Sciences des données pour les plus belles photos du monde

Contexte :
Depuis de nombreuses années la communauté des sciences des données s'intéresse à la récupération de données, à l'ajout d'informations, pour pouvoir extraire de nouvelles connaissances. C'est dans ce contexte que s'inscrit ce TER : il consiste à voir comment récupérer des informations sur le web (scraping), à analyser les informations et à les enrichir. Le contexte est celui du réseau social Reddit (https://www.reddit.com) et de la communauté (https://www.reddit.com/r/EarthPorn/) qui propose de magnifiques photos de la terre et l'objectif est de pouvoir positionner ces photos sur une carte automatiquement.

Il s'adresse à des personnes intéressées par les sciences des données et plus particulièrement aux données textuelles associées aux images afin de rechercher l'information utile pour localiser les lieux des images.

Travail à réaliser :
Dans un premier temps il conviendra de mettre en place une chaîne de traitements : il faudra développer une application qui permettra de récupérer des messages de la communauté Reddit [1]. A partir de ces données, il faudra analyser les textes associés afin de rechercher des indicateurs de localisation. De nombreux outils comme treetagger [2] existent et peuvent permettre d'aider à repérer les éléments dans la structure du texte. Vous utiliserez ensuite la base Geoname [3] qui vous permettra de retrouver les coordonnées GPS d'un lieu. Au final il conviendra d'afficher l'image sur une carte en utilisant OpenStreetMap [4]. Vous pourrez constater que trouver la localisation n'est pas une tâche facile car les textes sont souvent courts, Geoname peut retourner plusieurs résultats (e.g. il y a au moins 5 « Montpellier » dans le monde), etc. aussi à partir de la chaîne de traitements vous pourrez améliorer la reconnaissance des localisations et la sélection des endroits dans Geonames (mise en place d'heuristiques).

Prérequis :
- langage de programmation (Python)

Nombre d'étudiants : 3 à 5

Encadrant : Pascal Poncelet (contact : Pascal.Poncelet@lirmm.fr)

Références :
[1] Scaping Reedit : https://towardsdatascience.com/scraping-reddit-data-1c0af3040768
[2] Treetagger pour Python : https://treetaggerwrapper.readthedocs.io/en/latest/
[3] Geoname : https://www.geonames.org
[4] OpenStreetmap : https://www.openstreetmap.fr