Project Directory Structure Guide

May 30, 2025

Overview

This document outlines the folder structures and purposes of three main components of the project:

```
+-- 'code\_Scrap\_process\_Instagram_data'
+-- 'code\_Scrap\_process\_Ebay_data'
+-- Dashboards
```

1. Instagram Data Pipeline

Folder: code_Scrap_process_Instagram_data

```
code_Scrap_process_Instagram_data/
+-- code_Insta/
|   +-- data/
|   +-- devide.py
|   +-- download_images.py
|   +-- encode_image.py
|   +-- HDBSCAN_clusters.py
|   +-- images/
|   +-- model_train/
|   +-- models/
|   +-- run.py
|   +-- scrap_insta_data.py
|   +-- segment-anything/
|   +-- segment.py
|   +-- sentiment_analysis.py
|   +-- temp/
|   +-- requirements.txt
+-- data/
+-- myenv/
+-- Readme
+-- requirements.txt
```

Purpose: Complete pipeline to scrape Instagram data, process images and text, segment visuals, analyze sentiment, and train models.

Key Components:

- scrap_insta_data.py – Scrape Instagram data (images, captions, hashtags).

- download_images.py – Download and store images.

- encode_image.py – Resize, normalize, and encode images.

- HDBSCAN_clusters.py – Cluster data (image features or text).

- `segment.py` – Apply image segmentation.
- `sentiment_analysis.py` – Analyze textual sentiment.
- `run.py` – Main pipeline controller.
- `model_train/` – ML training resources.
- `segment-anything/` – Advanced segmentation tools.

2. eBay Fashion Data Pipeline

Folder: `code_Scrap_process_Ebay_data`

```
code_Scrap_process_Ebay_data/
+-- app/
+-- data/
|   +-- raw/
|   +-- processed_v1/
|   +-- processed_v2/
+-- models/
|   +-- categorize.py
|   +-- sentiment_analyser.py
+-- src/
|   +-- cleaner.py
|   +-- ebay_cloths.py
+-- notebooks/
|   +-- product_scraper.ipynb
|   +-- cleaning.ipynb
|   +-- data_enhancement.ipynb
|   +-- apply_ai_models.ipynb
+-- main/
|   +-- insert_data.py
|   +-- processing.py
```

Purpose: A modular pipeline for collecting, cleaning, analyzing, and enriching fashion product data, with future support for automation.

Key Components:

- `data/` – Raw and versioned processed datasets.
- `models/` – AI tools for product categorization and sentiment analysis.
- `src/` – Scripts for cleaning and eBay-specific data extraction.
- `notebooks/` – Jupyter notebooks for interactive experimentation.
- `main/` – Automation scripts for inserting and processing data.

3. Dashboards

Folder: `Dashboards`

Purpose: Power BI dashboard files used for visualizing insights from both Instagram and eBay data pipelines. These dashboards allow stakeholders to explore trends, sentiment summaries, and product categories in an interactive way.

Conclusion

This organized structure enables efficient development and deployment of social media and e-commerce data processing pipelines using Python, machine learning, and visualization tools.