Fichier explicatif du rendu de projet

Architecture du fichier de rendu :

- **assets**

  Contient le fichier css qui permet de faire le design du dashboard

- **data**
  - **Dashboard**
    - **kws**
      - **Word2VecMatrice**

        C’est ici que sont les csv qui permettent de créer la hitmap des keywords
      - **Word2VecVector**

        C’est ici que sont les csv qui permettent de créer le graphique des keywords
      - **ensemble des autres fichier**

        C’est ici que sont les autres fichier csv qui permettent de créer l’histogramme ainsi que le cloudword des keywords
    - **pays**
      - **Word2VecMatrice**

        C’est ici que sont les csv qui permettent de créer la hitmap des pays
      - **Word2VecVector**

        C’est ici que sont les csv qui permettent de créer le graphique des pays
      - **ensemble des autres fichier**

C’est ici que sont les autres fichier csv qui permettent de créer la carte ainsi que l’histogramme des pays

- **per**
  - **Word2VecMatrice**

    C’est ici que sont les csv qui permettent de créer la hitmap des personnes
  - **Word2VecVector**

    C’est ici que sont les csv qui permettent de créer le graphique des personnes
  - **ensemble des autres fichier**

    C’est ici que sont les autres fichier csv qui permettent de créer l’histogramme ainsi que le cloudword des personnes

- **TF**
  - **jsonTF**
    - **globalTF**

      Les différents fichier pour les kws trié avec la méthode de term frequency sur l’ensemble de chacun des corpus
    - **yearTF**

      Les différents fichier pour les kws trié avec la méthode de term frequency sur chaque année de chacun des corpus
- **TFIDF**
  - **jsonTFIDF**
    - **globalTFIDF**

      Fichier pour les kws trié avec la méthode TFIDF avec le corpus le plus petit en taille sur l’ensemble du corpus

- **jsonBases**

Tous les corpus initiaux permettant de traiter les différents problèmes (fournis initialement)

> 💡 Attention ce dossier n'existe pas sur github car il est vide
> Il faut rajouter manuellement les fichier car ceux-ci sont trop gros pour github

- **partie1**

  Programme python pour faire les premiers graphiques et premières analyses de données donnaient en introduction du projet

- **utiles**

  Programme python fournis initialement afin d'effectuer quelques premiers tests

- **dashboard.py**

  Fichier python dans lequel est créé le dashboard avec les différentes sous parties qui le compose

- **dashboardCreationKeywords.ipynb**

  Fichier permettant la création des différents csv servant pour le dashboard final qui représente les keywords.

- **dashboardCreationPays.ipynb**

  Fichier permettant la création des différents csv servant pour le dashboard final qui représente les pays.

- **dashboardCreationPeople.ipynb**

Fichier permettant la création des différents csv servant pour le dashboard final qui représente les personnes.

- **main.py**

  C'est ici que j'ai réaliser mes premières fonctions de tests et que j'ai enregistrer les premiers json qui me servirons par la suite pour la création des csv pour le dashboard final. J'ai utiliser ici les méthodes classiques, TF, TFIDF et Word2Vec.

- **mainJupyter.py**

  Fichier Jupyter, qui, tout comme le fichier main.py, m'a permis de réaliser mes premiers tests de graphiques avec plotly et dash et enregistrer certains csv qui me permettrons par la suite la création de nouveau csv pour le dashboard final.

Lancement du dashboard :

> *Pour lancer le dashboard il faut lancer le fichier dashboard.py et cela ouvrira un localhost afin de voir celui-ci.*

Contenu du dashboard

Une fois l'application dash ouverte vous pouvez voir un dropdown permettant de choisir l'article pour lequel on souhaite voir les données, ainsi que 3 tabs qui eux permettent de voir les informations, pour l'article choisit, soit des pays, des keywords ou encore des personnes.

Pour les pays :

- Partie Occurrences :
  - Une carte montrant les pays avec une échelle de couleurs en fonction du nombre d'article qui parle de se pays.

    - Un histogramme montrant les pays les plus cités avec le nombre d’articles qui les cites

- Partie Word2Vec :
    - Une hitmap montrant à quel points les 10 pays les plus cités sont utiliser ensemble ou non dans les texte, plus le chiffre se rapproche de 100 plus les pays sont souvent utiliser ensemble et proche dans les textes et plus il se rapproche de -100 plus les pays sont éloignés et ne sont pas utilisés ensemble dans les textes du corpus.
    - Les pays les plus cités sur un graphe, plus les 2 points représentant 2 pays sont proches plus cela veut dire que les pays sont souvent abordés ensemble dans les textes du corpus et inversement s’ils sont éloignés.

Pour les keywords:

- Partie Occurrences :
    - Un histogramme montrant les keywords les plus utilisées dans le corpus choisit.
    - Un wordcloud représentant les keywords les plus cités dans le corpus choisit, plus le mot est gros plus le mot est cités.
- Partie Word2Vec :
    - Une hitmap montrant à quel points les 10 keywords les plus cités sont utiliser ensemble ou non dans les texte, plus le chiffre se rapproche de 100 plus les keywords sont souvent utilisés ensemble et proches dans les textes et plus il se rapproche de -100 plus les keywords sont éloignés et ne sont pas utilisés ensemble dans les textes du corpus.
    - Les keywords les plus cités sur un graphe, plus les 2 points représentant 2 keywords sont proches plus cela veut dire que les keywords sont souvent abordés ensemble dans les textes du corpus et inversement s’ils sont éloignés.

Pour les personnes:

- Nous avons ici la même chose que pour les keywords mais avec les personnes dans les textes.