**Documentation du Script d'exploration des métadonnées ArXiv**

Ce script Python a pour objectif d'analyser des données d'articles académiques à partir d'un fichier JSON et de générer des statistiques et des visualisations à partir de ces données. Il utilise plusieurs bibliothèques Python telles que numpy, pandas, dask.bag et plotly.express pour accomplir ces tâches.

**Étape 1 : Lecture du fichier JSON**

- Le script commence par lire un fichier JSON contenant les données des articles ArXiv. Le chemin du fichier est spécifié dans la variable `lines`.

**Étape 2 : Transformation des données JSON en objets Python**

- Les lignes lues depuis le fichier JSON sont transformées en objets Python à l'aide de la bibliothèque json.

**Étape 3 : Analyse des données ArXiv**

*Nombre total de records:*

- Le script calcule et affiche le nombre total de records (articles) dans les données ArXiv.

*Fréquence des soumissionnaires (submitters):*

- Le script génère un histogramme des soumissionnaires les plus fréquents (top 10) et l'affiche.

*Fréquence des catégories:*

- Le script génère un histogramme des catégories les plus fréquentes (top 20) et l'affiche.

*Distribution des articles par année de soumission:*

- Le script extrait l'année de soumission la plus récente pour chaque article et crée un graphique de la distribution des articles publiés par année.

*Analyse des articles liés à l'Intelligence Artificielle (IA*):

- Le script filtre les articles appartenant à la catégorie "cs.AI" (Intelligence Artificielle).

- Il calcule le nombre total d'articles dans cette catégorie et affiche le résultat.

- Il génère un graphique de la distribution des articles IA par année de soumission.

*Analyse des auteurs:*

- Le script extrait les auteurs de chaque article et génère un histogramme des auteurs les plus fréquents (top 20).

**Prérequis**

Assurez-vous d'avoir les bibliothèques Python suivantes installées : numpy, pandas, dask.bag et plotly.express.

**Personnalisation**

- Vous pouvez personnaliser le chemin du fichier JSON en modifiant la variable `lines`.

- Vous pouvez ajuster les seuils pour les tops (par exemple, top 10 pour les soumissionnaires) en modifiant les paramètres dans les fonctions `frequencies` et `topk`.

- Vous pouvez également personnaliser les couleurs des graphiques en modifiant les paramètres `color\_discrete\_sequence` dans les appels à `px.line` et `px.bar`.

**Auteur**

* Hajji chakir Zouhair
* Chebbab Aya
* Houssam Meryem
* Msalek Aicha
* Rakhiss Bouchra
* Elouafi Abderrahmane