Projet fouille de données textuelles

1) Description du projet

Objectifs :

- choisir une problématique en lien avec la fouille de textes,
- récupérer un jeu de données,
- pré-traiter ces textes en y appliquant différentes méthodes d'analyse linguistique telles que celles abordées en cours pendant le semestre,
- proposer des représentations textuelles plus efficaces,
- choisir une méthode de fouille permettant d'exploiter ces représentations pour en extraire des connaissances.

Détails :

Choisir une problématique liée à la fouille de textes : trouver un cas d'étude où le traitement des données textuelles peut s'avérer pertinent. Vous pourrez vous inspirer pour cela des exemples présentés en cours ou proposés à la fin de ce document.
Choisir et récupérer un jeu de données : les données choisies devront nécessairement être textuelles. Plusieurs caractéristiques vous permettent de justifier le choix de vos données, comme notamment : leur volume, leur complexité (e.g. textes courts issus de des réseaux sociaux, textes mal orthographiés (e.g. corpus de sms), la pertinence de leur contenu par rapport à la tâche finale,…). Vous pouvez utiliser un jeu de données déjà existants ou si cela est possible créer le vôtre (en utilisant par exemple des techniques de scrapping). Dans tous les cas, il sera nécessaire de mettre en œuvre des programmes facilitant leur récupération, leur nettoyage ou leur formatage. Vous pourrez utiliser pour cela le langage des expressions régulières.
Pré-traiter ces données à l'aide de méthodes linguistiques : déterminer parmi toutes les notions présentées en cours, les traitements linguistiques les plus appropriés pour obtenir des représentations qui soient pertinentes : *i.e.* les plus à même de représenter le contenu des documents. Exemples de traitements linguistiques à appliquer : lemmatisation, extraction de certaines catégories grammaticales (noms, verbes, etc.), extraction de bi-grams ou de tri-grams, de groupes nominaux, d'entités nommées, suppression des informations inutiles, etc.
Appliquer une méthode de fouille de textes sur ces données : vous pouvez utiliser une méthode de fouille déjà implémentée en Python (voir par exemple le module Scikit-Learn pour les méthodes de machine learning classiques ou le module Keras pour les réseaux de neurones) sur la ou les représentation(s) obtenue(s) à l'étape précédente pour exploiter ces données et extraire des connaissances qui pourront s'avérer utiles en vue de la problématique traitée. Exemples de tâches possibles : classification (supervisée ou non), clustering : identification de thématiques ou de tendances (LDA, topic modeling, etc.), analyse de sentiments ou de polarité, moteur de recherche, etc.

Documents et fichiers attendus :

En complément de vos programmes et fichiers de données (jeu de données initial, fichiers de sortie), rédiger un document synthétisant le travail fourni dans le cadre de ce projet. Pour chaque étape du projet (choix de la problématique, des données, des traitements opérés et de la méthode de fouille appliquée), vous justifierez vos choix, et présenterez les principaux problèmes rencontrés et les solutions mises en œuvre pour les contourner. Enfin, vous pourrez en guise de conclusion donner votre ressenti sur l'intérêt des méthodes d'analyse textuelle dans le cadre d'un projet de fouille de données.

Organisation : travail en binôme
Date de remise : fin du semestre – **vendredi 20/12/2024 (20h)**

2) Techniques de fouille de données textuelles : exemples d'application et pistes de lectures

Exemple détaillé vu en cours (TD5) :

- Classification multi-classes de tweets selon leur polarité

Exemples de projets d'étudiants des années précédentes :

- Analyse de tweets extraits d'un journal en ligne et identification des tendances
- Analyse de polarité dans des articles de presse
- Classification (thématique) des différents discours présidentiels
- Classification non supervisée de documents issus d'un corpus de SMS
- Classification d'avis de clients (positifs/négatifs) : films, livres, restaurants, voyages, vêtements, *etc*.
- Clustering d'articles de presse et identification thématique
- Détection de spams
- Analyse de sentiments sur les avis clients d'un produit de cosmétique
- Analyse automatique de recettes de cuisine (prédiction de la difficulté), classification automatique de recettes de cuisine (par catégories, *etc*.)
- Moteur de recherche : à partir de recettes de cuisine (trouver les recettes à partir d'ingrédients)
- Moteur de recherche dans une collections de descriptifs de films
- Système de recommandation musicale : exploitation des paroles pour des suggestions personnalisées
- Fouille d'offres d'emploi
- *etc.*

Exemples d'applications et pistes de lectures :

- **classification** (apprentissage supervisé) :
  - détection de spams :
    - http://radimrehurek.com/data_science_python/
  - analyse (élémentaire) d'opinions (avis positif ou négatif) ou de sentiments :
    - https://larevueia.fr/nlp-avec-python-analyse-de-sentiments-sur-twitter/
    - https://www.academia.edu/14632498/text_mining_sentiment_analysis_Fr

- ***clustering*** (non supervisé) : identification de thèmes / topics / catégories :
  - https://lilbigdataboy.wordpress.com/2016/01/12/les-emils-de-sarah-palin-extraction-de-topics-avec-python/
  - http://brandonrose.org/clustering

- **identification de tendances, évolutions des mots et de leur usage dans le temps**
  - http://adilmoujahid.com/posts/2014/07/twitter-analytics/

- **autres applications** : résumé automatique (nuage de mots), moteur de recherche, extraction d'informations (détection de noms de produits, d'organisations …), comparaison de documents (*e.g.* analyse de discours) :

- *Etc.*

Exemples de données :

- Lexique d'émotions et de sentiments : http://advanse.lirmm.fr/feel.php
- Corpus de SMS (88000 messages en français). Description : http://88milsms.huma-num.fr
- Critique de films (anglais) : http://www.cs.cornell.edu/People/pabo/movie%2Dreview%2Ddata/ ou https://www.themoviedb.org