Collecte, Traitement, et Analyse de données de réseaux sociaux

On considère InPoDA, une plateforme pour l'analyse de données de réseaux sociaux. InPoDa est une plateforme fictive. Voici donc des hypothèses personnelles sur son fonctionnement :

Lorsqu'un utilisateur rédige une publication (un tweet), elle est publiée à travers un script python qui vérifie d'abord la structure de l'objet tweet, le nettoie en éliminant les caractères spéciaux (☺, etc.) au cas où on en a , et le stockera dans un fichier, appelé « zone d'atterrissage ». En pièces jointes un jeu de publications (tweets) que vous pouvez utiliser. N'hésitez pas à utiliser un autre jeu de tweets si vous le pensez nécessaire.

Notons que à chaque lecture d'une ligne du fichier de données fournies (une ligne pourra correspondre à un dictionnaire Python comme il s'agit de données json), il y aura un ajout d'un objet tweet au fichier intitulé « zone d'atterrissage ».

InPoDa propose un ensemble des opérations de traitement de données y compris:

- Identification de *l'auteur* de la publication
- Extraction de la liste de *hashtags* de la publication
- Extraction de la liste des *utilisateurs mentionnés* dans la publication
- Analyse de *sentiment de la publication* (le sentiment peut être positif ou bien négatif). Vous pouvez utiliser le module « *textblob* ».
- Identification du/des *topics* de la publication

D'autres opérations d'analyse de données sont également proposés par InPoDa :

- Top *K hashtags* (k est un paramètre passé par l'utilisateur)
- Top *K utilisateurs*
- Top *K utilisateurs mentionnés*
- Top *K topics*
- Le *nombre de publications* par *utilisateur*
- Le *nombre de publications* par hashtag
- *Le nombre de publications* par *topic*
- *L'ensemble de tweets d'un utilisateur spécifique*
- *L'ensemble de tweets* mentionnant *un utilisateur spécifique*
- *Les utilisateurs* mentionnant un *hashtag spécifique*
- *Les utilisateurs* mentionnés par un *utilisateur spécifique*

Pour avoir des résultats d'analyse de données à jours, l'analyste de données devrait être en mesure de déclencher l'exécution d'un processus qui consiste à récupérer les publications stockées dans la zone d'atterrissage et appliquer pour chaque publication l'ensemble de opérations de traitement de données décrites auparavant. Les résultats devraient être chargées dans un nouvel dataframe qui servira les opérations d'analyse de données décrites ci-dessus.

Travail à faire :

Le travail demandé consiste en quelques étapes principales, à savoir :

1. Élaboration d'un diagramme qui décrit le fonctionnement de InPoDa.
2. Prendre connaissance du sujet et plus particulièrement de la structure de données fournies dans le fichier de données
3. Utiliser quand nécessaire la programmation orientée objet pour encapsuler vos données et traitements.
4. Favoriser l'utilisation des expression régulières pour éviter les données malformées.
5. Utiliser les dictionnaires pour représenter un tweet. N'hésitez pas à re-modéliser le dictionnaire représentant un tweet de manière à mieux répondre aux opérations d'analyse de données demandées
6. Utiliser le module *matplotlib* pour visualier les résultats des opérations d'analyse de données

Modalités :

- Vous pouvez travailler en groupe de deux étudiant maximum
- Soumettez sur l'espace dédié sur e-campus votre compte rendu, un Jupyter Notebook y compris les explications nécessaires.
- L'évaluation se fait sur la base de votre compte rendu et par question/réponse par votre chargé de TD durant les deux dernières séances de TD.