EAS 503LEC JLI: Prog DB Data Sci

TOPIC MODELLING USING AG NEWS DATASET

Presented by:
Sai Swapnesh Pahi
Akanksha Kale

INTRODUCTION

Context:

- The era of vast digital content demands recommendation systems.
- Many articles lack explicit genre labels, complicating classification and recommendations.

Objective:

- Use topic modeling to infer hidden genres and enhance recommendation systems.
- Classify articles into genres like World, Sports, Business, and Sci/Tech.
- Create a scalable solution that utilizes text embeddings and efficient classification algorithms

DATASET DESCRIPTION

Dataset: AG News Dataset

- Size: 120,000 training samples, 7,600 test samples
- Classes: World, Sports, Business, Science/Technology
- Key Variables- Text, label

Relevance

- Suitable for topic modeling and news categorization.
- Offers diverse text content for robust analysis.

DATA PREPROCESSING

- Data Import:
  Loaded dataset and configured DistilBERT for embedding generation.
- Database Structure:
  Created tables for raw data and embeddings.
  Embedded data linked to text and labels.
- Batch Processing:
  Efficient embedding extraction using GPU.
- Dimensionality Reduction:
  Applied t-SNE for 2D visualizations

SQL DATABASE DESIGN

- Data Insertion and Population: Inserted data into raw_data and populated embeddings_data with embeddings.
- Efficient Embedding Extraction: Used batch processing with GPU acceleration for fast embedding extraction.
- Data Retrieval and Visualization: Executed SQL queries to retrieve and merge data for dimensionality reduction and scatter plot visualization.

ANALYSIS PROCESS

- Data Preparation:
  Loaded and shuffled the AG News dataset
  Created tables: raw_data for text and labels, embeddings_data for embeddings
- Text Embeddings
  Used DistilBERT to generate 768-dimensional embeddings
  Processed data in batches with GPU acceleration
- Dimensionality Reduction & Visualization
  Applied t-SNE to reduce embeddings to 2D for cluster visualization
- Classification & Evaluation
  Classified using Euclidean and Cosine distances
  Evaluated with confusion matrices and accuracy scores
- Outlier Detection
  Identified outliers based on distance from centroids
- Logistic Regression & PCA
  Trained logistic regression model
  Reduced dimensions using PCA to balance accuracy and efficiency
- Insights from Dimensionality Reduction
  Analyzed clustering patterns and their impact on accuracy

DATA ANALYSIS AND INSIGHTS

Analysis Techniques:

- Generated 768-dimensional embeddings using DistilBERT.
- Reduced dimensionality using PCA and t-SNE.

Results:

- Logistic regression achieved 91.2% accuracy with 768 features.
- Reduced to 10 features via PCA with minimal accuracy loss.
- Clustering visualizations showed clear separations between categories.

INTERACTIVE DASHBOARD

Features:

- t-SNE Plot: Interactive 2D scatter plot showcasing clusters.
- Model Comparisons: Dropdown menu for models with confusion matrices and accuracy metrics.
- Dimensionality Accuracy Plot: Visualizes the trade-off between features and model accuracy.
- User Input: Predicts genres for inputted text using trained models.

Purpose:

- Enhances user understanding of classification and clustering.
- Engages users through interactive exploration.

CONCLUSION

Achievements:

- Successfully classified news articles into distinct genres.
- Demonstrated robust clustering and dimensionality reduction techniques.
- Developed an interactive dashboard for data exploration.

Future Work:

- Extend to domain-specific datasets (e.g., legal or medical).
- Improve interpretability using techniques like SHAP values.

THANK
YOU