PARTIE DATA 1

EXPLORATION ET TRAITEMENT DES DONNES

Dans cette partie nous allons voir l'analyse et le travail réalisé sur le Dataset en expliquant les modification apporté ainsi son manipulation sur le code.

1.1 Jeux des données

1.1.1 La collection des données

Nous avons commencé d'abord notre collecte des données sur le site `https://www.gbif.org/fr/` mais nous avons rencontré des difficultés lors de la collecte des donnes en terme de la qualité des images ainsi le nombres insuffisants d'image pour chaque classe de plantes, ce qui ne nous permet pas de générer un model parfait de reconnaissance d'image. Pour cela nous avons cherché sur d'autre site, et nous avons trouvé le site `https://www.kaggle.com/abdallahalidev/plantvillage-dataset` qui contient des jeux de données partagées par des éditeurs d'ensemble de données et qui sont plus faciles à utiliser et bien organisés.

1.1.2 Exploration des données

Notre jeux de données se compose uniquement de fleurs de 5 classe nommés comme suivant :
1-Daisy
2-Dandelion
3-Rose
4-Sunflower
5-Tulip

Detail sur les jeux de données :

Pour les images des classes que nous avons choisi à entraîné, nous trouverons qu'il y a des informations perturbatrices par exemple, il y a quelque images qui présentes des insectes perturbants l'unicité de la fleur et ces photo ne représentants qu'une partie du jeux donnes, a

part ca notre jeux de donnes contient des images qui sont bien choisi pour générer un model de classification d'image.

//TODO Figure .. : Présence d'insectes perturbants l'unicité de la fleur

Modification apporté :

Dans notre jeux de données que nous avons collecté sur le site Kaggle, on trouve qui il y a des classes volumineuses que d'autres avec 984 pour le maximum (Tulip) et 733 pour le minimum (Sunflower), ainsi les photos ont des tailles très variés par exemple une majorité ont 200x200.Il était donc important d'ajuster notre Dataset afin d'avoir un standard avec lequel travailler et générer un model de meilleur performance.

Dans un premier temps Nous avons redimensionner tous les photos des classes à une dimension de 50x50 car il permet de conserver une bonne visualisation tout en nous permettant de travailler sur une matrice carrée bien plus petite, et donc de réduire le temps des calculs.D'autre part nous avons defini le meme nombre d'image pour chaque classe pour avoir un jeux de données uniforme.

Pour cela nous avons générer le fichier *resize.py* en python qui nous permet de modifier notre jeux de données utilisé pour générer le model.

1.1.3 Applications

TODO