1. Introduction

The purpose of this project is to analyze social and political events worldwide using the GDELT dataset. The dataset includes global events from 1979 to 2013, containing information such as event location, event codes, and the number of articles reporting on each event. Our primary aim was to identify event hotspots using clustering techniques. By identifying regions with high concentrations of events during specific time periods, decision-makers can better understand historical global trends and take proactive measures in the present and future.

2. Methodology

Data Processing: Loaded and processed an extensive dataset of 87 million rows using Apache Spark, removing missing values. Sampled 20% of the data for efficiency and split it into 80% training and 20% testing subsets.

Feature Engineering and Standardization: Selected longitude, latitude, and the number of articles as features for clustering. Standardized the features using StandardScaler to ensure uniform contribution across features.

Clustering Model: Employed K-means clustering to identify event hotspots. Based on the Elbow Method, the number of clusters was chosen to be 4. Evaluated the model with the Silhouette score, achieving a value of 0.5717, which indicates moderately well-separated clusters.

Visualization: Visualized a sample of clustering results on a global map to show event distribution and identify high-activity areas.

3. Code Implementation

Spark Session Initialization: Created a SparkSession with enhanced memory configurations to handle the dataset.

Data Cleaning: Removed rows with missing values for consistency.

Feature Extraction and Standardization: Combined features into a vector using VectorAssembler. Standardized these features to maintain consistency in their scales.

K-means Clustering: Applied K-means clustering to group the data into 4 clusters. Evaluated using the Silhouette score to assess clustering performance.

Testing and Visualization: Predicted clusters on test data. Visualized a sample of the predicted clusters to identify geographical trends.

Model Saving: Saved the trained model to disk for reuse.

4. Results and Findings

Silhouette Score: The obtained score of 0.5717 indicates moderate separation among clusters. The results suggest that some clusters are well defined, while others might have a slight overlap.

Cluster Centers: The cluster centers indicate the geographical regions and levels of activity represented by each cluster.

Visualization: The clustering map reveals hotspots in North America, Europe, and Asia.
The red color on the map indicates high-activity clusters in certain regions.

Key Insights: The clustering revealed regions with increased event activities, such as the United States, Europe, and parts of Asia. The analysis shows that these regions had frequent social and political events, providing insight for understanding historical trends.