Mini-Projet: Real time stock market prediction using sentiment analysis on social network

1. Introduction

Les réseaux sociaux sont devenus un moyen incontournable de partage, d'échange d'information (croyance, comportement, activités etc) et d'influence en temps réel par excellence. Ceci a engendré la génération d'une grande quantité de données dont le stockage, l'analyse et le traitement avec les moyens classiques n'est pas adapté voire impossible.

Objectif du projet :

Le but de ce projet est de réaliser une application big data pour le traitement de données massive en temps réel pour la prévision d'actions boursières (stock market). Deux types de sources de données seront nécessaires :

un réseaux social (ex : Reddit) vu la grande quantité d'information générée de façon continue.

Une source de données financière (Y**ahoo Finance, Bloomberg, Google trend, etc.)**

Les points importants à inclure dans le projet sont :

1. Identifier l'action boursière à étudier. Prendre en considération les actions les populaires sur Reddit (posts, commentaires, scores, subreddit, etc.). (ex : r/wallstreetbets)
2. Répertorier les données dans une base de données nosql de votre choix
3. Procéder à une étude théorique ((MA, RSI, MACD, volatilité, etc.), des indicateurs utilisés dans la prévision des actions boursière (étude théorique)
4. Analyse de sentiments des tweets en utilisant des bibliothèques spécifiques (vader, textblob, etc).
    - Créer un nouveau indice comme agrégation (ou bien majority voting) des résultats obtenus de ces outils
    - créer « social influence index » sur la base de : polarity of tweets + infleunce of their creator
    - Fusionner les données boursiers et sociales pour entamer la phase d'apprentissage
4. Créer deux modèles de prévision :
    - en utilisant un algorithme deep learning (LSTM, etc)
    - le choix des modèles doit être justifié.
5. Créer dashboard (Reddit, action boursière) pour le suivi des informations financières et sociales pertinentes sur l'action boursière choisie

2. Outils de travail

Kafka, Spark et BD Nosql (Mongo DB), Mlflow, DVC, Airflow, Power BI, etc.

Orchestration : Airflow, Kafka, Spark Streams pour pipeline temps réel.

Agents / frameworks : LangChain Agents, workflows AutoML, ou agents personnalisés qui appellent notebooks / pipelines Mlflow.

Monitoring & ML ops : MLflow pour tracking, DVC pour versionning des données, alertes via Slack/email.

3. Étapes du projet

1. Collecte et acquisition de données
2. Nettoyage et prétraitement des données
3. Analyse de sentiments/ influencers
4. Feature extraction (sentiment/influence index)
5. Entraînement/évaluation des modèles
6. visualisation temps réel (différentes vues)

4. Livrables du projet :

• Rapport de 12 pages max (https://fr.overleaf.com/latex/templates/ieee-conference-template/grfzhhncsfqn)

• Présentation 15 min

• Code-source (github)

5. Consignes :

- Le travail d'équipe est fortement recommandé (pentanome max)
- La mission de chaque membre de l'équipe doit être bien identifiée et equilibrée par rapport au autres membre.
- Dresser un planning prévisionnel avec les taches nécessaires pour l'accomplissement de projets avec le responsable.
- Décomposer les taches « separation of concern » pour favoriser le travail en parallèle
- Réaliser des documents d'interfaçage pour se mettre d'accord sur les points d'intersections entre les tâches
- Utiliser un outils (ex. Gira, trello, etc.) pour le Suivi de réalisation de projet. (Data Driven Scrum)