Social Media Trend Clustering

Tis project is about clustering trending topics from Twitter, Facebook, Instagram, and Reddit using machine learning models.

Project Team

Our dedicated team is comprised of:

- Istiaque Ahmed (ID: 1821433042)
- Dipanwita Mallick (ID: 201164642)
- S. Anam Ridwan Shanto (ID: 2022563042)
- Atkia Sadia Anika (ID: 2121860642)

Objective & Key Tasks

The primary objective of this project is to cluster trending topics from Twitter, Facebook, Instagram, and Reddit using machine learning models. This involves a systematic approach broken down into key tasks:

1 Data Scraping

Extract raw data from social media platforms using APIs and web scraping techniques.

2 Preprocessing

Clean and format the collected data to ensure it is suitable for analysis and model training.

3 Clustering

Apply machine learning models to identify and group similar trending topics.

Progress Update: Data Collection

Our data collection phase is underway, utilizing various sources and tools to gather comprehensive social media data.

Sources:

- Twitter (API)
- Facebook (web scraping)
- Instagram (APIs)
- Reddit (APIs)

Tools:

- **requests:** For fetching web page content.
- **BeautifulSoup:** For parsing HTML/XML data.
- **pandas:** For storing and managing datasets.

Tools & Technologies

We leverage a suite of powerful tools and technologies to execute this project effectively.

Python

Primary language for machine learning and data scraping.

Tweepy

Efficient access to the Twitter API for data retrieval.

NumPy

Supports numerical operations essential for data processing.

GitHub

Used for collaborative code management and version control.

Project Link: GitHub Repository

Project Timeline (12-13 Weeks)

Our project follows a structured timeline to ensure timely completion and systematic progress.

1-2	Data Collection	Scrape/API-call data from all platforms.
3-4	Preprocessing & CSV Export	Clean text, remove noise, and save as CSV.
5-6	Exploratory Data Analysis (EDA)	Visualize trends and patterns in data.
7-8	Feature Engineering & Model Selection	Extract features (TF-IDF) and pick models.
9-10	Model Training & Tuning	Train clustering algorithms (e.g., K-Means).
11-12	Visualization & Trend Analysis	Generate insights and cluster visualizations.
13	Final Report/Presentation	Compile results and present findings.

Future Steps: Model Training & Tuning

Following data preprocessing and EDA, the next critical phase involves the development and refinement of our machine learning models. This stage is crucial for accurately identifying and clustering trending topics.

Feature Engineering

Extract meaningful features from the preprocessed text data, such as TF-IDF (Term Frequency-Inverse Document Frequency), to prepare it for machine learning algorithms.

Model Selection

Evaluate and select appropriate clustering models, such as K-Means, DBSCAN, or hierarchical clustering, based on the characteristics of our data and project objectives.

Model Training

Train the chosen clustering algorithms using the prepared dataset to enable them to identify inherent patterns and groupings within the social media trends.

Model Tuning

Optimize model parameters through techniques like cross-validation and grid search to enhance clustering performance and accuracy.

Conclusion & Next Steps

This project is on track to deliver a robust solution for clustering social media trends. The systematic approach from data collection to model training and visualization ensures comprehensive analysis.

We are committed to providing valuable insights into social media dynamics. For more details and to track our progress, please visit our GitHub repository.

Visit GitHub