Fuzzy SQL

Release 1.0.0

Samer Kababji @ EHIL

Nov 07, 2022

CONTENTS

CHAPTER

ONE

INSTALLATION

The Fuzzy SQL package is currently private and can be installed by authorized personnel by:

```
(.venv) $ pip install git+ssh://git@github.com/skababji-ehil/fuzzy_sql.git#egg=fuzzy_sql
```

Check out *Usage* for further information.

The package includes the necessary dependencies.

CHAPTER

TWO

FUNCTIONS

fuzzy_sql.fuzzy_sql.**prep_data_for_db**(*csv_table_path: Path*, *optional_table_name='None'*, *is_child=False*, *metadata_dir='None'*, *nrows=None*) → tuple

Reads the input csv file and prepare it for importation into sqlite db for fuzzy-sql analysis. The file name (without extension) will be used as a table name in the database. All values are imported as strings. Any "'" found in the values (e.g. '1') is deleted. Any variable (columns) that include dots in their names will be replaced by underscores.

Parameters

- **csv_table_path** – The input file full path including the file name and csv extension.
- **optional_table_name** – This is an optional name of the table when imported into the database. The default 'None' will use the csv file name (without extension) as the table's name.
- **is_child** – A boolean to indicate whether teh input table is child or not. This will impact the generated metadata template. Enter 'False' if the input table is tabular or not a child.
- **metadata_dir** – The directory where the metadata file shall be saved. No metadata file is saved if the default value of 'None' is used.
- **n_rows** – The number of rows to be read from the input csv file. The default of None will read all the rows in the csv file.

Returns

The pandas dataframe in 'unicode-escape' encoding. The corresponding metadata dictionary. The dictionary is saved to the chosen path as provided in metadata_dir.

fuzzy_sql.fuzzy_sql.**import_df_into_db**(*table_name: str*, *df: DataFrame*, *db_conn: object*)

Imports the input dataframe into an sqlite database table. The data will NOT be imported if it already exists in the database.

Parameters

- **table_name** – The intended name of the table in the database.
- **df** – The input data
- **db_conn** – Database (sqlite3) connection object

class fuzzy_sql.fuzzy_sql.**RND_QRY**(*db_conn: object*, *tbl_names_lst: list*, *metadata_lst: list*, *params={'AGG_OPS': {'AVG': 0.5, 'MAX': 0.1, 'MIN': 0.1, 'SUM': 0.3}, 'CAT_OPS': {'<>': 0.25, '=': 0.25, 'IN': 0.15, 'LIKE': 0.15, 'NOT IN': 0.1, 'NOT LIKE': 0.1}, 'CNT_OPS': {'<': 0.1, '<=': 0.1, '<>': 0.1, '=': 0.2, '>': 0.1, '>=': 0.1, 'BETWEEN': 0.2, 'NOT BETWEEN': 0.1}, 'DT_OPS': {'<': 0.1, '<=': 0, '<>': 0.1, '=': 0.2, '>': 0.1, '>=': 0, 'BETWEEN': 0.2, 'IN': 0.1, 'NOT BETWEEN': 0.1, 'NOT IN': 0.1}, 'FILTER_TYPE': {'AND': 0.5, 'WHERE': 0.5}, 'JOIN_TYPE': {'JOIN': 0.5, 'LEFT JOIN': 0.5}, 'LOGIC_OPS': {'AND': 0.9, 'OR': 0.1}, 'NOT_STATE': {'0': 0.8, '1': 0.2}}*, *seed=False*)

Generates random queries for tabular and longitudinal datasets.

class fuzzy_sql.fuzzy_sql.**QRY_RPRT**(*dataset_table_lst*, *random_queries*)

CHAPTER

THREE

USAGE

INDEX