HOMEWORK 3 – BIG DATA PROJECT PROPOSAL

POOJA THAKUR ~*put2003*

ROHAN GORE ~*rmg9725*

Real-Time YouTube Trend Detection System

Introduction

The Real-Time YouTube Trend Detection System is a big data-oriented project designed to **identify emerging video trends** on YouTube within 24–48 hours of their inception. By leveraging APIs such as **YouTube Data API and Invidious API**, the system processes massive volumes of real-time data, including video metadata, comments, and engagement metrics, to detect trends early. It incorporates sentiment analysis to understand audience reactions and provides actionable insights for content creators, marketers, and analysts.

Suitability as a Big Data Project

This project qualifies as a big data initiative due to its alignment with the core characteristics of big data:

1. **Volume**: Processes large datasets from millions of videos daily, including metadata, comments, and engagement metrics.
2. **Velocity**: Handles real-time streams of data from APIs, ensuring insights are generated within minutes of data collection.
3. **Variety**: Integrates structured (video metadata), semi-structured (comments), and unstructured (thumbnails) data formats.
4. **Veracity**: Employs sentiment analysis to filter out noise and ensure the reliability of detected trends.
5. **Value**: Provides actionable insights for businesses and creators to optimize content strategies and marketing campaigns.

Technologies and Tech Stack

Layer	Technology/Tool	Purpose
Data Ingestion	YouTube Data API v3	Fetch metadata, comments, and engagement metrics
	Invidious API	Access trending videos without rate limits
Data Processing	Apache Spark	Batch and stream processing of large datasets
	Python (pandas)	Lightweight data manipulation
Storage	MongoDB	Store real-time processed data
	AWS S3	Archive historical data
Machine Learning	Hugging Face Transformers	Sentiment analysis on comments
Visualization	React + D3.js	Interactive dashboards for trend insights
Deployment	AWS Lambda + API Gateway	Cost-efficient serverless backend

4-Week Development Roadmap (40 Hours Total)

Week 1: Data Pipeline Setup (10 Hours)

- Configure YouTube Data API and Invidious API for data ingestion.
- Set up Apache Spark for batch and stream processing pipelines.
- Implement MongoDB for storing processed video metadata.

Week 2: Core Analytics Development (10 Hours)

- Develop trend detection logic using TF-IDF vectorization for video titles and tags.
- Implement time-decay scoring formula for identifying emerging trends:$\text{Trend Score} = \frac{\text{Views}}{\log\left(\text{Hours Since Upload} + 1\right)} \times \text{Sentiment Weight}$
- Integrate Hugging Face models for sentiment analysis on comments.

Week 3: Visualization & Alerting (10 Hours)

- Build a React-based dashboard with D3.js visualizations for trend timelines and heatmaps.
- Add alerting mechanisms via email or SMS when high-impact trends are detected.

Week 4: Deployment & Optimization (10 Hours)

- Deploy serverless backend on AWS Lambda with API Gateway for scalable access.
- Optimize Spark jobs for faster processing using caching and partitioning techniques.

Conclusion

The Real-Time YouTube Trend Detection System is an exemplary big data project that addresses the growing need for early trend identification in digital content ecosystems. By leveraging scalable technologies like Apache Spark, MongoDB, and AWS Lambda, this project ensures efficient handling of high-volume, high-velocity data streams while delivering actionable insights to users in real time.

⁂