Transaction Classification - Documentation

# Introduction

This project classifies banking transactions by parsing CSV files and categorizing merchants into predefined categories. It utilizes machine learning techniques like logistic regression and word embeddings for categorization, designed to work in a bank’s data pipeline.

# Features

- Extract merchant names from raw transaction data.  
- Categorize transactions using both lookup tables and machine learning models.  
- Supports continuous learning through an online classification algorithm.

# Setup

1. Set up the necessary directories in `initial\_setup/directories.py`.  
2. Download required data like merchant transactions, US Census city data, and GloVe word embeddings.  
3. Hand categorize seed data and preprocess word embeddings.  
4. Preprocess city data using `initial\_setup/cities\_into\_df.py`.

# How to Use

After setup:  
- Input: CSV files with transaction data.  
- Output: Categorized transactions in a CSV format.  
For further details, refer to the example in `test/online.py`.

# Running Tests

- Unit tests are located in `test/unit\_tests.py`.  
- Full system tests and data splitting for classification models are available in `test/gold\_standard.py`.

# Algorithm Overview

The transactions are first parsed to extract merchant names, followed by categorization:  
1. Lookup table for common merchants.  
2. Logistic regression model for less common transactions.  
3. Word embedding for merchant name similarity measurement.

# Parsing Details

The project uses regular expressions to extract merchant information from transaction strings.

# Categorization Process

The categorization is twofold:  
1. Lookup for well-known merchants.  
2. Logistic regression for unknown merchants using word embeddings and transaction amounts.

# Contributing

Feel free to submit issues or pull requests to improve the project. Refer to `CONTRIBUTING.md` for guidelines.

# License

This project is licensed under the MIT License.