Intelligent Search of most relevant offers in an e-commerce website using Latent Semantic Indexing (LSI).

Abstract:

The project aims to create a master dataset from three CSV files using MySQL Workbench and then perform semantic search using Latent Semantic Indexing (LSI) along with various text analytics techniques. The final dataset consists of five columns: OFFER, RETAILER, BRAND, BRAND_BELONGS_TO_CATEGORY, and IS_CHILD_CATEGORY_TO. The goal is to simplify query building and improve search results by incorporating LSI.

Tools and Libraries Used:
- MySQL Workbench: Used to combine and export the master dataset CSV file.
- Jupyter Notebook: Utilized for code execution and documentation.
- Libraries:
  - SpaCy: Known for its speed and accuracy, used for natural language processing.
  - Gensim: Efficient for working with large text corpora, employed for unsupervised machine learning tasks related to text.

Project Workflow:

1. Data Integration:
   - The three CSV files were combined using MySQL Workbench, employing an outer join to create the master dataset.
   - All the columns in the dataset are concatenated to a new OFFERS_new column avoiding null values and removing duplicates.

2. Data Preprocessing:
   - Data pre-processing is a crucial step in text analytics to prepare the data for analysis. It involves:
     - Removing unwanted characters, digits, and punctuation.
     - Tokenization: Splitting text into individual words or tokens.
     - Converting text to lowercase for consistency.
     - Lemmatization: Reducing words to their base form for better analysis.

3. Building Vocabulary:
   - The vocabulary of the corpus was constructed using Gensim. Each unique word in the text received an ID, and their frequency counts were stored. In Gensim, words are referred to as "tokens," and their index in the dictionary is called the ID.

4. Feature Extraction (Bag of Words):
   - A Bag of Words (BoW) model was employed to extract features from the text for machine learning. BoW involves:
     - Defining a vocabulary of known words.

- Measuring the presence of known words in documents.
- The `doc2bow` method of the dictionary was used to iterate through all words in the text, incrementing frequency counts for existing words and inserting new words into the corpus with a frequency count of 1.

5. Building Tf-Idf and LSI Model:
   - Tf-Idf (Term frequency-Inverse Document Frequency) was used to determine the importance of words in each document of the corpus. Tf-Idf helps identify the most relevant words.
   - The Tf-Idf model was then passed to the LSI (Latent Semantic Indexing) model, specifying the number of features to build. LSI helps identify semantic relationships between words and documents.

6. Semantic Searching:
   - After building the Tf-Idf and LSI models, the system is ready for semantic searching.
   - Semantic search involves querying the master dataset using natural language queries, and the LSI model helps in understanding the context and semantics of the queries.
   - Users can input search queries in plain language, and the system will return relevant results based on semantic similarity rather than relying solely on exact keyword matches.
   - This semantic search capability enhances user experience and improves the accuracy of search results, as it considers the meaning and context of words in the query.

Project Benefits:

- The project provides a powerful and efficient way to search and retrieve information from the master dataset.
- Semantic search using LSI allows for more flexible and context-aware queries, even when there is no exact keyword match.
- The data preprocessing steps ensure that irrelevant information is removed, improving the quality of search results.

Conclusion:

The project successfully combines data from multiple sources, preprocesses it for text analytics, and employs LSI to enable semantic search. By utilizing Gensim and SpaCy, it leverages efficient libraries for large-scale text processing. The project has the potential to be expanded and deployed for real-world applications, offering users a powerful and intuitive way to search and retrieve information from the master dataset.