Projet Plotly Python

L'objectif du projet de la fin de ce module est que vous manipuliez des données en Python puis de créer des visualisations pertinentes. Dans ce TD, vous allez manipuler les données en utilisant la librairie Plotly, librairie dont l'utilisation est très fortement recommandée pour le projet.

Les données

Les données sont un corpus d'articles de médias en ligne, prétraités et triés par date pour faciliter leur utilisation lors de la réalisation de visualisations. De chaque ont été extraits des mots clés et des entités, regroupés dans leur catégories respectives :

- kws (keywords) : les mots clés
- loc : les lieux
- org : les organisations
- per : les personnes
- mis : les autres entités détectées mais sans catégorie déterminée

Les mots clés et les entités ont été extraits grace à des méthodes de NLP, Natural Language Processing ou traitement automatique de la langue naturelle en français, mais à l'aide de méthode différentes.

Les mots clés ont été extraits à l'aide d'étiquetage morpho-syntaxique et de procédés de lemmatisation, des méthodes qui existent depuis plusieurs décennies mais qui sont de plus en plus précises. Tous les mots clés sont à leur forme nominale, c'est-à-dire que les noms et les adjectifs sont au masculin singulier si c'est possible.

Les entités sont extraites à l'aide de NER, Named Entity Recognition ou reconnaissance d'entités nommées en français. L'algorithme utilisé ici est un CNN, Convolutional Neural Network ou réseau neuronal convolutif en français.

Plotly

Plotly est une bibliothèque de fonctions pour faciliter la création de graphiques statistiques complexes et interactifs notamment en Python.

Cette librairie se divise en deux catégories : plotly.express (px) pour réaliser des graphique simplement et plotly.graph_objects (go) pour les graphiques plus avancés. Les données d'entrées pour les graphiques sont des dataframes de pandas, mais vous pouvez manipuler des listes et des dictionnaires que vous convertissez juste avant de faire votre graphique.

La documentation en ligne de Plotly est très riche et très accessible : https://plotly.com/python/

Les premiers graphiques

Pour cette partie, vous allez réaliser des graphiques sur les données de votre choix, mais plus le fichier sera gros et plus il vous faudra optimiser vos algorithmes.

1) Lancez le fichier print-structure.py pour voir la structure du fichier ; pour les petits fichiers, le formatage JSON de Mozilla Firefox peut vous aider à mieux comprendre.
2) Réalisez un barchart du nombre d'articles par mois dans le corpus https://plotly.com/python/bar-charts/#bar-chart-with-plotly-express
3) Sauvegardez ce graphique en html https://plotly.com/python/interactive-html-export/#saving-to-an-html-file
4) Réalisez un barchart des 10 mots clés les plus fréquents dans le corpus
5) Utilisez Dash pour afficher un barchat des 10 mots clés les plus fréquents avec un menu déroulant permettant de choisir l'année https://plotly.com/python/bar-charts/#bar-charts-in-dash

À vous de jouer

Maintenant, vous entrez dans le projet ; mettez-vous en binôme et faites les sorties les plus pertinentes. Vous êtes libre de menez le projet comme bon vous semble, mais voici des exemples de ce qui sera valorisé dans la notation :

- Dashboard clair
- Utilisation avancée de Plotly ou d'autres librairies de visualisation
- Utilisation de méthodes pertinentes (clustering sur les données, corrélation des mots clés, …)
- Croisement des données
- Data mining (Correspondence analysis des mots clés)
- Visualisations pertinentes pour une analyse du corpus
- Notes explicatives pour les visualisations les plus complexes

Vous pouvez vous concentrer sur quelques points ci-dessus si vous le voulez, tant que ces points sont bien approfondis.

Vous ne serez pas notés sur la qualité de votre code mais sur la qualité de vos graphiques et de votre réflexion. Vous pouvez donc ne pas commenter et documenter votre code, vous pouvez aussi faire des copiés-collés de codes trouvés sur Internet.