Mini-Projet: Real time social network analysis

1. Introduction

Les plateformes sociales jouent aujourd'hui un rôle majeur dans la formation de l'opinion publique, en particulier sur des sujets géopolitiques sensibles. Reddit, avec son fonctionnement basé sur des communautés thématiques (subreddits), constitue un espace privilégié pour analyser des discussions, opinions, débats et interactions en ligne.

Objectif du projet :

Ce projet consiste à développer une application Big Data capable de collecter, traiter et analyser en temps réel des données provenant de Reddit, en se concentrant sur les discussions liées au conflit Maroc–Polisario autour du Sahara.

L'objectif est exclusivement analytique et technique : étudier comment les utilisateurs interagissent, quels sujets émergent, quelles communautés se forment et quelles dynamiques influencent les conversations.

Les taches majeures à inclure dans le projet sont :

1. **Social network analysis :**
   - Identifier les principaux acteurs et communautés discutant du sujet sur Reddit
   - Extraire et modéliser les relations entre utilisateurs (réponses, citations, mentions)
   - Visualiser les graphes de réseaux
   - Analyser l'influence et les différents types de centralité
   - Détecter les communautés au sein du réseau (clusters, sous-groupes)
2. **Analyse de posts et commentaires**
   - Topic modeling : extraire automatiquement les thèmes dominants
   - Sentiment analysis : classifier la tonalité (positif, neutre, négatif) en créant votre propre modèle ML
   - User engagement analysis : votes, fréquence, récence, participation
   - Geospatial analysis (si applicable à partir des métadonnées non personnelles)
3. **Analyse de sentiments**

- L'analyse des sentiments doit être réalisée via un modèle à entrainer. Une comparaison peut être fait avec des modèles pré-entrainés ou une bibliothèques connues (VADER, TextBlob, etc)

4. **Dashboard et visualisation**

- Créer un dashboard interactif permettant :
- la visualisation temps réel (ou quasi temps réel) des données Reddit
- les graphes de réseau
- les thèmes dominants

- les indicateurs d'engagement
- les tendances temporelles

2. Outils de travail

Kafka : streaming temps réel

Spark / Spark Streaming : traitement massivement parallèle

NoSQL : MongoDB

Airflow : orchestration

Mongo atlas Visualisations /PowerBI/ Streamlit : dashboard

Pushshift API / Reddit API (PRAW) : collecte des données

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

• Présentation 15 min

• Code-source (github)

5. Consignes :

- Le travail d'équipe est fortement sollicité (quadrinome max)
- Chaque membre de l'équipe doit avoir une mission clairement définie et équilibrée par rapport au autres membres.
- Élaborer un planning prévisionnel comprenant les phases nécessaires à la réalisation du projet.
- Décomposer les taches « separation of concern » afin de favoriser le travail simultané
- Élaborer des documents d'interfaçage pour définir clairement les points d'intersections entre les tâches/membres
- Utiliser un outil de gestion de projet telque Gira, trello pour suivre la réalisation du projet selon la méthodologie Data Driven Scrum.