Extract – CSV, JSON, HTML, SQL, Spreadsheets

Transform – Data cleaning, summarization, selection, joining, filtering, aggregating. Use Python pandas or sqlite

Load – relational or non-relational database, local or in the cloud (using sql alchemy to PostgreSQL) You might need to install psycopg2

Create the DB and DB tables before you load the data into the db

* Cite data sources
* Extract data from the sources
* Transform the data – clean, join, filter, aggregating
* Load data into database

1. Proposal
   1. Create a database to help track book popularity, price and information about the author and book for the books published between year 2015 to 2020.
   2. Our database will help track popular books by rating and number of reviews.
2. Roles
   1. Satya web scraping and cleaning
   2. Lauren data cleaning
3. Sources
   1. <https://zenodo.org/record/4265096#.Y5vlCVHMJD->
   2. Web scraping from Wikipedia page for author information.
4. The process
   1. Extract
      1. Downloaded good books data csv from zenodo.org
      2. Web Scraping for the author information
   2. Transform
      1. Source 1

<https://zenodo.org/record/4265096#.Y5vlCVHMJD->

* + - * + The original dataset has around 50000 rows. We filtered the dataset for the books published in English and books published in latest 5 years in the dataset that is 2015 to 2020 for web scraping purpose as well as interest towards latest books.
        + There are about 3800 rows of data after filtering.
        + The original dataset has about 25 columns filtered down to 12 columns unnecessary columns were dropped.
        + title, author, rating, description, language, isbn, genres, pages, publishDate, publisher, numRatings, likedPercent
        + Author column has terms such as (Goodreads Author), (Translator), (Editor), (Foreword), (Pseudonym) and lots of non-alphabetical characters all of them were split by commas, brackets into separate columns then they were dropped to create main author column which has author name.
        + PublishDate column has dates that are in different formats such as 09/14/08 for which we don’t know the exact year of release, some of the rows have values that are not dates (published, expected publish dates) all of them were filtered and converted the remaining rows to one format (YYYY-MM-DD) and created the column called year published.
        + All the rows that have NA values were dropped.
        + \*\*Pages column was in string format converted to int.
        + \*\*Genres :
    1. Source

Web Scraping:

* + - * + Created the list of unique authors.
        + Scraped the Wikipedia page for the date of birth of the author.
        + Created a table for the author information.
  1. Load

1. Schema
   1. Visual from quick DB

Graphical user interface, application, chat or text message

Description automatically generated

1. Sample Queries
   1. Examples