

M2 - Économétrie Appliquée et Statistiques

IAE Nantes - Université de Nantes

**Projet SVM et réseaux de neurones**

**Quentin CHIFFOLEAU**

**Loïc Corven**

2021-2022

Sommaire

[I - Introduction 3](#_heading=h.gjdgxs)

[II – Présentation et préparation des données 3](#_heading=h.30j0zll)

[III – Méthodologie utilisée 3](#_heading=h.1fob9te)

[IV – Application et modélisation 3](#_heading=h.3znysh7)

[V - Conclusion 3](#_heading=h.2et92p0)

# I - Introduction

Wikipédia est la plus grande encyclopédie collaborative disponible sur le web. Créée par Jimmy Wales et Larry Sanger le 15 janvier 2001. Il s’agit d’une œuvre libre, en accès gratuit où n’importe qui peut diffuser des informations sous forme d’article. En quelques années, elle est devenue l’encyclopédie la plus fournie et la plus consultée au monde. Wikipédia est disponible dans 300 langues, celle la plus utilisée, l’anglais, compte plus de 6 millions d’articles en début 2020. Ces articles sont souvent agrémentés d’images comme support du texte.

C’est ici que se trouve l’objectif de notre projet Kaggle. En effet, le but est de prédire le titre ainsi que la description de la référence d’un dataframe composé d’images. Les données de ce projet sont scindées en deux fichiers : un dossier contenant les données d’apprentissage et un deuxième dossier contenant les données de test. S’agissant d’images, celles-ci prennent une place conséquente de plus de 300 Go.

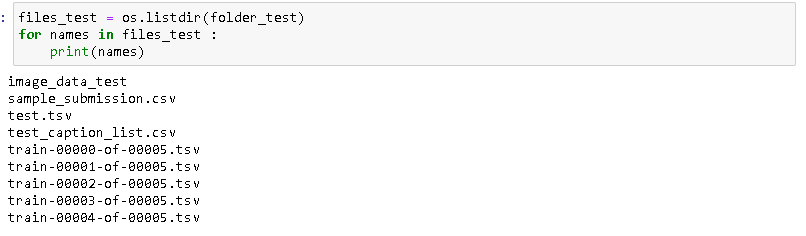
Pour résoudre ce problème, nous utiliserons des techniques de machine Learning. Il s’agit de techniques d’apprentissage des machines à l’aide de données d’apprentissages. Ainsi des modèles sont créés à partir de cet échantillon d’apprentissage et permettent de prédire des caractéristiques sur des échantillons test. Pour que cela fonctionne au mieux, il faut donc disposer d’un maximum de données d’apprentissages et que la diversité y soit très forte pour que le modèle soit le plus efficace face à un maximum de situation. Nous utiliserons ici des modèles d’apprentissage supervisé.

Dans une première partie, nous procéderons à une analyse descriptive de nos données, puis à la préparation de celles-ci pour l’application de nos modèles. Nous vous présenterons ensuite les méthodologies que nous avons utilisée pour répondre à la problématique. Dans un dernier temps, nous analyserons nos résultats et conclurons sur leurs pertinences.

# II – Présentation et préparation des données

Notre base se compose de deux dossiers : un d’apprentissage et un de test. En effet, les données d’image d’apprentissage étant tellement volumineuse, elles ont dû être séparées du reste des fichiers. Dans les dossiers image\_data\_test et image\_data\_train se trouvent les données images : l’url de l’image originale, image encodée en base 64 avec une résolution de 300 pixels et les urls des pages communes où l’on peut retrouver cette image.

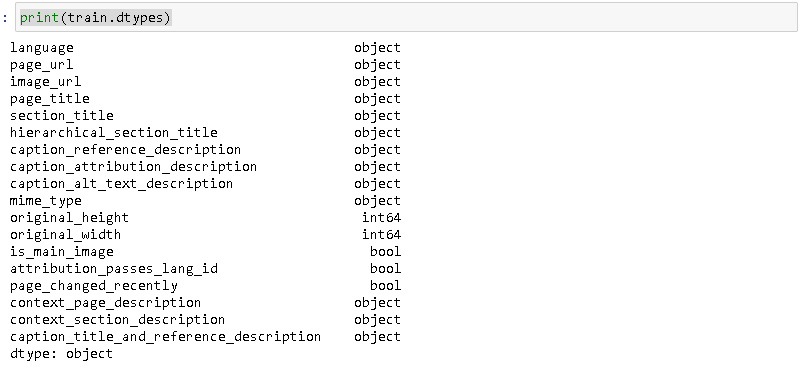
Dans le dossier test, nous disposons des fichiers listés ci-dessous :



Sample\_submission correspond au fichier où nos résultats de modèles seront rappatriés. Test correspond à la liste des urls d’images, chacune associée à un id auquel nous devons associer un titre et une référence de description présent dans le second fichier test\_caption\_list. Finalement, nous nous intéresserons aux différents fichiers train qui disposent de plus d’informations.

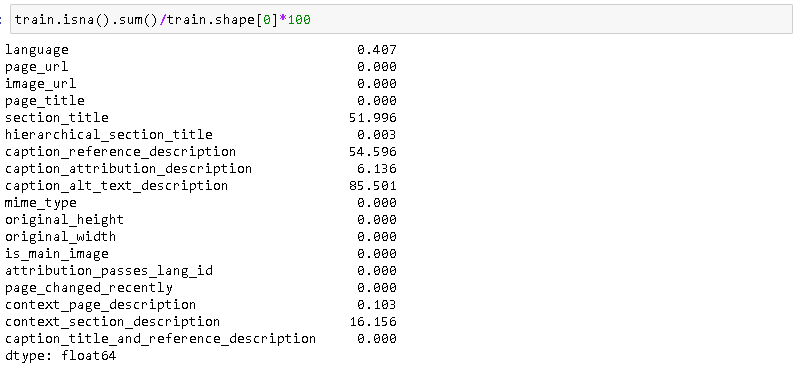
Ce fichier contient 18 colonnes que sont les suivantes :

* Language : la langue utilisée pour l’image
* Page\_url : l’url de la page où se trouve l’image
* Image\_url : l’url de l’image en question
* Page\_title : le titre de la page
* Section\_title : la section où se trouve la page
* Mime\_type : le type d’extension pour l’image
* Original\_height : la hauteur de l’image en pixels
* Original\_width : la largeur de l’image en pixels
* Is\_main\_image : binaire pour savoir s’il s’agit de l’image principale de la page
* Attribution\_passes\_lang\_id : binaire pour savoir si la langue de la page correspond à la description de l’image
* Page\_changed\_recently : binaire pour savoir si la page a été modifiée récemment
* Caption\_title\_and\_reference\_description : notre variable à définir pour test

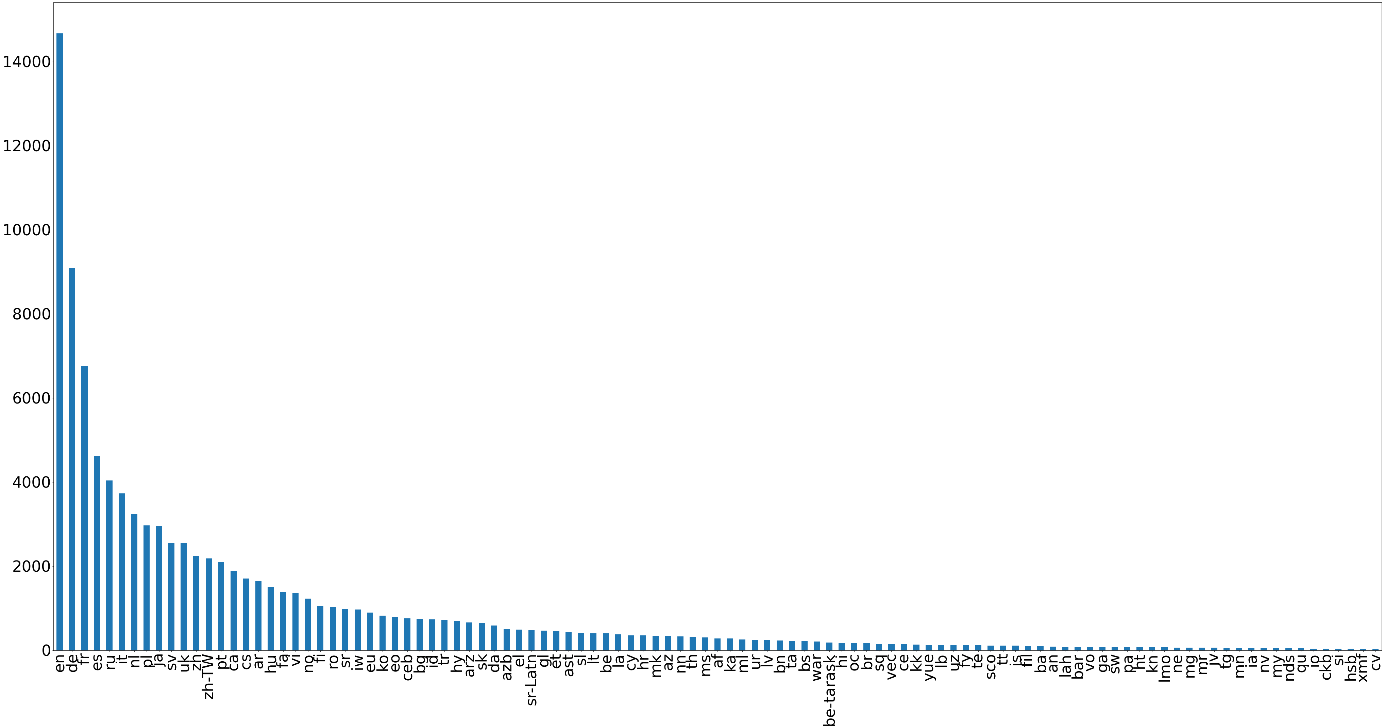
Il y a également 6 autres variables, qui correspondent à des références de pages, de sections et sous-sections. Nous regardons ensuite le type de l’ensemble de ces variables :

Comme montré ci-dessus, la plupart des variables sont des chaines de caractères, seuls la hauteur et la largeur de l’image sont des variables quantitatives et nous avons également trois variables binaires.

Nous regardons ensuite si nos données contiennent des valeurs nulles ou absentes à l’aide de la fonction. isnull(). Le test nous indique la présence de valeurs absentes. Nous regardons ensuite dans lesquelles se trouvent ces valeurs manquantes :

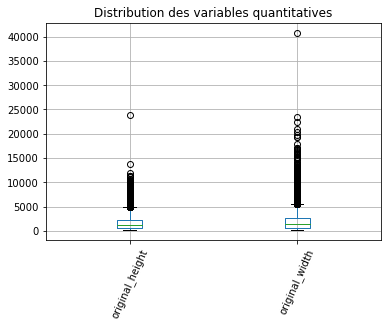


Comme indiqué ci-dessus en pourcentage du nombre d’observations, nous avons 8 variables qui contiennent des valeurs manquantes. Celles en contenant le plus sont les variables hierarchical\_section\_title, caption\_reference\_description, caption\_alt\_text\_description et context\_section\_description avec respectivement 51.99, 54.59, 85.50 et 16.15 % de valeurs manquantes.

Afin d’avoir une idée des langues présentes, nous fabriquons un graphique en batons en fonction de la langue utilisée par chaque page :

Nous remarquons ainsi que le nombre de langues est très élevé avec l’anglais qui prédomine sur l’ensemble des autres. Notons que les 6 langues les plus présentes sont : l’anglais, l’allemand, le français, l’espagnol, le russe et l’italien.

Finalement nous procédons à la création de boxplots pour nos deux variables quantitatives :

On remarque que les images ont globalement la même taille. Cependant, nous supposons la présence d’outliers dans la partie haute de chaque boxplot. Cela paraît logique, signifiant que la plupart des images ont la même résolution et que seulement un nombre faible dispose d’une résolution élevée (hauteur et largeur élevées).

# III – Méthodologie utilisée

# IV – Application et modélisation

# V - Conclusion