Unique Address Identification

Overview:

This project is designed to process address data. The system is divided into several tasks, each handling a specific part of the address processing pipeline. The primary goal is to extract, clean, validate, and identify unique addresses, grouping them by loan number.

Components

Task 1: Extract Loan Addresses (`task1.extract_loan_address`)

- **Purpose**: Extracts the loan number and address from an Excel file named 'main.xlsx' (source of loan number and addresses).
- **Output**: A new Excel file, 'processed_addresses.xlsx', containing columns for loan number and address.

Task 2: Extract and Add Pin code (`task2.process_address_1`)

- **Purpose**: Extracts pin code from the addresses.
- **Method**: Uses regular expressions to find a sequence of 6 digits representing the pin code.
- **Output**: Updates 'processed_addresses.xlsx' with a new column for pin code.

Task 3: Extract and Add State (`task3.process_addresses_2`)

- **Purpose**: Identifies and adds the state for each address.
- **Method**: Compares words in the address with a predefined list of Indian states and union territories.
- **Output**: The Excel file is updated with a new column for states.

Task 4: Extract and Add District (`task4.process_addresses_3`)

- **Purpose**: Identifies and adds the district for each address.
- **Method**: Searches for district names within the address string.
- **Output**: The 'processed_addresses.xlsx' file gets a new column for districts.

Task 5: Validate and Clean Addresses (`task5.task5_execute`)

- **Purpose**: Validates and cleans each address.
- **Method**: Addresses are cleaned of special characters and duplicates. Addresses are validated based on length and word count.
- **Output**: The Excel file is updated with a column indicating whether each address is valid or invalid.

Task 6: Identify Unique Addresses (`task6.task6_execute`)

- **Purpose**: Identifies unique addresses in the dataset.
- **Method**: Normalizes addresses and uses fuzzy matching to identify uniqueness.
- **Output**: The updated Excel file includes normalized addresses and a unique address identifier.

Workflow

1. **Extract Loan Addresses**: Starts with extracting loan numbers and addresses.
2. **Process Address for Pin code**: Adds pin code to each address.
3. **Process Address for State**: Includes state information.

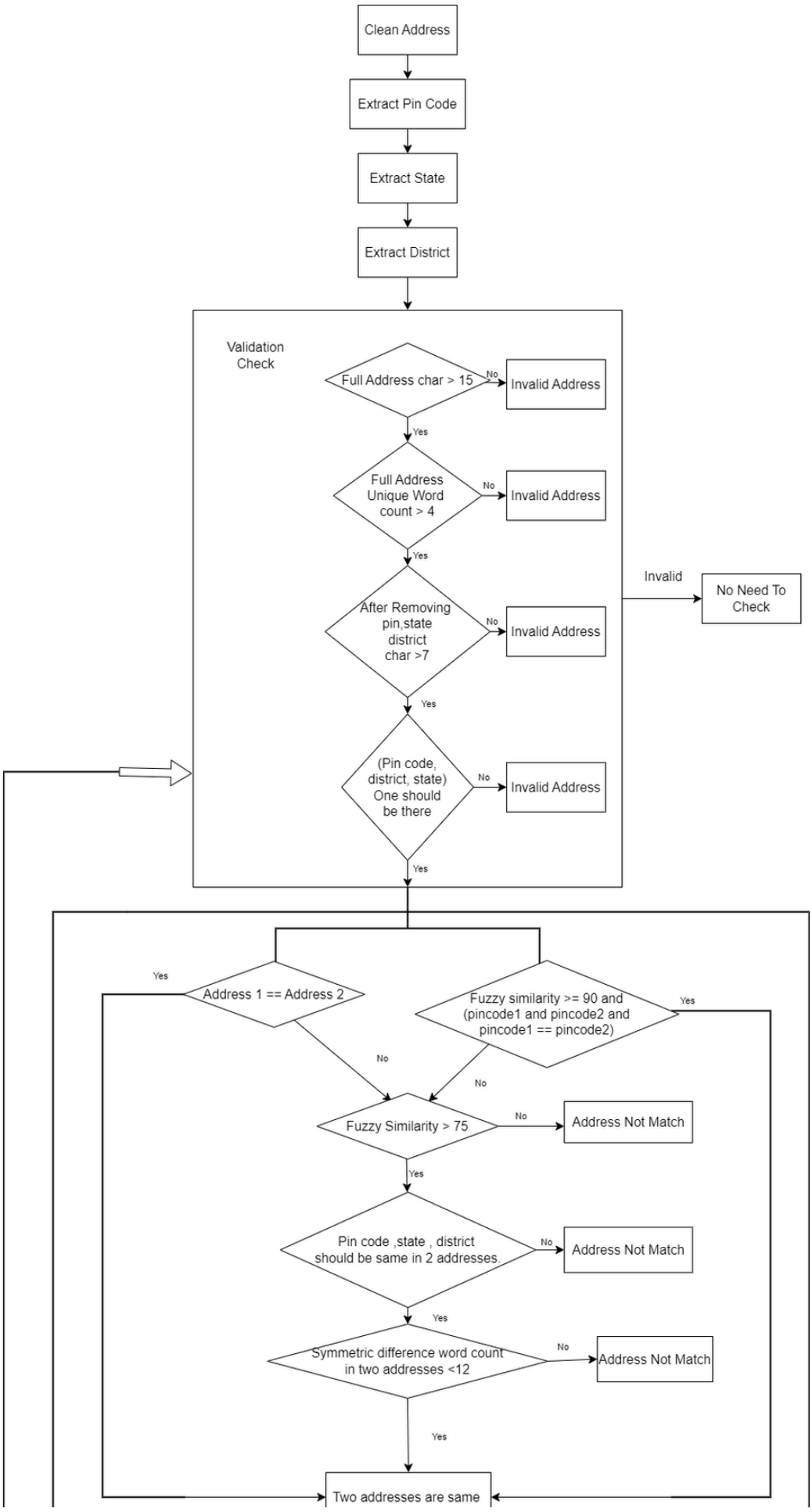
4. **Process Address for District**: Appends district details.
5. **Validate and Clean Addresses**: Cleans and validates each address.
6. **Identify Unique Addresses**: Marks unique addresses in the dataset.

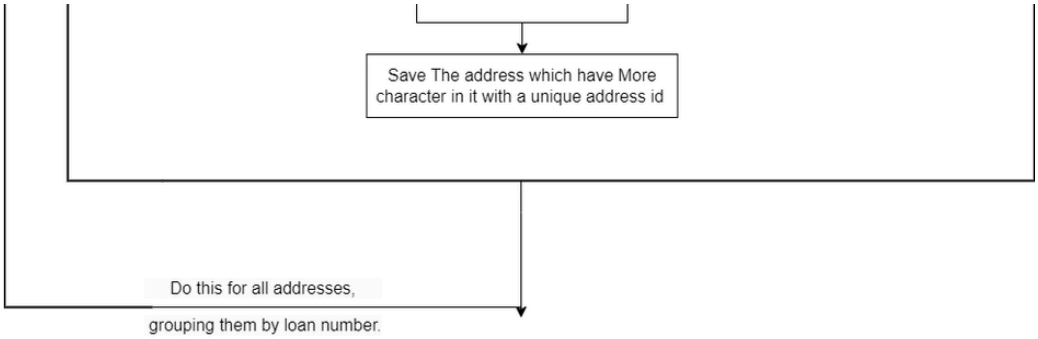
Execution

To run the entire pipeline, execute the main function. This function sequentially calls each task, ensuring that the dataset is processed step-by-step, culminating in a comprehensive dataset with unique addresses identified and grouped by loan number.

System Requirements

- Python 3.11
- Pandas Library
- NLTK Library
- Fuzzy Wuzzy Library


Clean Address
Extract Pin Code
Extract State
Extract District
Validation Check
Full Address char > 15
No
Invalid Address
Yes
Full Address Unique Word count > 4
No
Invalid Address
Yes
After Removing pin,state district char >7
No
Invalid Address
Yes
(Pin code, district, state) One should be there
No
Invalid Address
Yes
Invalid
No Need To Check
Address 1 == Address 2
Yes
Fuzzy similarity >= 90 and (pincode1 and pincode2 and pincode1 == pincode2)
Yes
No
No
Fuzzy Similarity > 75
No
Address Not Match
Yes
Pin code ,state , district should be same in 2 addresses.
No
Address Not Match
Yes
Symmetric difference word count in two addresses <12
No
Address Not Match
Yes
Two addresses are same



Flowchart of Unique Address Identification Algorithm