**NLP – Final Project**

**Analysis of Unrecognized User Requests in Goal-Oriented Dialog Systems**

**Submitters:**

**Shachar Levy – 318974557**

**Maayan Mashhadi – 318702230**

**Introduction:**

In this document, we explore a text clustering and title generation approach specifically tailored for handling unrecognized requests. Unrecognized requests often contain unstructured text data, making it challenging to categorize and analyze them efficiently. By clustering similar requests and generating titles for the clusters, we aim to organize and provide insights into these unrecognized requests.

**Understanding the Code:**

The provided code consists of several components:

Library Import: Necessary libraries such as json, numpy, pandas, and modules from sentence\_transformers and nltk are imported.

Constants and Parameters: Constants like MODEL\_NAME and parameters for clustering (parameters) are defined.

Sentence Transformer Model: A Sentence Transformer model is initialized using the specified MODEL\_NAME.

**Section 1: Clustering of Requests**

**Approach**

The clustering of requests involves grouping similar requests together based on their semantic similarity. The process is implemented using dynamic means clustering, a technique that iteratively updates cluster centroids and assigns data points to clusters until convergence is achieved.

**Implementation Details**

**Data Preprocessing:**

The script loads textual data from a CSV file using the Pandas library.

Text preprocessing steps include removing special characters, converting text to lowercase, and tokenizing sentences.

Embedding Generation:

Textual data is encoded into numerical vectors using a pre-trained Sentence Transformer model.

These embeddings capture the semantic meaning of sentences in a dense vector space.

**Dynamic Means Clustering:**

Centroids are initialized as empty lists, and cluster assignments are tracked using NumPy arrays.

The algorithm iterates through the data points, updating cluster centroids and assignments based on the Euclidean distance between data points and centroids.

Convergence is achieved when cluster assignments remain unchanged between iterations or when the maximum number of iterations is reached.

**Cluster Representation:**

Clusters are represented as lists of sentences grouped by their assigned centroids.

Unclustered sentences, i.e., those that do not meet the minimum cluster size requirement, are stored separately.

**Section 2: Generating Titles**

**Approach**

Title generation involves creating descriptive titles for each cluster to summarize the common theme or topic represented by the clustered requests. Titles are generated based on the most frequent 3-grams (sequences of three consecutive words) within each cluster.

**Implementation Details**

**Text Processing:**

Stop words and punctuation characters are removed from sentences to focus on meaningful content.

Tokenization and generation of 3-grams are performed using the Natural Language Toolkit (NLTK) library.

**Title Extraction:**

The most frequent 3-gram within each cluster is identified using the Counter class from the collections module.

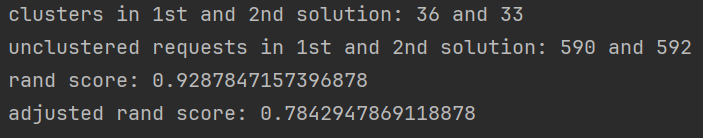
If no significant 3-gram is found, a default title is assigned.

**Output Formatting:**

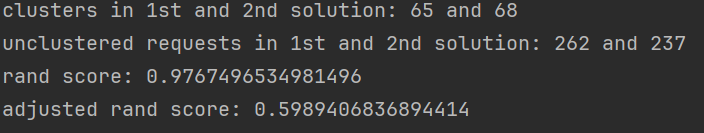
Titles and clustered requests are structured into a JSON format for easy readability and further analysis.

**Evaluation of the clustering outcome against the provided solution – RI, ARI**

**Example result for covid19-unrecognized-requests:**

****

**Example result for banking-unrecognized-requests:**

****