



Projet NLP Text Mining

Analyse des avis [Trip Advisor](#)
concernant des restaurants lyonnais

Objectif du projet (1)

- Choisir une sélection de restaurants lyonnais à traiter sur TripAdvisor (ex. [Brasserie Georges](#)) (ex. une dizaine [?] avec un nombre d'avis conséquent)
- On souhaite récupérer les informations sous ses différentes dimensions (localisation, présentation, services, note(s) globale(s), ...).
- Et récupérer les avis laissés par les clients (date, note, etc.). En ciblant en particulier le corpus des commentaires (avis).
- L'ensemble des informations sont parsées, structurées et stockées dans un entrepôt de données.
- Deux niveaux d'analyse : [intra-restaurant](#) et [inter-restaurants](#), en intégrant une dimension géographique dans l'étude, peut-être croisées avec des données « open data » (ex. situation géographique, nombre de restaurants dans le périmètre, transports / parking, informations socio-économiques...)

Objectif du projet (2)

- Analyse des commentaires (techniques de NLP), mise en relation avec les caractéristiques des restaurants, avec les notes, etc.
- Analyse comparative des restaurants, sous ses différents aspects dont les commentaires déposés par les clients.
- Créez une application web interactive (!) PYTHON / DASH, BOKEH, STREAMLIT, ou autre, destinée à guider l'exploration et l'analyse du corpus.
-  On doit pouvoir ajouter dynamiquement un nouveau restaurant à étudier, dont les informations sont automatiquement stockées dans la B.D. sans intervention manuelle.
-  Toutes autres fonctionnalités qui vous paraissent pertinentes (ex. résumé automatique des avis, etc.). A vous de voir tant que vous le justifiez.

Spécifications techniques

- Les informations seront « aspirées » à l'aide de techniques de « web scraping » (pas manuellement donc) (ex. BeautifulSoup, Selenium, etc.). Complétées avec des API ? Les sources et la procédure utilisée doivent être décrites en détail dans le rapport.
- Cette procédure est destinée à alimenter une base de données, laquelle doit être modélisée sous la forme d'un entrepôt (table de faits, dimensions). La base est stockée dans un SGBD libre (ex. [MySQL](#), [SQLite](#) [[ne nécessite pas un serveur](#)], etc.).
- L'application PYTHON doit s'articuler directement sur la base de données.
- L'analyse doit intégrer une dimension géographique, on s'attend à voir des représentations cartographiques interactives dans l'application.
- L'application doit être aussi dynamique que possible, les graphiques interactifs (ex. plotly, etc.) seront appréciés (pas de pages avec des graphiques statiques).
- Le tout doit être déployé via une image « docker ». L'utilisateur doit seulement avoir à récupérer l'image et lancer le conteneur (via « Docker Desktop pour Windows » par ex.)

A rendre

- Transcrire la démarche et les conclusions dans un rapport PDF. Il doit être rédigé en LaTeX. Il décrit les problématiques mises en place, les stratégies de scrapping et d'extraction d'informations, la structure de la base, l'architecture de l'application, et les principales fonctionnalités et analyses proposées. Ainsi que les conclusions que l'on peut en tirer.
- Un **tutoriel vidéo commenté** en deux parties (ou 2 vidéos). **(1)** Procédure d'installation de l'application et son démarrage **(2)** Montrer et commenter les différentes fonctionnalités de l'application et les analyses qui en découlent.
- A mettre sur un drive : le rapport en PDF, le corpus utilisé (la base de données, avec des indications pour y accéder hors ligne), tout le code source Python utilisé, en particulier celui de l'application, l'image docker prête à déployer, la (les) vidéo(s).

Critères d'évaluation

- Qualité et clarté du rapport
- Intérêt des problématiques développées
- Pertinence des analyses et des résultats
- Qualité, interactivité, dynamisme de l'application visuelle interactive
- Qualité et organisation du code Python, architecture de la base
- Une soutenance est prévue (45 mn) : présentation, démonstration de l'application, questions-réponses.
- A réaliser en groupes de 4 ou 3 étudiants (à tirer au sort).

Calendrier

- Diffusion du sujet : mardi 26 novembre 2024
- Retour attendu : dimanche 12 janvier 2025 au soir
- Soutenance : semaine du 20 janvier 2025 (à préciser)
- Mettre votre travail (le tout) sur un drive. M'envoyer le lien à :
 - ricco.rakotomalala@univ-lyon2.fr
 - Sujet : [SISE – NLP Text Mining] Noms des étudiants