**Notice**

Cette notice décrit les variables utilisées présentes dans le fichier Data.xlsx ainsi que les variables créées lors de l’exécution du code python du notebook Exploration.ipynb et sauvegardées dans le fichier Data\_analyse.csv généré par le code.

***Définition des variables :***

**Eleve\_ID :** identifiant individuel de chaque élève

**genre :** 0 = garçon

1 = fille

**classe :** 1 = Marseille 7ème1

2 = Marseille 7ème 2

3 = La Ciotat

**texte** : chemin vers le fichier .docx contenant le texte original

**personnages** : genre des personnages (enfants) présent dans la rédaction

0 = inconnu

1 = garçon

2 = fille

Étant donné qu’il y a deux enfants dans l’histoire, les cellules peuvent prendre l’une des six valeurs suivantes :

00 : genre inconnu pour les deux enfants

10 : un garçon + un enfant dont le genre est inconnu

11 : deux garçons

20 : une fille + un enfant dont le genre est inconnu

21 : une fille + un garçon

22 : deux filles

***Définition des variables créées lors de l’exécution du code :***

**texte\_extrait** : Textes originaux extraits des fichiers word

**texte\_nettoye** : Textes originaux tronqués du premier paragraphe commun à l’ensemble des copies

**nb\_mots** : Nombre de mots présents dans texte\_nettoye

**texte\_sans\_stopwords** : Textes lemmatisé auxquels les mots vides ont été retirés

**texte\_lemmatise** : Version lemmatisé de la version du texte nettoyé

**nb\_mots\_uniques** : Nombre de lemmes différents dans le texte

**TTR** : Ratio Type/Token, définit comme de nombre de lemmes différents divisé par le nombre de mots total dans chaque texte. Il représente la richesse lexicale du texte proportionnellement à sa longueur. Plus le TTR est proche de 1, plus le texte utilise proportionnellement des mots issus de champs lexicaux différents.

**texte\_sans\_stopwords** : Textes lemmatisé auxquels les mots vides ont été retirés

**top\_20\_mots** : Liste des 20 lemmes (à l’exception des mots vides) les plus utilisés dans chaque texte.