PENS Dataset: Trajectory-Based KL Divergence for User and Document Representations

Student Name: Hemish Savaliya
Enrollment ID: **202001206**
B.Tech. Project (BTP) Report
BTP Mode: ***On Campus***
Dhirubhai Ambani Institute of ICT (DA-IICT)
Gandhinagar, India
202001206[at] daiict.ac.in

Mentor's Name: Prof. Surish Dasgupta
Dhirubhai Ambani Institute of ICT (DA-IICT)
Near Indroda Circle Gandhinagar 382007, India
Sourish_dasgupta[at] daiict.ac.in

Abstract- Abstract

This study leverages KL divergence to evaluate the alignment between user and document representations within a trajectory-based interaction dataset. The user representations were derived using Markov Chain models on action trajectories, while document representations were constructed from embeddings of generated summaries, aggregated and weighted based on trajectory lengths. Both representations underwent normalization to ensure valid probability distributions, enabling KL divergence computation for each trajectory.Through comprehensive analysis, key insights Divergence Patterns, Top and Least Divergent Cases , Visualizations were extracted

The workflow integrates dimensionality reduction (PCA), trajectory-specific weights, and probabilistic softmax transformations to capture both semantic and probabilistic nuances of user-document interactions. This robust pipeline offers a scalable framework for evaluating personalization and recommendation systems, providing actionable insights for enhancing alignment between user preferences and content delivery.

I. PROBLEM STATEMENT

The primary purpose of this project is to evaluate and enhance the alignment between user preferences and document content within a trajectory-based interaction dataset. By leveraging advanced techniques, including KL divergence, Markov Chains, and probabilistic embeddings, the project aims to:

- **Quantify Alignment:**
  Use KL divergence as a metric to measure how well document representations align with user behavior patterns, providing a data-driven approach to assess future implementations.
- **Category Diversity**: Quantify the variation in the type of content consumed by workers based on category embeddings.
- **Understand User Behavior:** Model user actions through Markov Chains to capture the sequential nature of interactions, enabling deeper insights into user preferences and their evolution over time.
- **Optimize Representations:** Develop robust user and document representations by combining dimensionality reduction (PCA), trajectory weighting, and probabilistic softmax transformations to retain semantic meaning and improve computational efficiency.
- **Identify Systemic Patterns:** Detect outliers, clusters, and trends in user-document interactions using innovative visualizations, aiding in the discovery of potential areas for improvement in recommendation or personalization systems.
- **Enhance Personalization Systems:** Provide actionable insights for improving the alignment between user preferences and content delivery, thereby boosting the effectiveness of personalization strategies.

This project not only establishes a comprehensive framework for analyzing user-document alignment but also provides key methodologies that can be applied to a wide range of personalization and recommendation system evaluations.

II. INTRODUCTION

This project focuses on analyzing the alignment between user behavior and document content by employing KL divergence as the core metric for comparison. The analysis begins by modeling user preferences using Markov Chains, which effectively capture sequential behavior patterns, and by deriving document representations from user-generated summary embeddings. These embeddings are weighted based on the significance of their corresponding trajectories to ensure that each document's representation reflects its importance in user interactions.

Dimensionality reduction is achieved through PCA, preserving semantic richness while enhancing computational efficiency. Probabilistic representations are generated for both user preferences and document embeddings, enabling a meaningful comparison via KL divergence for each trajectory.

The resulting KL divergence values are visualized using two innovative approaches: Row-wise KL Divergence with Highlights, which marks the top and bottom divergent rows for clarity, and Scatter Plot of KL Divergence with Outliers Highlighted, which captures patterns and anomalies across the dataset. These visualizations provide actionable insights into how well user preferences align with document representations, identifying cases of strong and weak alignment.

This work not only provides a comprehensive analysis of personalization effectiveness but also sets a foundation for optimizing recommendation systems by leveraging detailed interaction data and behavioral modeling.

III. ABOUT THE DATASET

1. User Interaction Dataset (synthetic-original-augmented-D2-1.csv)

Context:

This dataset is the foundation for understanding user behavior and interaction patterns with news posts. It provides detailed insights into user trajectories, including actions such as clicks, skips, and summary generations, which form the basis for modeling user preferences and behaviors.

Content:

- **UserID**: Unique identifier for each user in the dataset.
- **Docs**: A list of NewsIDs and SummIDs representing all news posts accessed by the user, grouped by their actions.
- **Action**:
    - The sequence of user interactions with news posts:
        - **click**: Indicates that the user clicked on a news post.
        - **skip**: Indicates that the user skipped a news post.
        - **gen_summ**: The user started generating a summary of all previously clicked news posts (excluding skipped posts).
        - **summ_gen**: Indicates the completion of a summary and provides a corresponding SummID.
- **Summaries**: The count of summaries generated by each user.

Important Notes:

- The sequence of actions is perturbed, meaning gen_summ might appear before summ_gen in some rows due to data perturbation.

Document Metadata Dataset (summ.csv)

Context:

This dataset provides detailed metadata for the summaries generated by users. It serves as the link between user interactions and the semantic representations used for document modeling.

Content:

- **SummID**: Unique identifier for each summary.
- **NewsID**: The NewsID of the primary news post associated with the summary.
- **UserID**: Unique identifier of the user who generated the summary.
- **Summary**: The actual textual summary generated by the user, capturing the core content of the summarized news posts.

Context: Final Dataset (embed_df)

The embed_df dataset integrates user interaction data and semantic representations derived from summary embeddings. It serves as the primary dataset for analyzing the alignment between user behavior and document content using KL divergence.

Content:

- **UserID**: Unique identifier for each user.
- **Docs**: A list of all news posts accessed by the user within a trajectory/interaction, represented by their corresponding NewsIDs and SummIDs.
- **Action**:
    - The sequence of user actions within each trajectory/interaction:
        - **click**: Indicates that the user clicked on a news post.
        - **skip**: Indicates that the user skipped a news post.
        - **gen_summ**: The user initiated the generation of a summary for all previously clicked posts (excluding skipped ones).
        - **summ_gen**: Denotes the completion of a summary and corresponds to a unique SummID.
- **Summaries**: The total number of summaries generated by the user within the trajectory/interaction.
- **SummID**: A list of all unique summaries (S1, S2, ..., SN) generated by the user in the trajectory/interaction, recorded in the order they are generated.
- **Mapped_Embeddings**: A list of embeddings corresponding to each SummID within the trajectory/interaction. These embeddings are derived from the Summary column in summ.csv and represent the semantic content of the summaries.

IV. TOOLS AND TECHNOLOGIES

I. Libraries and Frameworks :

- **Pandas:** Used for data manipulation, cleaning, and structuring the embed_df dataset.
- **NumPy:** Utilized for numerical operations, vector computations, and reshaping data structures.
- **SciPy:** Specifically, the scipy.stats.entropy function was used for KL divergence computation.
- **Scikit-learn(for):**
    - **Dimensionality Reduction:** Principal Component Analysis (PCA) to reduce embedding dimensions while preserving semantic meaning.
    - **Clustering:** DBSCAN and other clustering algorithms for analyzing user behavior and document patterns.
    - Preprocessing user features for clustering and analysis.
- **Seaborn and Matplotlib(For visualizing):**
    - KL divergence trends (line plots, scatter plots).
    - Distributions (histograms, KDE plots).
    - Heatmaps for trajectory and user-document relationships..

II. Machine LearModels:

- **DistilBERT (Distilled BERT):**Used for generating summary embeddings from the Summary column in summ.csv.These embeddings capture the semantic meaning of summaries, serving as the document representations in the KL divergence computation.

III. Analytical Techniques:

- **Markov Chains:** Applied to model user trajectories and capture sequential user actions (click, skip, gen_summ, summ_gen) as behavioral probabilities.

- **Principal Component Analysis (PCA):** Used to reduce high-dimensional summary embeddings into a compact form for efficient processing.
- **Softmax Transformation:** Normalized reduced embeddings into valid probability distributions for KL divergence computation.
- **KL Divergence:** The primary metric to compare user preferences (behavioral probabilities) with document representations (semantic embeddings).

V. KL DIVERGENCE FORMULA

➢ **Formula Overview**

The **Kullback-Leibler (KL) Divergence** is a measure of how one probability distribution P (user representation) differs from a second probability distribution Q (document representation). Mathematically, it is expressed as:

$$D_{kl}(P||Q) = \sum_i P(i) log \frac{P(i)}{Q(i)}$$

Where :

User Representation (P):

- Derived from Markov Chains capturing user action probabilities in trajectories.

Document Representation (Q):

- Derived from PCA-reduced summary embeddings, normalized using Softmax.

Step-by-Step Calculation:

➢ For each trajectory:
  - Compare the $P(i)$ (user probabilities) against $Q(i)$ (document probabilities).
  - Apply the formula:

$$D_{kl}(P||Q) = \sum_i P(i) log \frac{P(i)}{Q(i)}$$

Interpretation:

$D_{kl}(P||Q) = 0$: Perfect alignment (rare case).

$D_{kl}(P||Q) > 0$: Misalignment, with larger values indicating worse personalization.

Purpose of the Formula

- KL divergence quantifies the mismatch between user preferences (P) and document content (Q).
- A lower divergence indicates better alignment (user preferences are well represented by document embeddings).
- A higher divergence highlights potential misalignments, suggesting areas for improving personalization.
- Helps assess how effectively a recommendation system adapts document content to user behavior.

VI. IMPLEMENTATION DETAILS

1. **Dataset Loading:** The embed_df dataset, containing UserID, Docs, Action, Summaries, SummID, and Mapped_Embeddings, is loaded and inspected for missing values or inconsistencies, ensuring quality and readiness for processing.
2. I**nteraction Segmentation:** User actions are segmented into Main Trajectories, capturing direct sequences leading to `summ_gen`, and Upto Trajectories, covering actions leading up to `gen_summ`. These segments provide immediate and contextual behavioral insights.
3. **Markov Chains:** Represents user behavior by modeling probabilistic transitions between actions. These transitions are flattened into vectors for seamless comparison with document representations.
4. **Document Representations:** Derived from Mapped_Embeddings to encapsulate semantic content. PCA reduces dimensionality, and Softmax normalization ensures valid probability distributions for embedding comparisons.
5. **Weighted Trajectories:** Assigns weights to trajectories based on their length or significance,

prioritizing meaningful user interactions and aligning them with document representations.

6. **KL Divergence:** Measures the alignment between user behaviors (Markov Chains) and document embeddings. This provides insight into how well documents match user preferences, with results stored for analysis.
7. **Barplot Visualization:** Uses the first 100 rows of the dataset to create a barplot, clearly depicting worker-level KL divergence scores. This method replaces clustering approaches, offering a more straightforward interpretation of alignment patterns.

This modular pipeline provides a clear and actionable framework for analyzing user preferences and evaluating document alignment using KL divergence and worker-level diversity visualization.

VII. INSIGHTS

Comprehensive Analysis and Insights from KL Divergence Visualizations

1. Max, Min, and Mean KL Divergence Values

- **Max KL Divergence: 2.411848680806652**

  Indicates the highest divergence between the user representation and document representation. Likely represents scenarios where the user's preferences are highly misaligned with the document's content.

- **Min KL Divergence: 0.047150734921773116**

  Represents cases of near-perfect alignment between user preferences and document content.

- **Mean KL Divergence: 0.3891932324499312**

  Suggests that, on average, users have moderate divergence, indicating room for improvement in content personalization.

2. Top and Least Divergent Rows

- **Top Divergent Rows:**
  - Consistently show user representations like [1.0, 0.0, 0.0, 0.0, ...], indicating very specific user preferences (clicks or skips focused on one action).
  - Document representations are evenly distributed, suggesting these documents are generalized rather than tailored.
  - **Insight:** Personalization for these users can be improved by emphasizing their dominant preferences in the recommendations.

- **Least Divergent Rows:**
  - Have user representations with diverse probabilities (e.g., [0.1, 0.1, 0.1, ...]), indicating balanced behavior across actions.
  - Document representations closely align with these preferences, showing successful personalization.
  - **Insight**: The system works well for users with generalized preferences.

3. Line Plot (Row-wise KL Divergence with Highlights)

Trend Analysis:

High variability across rows, with periodic peaks indicating misaligned recommendations.

Highlighted Points:

Red (Top Divergences): Peaks corresponding to maximum divergence.

Green (Low Divergences): Dips indicating aligned user-document preferences.

Insight:

Outlier points suggest the need for dynamic recommendation strategies based on user profiles.

Potential to focus on top divergent rows to reduce misalignment.

4. Scatter Plot (KL Divergence with Outliers Highlighted)

Trend Analysis:

Dense concentration of KL divergence values near the lower end, indicating most rows have acceptable alignment.

Scattered outliers at higher divergence values show less frequent but significant misalignments.

Highlighted Points:

Red (Top Divergences): Clearly visible as outliers,

providing critical cases for analysis.

Green (Low Divergences): Represent success cases in personalization.

Insight:

The visualization shows the system's overall robustness, with few extreme misalignments. Focus on red outliers to improve the system further.

5. Key Insights

Alignment Patterns:

The majority of rows have KL divergence values below 1, indicating moderate alignment between user preferences and document content.

Outliers:

Outlier rows with KL divergence above 2 are critical for improving recommendation systems.

User Segmentation:

Users with diverse preferences (low KL divergence) benefit more from the system, while those with highly specific preferences face misalignments.

System Performance:

The system works effectively for generalized user behavior but struggles with extreme or specific preferences.

VIII. CONCLUSIONS

This analysis provided key insights into user behavior through KL Divergence, Interaction Count, and Click-Skip Ratio. These metrics highlighted user engagement patterns and their alignment with semantic summaries, enabling actionable recommendations for personalization strategies.

KL Divergence quantified alignment between user actions and content. Low divergence reflected clear preferences, while high divergence indicated exploratory or inconsistent behavior, emphasizing the need for tailored content delivery.

Interaction Count and Click-Skip Ratio revealed focused, exploratory, selective, and indecisive user patterns, offering a nuanced view of engagement levels. The trajectory-based methodology proved scalable, helping categorize users for targeted interventions.

In summary, this framework supports refining personalization systems and enhancing user engagement through data-driven insights.