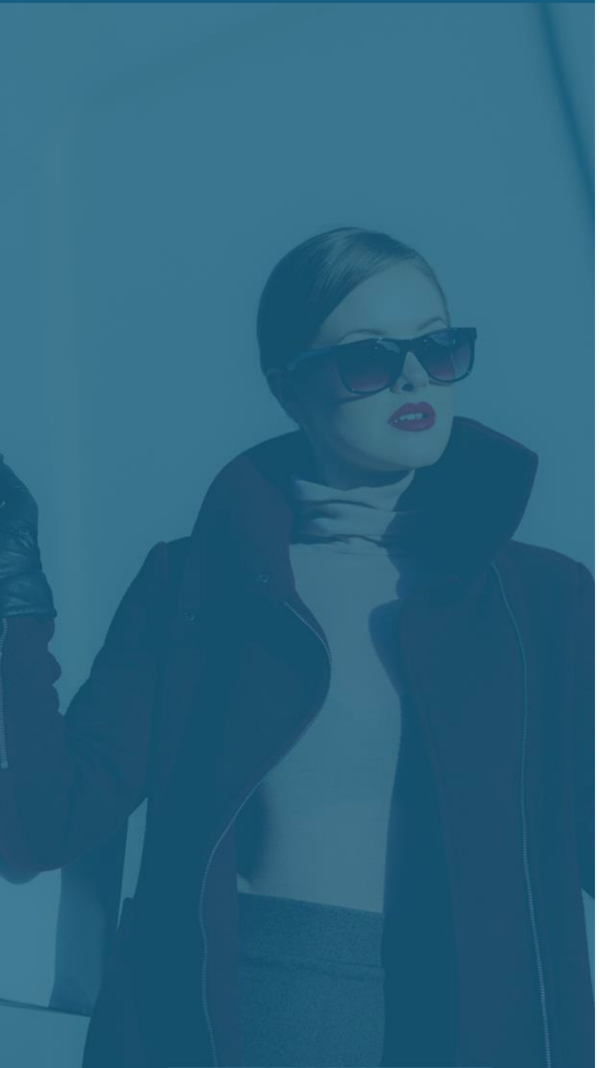
2024 Fall

Fashion Recommender System Based on Texts

Introduction to Recommender Systems Term Project Proposal

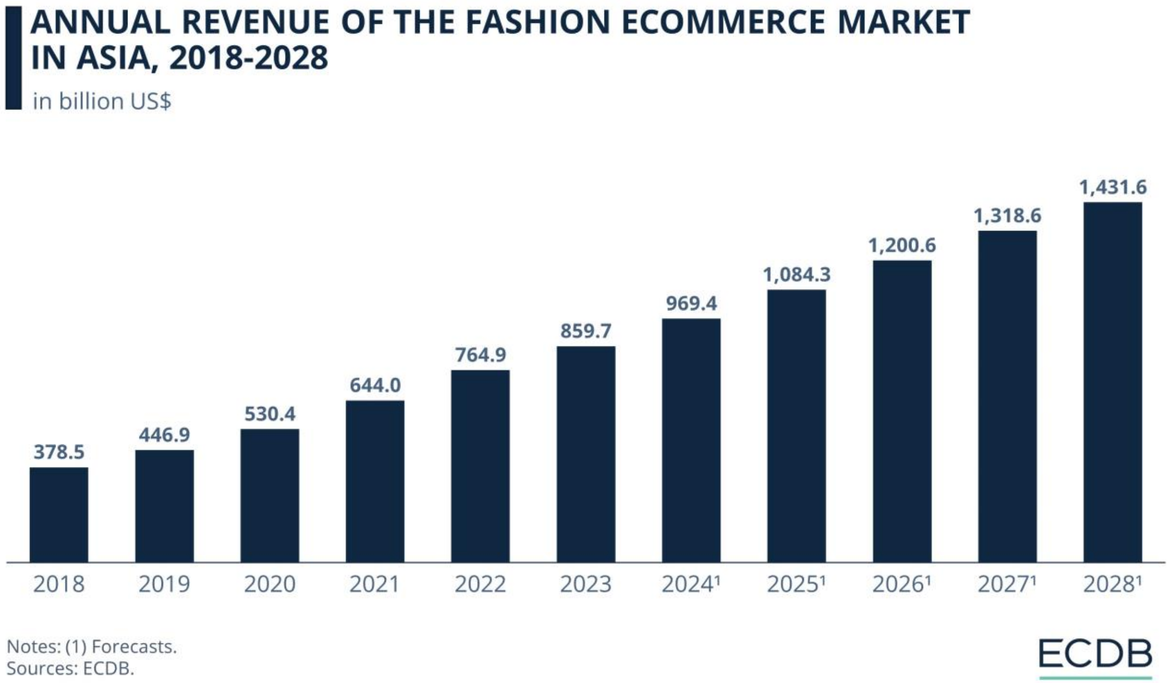
20171664 Lee Sangyun
20200977 Han Youngtae

Table of Contents



1. Background & Motivation

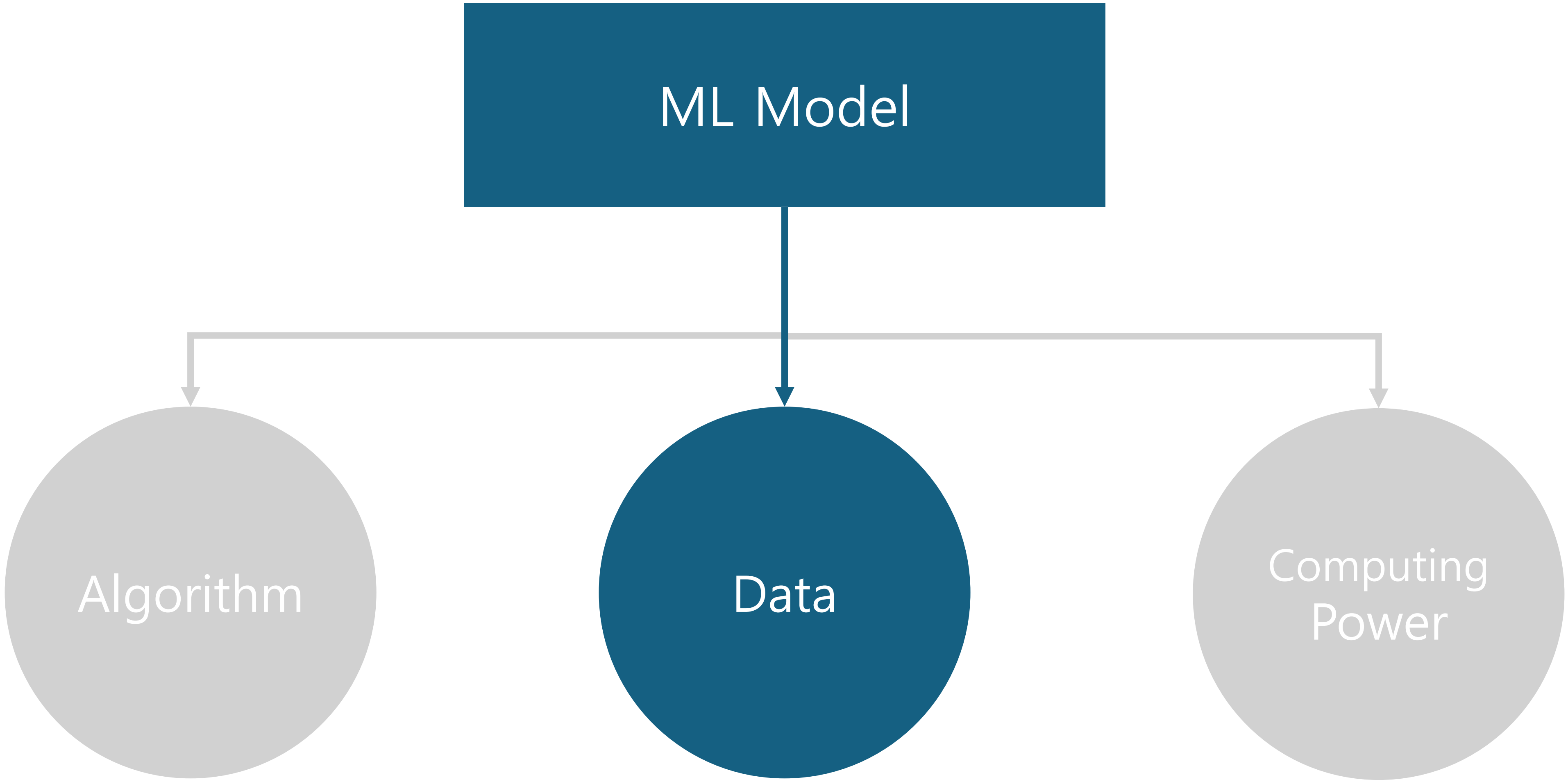
Fashion Domain



The global fashion e-commerce market has rapidly grown.
In addition, the market is expected to grow faster in the future.

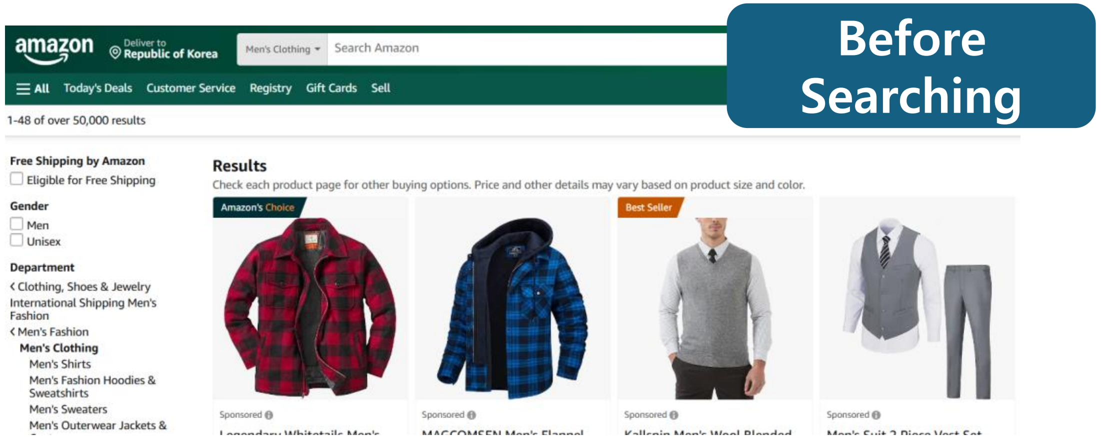
1. Background & Motivation

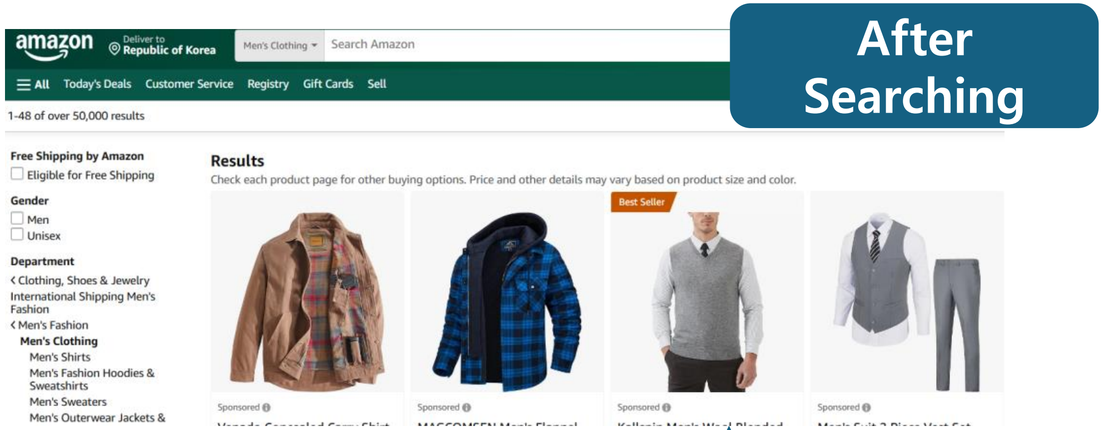
Components of ML

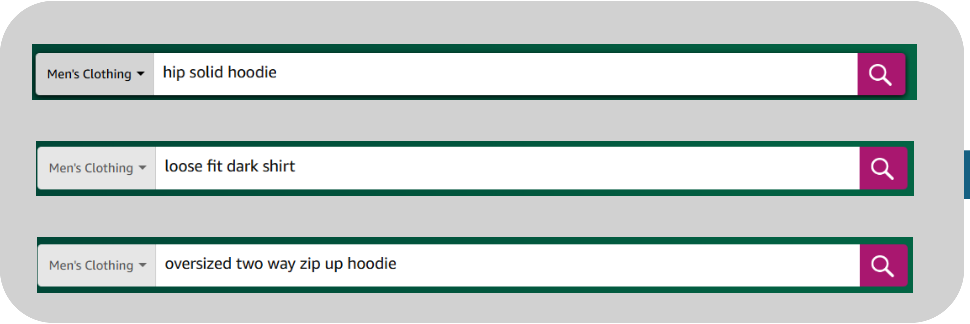


2. Problem & Goal of the Project

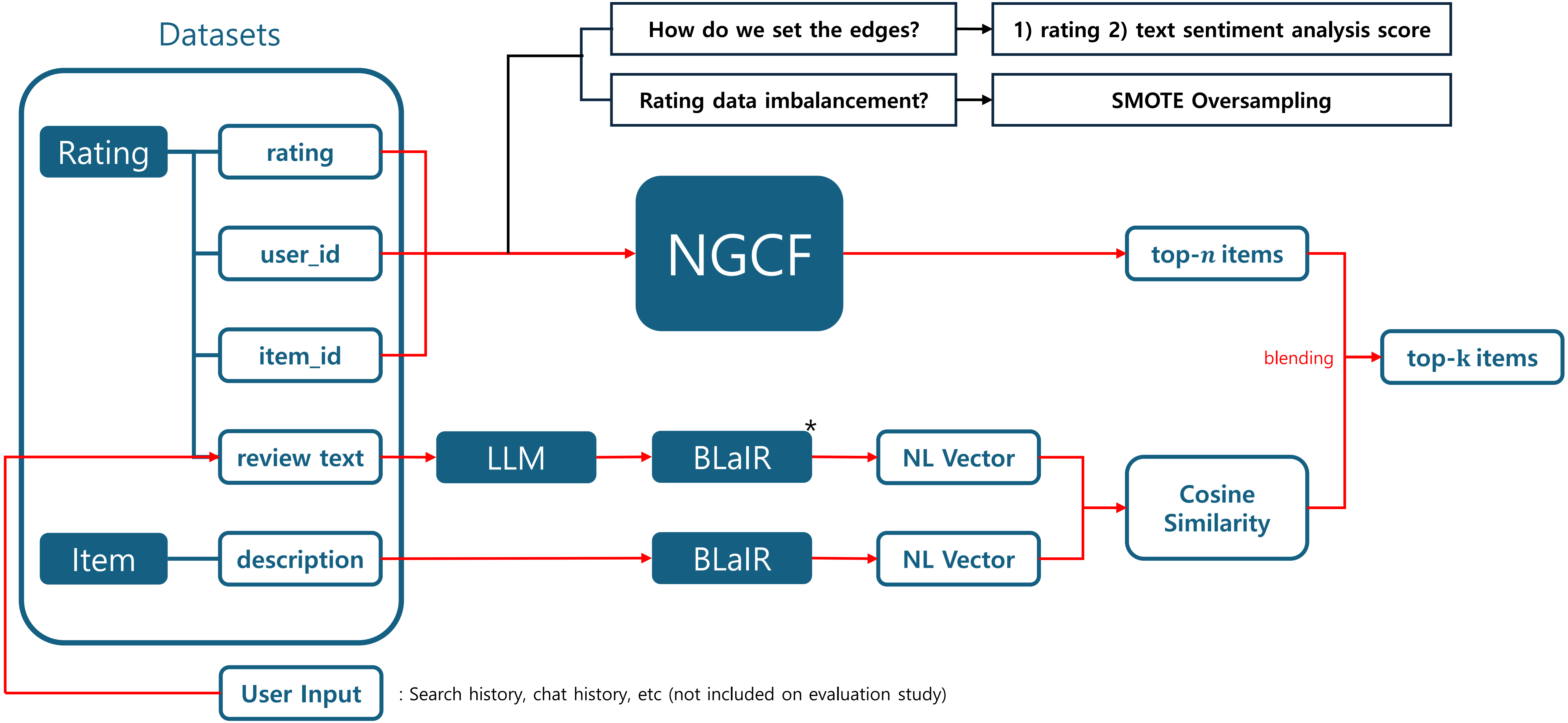
Problem & Goal







3. Model Pipeline



: Search history, chat history, etc (not included on evaluation study)

*Bridging Language and Items for Retrieval and Recommendation (arXiv, 2024)

3. Model Pipeline

LLM Processing Example

Input :
My son is a big Ninjago fan and these are perfect! The color is vibrant and hasn't faded in the wash. It also is made well and the pocket inserts and filters made it perfect! It looked a bit bigger than expected but once the filter was it, it was perfect to show off the entir e design and was a good size for my 6 year old who is closer to the size of an 8/9 year old.

LLM

Prompt Engineering

```
{"role": "system", "content": "You are a assistant for reorganizing texts. The inputs are about fashion item reviews."},
{"role": "system", "content": "You reorganize the text focusing on the feature of the fashion items."},
{"role": "system", "content": "Describe the features that the user likes and wants. Exclude emotional modifiers."},
{"role": "system", "content": "Write it in a continuous text format, in the first person, in a stiff tone."},
{"role": "system", "content": "Write like 'I like something', 'I want something', 'I expect something'."},
{"role": "system", "content": "Delete the descriptions that are unrelated to the item's character or feature"},
⋮
```

Output :
I like the vibrant color that hasn't faded in the wash. I want a well-made product with pocket inserts and filters. I expect the product t o be a good size for my 6-year-old child who is closer to the size of an 8/9-year-old.

4. Dataset & Evaluation

Datasets

amazon

Amazon Reviews'23

Evaluation

Recall@k

MAP@k

5. Project Schedule

Timeline

November

December

Data Preprocessing

Model Building

Evaluation Study
Ablation Study

Member's role

Lee Sangyun	NGCF Modeling, evaluation
Han Youngtae	Data Preprocessing, NLP with BLaIR, LLM, evaluation

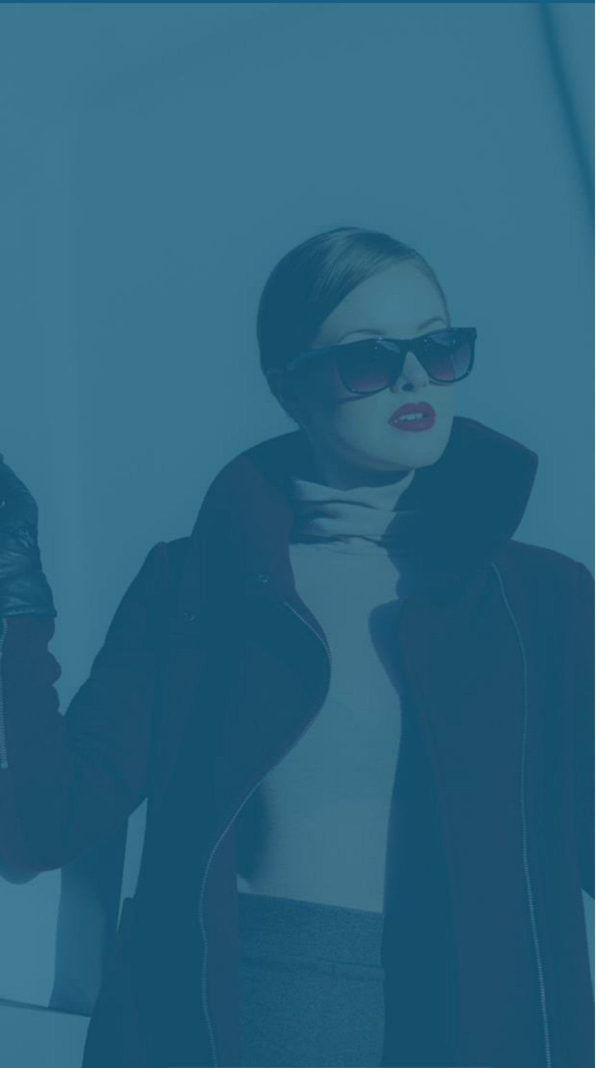
2024 Fall

Fashion Recommender System Based on Texts

Introduction to Recommender Systems Term Project Final Presentation

20171664 Lee Sangyun
20200977 Han Youngtae

Table of Contents



1. Research Environment

Nvidia Quadro RTX 6000

```
work@main1[Member_Server]:~/default/DO_NOT_ACCESS$ lspci | grep -i nvidia
06:00.0 3D controller: NVIDIA Corporation TU102GL [Quadro RTX 6000/8000] (rev a1)
2f:00.0 3D controller: NVIDIA Corporation TU102GL [Quadro RTX 6000/8000] (rev a1)
86:00.0 3D controller: NVIDIA Corporation TU102GL [Quadro RTX 6000/8000] (rev a1)
d8:00.0 3D controller: NVIDIA Corporation TU102GL [Quadro RTX 6000/8000] (rev a1)
```

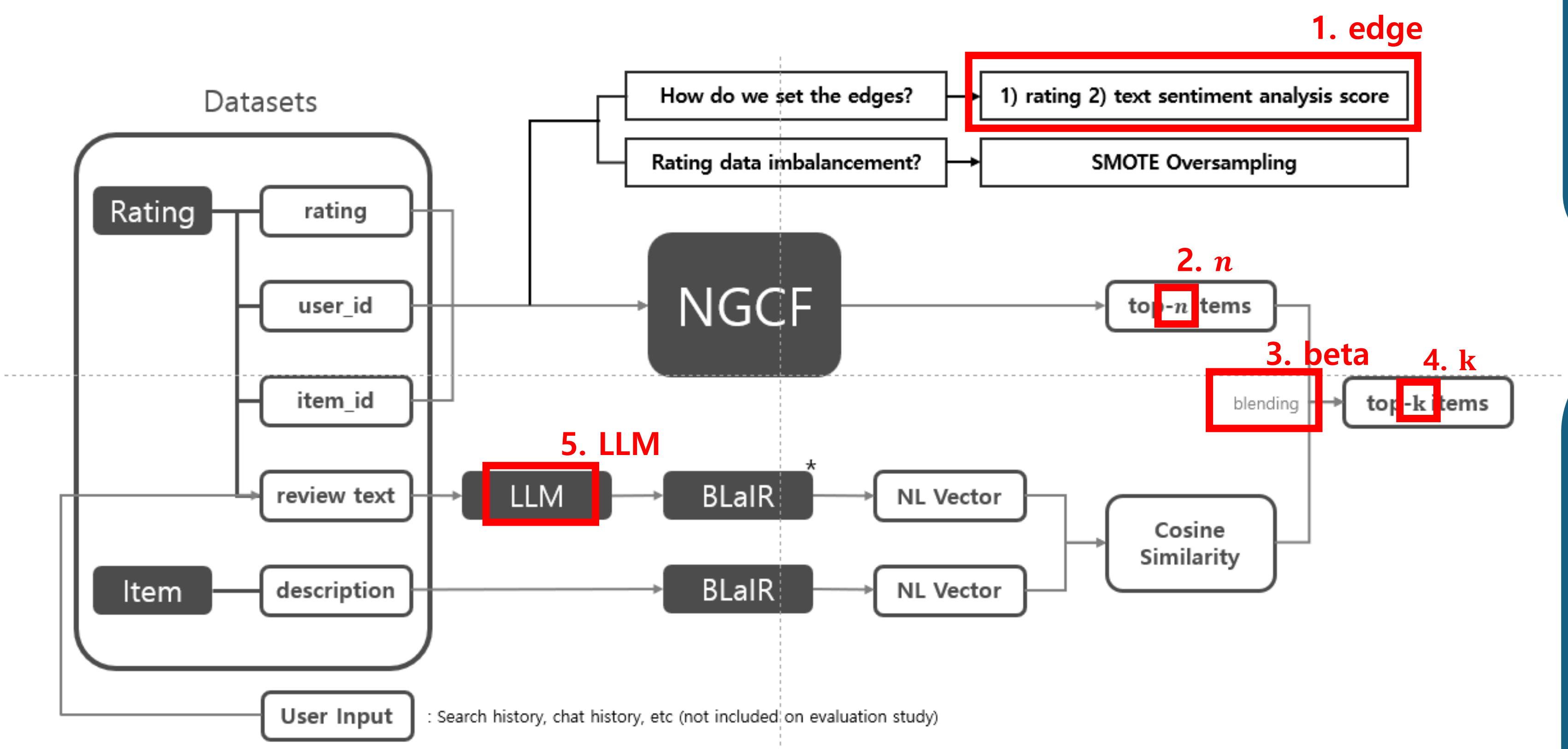
2. Experiments

Preprocessing

1	Drop unnecessary columns and NaNs
2	Drop users and items that the number of purchase is less than 3
3	Sentiment Analysis: positive(0~1), negative(-1~0)
4	Train / Test split -> 0.8 / 0.2

2. Experiments

Parameters



Baseline

1. edge = [rating, sent, blend]
2. k = [50, 100]

3 x 2 = 6 models

Ours

1. edge = [rating, sent, blend]
2. n = [100, 500, 1000]
3. beta = [0.25, 0.5, 0.75]
4. k = [50, 100]
5. LLM = [with, w/o]

3 x 3 x 3 x 2 x 2 = 108 models

3. Experiments Results

Evaluation

(unit: %)

edge	n	Recall@k						MAP@k					
		β=0.25		β=0.5		β=0.75		β=0.25		β=0.5		β=0.75	
		k=50	k=100	k=50	k=100	k=50	k=100	k=50	k=100	k=50	k=100	k=50	k=100
rating	100	1.1793	1.4976	1.1606	1.4976	1.1606	1.4976	0.0755	0.0530	0.0589	0.0447	0.0497	0.0401
	500	3.2011	4.1558	3.0700	4.0996	3.0326	4.0809	0.1949	0.1393	0.1457	0.1140	0.1154	0.0984
	1000	4.5676	5.9716	4.3804	5.8967	4.1932	5.8405	0.2700	0.1948	0.2034	0.1602	0.1587	0.1370
sent	100	1.7971	2.0217	1.7784	2.0217	1.6848	2.0217	0.1144	0.0788	0.9942	0.0713	0.0724	0.0575
	500	3.5380	4.6799	3.2572	4.3429	1.9281	3.3321	0.1903	0.1418	0.1410	0.1129	0.0777	0.0682
	1000	4.2868	6.0090	3.7626	5.2415	1.9468	3.4259	0.2348	0.1761	0.1562	0.1275	0.0779	0.0688
blend	100	1.0483	1.2729	0.8986	1.2729	0.5616	1.2729	0.0499	0.0383	0.0265	0.0260	0.0149	0.0185
	500	2.6769	4.3617	1.3104	2.9015	0.5616	1.8532	0.1038	0.0927	0.0302	0.0393	0.0149	0.0200
	1000	3.1823	5.5410	1.3104	2.9951	0.5616	1.8532	0.1193	0.1092	0.0302	0.0398	0.0149	0.0200

[Best Parameters] edge: (rating or sent), n: 1000, beta: 0.25

3. Experiments Results

Result by parameters

Baseline – by edge

edge	recall	map
blend	0.842381	0.012696
rating	1.244852	0.033058
sent	1.544365	0.038745

Ours – by n

n	recall	map
100.0	1.424768	0.054987
500.0	3.132410	0.102243
1000.0	3.942552	0.127693

Ours – by edge

edge	recall	map
blend	1.909397	0.043203
rating	3.248832	0.120353
sent	3.085282	0.110818

Ours – by beta

beta	recall	map
0.25	3.388244	0.132048
0.50	2.844336	0.090389
0.75	2.267149	0.062485

3. Experiments Results

Ablation Study

(unit: %)

edge	Ablation Settings			Recall@50	Recall@100	MAP@50	MAP@100	Recall@100 Delta
	NGCF	LLM	BLaIR					
rating	✘	✘	✘	0.037439	0.037439	0.000475	0.000664	-
	✔	✘	✘	0.992138	1.497567	0.034915	0.031201	39x
	✔	✘	✔	4.567578	5.971546	0.270034	0.194770	158.5x
	✔	✔	✔	4.567578	5.971546	0.270034	0.194770	158.5x
sentiment	✘	✘	✘	0.037439	0.037439	0.000475	0.000664	-
	✔	✘	✘	1.067016	2.021715	0.040268	0.037221	53x
	✔	✘	✔	4.286784	6.008985	0.234762	0.176141	159.5x
	✔	✔	✔	4.286784	6.008985	0.234762	0.176141	159.5x

4. Conclusion and Limitations

Conclusion

1. Our model showed approximately 3-4 times better performance than the baseline and about 160 times better than random recommendation in terms of recall@100.

2. The noise removal process using LLM was unnecessary.
   The natural language embedding model had already learned enough to handle the noise.

Limitations and further study proposal

1. The dataset was too sparse because we used explicit feedback dataset.

2. We tested only one natural language embedding model (BLaIR).

3. Among diverse user inputs, we could only use reviews.
   i.e. we could NOT use other data like search histories.