
Report on `task merged.ipynb`

Objective: The primary objective of this notebook is to merge two datasets, perform text translation, and calculate text similarity.

Data Loading and Preprocessing:

- The notebook begins by importing several libraries, including `pandas`, `numpy`, `re`, `nltk` (with `stopwords` and `word_tokenize`), `seaborn`, `plotly.express`, `matplotlib.pyplot`, `collections.Counter`, `deep_translator`(specifically `GoogleTranslator`), `sklearn.feature_extraction.text` (with `TfidfVectorizer`), and `sklearn.metrics.pairwise` (with `cosine_similarity`).
- It loads data from an Excel file named "Data for Task 2.xlsx" into a pandas DataFrame named `df`.
- `df.info()` and `df.head()` are used to inspect the structure and content of the DataFrame.
- A language detection function `detect_language` is defined using `GoogleTranslator` to identify the language of the text in the `text` column.
- A translation function `translate_text` is defined to translate non-English text in the `text` column to English using `GoogleTranslator`.
- The `detect_language` and `translate_text` functions are applied to the `text` column to create a new column `Translated_text`.

Text Similarity Calculation:

- The `Translated_text` column is filled with empty strings where it has null values.
- `TfidfVectorizer` is used to convert the text data in `Translated_text` into a TF-IDF matrix, which represents the importance of words in each document.
- `cosine_similarity` is then used to calculate the cosine similarity between the TF-IDF vectors, resulting in a similarity matrix.
- The shape of the similarity matrix is printed.

Data Merging:

- The notebook loads the "task1.csv" file (generated from `task1.ipynb`) into a DataFrame named `df1`.

- It is explicitly mentioned in the notebook that a primary key was not found in the "Data for Task 2.xlsx" dataset.
- To proceed with merging, both `df1['VIN']` and `df['Primary Key']` columns are converted to string type to ensure compatibility.
- An outer join is performed on `df1` and `df` using the `VIN` column from `df1` and the `Primary Key` column from `df` as join keys. The `how='outer'` argument ensures that all rows from both DataFrames are included in the result. Suffixes are added to differentiate columns from the two DataFrames.
- The shape of the merged DataFrame (`merged_df`) and the first few rows are printed.

Output:

- The merged DataFrame `merged_df` is saved to a CSV file named "Final.csv".

Summary:

The `task2merged.ipynb` notebook focuses on text processing and data integration. It translates text from various languages into English, calculates text similarity using TF-IDF and cosine similarity, and merges this data with the output of the first notebook (`task1.csv`) using an outer join. The key challenge and the solution are clearly stated in the notebook, where the absence of a clear primary key in the second dataset was handled by using an outer join on the `VIN`and `Primary Key` columns. The final output is a merged dataset that combines the processed text data with other relevant information.

Sources