Contexte

Mentionné durant le cours, le **Topic Modeling** est **un sous-champ du Natural Language Processing** visant à **extraire les sujets de discussions principaux** d'un corpus de texte. On appelle un sujet de discussion **un groupe de mots ou un groupe de phrases** partageant une thématique spécifique.

Cette problématique est donc un sujet **non-supervisé** qui s'apparente à du **clustering**. De fait, de nombreuses méthodes de clustering existent – K-means, DBScan, Agglomerative clustering, etc. Cependant les données textuelles possèdent des particularités qui ont résulté en la création d'algorithmes spécifiques tirant profit de ces dernières.

Il est possible d'appliquer des méthodologies de Topic Modeling à n'importe quelle source de textes : commentaires postés sur les réseaux sociaux, articles scientifiques, pages Wikipédia, review de produits, etc. Le type de corpus analysé conditionne fortement la méthodologie à utiliser (vous verrez probablement que plus les textes sont courts, plus la tâche devient ardue).

Dans le cadre de ce projet, vous travaillerez sur le jeu de données ***News Category Dataset*** utilisé dans le TD NLP #2 - Data Pipeline, contenant 200K headlines de journaux en anglais.

Sujet

L'objectif de ce projet est d'étudier les performances d'algorithmes de Topic Modeling sur les données mentionnées ci-avant. Pour ce faire, vous devrez :

1. Analyser le corpus de texte pour en extraire ses caractéristiques spécifiques (taille moyenne, types de mots utilisés, mots les plus fréquents, stopwords, etc.).
2. Sélectionner **3 méthodologies de Topic Modeling / Clustering** vous semblant en phase avec les données à traiter
3. Définir une ou plusieurs métriques permettant de mesurer la qualité de vos modèles
    a. Note : vous pouvez utiliser les catégories des headlines fournies dans le jeu de données. Cependant, il s'agit ici d'un problème de clustering avant tout, il faut donc que vous mettiez également en avant des métriques prenant en compte cela.
4. Réaliser les tests comparatifs de chacun des modèles que vous avez sélectionné
5. Conclure sur la meilleure méthodologie à utiliser dans votre cas et préciser les pistes d'améliorations de votre analyse

Pensez à justifier vos choix !

Modalités de rendu

- **Taille des équipes :** 3 personnes
- **Format de rendu :** Notebook Jupyter présentant les résultats de l'étude
    - Import à mettre dans la première cellule du projet
    - A déposer sur Sharepoint à ce lien
    - **Nom du fichier :** Nom1_Nom2_Nom3_ProjectTopicModeling.ipynb

- **Critères d'évaluation :**
    - Qualité de l'étude des caractéristiques du corpus et de la sélection des méthodologies à tester /5
    - Qualité des métriques sélectionnées ou/et créées /5
    - Qualité de l'analyse des différentes méthodologies /5
    - Qualité de la conclusion finale /5

- **Date de rendu** : 26/11/2021 -> (si cela pose problème vis-à-vis de vos examens, faites le moi savoir)

Ressources

Vous trouverez dans le dossier Teams lié à ce cours trois documents pouvant vous aider pour ce projet, à savoir :

- La version finale du cours auquel vous avez assisté, disponible ici
- Une brève présentation sur le Topic Modeling présentant différents types de méthodologies, disponible ici
- Le TD de NLP Data Pipeline corrigé qui vous servira de base pour ce projet, disponible ici

Pourriez-vous s'il vous plaît ajouter les équipes à l'excel disponible à ce lien une fois celles-ci établies ?

Si vous avez la moindre question n'hésitez pas à me contacter par retour de mail !