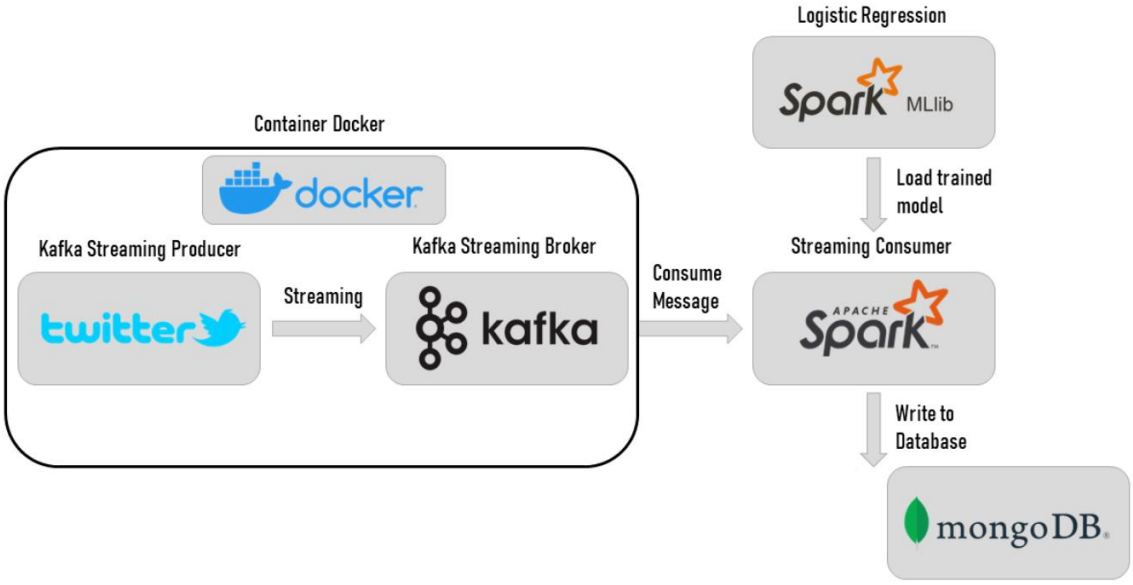
Rapport

Prédiction de sentiments en temps réel des flux de données de réseau social Twitter



Encadré par : Yasyn EL YUSUFI

Fait par : EL MEZIANE Chaïma
Ohmad Abdessamade
Zerbet Yasmine

Année Universitaire : 2023/2024

1. Introduction

Ce projet vise à développer une application Web basée sur Apache Kafka et Spark Streaming pour l'analyse des sentiments en temps réel des tweets. L'objectif est de prédire le sentiment (positif, négatif, neutre ou non pertinent) d'un tweet donné.

2. Architecture et Technologies Utilisées

a. Architecture

L'architecture du projet se compose des éléments suivants:

1. **Flux de données Twitter:** Les tweets sont collectés d'un fichier csv «twitter_validation.csv».
2. **Apache Kafka:** Kafka servira de plateforme de streaming pour traiter les tweets entrants. Les tweets sont publiés dans un topic Kafka nommé "tweets".
3. **Apache Spark Streaming:** Spark Streaming est utilisé pour traiter les tweets du topic Kafka. Le traitement a impliqué les étapes suivantes:
    - **Prétraitement:** Les tweets sont nettoyés et prétraités pour extraire des caractéristiques pertinentes.
    - **Entraînement du modèle:** Un modèle d'apprentissage automatique supervisé (Régression logistique) sera entraîné sur un ensemble de données de tweets étiquetés (sentiment connu).
    - **Prédiction:** Le modèle entraîné sera utilisé pour prédire le sentiment des nouveaux tweets.
    - **Enregistrement des résultats:** Les prédictions de sentiment seront enregistrées dans une base de données MongoDB.
4. **Application Web:** Une application Web est développée en utilisant Angular pour visualiser les données de test et les résultats de l'analyse des sentiments.

b. Outils et technologies

Les outils et technologies utilisés dans ce projet incluent:

- **Python:** Pour le développement des scripts de traitement des données, l'entraînement des modèles de machine learning et l'interaction avec les différentes technologies.
- **Docker:** Pour containeriser les différentes parties de l'application, spécialement la partie kafka et zookeeper, assurant une portabilité et une scalabilité aisées.
- **Apache Kafka:** Pour le streaming de données en temps réel.
- **Apache Spark (PySpark):** Pour le traitement des flux de données et l'entraînement des modèles de machine learning.
- **MongoDB:** Pour stocker les résultats des prédictions de sentiments.
- **Angular:** Pour développer l'interface utilisateur de l'application Web.
- **ExpressJS:** Pour créer l'API backend qui interagit avec MongoDB.
- **NLTK:** Pour le prétraitement des données textuelles (tokenization, filtrage des stop words, lemmatisation).
- **Bootstrap:** Pour le développement de l'interface utilisateur responsive et esthétique.
- **Matplotlib:** Pour la visualisation des données et des résultats d'analyse.

3. Implémentation

a. Spark et Entraînement des Modèles

Lecture des Données

- Les données sont chargées à partir du fichier « twitter_training.csv » en utilisant Pyspark.

- Une session Spark est créée pour gérer le traitement des données.

Prétraitement des Données

- Les données sont nettoyées et préparées pour l'analyse. Les étapes incluent la tokenization, le filtrage des stop words, et la lemmatisation avec NLTK.

- Ces étapes permettent de transformer le texte brut des tweets en un format utilisable pour l'entraînement des modèles de machine learning.

Sélection et Entraînement des Modèles

- Un modèle de machine learning supervisés tels que la régression logistique est utilisé.

- Les données prétraitées sont transformées en features, puis le modèle est entraîné sur la base « twitter_training.csv ».

Évaluation et Sauvegarde du Modèle

- Les modèles sont évalués pour sélectionner le plus performant, le meilleur modèle est enregistré sous format spark pour une utilisation ultérieure dans la prédiction en temps réel.

b. Kafka

Mise en Place des Brokers, Topics et Partitions

- Kafka est configuré avec les brokers, topics, et partitions nécessaires pour le traitement des flux de données Twitter à partir du fichier «twitter_validation.csv ».

- Les producteurs et consommateurs sont configurés pour publier et lire les messages des topics Kafka.

Utilisation de Kafka Streams

- Kafka Streams est utilisé pour lire les données de Twitter à partir du fichier «twitter_validation.csv ».

- Les données entrantes sont traitées en utilisant le modèle de machine learning pré-entraîné pour prédire les sentiments.

- Les résultats des prédictions sont sauvegardés dans une base de données MongoDB.

c. Front-End et le Back-End

Création de l'API

- Utilisation d'ExpressJS pour développer l'API backend qui gère les requêtes et les prédictions en temps réel.

- L'API interagit avec la base de données MongoDB pour stocker les résultats des prédictions.

Interface Utilisateur

- Développement de l'interface utilisateur en utilisant Angular et Bootstrap.

- Une landing page et un tableau de bord sont créés pour visualiser les résultats des analyses de sentiments de manière agréable et interactive.

- L'interface affiche les tweets, les prédictions de sentiments, et des graphiques de visualisation des données.

Intégration

- Le front-end est intégré au back-end pour permettre une interaction fluide entre l'utilisateur et l'application.

- Les résultats des prédictions en temps réel sont affichés de manière dynamique sur l'interface utilisateur.

4. Résultats et Discussion

Les résultats de ce projet montrent que l'utilisation de technologies de traitement des flux en temps réel et de machine learning est efficace pour l'analyse des sentiments sur les réseaux sociaux. Le modèle de régression logistique a montré des performances satisfaisantes dans la prédiction des sentiments des tweets. Les résultats des prédictions ont été visualisés en temps réel sur l'application Web, offrant une interface utilisateur interactive et informative. L'intégration réussie des différentes technologies a permis de créer une solution scalable pour l'analyse en temps réel des sentiments des tweets.

5. Conclusion

Ce projet démontre l'efficacité de l'utilisation de technologies de traitement des flux en temps réel et de machine learning pour l'analyse des sentiments sur les réseaux sociaux. L'application développée offre une solution complète et scalable pour la prédiction des sentiments des tweets en temps réel, ouvrant de nombreuses opportunités pour l'analyse des données sociales.

6. Référence

- Source de données : Twitter Sentiment Analysis (kaggle.com).
- Documentation des technologies utilisées : Apache Kafka, Pyspark, NLTK, MongoDB, Docker.