

CS410 Course Project Report

LLM-Augmented Review-Aware Search for Amazon Appliances

Course: CS410 – Text Information Systems

Semester: Fall 2025

Team members:

- Tina Harter (tharter2)
- Dezhao Li (dezhaol2)
- Ziyuan Gu (ziyuang3)

Overview

Our project builds a small information retrieval system for the **Amazon Appliances** collection. The goal is to let a user:

1. Ask a natural language query such as
"I need a quiet compact dishwasher under \$400 for a small apartment."
2. Get a ranked list of relevant products.
3. Pick one product and quickly understand its user reviews through:
 - rating statistics and star distribution, and
 - an aspect-based summary of opinions.

The system combines traditional IR (BM25 + Pyserini) with a local LLM (Ollama) and is exposed as an interactive command-line tool.

Data and Preprocessing

We use two course datasets:

- `meta_Appliances.jsonl` – product metadata
- `Appliances.jsonl` – user reviews

Two scripts in `utils` clean the data:

- `clean_meta_appliances.py`
 - Normalizes IDs, titles, categories and descriptions.
 - Keeps only useful fields (product_id, title, brand, price, average_rating, rating_number, description).
 - Writes a compact `meta_Appliances_cleaned.jsonl` for indexing.
- `clean_appliances_reviews.py`
 - Keeps only verified reviews and removes very short ones.
 - Normalizes review title/body and stores rating, text, helpful votes, etc.

- Outputs `Appliances_cleaned.jsonl`.

These cleaned files are the basis for retrieval and review analysis.

System Design

The code is organized as:

- `utils/retriever.py` – corpus preprocessing and Pyserini BM25 retriever (+ LLM-based compression).
- `utils/llm.py` & `utils/config.py` – Ollama integration and configuration.
- `utils/review_loader.py` – load all reviews for a given `product_id`.
- `utils/state.py` – LangGraph state definition.
- `agent.py` – graph with retrieval and answer-generation nodes.
- `main.py` – two-stage CLI.

Indexing & Retrieval

We preprocess the cleaned metadata into a Pyserini-compatible corpus and build a Lucene BM25 index. A custom `PyseriniBM25Retriever` returns `Document` objects with both text and metadata (id, title, price, ratings). We wrap it in a `ContextualCompressionRetriever` so the LLM sees only the most relevant snippets.

LLM usage

A local Ollama model (e.g., `llama3.2`) is used for:

1. Compressing retrieved documents.
2. Generating a ranked product list for the user query.
3. Summarizing reviews for a single product.

User Interaction

`main.py` implements a two-stage interaction:

Stage 1 – Product search

```
Please enter your question (or type 'exit' to quit): I need a quiet compact dishwasher under $400 for a small apartment.
Answer: 1. Whirlpool WDF518SAAM 18" Stainless Steel Full Console Dishwasher - Energy Star
  - ID: B009B2JHXE
  - Rating: 1.0 stars (1 reviews)
  - Description: Compact stainless steel tall tub dishwasher that offers big features and convenience all in one.
2. Portable Countertop Dishwasher, 5-Liter Compact Portable Countertop Dishwasher 360° Streak-Free Deep Cleaning for Small Apartments, Dorms and RVs
  - ID: B08RRY15GC
  - Rating: 2.6 stars (4 reviews)
  - Description: Compact portable dishwasher ideal for small apartments, RVs, and compact spaces.
3. IQUIRE Portable Dishwasher, 7 Washing Programs, PTC Air-Dry, Anti-Leakage, Fruit & Vegetable Soaking, Topload Compact Dishwasher for Apartments, Dorms and RVs
  - ID: B09ZHFVGV
  - Rating: 3.9 stars (25 reviews)
  - Description: Compact dishwasher with big LED touch button for easy operation.
4. DELLA Mini Compact Countertop Dishwasher 6 Place Settings Portable For Small Apartment Home Kitchen, White
  - ID: B07GBGFNV1
  - Rating: 4.0 stars (36 reviews)
  - Description: Della mini compact countertop dishwasher with small setting capacity plate for office RV condo apartment kitchen.
5. DELLA Compact Mini Dishwasher with 6 Wash Cycles Small Setting Capacity Plate for Office RV Condo Apartment Kitchen, Silver
  - ID: B07GBHS24F
  - Rating: 3.9 stars (26 reviews)
  - Description: Compact dishwasher under $400 for a small apartment.
```

1. User enters a free-text query.
2. The LangGraph agent calls the retriever and LLM. The LLM returns a short ranked list, e.g.:

1. Product Title A - ID: B0XXXXXXX

2. Product Title B - ID: B0YYYYYYY

3. We parse the list numbers and store a mapping from rank \rightarrow (product_id, title).

Stage 2 – Single-product review exploration

```
Enter a product number (1, 2, 3...) to retrieve its reviews, or type 'back' to ask a new question: 1
Searching for reviews of product ID: B01FC1HG5S in dataset/Appliances_cleaned.jsonl...

--- REVIEWS FOR: Best Ever Reusable K Cup and K Carafe for Keurig 2.0 - K200, K300, K400, K500 Series. (ID: B01FC1HG5S) ---

--- REVIEW STATS ---
Total reviews with rating: 36
Average rating: 4.31 / 5
5★: 25 ( 69.4%) #####
4★: 4 ( 11.1%) ##
3★: 3 ( 8.3%) #
2★: 1 ( 2.8%)
1★: 3 ( 8.3%) #
--- END STATS ---

Review 1 (Rating: 5.0):
Title: Works with Series 2!
Text: Key issue ..does it work with new Series 2 .. And Yes! Fully with out a question. If I had a question.. What brand are people using to get a stron
.

Review 2 (Rating: 5.0):
Title: Perfect replacement and alternative for K cups!
Text: Works great! The tops come off pretty easily so be prepared for that, but these definitely help you fool the K2.0 to thinking you are using real k
.

Review 3 (Rating: 5.0):
Title: It works
Text: Works great with the K250 2.0. I purchased 2 and my mom loves hers as well...

Review 4 (Rating: 3.0):
Title: Works half way
Text: I purchased this set not know what model my kurig is. I thought it was compatible but only the small single purple cup works in mine. The larger ora
.

Review 5 (Rating: 5.0):
Title: Works great in my Keurig 2.0 K250
Text: I have a Keurig 2.0 K250 model. This filter has been working without any problems for over a month, and I would highly recommend it, since you get t
.

Showing 5 of 36 total reviews found.
```

When the user types a number (e.g., 1):

1. We load all reviews for the corresponding product_id using review_loader.py. We compute rating statistics and star distribution:

```
--- REVIEW STATS ---
Total reviews with rating: 36
Average rating: 4.31 / 5
5★: 25 ( 69.4%) #####
4★: 4 ( 11.1%) ##
3★: 3 ( 8.3%) #
2★: 1 ( 2.8%)
1★: 3 ( 8.3%) #
--- END STATS ---
```

2. We print a few raw reviews (rating, title, truncated text) for transparency.

3. We call the LLM with an **aspect-based summary prompt**. The output has the structure:

- Overall Verdict
- Pros / Cons
- Performance/Cleaning
- Noise
- Ease of Use
- Build Quality & Durability
- Price/Value for Money

```
----- GENERATING OVERALL SUMMARY -----

**Verdict:** This compact dishwasher offers efficient cleaning and various washing programs, but its reliability is questionable.

**Pros:**
* Portable and compact design suitable for small kitchens, apartments, dorms, and RVs.
* Variety of 7 washing programs, including Fruit & Vegetable Soaking, which is a unique feature.

**Cons:**
* Reliability issues with error codes (E5) that can be frustrating to troubleshoot and resolve.
* Some users report compatibility issues with certain faucets or water supplies.

**Aspect Breakdown**

- **Performance/Cleaning**
  - Summary: Users praise the dishwasher's efficiency in cleaning dishes, but some mention occasional issues with hot water temperature during cycles.

- **Noise**
  - Summary: There is no notable mention of excessive noise or quietness in reviews.

- **Ease of Use**
  - Summary: While users appreciate the variety of washing programs and easy operation, some report difficulties with error codes (E5) that require troubleshooting.

- **Build Quality & Durability**
  - Summary: Users' concerns about reliability issues and potential compatibility problems raise questions about the dishwasher's long-term durability.

- **Price/Value for Money**
  - Summary: Some users feel that the product offers good value for its price, while others seem to have had a disappointing experience.

----- END SUMMARY -----
```

This helps the user quickly see, for example, whether a dishwasher is actually quiet and how reliable it is over time.

Conclusion

In this project we built a small but complete IR + text mining system for the Amazon Appliances collection. From raw JSONL data, we cleaned and indexed product metadata, used BM25 + Pyserini for retrieval, and integrated a local LLM to rank products and generate aspect-based summaries. The two-stage CLI with rating statistics and star distributions lets a user go from a vague natural language need to concrete product options and a quick understanding of real user opinions.

Doing this project made the ideas from CS410 much more concrete. We had to apply document representation and indexing in practice, tune retrieval behavior, and deal with noisy, real-world data. At the same time, integrating an LLM showed us how “classic” IR and modern language models can work together: retrieval finds relevant items, while the LLM helps summarize and explain them. Overall, this gave us hands-on experience building an end-to-end system and a deeper, more practical understanding of the course material.