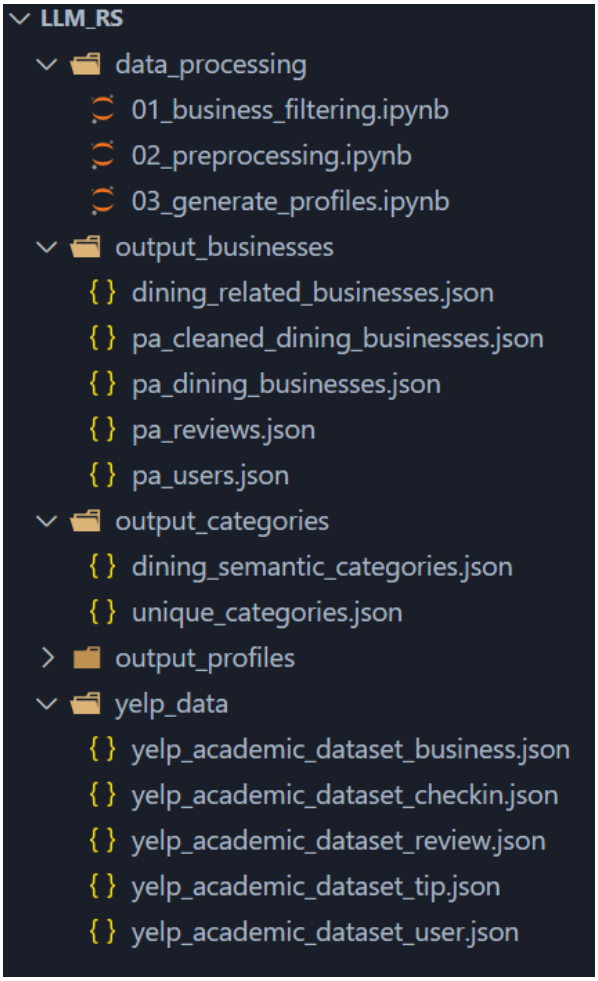
2.12-2.18 Weekly Record (Technical)

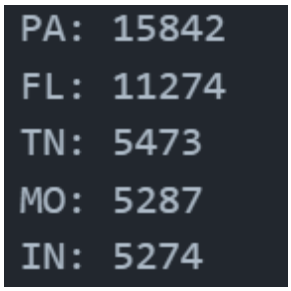
Restructure



Filter Dining Businesses

✅ (What has been done before) **Extract Unique Categories- Identify Dining Categories with Transformer- Filter All Dining Businesses**

✅ **Focus on one area- Select the best candidate**



Scale: PA>FL

Review Counting, Rating Distribution, Business Diversity: PA ≈ FL

✅ Filter Dining Businesses Located in **PA**

Process PA_Dining Businesses Data

✅ **Clean PA Data:**

Remove **closed businesses**. //Check whether 'is_open=0' signifies temporary closure or permanent closure.

Remove businesses with **low average ratings (stars < 2.0)**. //Try increasing the threshold and note how many businesses are filtered out.

Remove businesses with **zero reviews**. //Try increasing the threshold and note how many businesses are filtered out.

Q: Standard?

Stars Distribution: {4.0: 3977, 4.5: 2562, 3.0: 2405, 3.5: 3518, 2.5: 1482, 5.0: 590, ***1.5: 368, 2.0: 867, 1.0: 73****}*

Avg Review Count (Open): 73.41

✅ **Filter PA_Reviews & PA_User**

Only related to PA

Discard unnecessary attributes like `friends`, `compliments`, etc to save costs

Prepared for following profile generation

User & Item Profile Generation

Item Profile

Source	Field	Value Analysis
business	name	Essential – Represents basic business information
	categories	Essential – Reflects the type of restaurant (e.g., Chinese, BBQ) // 1. Consider additional ways to utilize this structured data beyond just providing descriptive information 2. Explore links between categories (e.g., Pizza & Italian)
	stars	Important structured data – Indicates overall quality, can serve as a **weight** // label, supervised learning
	review_count	Moderately important – Reflects popularity, can serve as a **weight**
	attributes	Optional – Sometimes describes distinctive restaurant features (e.g., parking, delivery)
review	**text**	Essential – High-quality textual description, captures user experience
	stars	Moderately important – Reflects individual user opinions about the business
tip	**text**	((What has been done before: verify that it can) Supplement customer advice or quick suggestions

User Profile

Source	Field	Value Analysis
user	review_count	Moderately important structured data – Reflects user activity level, can serve as a **weight**
	average_stars	Moderately important – Indicates user's overall rating tendency, can serve as a **weight**
	yelping_since	Less important – Reflects user's tenure and experience

review	**text**	Essential – High-quality text, reflects user preferences and habits
	stars	Moderately important – Indicates user's specific evaluation of different businesses

✅ Approach: embedding plus structured data weighting

For example:

Create a profile text for each business by concatenating: name+categories+Top-5 recent review texts + attributes

Encode this profile text into embedding vector using Transformer

Augment the embedding with two numerical weights

Each item profile consists of:

```
{
   "business_id": b_id,
   "embedding": embedding,
   "weight_stars": weight_stars,  # Float in [0, 1]
   "weight_review_count": weight_review_count  # Float in [0, 1]
}
```

Reviews capture personalized experiences and preferences, while structured data such as categories and attributes provide factual information about the businesses.

❌ Computational Limitations

Concatenation - not a problem

Transformer Text Encoding - time intensive

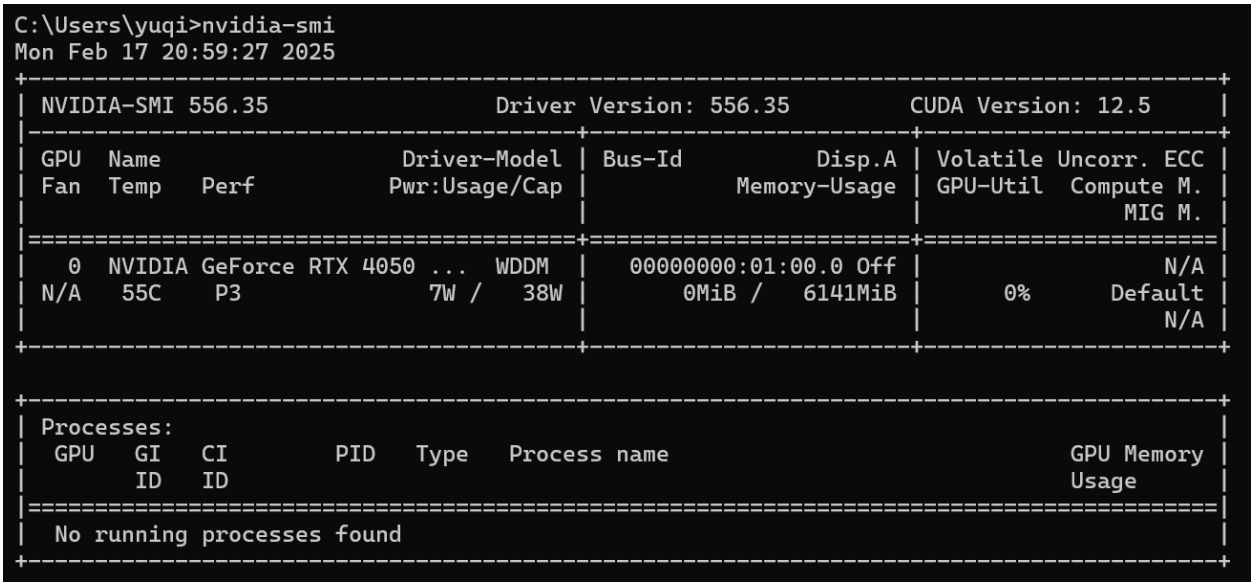
```
from sentence_transformers import SentenceTransformer

model = SentenceTransformer('all-MiniLM-L6-v2')
print(model.get_sentence_embedding_dimension())
```

✓ 56.9s

```
384
```

CPU might not handle it

Solutions: GPU/ batch processing (TODO)



```
C:\Users\yuqi>nvidia-smi
Mon Feb 17 20:59:27 2025
+-----------------------------------------------------------------------------------------+
| NVIDIA-SMI 556.35                 Driver Version: 556.35         CUDA Version: 12.5     |
|-----------------------------------------+------------------------+----------------------+
| GPU  Name                  Driver-Model | Bus-Id          Disp.A | Volatile Uncorr. ECC |
| Fan  Temp   Perf          Pwr:Usage/Cap |           Memory-Usage | GPU-Util  Compute M. |
|                                         |                        |               MIG M. |
|=========================================+========================+======================|
|   0  NVIDIA GeForce RTX 4050 ...  WDDM  |   00000000:01:00.0 Off |                  N/A |
| N/A   55C    P3              7W /   38W |       0MiB /   6141MiB |      0%      Default |
|                                         |                        |                  N/A |
+-----------------------------------------+------------------------+----------------------+

+-----------------------------------------------------------------------------------------+
| Processes:                                                                              |
|  GPU   GI   CI        PID   Type   Process name                              GPU Memory |
|        ID   ID                                                               Usage      |
|=========================================================================================|
|  No running processes found                                                             |
+-----------------------------------------------------------------------------------------+
```

Q: Computational power issue?

Use a small portion of the dataset to validate the feasibility of the approach on my own machine first, then request GPU resources at the university

Next:

Solve issues until now+ Technical Approach