Sorbonne Data Analytics : Projet Text Mining

Suivi du #Ukraine

Description du projet

L'objectif de ce projet est de produire une analyse des tweets en français sur la guerre russo-ukrainienne de 2022. Dans ce projet vous devrez produire une analyse descriptive des tweets collectés et utiliser des modèles de Machine Learning pour essayer de faire émerger les grandes thématiques évoquées et suivre leur évolution.

Vous pouvez réaliser ce projet par groupe de 4 au maximum.

Les étapes du projet

Pour ce projet nous vous recommandons de suivre les étapes suivantes :

1. L'extraction des tweets – 10 %

Ci-dessous les quelques règles pour l'extraction :

- **La période de collecte :** février 2022 à aujourd'hui
- **La langue :** collecter uniquement les tweets en français
- **Le périmètre :** collecter les tweets contenant le #Ukraine

La collecte des tweets devra se faire en Python, vous pouvez utiliser l'API v2 de twitter pour des motifs académiques. Vous pouvez vous référer aux articles ci-dessous :

- https://datascienceparichay.com/article/python-get-data-from-twitter-api-v2/
- https://towardsdatascience.com/an-extensive-guide-to-collecting-tweets-from-twitter-api-v2-for-academic-research-using-python-3-518fcb71df2a
- https://developer.twitter.com/en/docs/twitter-api/getting-started/about-twitter-api

2. Analyse exploratoire – 30 %

Pour ce projet vous devrez mettre en place une analyse exploratoire des données.

- Vous devrez commencer par **affiner votre périmètre** : conserver ou non les retweets (il faudra justifier votre choix). Vous pouvez également vérifier et exclure les quelques tweets qui auront été collectés en anglais malgré vos filtres (çela arrive fréquemment)
- Il vous faudra ensuite **monitorer l'évolution du #Ukraine** au cours du temps : nombre de tweets (line plot) et être capable de l'expliquer en fonction de vos connaissances du sujets (date clés, etc…)
- Faire une **analyse des #** présents dans les tweets (comptage, nuages de mots, …). Analyser les # permet d'avoir une vue d'ensemble des sujets évoqués.
- Vous pourrez reproduire les analyses ci-dessus uniquement sur le périmètre **des comptes officiels** (par exemple ceux du gouvernement)

3. Le pré processing - 60 %

Une des **étapes clé** de votre projet est le pre processing de vos tweets. Pour ce faire vous devrez mobiliser les techniques vues en cours : tokenisation, stemmatisation, gestion des stop words, …

Il est attendu que vous produisiez deux versions normalisées de vos tweets :

- Une version **stemmatisée**
- Une version **lemmatisée**

Vous arbitrerez entre ces deux versions en fonction de vos objectifs (visualisation, modélisation, …)

BONUS : Vous pouvez analyser les proximités entre les mots de votre corpus en utilisant les embeddings (plongements de mots) et les visualiser à l'aide d'un TSNE

Ressources :

- https://radimrehurek.com/gensim/models/word2vec.html
- https://arxiv.org/pdf/1301.3781.pdf
- https://www.kaggle.com/code/jeffd23/visualizing-word-vectors-with-t-sne/notebook
- https://towardsdatascience.com/t-sne-clearly-explained-d84c537f53a

4. La modélisation – 80 %

L'objectif de ce projet est d'utiliser des techniques de NLP pour faire émerger les sujets présents dans votre corpus de manière non supervisée, on parle alors de Topic Modeling.

- La première étape est de **vectoriser vos documents**. Une approche de type Bag Of Word est recommandée (CountVectorizer, TfidfVectorizer). Une attention particulière doit être apportée au choix des hyper paramètres pour contrôler la taille et la qualité de vos représentations.
- Vous devrez ensuite **utiliser l'algorithme LDA** (Latent Dirichlet Allocation) pour faire émerger vos topics. C'est un algorithme non supervisé, ce sera donc à vous de définir le nombre de sujets. Plusieurs modules de Python peuvent être utilisés : *Scikit-Learn, Gensim, Tomotopy*.
- **BONUS :** Une approche complémentaire peut également être utilisée pour comparer les résultats en utilisant une ACP pour réduire la dimension des Bag Of Word suivi d'un Kmeans.
- **Il est attendu une analyse fine des résultats.** Quel que soit l'algorithme utilisé, vous devrez analyser les sujets retenus. C'est-à-dire leur affecter un nom en fonction de leurs mots les plus discriminants et les affecter à chacun de vos tweets. Ainsi vous pourrez analyser la répartition de vos tweets par sujet et reprendre une partie de votre analyse exploratoire pour les différents sujets (répartition dans le temps, nuages de mots, …)

Remarque : Lorsque vous utilisez des modèles non supervisés, il est parfois compliqué d'évaluer les résultats obtenus, vous devrez donc avoir un regard critique sur vos sujets.

Astuce : Pour la LDA, Une bonne technique pour apprécier la qualité de vos sujets est d'affecter le sujet majoritaire pour chacun de vos tweets, de trier par

contribution décroissante et d'analyser à quelle moment la pertinence de vos sujets se dégrade.

Ressources :

- https://towardsdatascience.com/light-on-math-machine-learning-intuitive-guide-to-latent-dirichlet-allocation-437c81220158
- https://bab2min.github.io/tomotopy/v0.12.1/en/#what-is-tomotopy
- https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.LatentDirichletAllocation.html

5. BONUS : Création d'un Dashboard de restitution – 100 %

Pour ce projet, en bonus vous pouvez **créer un Dashboard interactif** pour restituer les résultats de vos analyses.

Vous pouvez construire votre tableau de bord à l'aide du module *Streamlit* et le déployer gratuitement

Pour cette partie laissez parler votre imagination !

Ressources : https://streamlit.io/

Les rendus

- Vous avez jusqu'au dimanche **31 octobre minuit** pour rendre le projet
- Vous devrez nous remettre **le notebook commenté contenant vos analyses exploratoires et vos modélisations** ainsi que la version **html (avec le code exécuté et sans erreur)** du notebook
- Le lien vers votre Dashboard, qui devra rester accessible tout au long du SDA
- La semaine suivant le rendu sera consacrée à la préparation de votre soutenance (création de vos slides)