Projet Big data: Analyse de Sentiment dans une Bibliothèque Numérique

Modalités:

Par groupe de 4, vous devrez analyser le problème avant de vous répartir les tâches et de vous lancer dans les développements.

Contexte :

L'essor des bibliothèques numériques a ouvert de nouvelles possibilités pour l'analyse et l'exploitation de vastes ensembles de données textuelles. Dans ce contexte, notre projet se concentre sur l'analyse de sentiment dans une bibliothèque numérique. Nous travaillons avec une institution possédant une bibliothèque numérique contenant une multitude de textes numérisés, allant des livres aux articles de revues.

Objectif :

L'objectif principal de ce projet est de développer un système Big Data et d'Intelligence Artificielle pour analyser les textes numérisés de la bibliothèque, en mettant particulièrement l'accent sur l'analyse de sentiment. Nous visons à utiliser des technologies Big Data telles que Scala et Spark pour gérer le volume important de données textuelles et à mettre en œuvre des modèles d'IA capables de classifier le sentiment exprimé dans ces textes.

Déroulement du Projet :

1. Collecte des Données : Les textes numérisés seront déjà disponibles dans un format exploitable, provenant de la bibliothèque numérique. Cependant, pour assurer une efficacité maximale dans le traitement Big Data, une structuration adéquate des données peut être nécessaire. Cela pourrait impliquer la mise en place de pipelines de collecte pour organiser les données en fonction de critères tels que l'auteur, le titre, la date de publication, etc.
2. Prétraitement des Données : Le prétraitement des données est une étape critique pour garantir la qualité des données avant l'analyse. Cela pourrait inclure la normalisation du texte pour uniformiser les formats, la suppression des balises HTML pour ne conserver que le texte brut, la correction des erreurs de numérisation pour garantir l'intégrité des données, etc. Ces processus de nettoyage et de prétraitement sont essentiels pour garantir la fiabilité des résultats de l'analyse de sentiment ultérieure.
3. Analyse de Sentiment : Une fois les données prétraitées, l'analyse de sentiment peut commencer. Cette étape implique l'entraînement de modèles de classification de sentiment sur les textes prétraités. Ces modèles sont capables de déterminer si un texte exprime un sentiment positif, négatif ou

neutre. L'utilisation de techniques d'IA avancées peut être explorée pour améliorer la précision de cette classification.

4. Traitement Big Data : Scala et Spark seront les principaux outils utilisés pour le traitement parallèle des données textuelles numérisées. Spark offre des fonctionnalités puissantes pour le traitement distribué, ce qui permet de gérer efficacement le volume important de données de la bibliothèque numérique. Les étudiants devront développer des scripts Scala pour manipuler les données et les traiter en utilisant les fonctionnalités de Spark, telles que les transformations et les actions.

5. Intégration d'Intelligence Artificielle : Outre les modèles de classification de sentiment, l'intégration de techniques d'IA avancées peut être explorée pour améliorer la précision de l'analyse de sentiment. Cela peut inclure l'utilisation de réseaux neuronaux profonds ou de modèles de traitement du langage naturel pré-entraînés pour capturer des nuances subtiles dans le langage humain.

6. Visualisation des Résultats : Une fois l'analyse de sentiment effectuée, il est essentiel de fournir une interface utilisateur interactive pour visualiser les résultats de manière conviviale. Cela implique le développement d'une interface utilisateur front-end intégrée à un serveur back-end pour fournir les résultats de l'analyse. Les visualisations dynamiques telles que des graphiques, des nuages de mots et des diagrammes de tendances peuvent être utilisées pour aider les utilisateurs à explorer et à comprendre les données analysées.

Livrables :

- Code source de l'application, y compris les scripts Scala et Spark, ainsi que le code front-end et back-end.
- Présentation orale pour présenter les résultats et les découvertes aux autres groupes.
- Diagramme d'architecture des différents services du projet.

Ressources:

lien vers les livres :https://drive.google.com/file/d/16KCjV9z_FHm8LgZw05RSuk4EsAWPOP_z/view

Créer un projet de developpement scala: https://docs.scala-lang.org/scala3/book/tools-sbt.html

Introduction au sentiment analysis : https://towardsdatascience.com/sentiment-analysis-concept-analysis-and-applications-6c94d6f58c17

Spark MLLib: https://spark.apache.org/docs/latest/ml-guide.html Pyspark MLLib : https://spark.apache.org/docs/latest/api/python/reference/pyspark.mllib.html

Un exemple d'analyse de sentiment avec spark : https://towardsdatascience.com/sentiment-analysis-on-streaming-twitter-data-using-spark-structured-streaming-python-fc873684bfe3