**Collecting Datasets Through Web Scraping**

**1. Objectives**

The main goal of this documentation addresses the procedures needed to extract product reviews from AliExpress using web scraping methods. The data collection through web scraping activity leads to results that shall support subsequent research in the thesis. The specific objectives are:

1. **Data Collection**: The data scraping operation targets product reviews from AliExpress for various products including Crossbody Bags alongside Car Phone Holders dashcams and Smartwatches.
2. **Dataset Preparation**: The collected data will undergo preparation as a CSV format which makes it accessible for analysis purposes**.**
3. **Automation**: The process needs automation because it will enable scraping of multiple product URLs and obtain numerous reviews for every product.
4. **Data Integrity**: The collected data should be both accurate and relevant and complete for research objectives.

**2. Introduction**

The significance of customer reviews encompasses the comprehension of consumer behavior together with the inspection of product performance and detection of market trends in modern e-commerce business. User-generated reviews about a wide product selection exist extensively on AliExpress because it is one of the world's largest online retail platforms. The reviews yield vital information about how satisfied customers are along with the quality of products and necessary improvement sections.

The documentation details how to extract AliExpress product reviews through **Appify.com**'s web scraping library framework. Autonomous data collection occurs through Python with **ApifyClient** library support to achieve efficient and extensive data scraping. CSV files serve as storage units for data that gets collected where further analysis takes place.

This examination investigates a wide range of products that cover fashion, electronics, home accessories along with automotive products. Various product categories bring together a large dataset which enables strong conclusions in the thesis research.

**3. Theory**

The process begins with collecting product reviews from AliExpress using the Apify.com web scraping library. A diverse range of best-selling products is considered, including:

* Crossbody Bags
* Car Phone Holders
* Dashcams
* Portable Car Vacuums
* Electric Toothbrushes
* Sofa Covers
* Broom Holders
* Dog Collars
* Wallets
* Phone Cases
* Security Cameras
* Bluetooth Earbuds
* Smartwatches
* Kitchen Accessories
* Fitness Equipment
* Beauty and Health products
* Automotive accessories
* Pendrive, Fairy Light and LED Lamps

Total **110** products review is scraped from these categories.

**3.1 Tools and Libraries**

The following tools and libraries were used for the scraping process:

* **Python**: The primary programming language used for scripting the scraping process.
* **Notebook**: Google Colab notebook was used for the python environment.
* **ApifyClient**: A Python library provided by Appify.com to interact with the Apify platform and scrape data from AliExpress.
* **CSV Module**: Used to store the scraped data in a structured CSV file.

The screenshot of necessary code of the scrapping process for collecting datasets are given below:

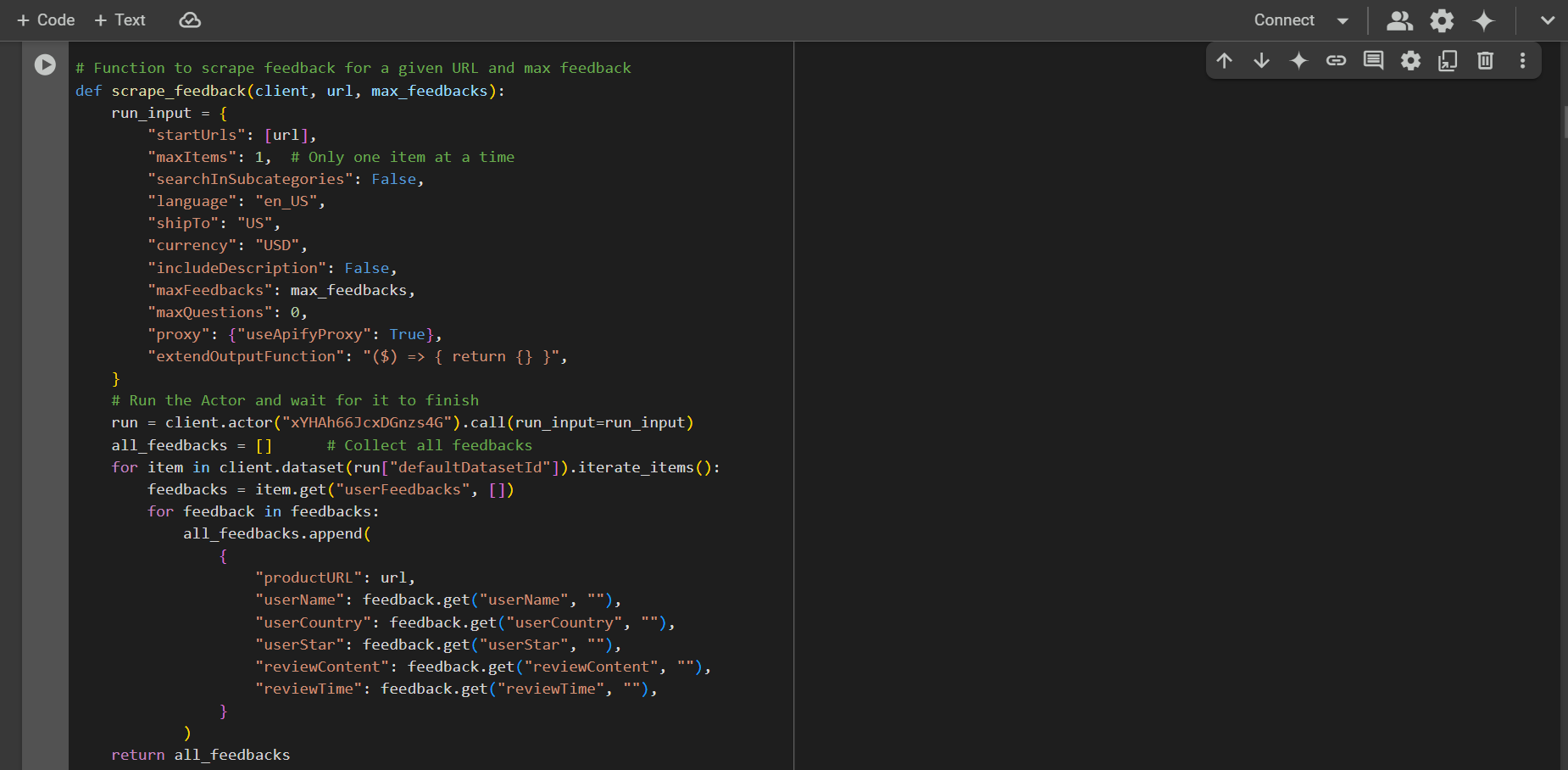


Fig-1: Function to scrape feedback for a given URL and max feedback

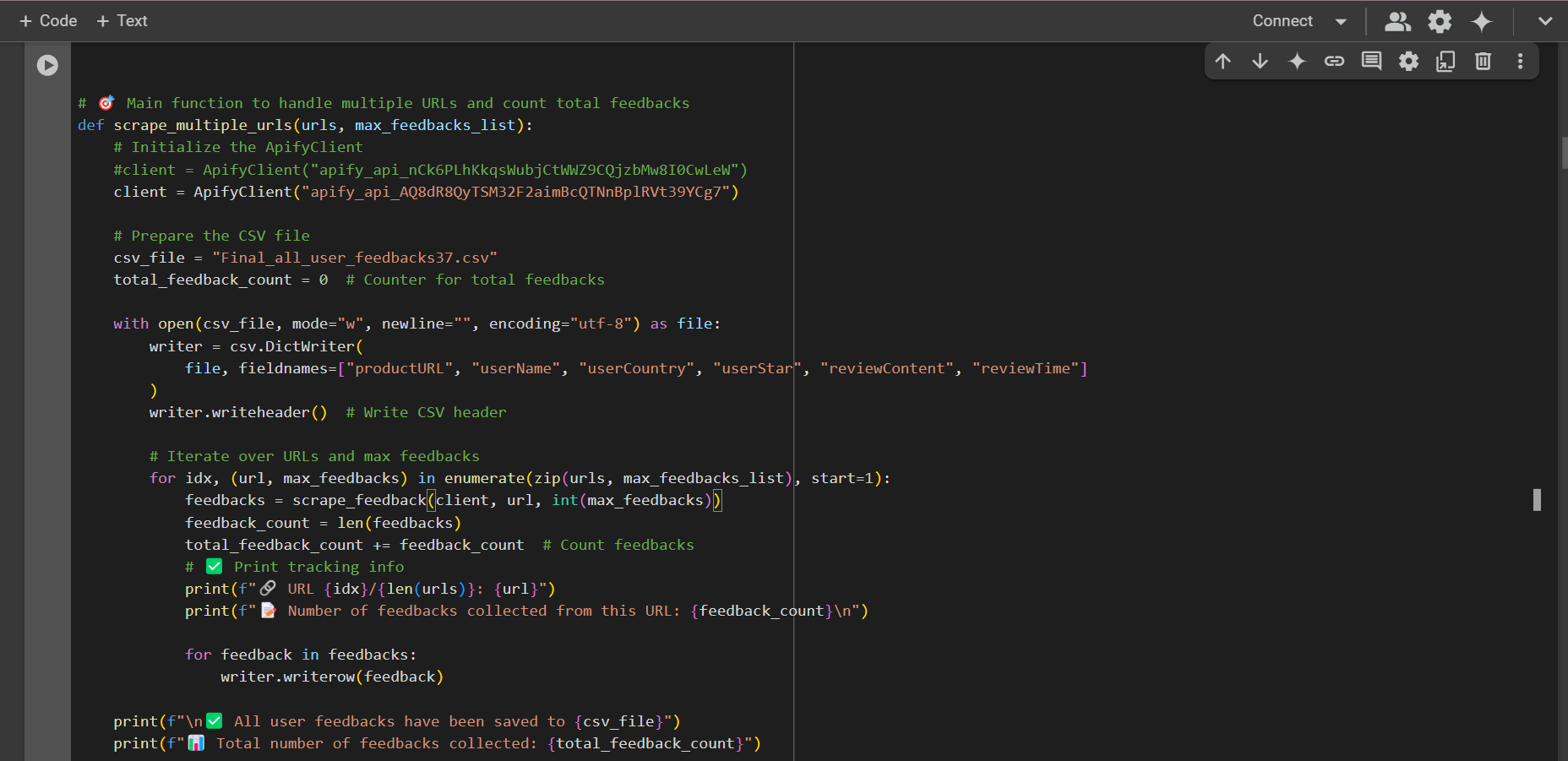


Fig-2: Main function to handle multiple URLs and count total feedbacks

**4. Results**

The pre-processing operation established a structured format of clean data from raw scraped reviews which made them ready for subsequent analysis. Here are the key outcomes:

1. **Dataset Size Reduction**:
   * Originally the raw dataset contained **22,915** reviews. The English review filter resulted in the reduction of **17,586** rows to make the data relevant for analysis.
2. **Language Detection**:
   * The *langdetect* library detected the languages found in each review. The analysis retained only reviews which the system recognized as English to maintain a uniform dataset.
3. **Text Preprocessing**:
   * The reviewContent text received the following preprocessing method:
     + Convert text to lowercase.
     + Remove punctuation and extra whitespace.
     + Tokenize the text and remove English stopwords.
     + Apply lemmatization to normalize words (e.g., "running" → "run").
4. **Final Dataset**:
   * The preprocessed data stored as processed\_user\_reviewContent.csv contained the following columns for the CSV file:
     + productURL, userName, userCountry, userStar, reviewContent, reviewTime, and language.
5. **Efficiency**:
   * Advanced quantitative analysis of sentiment and topics within the dataset becomes feasible after dataset preprocessing while eliminating the need for extra cleaning operations.

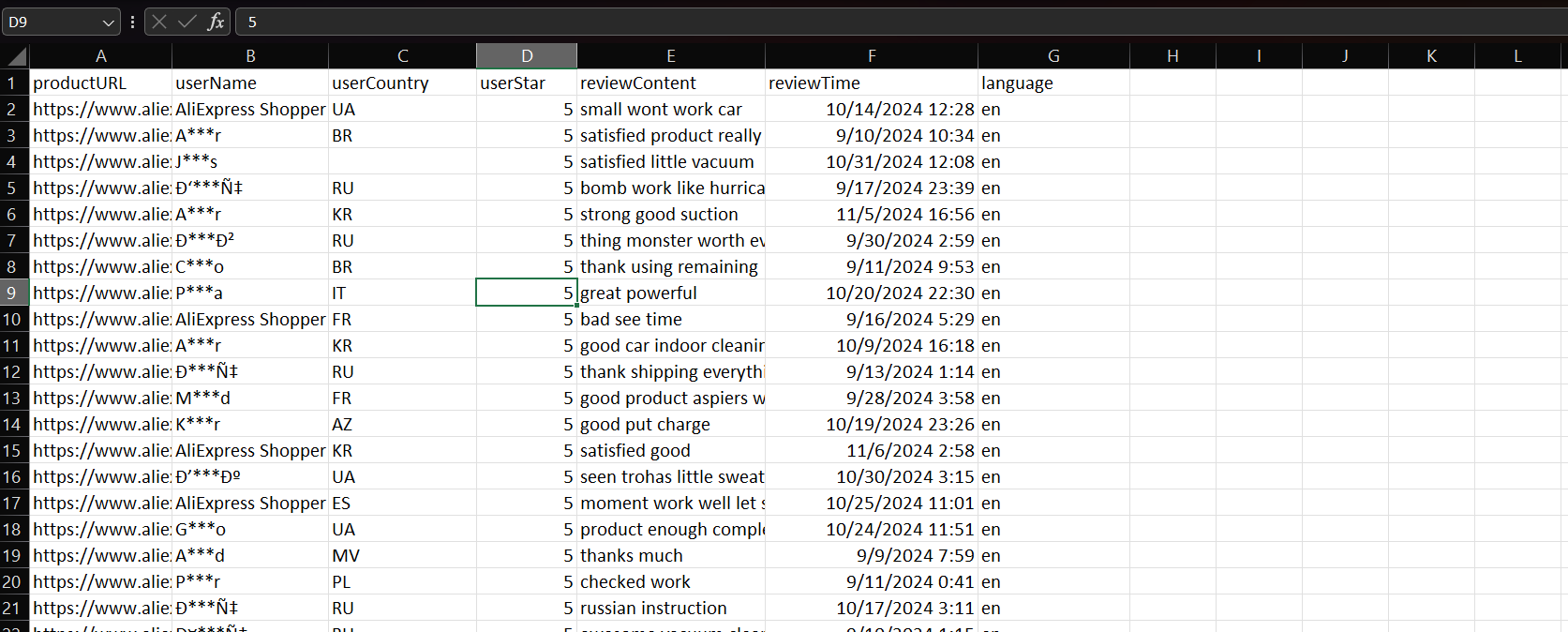


Fig-3: Preprocessed Datasets

This preprocessing step has significantly improved the quality and usability of the dataset, making it suitable for the next stages of the thesis.

**5. Conclusion**

Web scraping enabled the successful acquisition of product reviews from AliExpress. The automated scraping approach delivers scalable operation which enables researchers to gather extensive datasets from various product sections. The well-organized CSV file produced by this method allows subsequent analysis in this thesis.