M2 Data mining – Année 2021-2022
Big data management
Projet Lac de données
J. Darmont (https://eric.univ-lyon2.fr/jdarmont/), 03/02/22

Objectif : Mettre en œuvre un mini-lac de données textuelles

1. Données

- Choisir un corpus de textes, par exemple sur la plateforme Kaggle ou encore dans la liste de corpus de Wikipedia. Si des données ou des métadonnées structurées ou semi-structurées y sont associées, les inclure dans le lac.
- Prévoir un mode de stockage pour les données (système de fichiers d'un ordinateur, HDFS…) et les stocker effectivement.

2. Métadonnées

- Définir un modèle de métadonnées simple (métadonnées intra et inter-objets ou goldMEDAL, par exemple) qui vous sera nécessaire pour interroger le lac de données textuelles.
- Prévoir un mode de stockage pour les métadonnées, par exemple Apache Atlas, Neo4J et/ou MongoDB.
- Instancier les métadonnées dans les outils de stockages définis à l'étape précédente. Apache Tika peut aider pour les métadonnées intra-objet.
- Les métadonnées globales doivent au moins inclure un index inversé des termes des documents textuels. Il peut être généré par exemple à l'aide d'Elasticsearch ou de Solr.

3. Analyses

- Nettoyer/transformer les données si nécessaire. Inclure les résultats et les transformations effectuées dans les métadonnées.
- Proposez des analyses relatives aux données textuelles du lac. Les outils de BI utilisés en TD peuvent être de nouveau employés, ainsi que tous les outils utiles en data science.

4. Rendu

Rapport synthétique :

- Introduction/présentation des données
- Description des étapes 1 à 3
- Conclusion, problèmes rencontrés, perspectives
- Code en annexe si nécessaire

Rapport à rendre à jerome.darmont@univ-lyon2.fr

- Non-alternante·s : 13/02/2022
- Alternant·es : 04/04/2022

Possibilité de travailler en binôme dans les deux cas