Documenter les insights obtenus à partir de l'analyse des n-grammes :

J’ai nettoyé et prétraité les données, généré des n-grams de différentes tailles, et analysé la fréquence de ces n-grams pour identifier les plus fréquents.

1. **Nettoyage et pré traitement** :
    - Conversion des textes en minuscules.
    - Remplacement des lettres accentuées par des lettres normales
    - Suppression des caractères spéciaux et des chiffres.
    - Tokenisation de texte (diviser la chaine de caractères en une liste des mots individuels) .
    - Suppression des mots vides (stopwords “french”).
2. **Génération des n-grams** :
    - Création de n-grams de tailles : 1 à n à partir des textes prétraités.
3. **Analyse de fréquence** :
    - Comptage de la fréquence de chaque n-gram.
    - Identification des n-grams les plus fréquents pour chaque taille.

Résultats et Interprétations:

J'ai pris 4 exemples de n-grammes (1,2,3 et 8 grammes) pour présenter leurs résultats et leur interprétation.

1-grams les plus fréquents

1-gram	Fréquence
donnees	140
data	90
science	80

D’après 1-grams : Les mots « donnees », « data» et « science » figurent en grande partie dans les textes . indiquant que le corpus concerne le domaine de la science des données.

2-grams les plus fréquents

2-grams	Fréquence
data science	80
machine learning	19

- Ces 2 bigrammes, « data science » et « machine learning », sont parmi les plus fréquents. Suggèrent que le corpus se concentre principalement sur la data science, avec des discussions sur les algorithmes, le machine learning .

3-grams les plus fréquents

3-grams	Fréquence
algorithmes machine learning	19
machine learning necessitent	19
grandes quantites donnees	19
outils visualisation comme	19
comme matplotlib seaborn	19

Ces 3-grammes les plus fréquents donnent un aperçu des thèmes récurrents dans le corpus, tels que les algorithmes de machine learning, les exigences en données, et l'utilisation d'outils de visualisation comme Matplotlib et Seaborn.

8-grams les plus fréquents

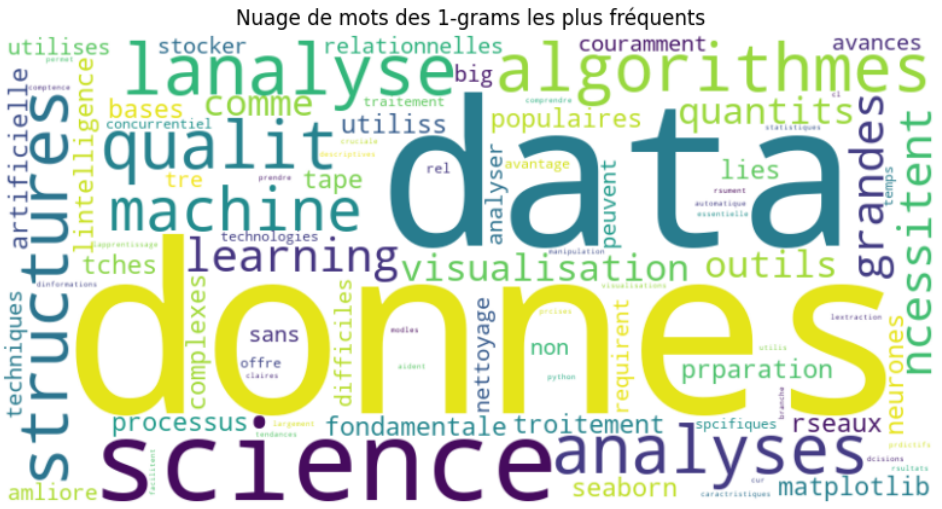
8-grams	Fréquence
outils visualisation comme matplotlib seaborn populaires data science	19
donnees non structurees peuvent etre difficiles analyser sans	19

non structurees peuvent etre difficiles analyser sans techniques	19

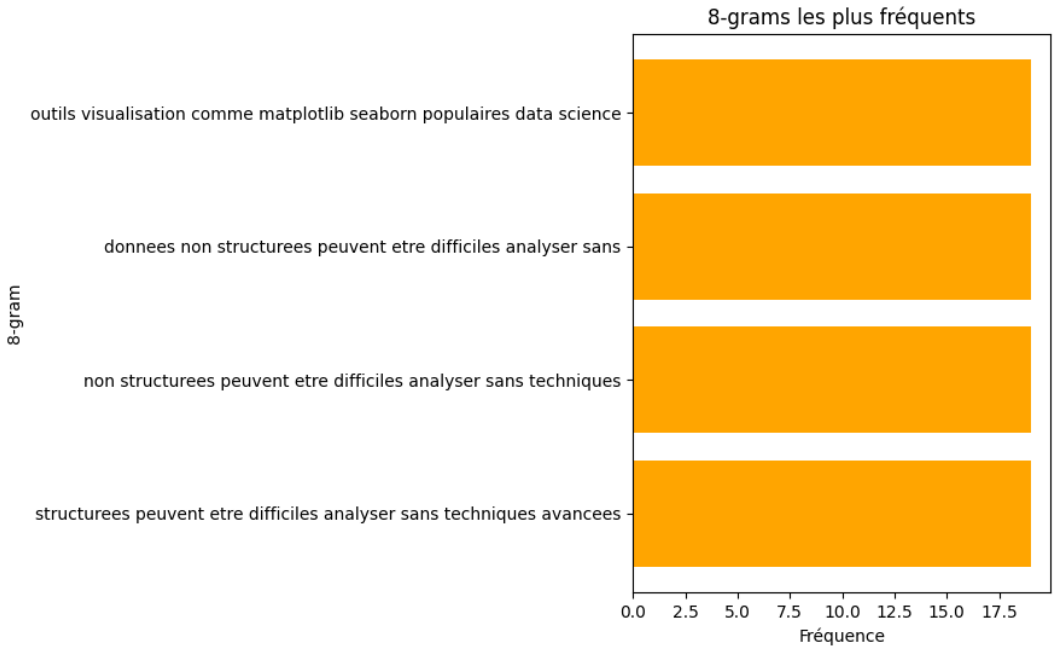
Les 8-grammes fréquents soulignent des aspects clés de l'analyse de données, y compris l'utilisation répandue d'outils de visualisation, les défis rencontrés lors de l'analyse des données structurées et non structurées, ainsi que l'importance des techniques avancées dans ce domaine.

Visualisation

Exemple pour 1-gram： Un nuage de mots



Exemple pour 8-grams



Résumé

Une analyse n-gramme a permis d'identifier les termes les plus fréquents, ainsi que les combinaisons de termes, dans les textes. Cela montre les détails précieux sur le contenu du texte et les sujets susceptibles d'intéresser les personnes impliquées dans la science des données.