer vos données en déterminant celles qui sont des Vrais Positifs, c'est-à-dire que le label qui leur est distribué est bien le bon (Paris = LOC, Marine = PERS …), les Faux Positifs (les labels attribués ne sont pas les bons : Charlotte = ORG, Limoges = Pers, Aller = LOC)
   2. Proposer un programme qui permet de représenter de manière quantitative ces résultats. Pour ce faire, vous pouvez utiliser un plot / représentation graphique en forme de camembert, par exemple.