# **ReadMe: Reference Extraction**

## **Overview**

This script processes a set of text files to extract footer notes, clean and format them, and then derive useful references and additional data. The output is saved in CSV files at each step for further analysis and record-keeping.

## **Prerequisites**

Ensure you have the following Python libraries installed:

* re: For regular expressions.
* os: For operating system interactions.
* csv: For reading and writing CSV files.
* spacy: For natural language processing and entity recognition.
* pandas: For data manipulation and analysis.

Install missing libraries with:

‘pip install spacy pandas’

Additionally, download the spaCy English language model:

‘python -m spacy download en\_core\_web\_sm’

## **Execution Steps**

### **1. Footer Note Extraction**

This step extracts footer notes from a folder of text files and saves them in a CSV file.

* Input: A folder path containing text files with potential footer notes.
* Output: A CSV file named 1.csv with extracted footer notes.

### **2. Remove Tables from Footer Notes**

This step cleans the extracted footer notes by removing unnecessary patterns, tables, and other clutter.

* Input: The CSV file from Step 1 (1.csv).
* Output: A cleaned CSV file (2.csv).

### **3. Extract Clean Footer Notes**

This step processes the cleaned footer notes to ensure proper formatting and distribution of paragraphs.

* Input: The cleaned CSV file from Step 2 (2.csv).
* Output: A formatted CSV file (FooterNotes.csv).

### **4. Extract References**

In this step, references are extracted and analyzed from the clean footer notes using entity recognition and custom rules.

* Input: The CSV file from Step 3 (FooterNotes.csv).
* Output: A CSV file with extracted references (References.csv).

## **Running the Script**

### **Setting Up the Input Parameters**

* Folder Path: The folder containing the text files to be processed.
* Output CSV File Names: Names of the CSV files where intermediate and final results are saved.

### **Executing the Script**

Run the script to:

* Extract and clean footer notes from text files.
* Remove redundant tables and patterns.
* Distribute paragraphs properly.
* Extract references from the cleaned footer notes.

### **Accessing the Output**

After running the script, the outputs include:

* Extracted Footer Notes: Saved in 1.csv.
* Cleaned Footer Notes: Saved in 2.csv.
* Formatted Footer Notes: Saved in FooterNotes.csv.
* Extracted References: Saved in References.csv.

## **Troubleshooting and Common Errors**

* Error Reading Files: Ensure the folder path is correct and contains the expected text files.
* Pattern Matching Errors: Double-check the regular expression patterns to ensure they are correct.
* CSV Write Errors: Ensure the output paths are correct and writable.