**ETL Pipeline Documentation: China Tariff Data Processing**

1. **Overview**

This ETL pipeline extracts tariff, VAT, and excise duty data from a Chinese government PDF (source: <https://gss.mof.gov.cn/gzdt/zhengcefabu/202412/P020250127378158317243.pdf> ), cleans and transforms it, and outputs structured CSV files for cross-border trade compliance analysis.

1. **Pipeline Components**

* Extract

Input: PDF file (P020250127378158317243.pdf)

Tools: tabula-py (Python wrapper for Apache PDFBox)

Process:

* Extract all tables using tabula.read\_pdf() with lattice=True for better parsing.
* Filter tables containing target headers (e.g., 税则号列 [HS Code], 最惠国税率(%) [MFN Rate]).
* Combine valid tables into a single DataFrame and export to \_combined\_china.csv.

**Key Functions:**

tables = tabula.read\_pdf("P020250127378158317243.pdf", pages="all", lattice=True)

filtered\_tables = [table for table in tables if all(header in table.iloc[0] for header in target\_headers)]

* Transform

Objective: Clean and standardize data for ML/analytics.

1. Column Renaming & Null Handling

* Rename ambiguous columns (e.g., Unnamed: 7 → new\_col1).
* Add placeholder columns (guard\_col1, guard\_col2) for validation.

2. Data Shifting Logic

* HS Code Parsing: Extract trailing numeric values (e.g., 6∆0) and shift them to MFN\_Rate if Preferential\_Rate is empty.
* Chinese Text Handling: Detect and remove Chinese characters from numeric columns.

- Special Cases:

- Shift values right if Preferential\_Rate(Agreement rate)\_1 contains "MO".

- Remove rows with standalone "R" characters or ex prefixes in HS codes.

Key Functions:

* def extract\_last\_value(value): # Extracts trailing numeric/∆ values
* def shift\_row(row): # Shifts values right if Chinese text detected
* def remove\_chinese\_chars(text): # Regex to strip non-Latin characters

1. **Final Cleansing**

- Drop rows without alphanumeric characters.

- Split concatenated tariff rates (processed via Excel TEXTJOIN) into discrete columns.

- If a country is repeated, it is denoted with raise to the power of R. If an R is standing alone, it is raise to the power of the country before or after it before the nearest comma.

- All ‘?’ means ‘Δ’

* **Load**

Outputs:

1. Intermediate Files:

* review\_data.csv (post-initial cleaning)
* review\_data4.csv (post ex row removal)

2. Final Output:

* clean\_7\_1.csv (fully processed data).

3. Usage

* Dependencies:

pip install tabula-py pandas numpy

* Run Pipeline:

python china\_tarrif\_ETL\_script.py

* Output Validation:

Check clean\_7\_1.csv for final structured data.

4. Error Handling

* Missing Headers: Skips tables without target columns.
* Encoding Issues: Uses utf-8-sig for CSV exports.
* Null Values: Preserves NaN for empty fields.

Maintained by:Ogbonna Prince

Last Updated: 29th March, 2025