Projet Text Mining

Analyse régionale des offres d'emploi

Objectif du projet

- Choisir un corpus à traiter, extrait des sites d'offres d'emploi accessibles en ligne (ex. emploipublic.fr, emploi-territorial.fr, APEC, indeed, etc.)
- On souhaite analyse le corps (texte) des annonces. Mettez le focus sur un type d'emploi ou de compétences.
- Transcrire la démarche, la structure de la base de données, l'architecture du programme, et les résultats dans un rapport.
- Une application web interactive (!) R-Shiny destinée à guider l'exploration et l'analyse du corpus est demandée.
- Une soutenance est prévue : présentation, démonstration de l'application, questions-réponses.
- A réaliser en groupes de 3 étudiants (à tirer au sort).

Spécifications techniques

- Autant que possible, les offres d'emplois seront « aspirées » à l'aide de techniques de « web scraping » (pas manuellement donc) (ex. rvest, etc.). 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 (faits, dimensions). La base est stockée dans un SGBD libre (ex. MySQL, SQLite, etc.).
- L'application R-Shiny doit s'articuler directement sur la base de données.
- L'analyse doit intégrer une dimension régionale / territoriale, 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.

A rendre

- Un rapport au format PDF. Il doit être rédigé en LaTeX.
- Il décrit la nature du corpus utilisé, les problématiques mises en place, transcrire la démarche, les stratégies d'extraction des données, la structure de la base, l'architecture de l'application, et les principales analyses proposées. Ainsi que les conclusions que l'on peut en tirer.
- Un tutoriel vidéo en deux parties. **(1)** Décrire la procédure d'installation du SGBD, l'importation des données, l'installation de l'application R-Shiny. Je dois pouvoir lancer moi-même l'application d'après les éléments que vous me fournirez ! **(2)** Montrer et commenter les différentes fonctionnalités de l'application R-Shiny, et les analyses qui en découlent.
- A mettre sur un drive : le rapport en PDF, le corpus utilisé (la base), le code source R de l'application Shiny, les fichiers d'installation (encore mieux si c'est un SETUP – script ou autre…).

Critères d’évaluation

- Qualité et clarté du rapport
- Intérêt du corpus et des problématiques développées
- Pertinence des analyses et des résultats

- Qualité, interactivité, dynamisme de l’application Shiny

- Qualité et organisation du code R, architecture de la base

Calendrier

- Diffusion du sujet : jeudi 16 décembre 2021
- Retour attendu : lundi 31 janvier 2022 au soir
- Soutenance : semaine du 7 février 2022

- Mettre votre travail (le tout) sur un drive. M'envoyer le lien à :
  - ricco.rakotomalala@univ-lyon2.fr
  - Sujet : [SISE – Text Mining] Noms des étudiants