iNeuron.ai

Project Report

Big Data Internship Report

Social media community using optimized clustering algorithm

Submitted by

Arbash Hussain

1. introduction

1.1 project background

in the era of social media, understanding user behavior is crucial for targeted marketing, improving user experience, and enhancing community engagement. this project focuses on clustering users in a social network based on their activity data using a big data approach. by applying clustering techniques, we can identify distinct groups of users with similar behaviors.

1.2. Objectives

- to generate and process synthetic user activity data for a social network.
- to perform clustering on users based on their activity patterns using big data tools and techniques.
- to visualize the resulting clusters to gain insights into user behavior.

1.3. Scope

the scope of this project includes data ingestion, transformation, clustering, and visualization. the project utilizes apache spark for handling large datasets, scikit-learn for clustering, and plotly for visualizations, with a web interface developed using flask and flask-socketio for real-time interactions.

2. Methodology

2.1 data ingestion

synthetic data is generated using the `faker` library. the data includes user ids, post ids, timestamps, content, and interaction counts. this data simulates user activities such as posting, viewing, and commenting.

- **tools used**: faker, apache spark
- **data attributes**:
  - `user_id`: unique identifier for each user.
  - `post_id`: unique identifier for each post.
  - `post_timestamp`: timestamp of post creation.
  - `comment_timestamp`: timestamp of comment creation.
  - `post_content`: content of the post.
  - `comment_content`: content of the comment.

  - `view_count`: number of views for the post.
  - `comment_count`: number of comments on the post.

2.2 data transformation

data is aggregated to calculate user activity metrics such as unique posts, total views, and total comments. features are scaled, and principal component analysis (pca) is applied to reduce dimensionality.

- **tools used**: apache spark, pca
- **transformation steps**:
  - aggregate data by `user_id`.
  - calculate `comment_to_view_ratio`.
  - scale features using `standardscaler`.
  - reduce dimensions with pca to extract two principal components.

2.3 model training

The k-means clustering algorithm is applied to group users based on their activity patterns. the elbow method is used to determine the optimal number of clusters.

- tools used: scikit-learn
- modeling steps:
  - calculate wcss for different numbers of clusters.
  - determine the optimal number of clusters using the elbow method.
  - train the k-means model and assign cluster labels to users.

2.4 visualization

The clustered data is visualized using plotly, creating an interactive scatter plot to display user clusters.

- tools used: plotly
- visualization steps:
  - plot pca features on a 2d scatter plot.
  - color code points based on cluster assignments.
  - export the plot as an interactive html file.

2.5. Web application

A web interface is created using flask and flask-socketio to run the pipeline and display results with real-time updates.

- tools used: flask, flask-socketio
- application features:
  - endpoint to initiate the pipeline.
  - real-time updates during pipeline execution.
  - display of clustering results.

3. Results

- **Data ingestion**: successfully generated and ingested synthetic data for 1,000 users and 10,000 posts.
- **Data transformation**: reduced dimensions from 4 features to 2 principal components.
- **Model training**: identified optimal clusters and assigned cluster labels to users.
- **Visualization**: created an interactive plot displaying user clusters.
- **Web application**: provided a functional interface for running and visualizing the pipeline.

4. Conclusion

The project successfully implemented a pipeline to cluster users in a social network based on their activity data. the use of big data tools like apache spark enabled efficient data processing, and the k-means algorithm effectively grouped users into meaningful clusters. the visualization provided insights into user behavior, which can be leveraged for various applications such as targeted marketing and user segmentation.

5. future work

- real data integration: apply the pipeline to real social network data for more accurate clustering results.
- advanced clustering techniques: explore other clustering algorithms like dbscan or hierarchical clustering for comparison.
- feature expansion: incorporate additional features such as user demographics and network connections.
- scalability: optimize the pipeline for larger datasets and real-time processing.

6. References

- https://spark.apache.org/docs/latest/
- https://scikit-learn.org/stable/documentation.html
- https://faker.readthedocs.io/en/master/
- https://plotly.com/python/
- https://flask.palletsprojects.com/en/2.0.x/
- https://flask-socketio.readthedocs.io/en/latest/