**Classification de personnages dans la trilogie originale Star Wars :**

**Remarques générales :**

* **On n’utilise pas de mots en anglais (pattern, loop, etc ..)**
* **Soyez pédagogues. Vous devez aider le lecteur à entrer dans le sujet grâce à un maximum d’informations.**

**Introduction :**

La saga Star Wars est constituée de 3 trilogies, qui font intervenir de nombreux personnages, que l'on peut distinguer selon leur espèce, leur genre, ou encore leur appartenance à un groupe. Pour chacun des trois épisodes (non, vous avez 7 épisodes) de la trilogie originale (pas seulement, donc il faut détailler), nous avons pour données les dialogues (dire que c’est les scripts complets avec dialogues et didascalies) (dire qu’il y a un fichier par film) ainsi que des caractéristiques concernant les personnages présents (dire que c’est un fichier différent des films).

**(Problématique : elle fait partie de l’introduction)**

L'objectif de ce travail est de mettre en évidence les principaux indicateurs (âge, genre, champs lexicaux utilisés) qui influencent le nombre d'intervention et le temps de parole des personnages (est-ce que vous avez fait quelque chose pour répondre à ces questions ? Si oui, il faudra les décrire, sinon, il faudra enlever cette phrase). On pourra en déduire si la répartition du temps de parole est équilibrée ou pas (est-ce que vous avez fait quelque chose pour répondre à cette question ? Si oui, il faudra les décrire, sinon, il faudra enlever cette phrase). Enfin, on essayera de répondre à la question suivante : d'après l'analyse des données (quelles sont des données ?) de la première trilogie (non, comme on l’a dit plus haut, vous avez 7 épisodes), aurait-on pu prédire que le personnage principal de la dernière trilogie était une femme ? (Dire en quelques phrases précises les grosses étapes que vous avez faites pour répondre à cette question)

**Pré-traitement des données :**

La première étape de ce projet sera de prendre en main les données (brutes) des scripts de la saga, et d'opérer les premiers traitements permettant de les transformer en données exploitables automatiquement. (Cette phrase n’arrive pas au bon moment, il faut d’abord décrire les données, peut-être ajouter un screen, et dire pourquoi ces données n’étaient pas automatiquement exploitables)

Ici notre base de données (est ce que vous parlez d’un seul film ou de tous les films ? Est-ce que les fichiers de films ont tous le même format ? Est ce que vous avez fait le pré-traitement sur tous) n’est pas un dataframe (précisez la structure que vous avez et la structure d’un dataframe), il s'agit d'un texte composé de noms des personnages, de dialogues et de didascalies. On s'intéresse uniquement aux personnages et à leurs dialogues, donc on va d'abord faire le tri afin d'obtenir des données structurées faciles à exploiter. (qu’entendez-vous par « facile à exploiter » ? Dire quelle structure doivent avoir vos données pour être exploitées)

Nous avons remarqué que les discours des personnages sont décalés (le mot correct est « indentés ») de 25 espaces, suivis d'un espace, c'est pour cela qu'on a défini 3 groupes dans la fonction re.compiler. (Décrire : Pourquoi avez vous cherché cette indentation ? Dire ce que vous cherchiez en faisant ca ? Ensuite dire comment l’avez fait !) (décrire: d’où sort la fonction re.compiler ? D’un package ? Que fait-il globalement, et que fait cette fonction en particulier? Quelles sont ses entrées et ses sorties ?) On a ensuite appliqué la fonction pattern.finditer (décrire : qui est cette fonction, quelles sont ses entrées et ses sorties) à notre texte de base (définir « texte de base » please ?) pour demander le groupe 2 (décrire : c’est quoi le groupe 2 ?) en sortie, ce qui nous a permis d'obtenir la partie texte qui nous intéresse notamment (est-ce que ça renvoie d’autres choses ? Sinon, il ne faut pas utiliser le mot « notamment » mais décrire toutes les sorties) les noms des personnages suivis de leurs discours. Les librairies pandas et numpy sont très pratiques pour ces manipulations. (De quelles manipulations parlez vous ? Précisez de quel langage sont ces deux librairies et quels types de fonctions / objets vous vous êtes servis.)

Pour cela, on a créé un ‘pattern’ (dire plutôt motif régulier que pattern) pour chercher toutes les chaines de caractères qui y correspondent. Par une loupe (on écrit « boucle » en français, et on évite absolument les mots anglais dans un texte en français, c’est une convention de rédaction) nous avons créé une liste de toutes les chaines de caractères qui correspondent à notre pattern. Dans un premier temps, nous avons supprimé mannuellement les 11 premières lignes qui nous n'intéressent pas car elles correspondent à des didascalies (Précisez pourquoi vous avez enlevés ces 11 lignes ? Manuellement ou pas ? En quoi sont-elles différentes des autres didascalies ? Leur indentation est-elle différente ?) . Nous avons remarqué que les prénoms des personnages ont une structure spéciale, ils sont précédés par 13 ou 14 espaces (13 ou 14 espaces ? 13 ou 14 espaces de plus que les 25 ? Les noms des personnages ont-ils d’autres caractéristiques que vous avez utilisés ? ), nous avons donc créé un 'pattern' pour les récupérer. Nous avons ensuite regroupé les prénoms des personnages dans une liste. Tous les éléments qui ne sont pas dans cette liste sont des répliques des personnages, ce qui nous a permis de récupérer les répliques des personnages. (vous êtes sur que ce ne sont que des répliques de personnages ? Comment avez vous trier les répliques des personnages et les didascalies ?)

Pour la concaténation des paroles (On dit plutôt « dialogues ou répliques» que « paroles ». Précisez : pour la concaténation des dialogues, est-ce que c’est la concaténation sur un seul film, ou sur plusieurs films ?), nous avons utilisé la fonction join() qui permet la concaténation des chaines de caractères dans une liste. Ainsi on obtient un tableau composé des noms de personnages et leurs dialogues correspondant, il s’agit d’un tableau de 2 colonnes et 1002 lignes (1002 lignes, c’est pour un seul film ? Si oui précisez lequel. Bien rappeler que dans ce tableau les personnes apparaissent pas une seule fois selon leurs dialogues). (Précisez que ce tableau est le format que vous cherchiez pour exploiter vos données) Afin d’automatiser cette démarche et pouvoir l’appliquer aux autres bases de données (Non. Vous avez une seule base de données, composée de (combien?) fichiers scripts brutes ou pas. Vous devez impérativement préciser tout ça.) nous avons créé une fonction qui fait tout le travail (qui peut s’appliquer aux autres fichiers, via des paramètres que vous devez décrire) car il est très pénible, voire impossible, de supprimer toutes les lignes inutiles manuellement. (Dites dès le début que c’est votre fonction que vous allez décrire dans ce paragraphe, et que vous faites cette fonction pour l’appliquer à plusieurs fichiers.)