IMDB Data Conversion Documentation

This project demonstrates the conversion of data files from the TSV (Tab-Separated Values) format to the CSV (Comma-Separated Values) format using Python and the Pandas library. The data used in this project is sourced from the IMDB database.

Project Overview

The primary objective of this project is to convert IMDB data files in TSV format to CSV format for easier usage in data analysis and processing workflows. The script processes multiple TSV files and saves the converted data into corresponding CSV files.

Data Source

The data files used in this project are part of the IMDB dataset and include the following:

- title.ratings.tsv
- title.crew.tsv
- title.episode.tsv
- title.basics.tsv
- title.akas.tsv
- title.principals.tsv
- name.basics.tsv

These files were downloaded from the IMDB Datasets website. The dataset provides extensive information about movies, TV shows, and their associated metadata.

Understanding TSV and CSV Formats

TSV (Tab-Separated Values)

- **Format**: Data values are separated by tabs (\t).
- **Usage**: Preferred for datasets where data fields might contain commas, as it avoids conflicts.
- **File Example**:
- tconst rating votes
- tt0000001 5.6 1600

  tt0000002 6.1 700

CSV (Comma-Separated Values)

- **Format**: Data values are separated by commas (,).
- **Usage**: Widely supported by most data analysis tools, databases, and software.
- **File Example**:
- tconst,rating,tvotes
- tt0000001,5.6,1600

tt0000002,6.1,700

Python Script

Below is the Python script used for the conversion process:

```
import pandas as pd


# List of full paths to the TSV files (replace these with your actual file paths)
tsv_files = [
    "C:/Users/Yashwanth/Downloads/Unzip/title.ratings.tsv",
    "C:/Users/Yashwanth/Downloads/Unzip/title.crew.tsv",
    "C:/Users/Yashwanth/Downloads/Unzip/title.episode.tsv",
    "C:/Users/Yashwanth/Downloads/Unzip/title.basics.tsv",
    "C:/Users/Yashwanth/Downloads/Unzip/title.akas.tsv",
    "C:/Users/Yashwanth/Downloads/Unzip/title.principals.tsv",
    "C:/Users/Yashwanth/Downloads/Unzip/name.basics.tsv"
]


# Loop through each TSV file and convert to CSV
for tsv_filename in tsv_files:
    # Read the TSV file
    df = pd.read_csv(tsv_filename, sep="\t")


    # Generate the CSV file path (replace .tsv with .csv)
    csv_filename = tsv_filename.replace(".tsv", ".csv")


    # Save as CSV file
    df.to_csv(csv_filename, index=False)


    print(f"Converted {tsv_filename} to {csv_filename}")
```

Steps to Execute

1. **Install Pandas Library**: Ensure you have Pandas installed in your Python environment. You can install it using:

```
pip install pandas
```

2. **Download the IMDB Data**: Download the required TSV files from the IMDB Datasets.
3. **Set Up File Paths**: Update the tsv_files list in the script with the actual paths to your downloaded TSV files.
4. **Run the Script**: Execute the Python script to convert the TSV files to CSV format.
5. **Check Output**: The converted CSV files will be saved in the same directory as the original TSV files.

Data Source

The data files were downloaded from the IMDB Datasets.

About TSV and CSV Formats

- **TSV (Tab-Separated Values)**: Uses tab (\t) as a delimiter. Ideal for avoiding issues with embedded commas in data.
- **CSV (Comma-Separated Values)**: Uses commas (,) as a delimiter. Widely used and supported format for data analysis.

How to Use

1. Clone this repository.
2. Replace the file paths in the tsv_files list with your local file paths.
3. Run the imdb_data_conversion.py script.
4. The CSV files will be generated in the same directory as the original TSV files.

Dependencies

- Python 3.x
- Pandas library (pip install pandas)

License

This project uses data from IMDB, available under their terms of use.