What's in My Scent?
-STS based perfume recommendation system

Final Project Report
by Dohoon Kim

Table of contents

Introduction: Background

“후각이라는 감각은 오랫동안 차별과 냉대를 받아왔고, 그 언어적인 표현에 있어서도 발달하지 못했다“-(서종석 2012)

We use graph neural networks (GNN) to generate a Principal Odor Map (POM) that preserves perceptual relationships and enables odor quality prediction for novel odorants.-(Brian K.Lee et al. 2022)

Introduction: Objective

Input:
Ambiguous expression
about fragrance

Sentence
Transformer

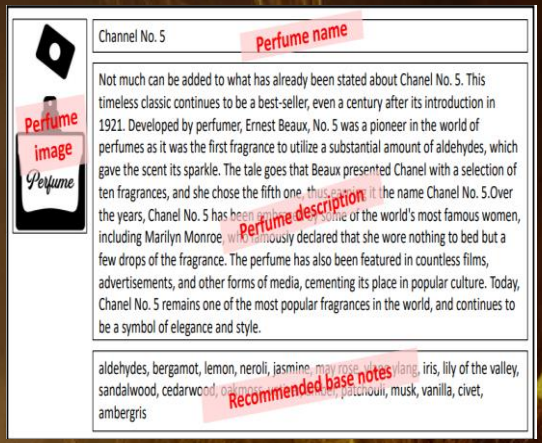
Output:
Clear Products

Training: Original Data

FRAGRANTICA

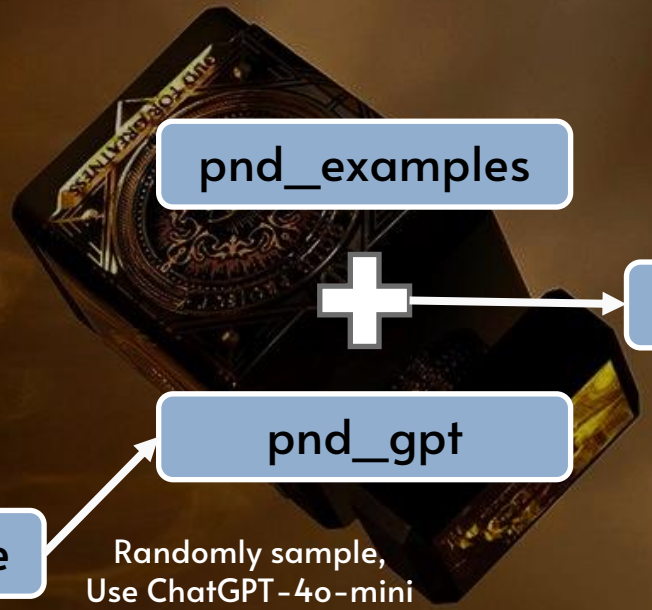
kaggle™

olgagmiufanal/fragrantica-com-fragrance-dataset/fra_cleaned.csv



Perfume and description(PND) (Jooyoung Kim et al.)

Training: Data Preproduction



fra_cleaned

Add description, integrated notes, regularize

fragrantica_database

Use as vector DB for RAG

pnd_examples

pnd_gpt

Randomly sample, Use ChatGPT-4o-mini generate description gorresponding to notes

prediction_train

Training FastText

corpus

Positive pairs(1.0)
Negative pairs(0.0)

training_pairs

Data Preproduction:Database

1. Data Preproduction:

```
def data_prep(self):
    self.data['Rating Value'] = self.data['Rating Value'].astype(str).str.replace(',', '.').astype(float)
    self.data['Perfume'] = self.data['Perfume'].astype(str).str.replace('-', ' ')
    self.data['Brand'] = self.data['Brand'].astype(str).str.replace('-', ' ')
    self.data['Year'] = self.data['Year'].fillna(0).astype(int)
    self.data.replace({'unknown', 'Unknown', 0, 0.0}, np.nan, inplace=True)
    self.data['Year'] = self.data['Year'].fillna(0).astype(int)
    return self
```

Data Preproduction:Database

2. Data Filtering:

```
def data_filtering(self):
    self.data = self.data.drop(self.data[(self.data['Rating Value'] < 3) |
                                          (self.data['Rating Count'] < 50) |
                                          (self.data['Year'] < 2000)].index)
    brand_counts = self.data['Brand'].value_counts()
    valid_brands = brand_counts[brand_counts >= 3].index
    self.data = self.data[self.data['Brand'].isin(valid_brands)].reset_index(drop=True)
    return self
```

Data Preproduction:Database

3. Data Augmentation:

```
def generate_description(self):
    self.data['description'] = self.data.apply(
        lambda row: (
            f"{row['Perfume'].title()} by {row['Brand'].title()} is a {row['Gender']} fragrance featuring top notes of {row['Top']}, "
            f"middle notes of {row['Middle']}, and base notes of {row['Base']}. "
            f"The main accords are {', '.join(filter(pd.notna, [row['mainaccord1'], row['mainaccord2'], row['mainaccord3'], row['mainaccord4'], row['mainaccord5']]))}. "
            f"Released in {row['Year']} from {row['Country']}, this fragrance has a rating of {row['Rating Value']} out of 5 from {row['Rating Count']} votes. "
            f"{('Crafted by perfumer ' + row['Perfumer1'].title() + '.') if pd.notna(row['Perfumer1']) else ''} {(' and ' + row['Perfumer2'].title()) if pd.notna(row['Perfu
        ),
        axis=1
    )
    return self
```

Data Preproduction:pnd_gpt

1. Randomly sample 300+ items from “fragrantica_database.csv”
2. Generate description of each items using “GPT-4o-mini”
3. Create “pnd_gpt” as same format of “pnd_examples”

```
for i in range(len(df_sample_1)):
    prompt_template = f"""You are a fragrance expert tasked with summarizing perfume descriptions based on\
    given data and additional information from the Fragrantica website. \
    Your task is to generate a concise and engaging description of each perfume, \
    focusing on its mood and user reviews. Follow these guidelines:\

    1. **Content**:
      - Focus on the overall mood, olfactory characteristics and users assessment of the perfume.\
      - Summarize user reviews to highlight key impressions (e.g., "elegant," "fresh," "long-lasting").\
      - You can find fundamental features of the perfume from the given description:{df_sample['description'][i]} but
      do not directly copy sentences from that.\
      - You have to retrieve user reviews from the website URL: {df_sample['url'][i]}.
      - Do not explicitly mention the Top, Middle, or Base notes, but imply them through descriptive language\
      (because your description will be used for fine-tuning a model which predicts olfactory notes from the description).

    2. **Tone**: Use professional yet engaging language suitable for a fragrance description.

    3. **Example**:
      - "The first new feminine Creed fragrance for five years (the last was 2000 Fleurs), and the first to be debuted in the USA.
      The scent is inspired by Olivier Creed's travels on the high seas, 'an element of nature that connects all humanity' and according to Creed,
      the bottle evokes the shoulders of a feminine figure as she rests upon white sands caressed by the ocean's gentle current.
      A silver ribbon at the neck conveys sunshine dancing on the Aegean sea."
    """
```

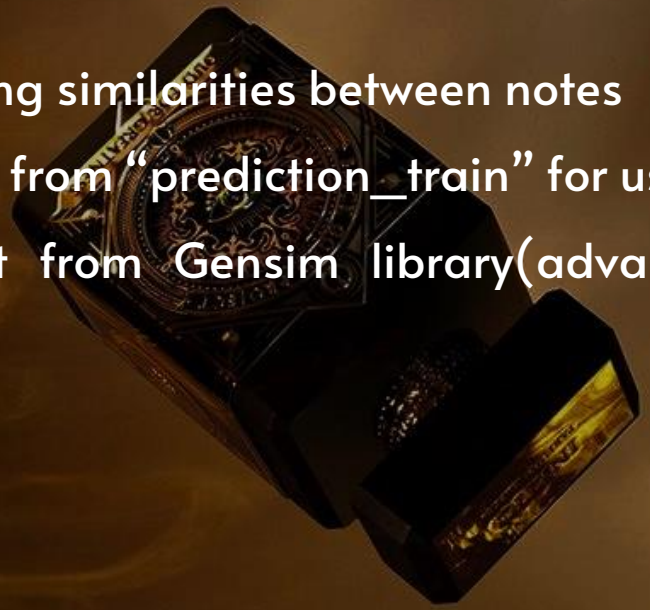
Data Preproduction:prediction_train

1. Merge “pnd_gpt” and“pnd_examples”
2. Standardize overlapping notes using “GPT-4o-mini”
3. Result:

	description	notes
0	Eau D'Italie is an olfactory poem that transpo...	['bergamot', 'blackcurrant', 'incense', 'clay'...
1	Rosa Gallica by Brecourt is an exquisite unise...	['incense', 'pink pepper', 'rose', 'myrrh', 'e...
2	**Rose Blush Cologne 2023 by Jo Malone London*...	['lychee', 'basil', 'rose', 'musk']
3	**Qimmah For Women by Lattafa Perfumes**\n\nIm...	['almond', 'coffee', 'tuberose', 'jasmine', 't...
4	**Alba Di Seoul by Santa Maria Novella** is a ...	['green notes', 'pine', 'woody notes', 'orient...
...	...	...
379	Halfeti draws inspiration from the lavish good...	['grapefruit', 'bergamot', 'green notes', 'arm...
380	Created by perfumer Julie Pluchet working with...	['lime', 'mimosa', 'tuberose', 'hyacinth', 'la...
381	No flowers in this scent! Just a whole load of...	['bergamot', 'strawberry', 'dewberry', 'honey'...
382	This very distinct sweet fragrance is one of t...	['bergamot', 'herbal notes', 'lavender', 'pepp...
383	Inspired by the clean fresh scent of a soap fr...	['bergamot', 'neroli', 'petitgrain', 'orange f...

384 rows × 2 columns

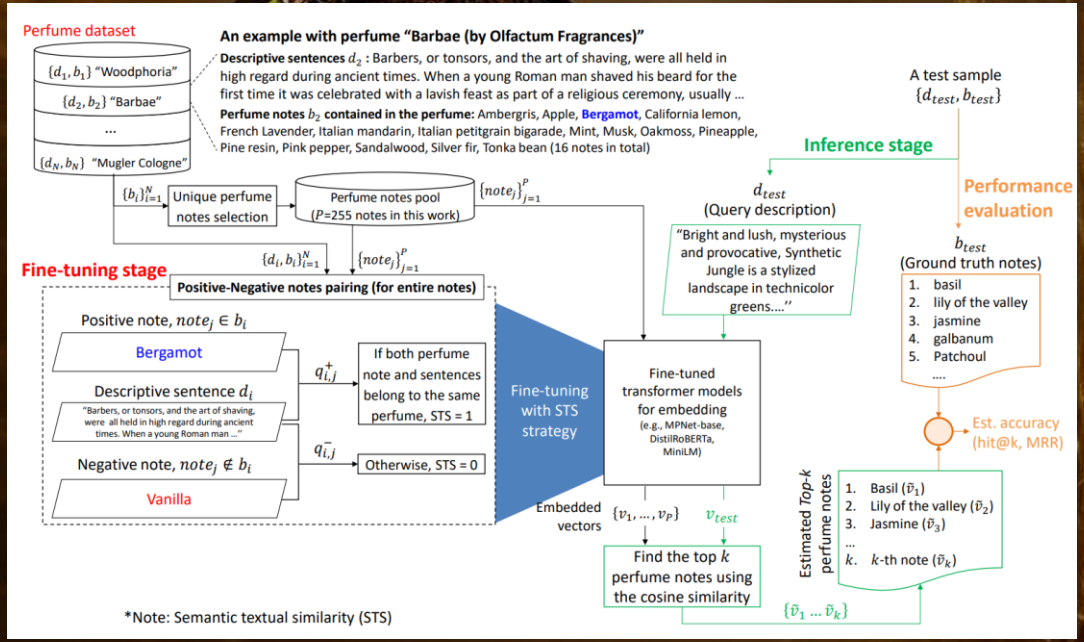
Training: FastText



1. For use of calculating similarities between notes
2. Extract description from “prediction_train” for using as “corpus”
3. Training FastText from Gensim library(advanced version of Word2Vec)

Training: Sentence Transformer

1. Apply and develop fine-tuning idea from Jooyoung Kim et al.



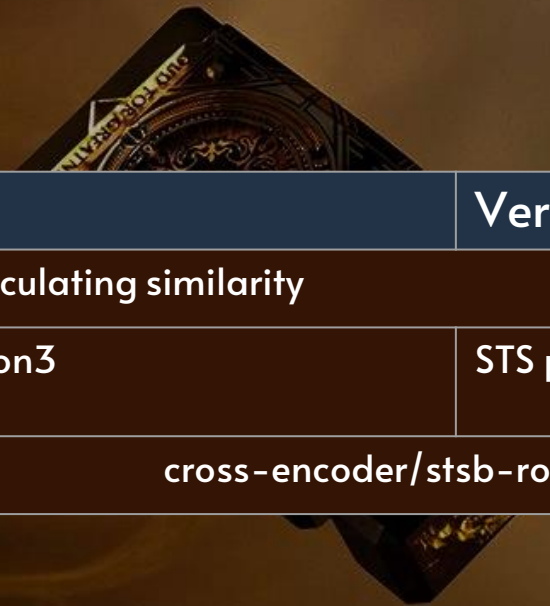
Training: Sentence Transformer

Description	Notes
Heroic Man by La Rive is a bold and captivating fragrance that defines modern masculinity with its warm and spicy character. Released in 2022, this scent embodies an adventurous spirit that exudes confidence and sophistication. Users describe it as both aromatic and invigorating.....	"['cardamom', 'pink pepper', 'mint', 'violet leaf', 'cinnamon', 'melon', 'pineapple', 'nutmeg', 'lavender', 'sage', 'vanilla', 'chestnut', 'cedar', 'amberwood', 'guaiac wood']"
Positive Pair	('Heroic Man by La Rive is a bold and captivating fragrance that defines modern masculinity with its warm and ...' , 'cardamom' , 1.0)
Negative Pair	('Heroic Man by La Rive is a bold and captivating fragrance that defines modern masculinity with its warm and ...' , 'bergamot' , 0.0)
Goal: Training transformer in a manner of predicting notes from Natural Language	

Training: Sentence Transformer

	Version1	Version2	Version3	Version4
Positive Notes (N notes)	combine N notes altogether	make N pairs with each notes	combine N notes altogether	make N pairs with each notes
Negative Notes (N notes)	combine randomly selected N notes except positive notes	N pairs of one of least similar notes	combine least similar N notes	All pairs of notes except positive notes
P:N ratio	1:3	1:1	1:3	1:many
Feature		FastText	FastText	Original Method
Model	sentence-transformers/all-MiniLM-L6-v2			

Training: Cross Encoder



	Version3	Version4
Goal	To optimize calculating similarity	
Training Set	STS pairs Version3	STS pairs Version4
Model	cross-encoder/stsb-roberta-large	

Result: Sentence Transformer

```
original:  {'pearson_cosine': 0.49687206624610303, 'spearman_cosine': 0.4539509857075771}
v1:  {'pearson_cosine': 0.8425746761744255, 'spearman_cosine': 0.718974393548417}
v3:  {'pearson_cosine': 0.9339541699697309, 'spearman_cosine': 0.733406361302126}
original:  {'pearson_cosine': 0.049094796815849724, 'spearman_cosine': 0.06253949726370187}
v2:  {'pearson_cosine': 0.9221809753640012, 'spearman_cosine': 0.8309208495457832}
v4:  {'pearson_cosine': 0.36641281050343105, 'spearman_cosine': 0.20018342620535076}
```

Accuracy by Cosine Similarity: v3 > v2>v1>v4

Result: Sentence Transformer

	Real Notes	Predicted Notes (vo)	Predicted Notes (v1)	Predicted Notes (v2)	Predicted Notes (v3)	Predicted Notes (v4)
0	rose	flowers	floral notes	sandalwood	pink rose	musk
1	cherry blossom	cherry blossom	cherry blossom	saffron	bergamot	bergamot
2	pink peony	floral notes	bergamot	ambergris	rock rose	amber
3		apple blossom	pink rose	patchouli	cardamom	vanilla
4		lady of the night flower	lady of the night flower	cardamom	cherry blossom	jasmine
5		night blooming jasmine	grapefruit blossom	civet	lavender	patchouli
6		orange blossom	lavender	musk	night blooming jasmine	sandalwood
7		grapefruit blossom	cardamom	hinoki wood	lady of the night flower	rose

- Model Performance Metrics:

	Model	Matched Count	Accuracy (%)
0	vo	1	33.33
1	v1	1	33.33
2	v2	0	0.00
3	v3	1	33.33
4	v4	1	33.33

Result: Sentence Transformer

	Real Notes	Predicted Notes (vo)	Predicted Notes (v1)	Predicted Notes (v2)	Predicted Notes (v3)	Predicted Notes (v4)
0	ambranum	crystal amber	bergamot	sandalwood	pink rose	musk
1	citrus notes	leather	lavender	guaiac wood	bergamot	bergamot
2	frankincense	black amber	amber	damask rose	cardamom	amber
3	patchouli	amber	cardamom	cardamom	lavender	jasmine
4	black pepper	lavender	leather	pink rose	black cardamom	vanilla
5	cypriol	cosmos flower	pink rose	agarwood	rock rose	sandalwood
6	bay leaf	pink rose	black amber	passionfruit	lady of the night flower	patchouli
7	sandalwood	flowers	crystal amber	rose	night blooming jasmine	rose
8	periploca					

- Model Performance Metrics:

	Model	Matched Count	Accuracy (%)
0	vo	0	0.00
1	v1	0	0.00
2	v2	1	11.11
3	v3	0	0.00
4	v4	2	22.22

Result: Sentence Transformer

Model Predictions Comparison:

	Real Notes	Predicted Notes (vo)	Predicted Notes (v1)	Predicted Notes (v2)	Predicted Notes (v3)	Predicted Notes (v4)
0	amber	cherry blossom	grapefruit blossom	bergamot	pink rose	musk
1	vanilla	floral notes	lemon blossom	ambergris	lemon blossom	amber
2	lychee	silk tree blossom	cherry blossom	damask rose	lavender	bergamot
3	patchouli	lemon blossom	floral notes	black vanilla husk	bergamot	vanilla
4	cherry blossom	apple blossom	apple blossom	hyacinth	rock rose	jasmine
5	bergamot	flowers	lavender	blue hyacinth	orange blossom	patchouli
6	pear	orange blossom	silk tree blossom	labdanum	apple blossom	sandalwood
7		pear blossom	orange blossom	sandalwood	cherry blossom	rose

Result: Sentence Transformer

Model Performance Metrics:

	Model	Matched Count	Accuracy (%)
0	vo	1	14.29
1	v1	1	14.29
2	v2	1	14.29
3	v3	2	28.57
4	v4	4	57.14

Result: Sentence Transformer

	Real Notes	Predicted Notes (vo)	Predicted Notes (v1)	Predicted Notes (v2)	Predicted Notes (v3)	Predicted Notes (v4)
0	pink pepper	silk tree blossom	bergamot	biovanilla	bergamot	musk
1	rosewood	night blooming jasmine	lavender	labdanum	pink rose	bergamot
2	neroli	cosmos flower	cardamom	sandalwood	lavender	amber
3	guaiac wood	pear blossom	honey jasmine	guaiac wood	cardamom	vanilla
4	patchouli	black vanilla husk	vanilla	saffron	rock rose	jasmine
5	vetiver	lavender	night blooming jasmine	gurgum wood	night blooming jasmine	patchouli
6	saffron	lemon blossom	natural musk	black vanilla husk	black cardamom	sandalwood
7	agarwood	apple blossom	clary sage	musk	pink pepper	rose
8	sandalwood					

- Model Performance Metrics:

	Model	Matched Count	Accuracy (%)
0	vo	0	0.00
1	v1	0	0.00
2	v2	3	33.33
3	v3	1	11.11
4	v4	2	22.22

Result: Sentence Transformer

Model Predictions Comparison:

	Real Notes	Predicted Notes (vo)	Predicted Notes (v1)	Predicted Notes (v2)	Predicted Notes (v3)	Predicted Notes (v4)
0	woody notes	leather	bergamot	gurgum wood	bergamot	musk
1	amber	lavender	lavender	saffron	pink rose	bergamot
2	incense	silk tree blossom	amber	sandalwood	lavender	amber
3	agarwood	mahogany	black amber	guaiac wood	rock rose	vanilla
4	bergamot	amber	amberwood	black vanilla husk	cardamom	jasmine
5		amberseed	clary sage	ambergris	black cardamom	sandalwood
6		guaiac wood	black vanilla husk	labdanum	silk tree blossom	patchouli
7		black amber	silk tree blossom	bergamot	night blooming jasmine	cedar

Model Performance Metrics:

	Model	Matched Count	Accuracy (%)
0	vo	1	20.0
1	v1	2	40.0
2	v2	1	20.0
3	v3	1	20.0
4	v4	2	40.0

->lack of diversity, overfitted

Result: Cross Encoder

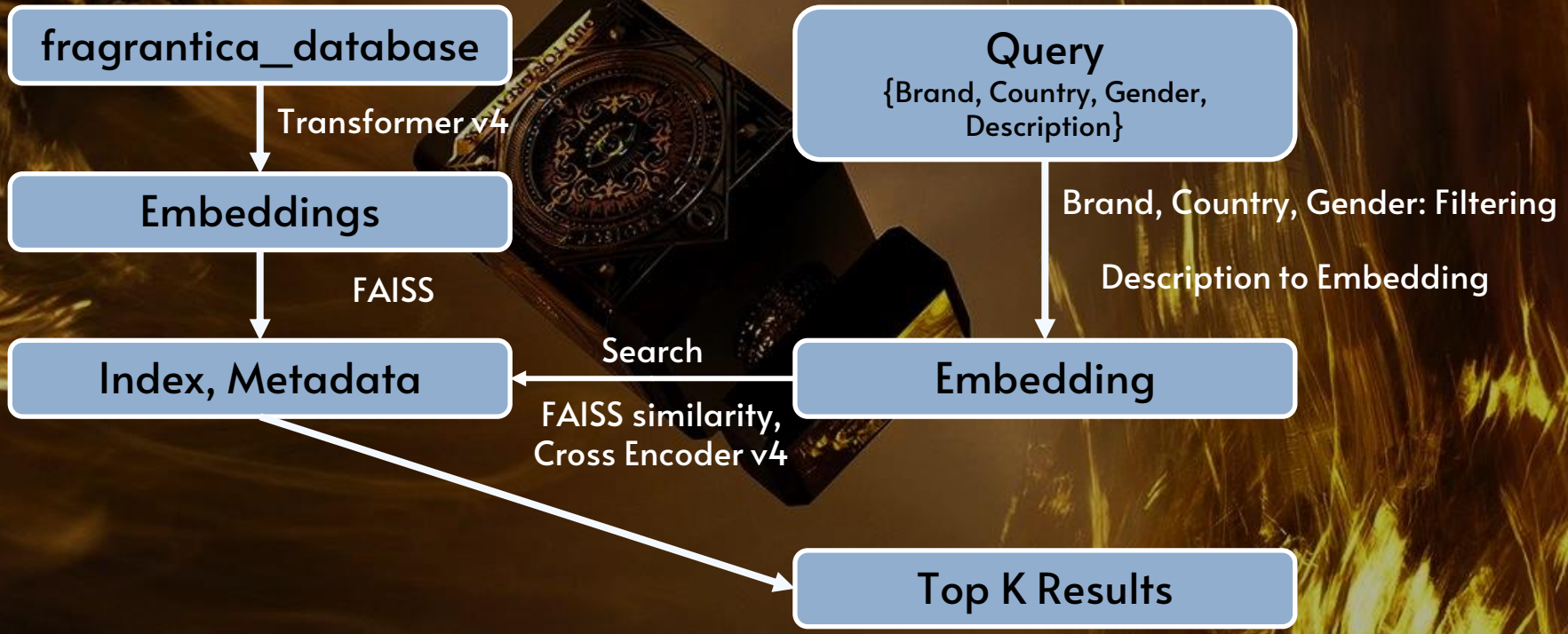
```
from sentence_transformers.cross_encoder.evaluation import CECorrelationEvaluator

cross_model_original = CrossEncoder('cross-encoder/stsb-roberta-large', num_labels=1)
ce_evaluator = CECorrelationEvaluator.from_input_examples(val_examples)
ce_evaluator(cross_model_original)
```

```
0.28732735800566295
```

```
[ ] ce_evaluator = CECorrelationEvaluator.from_input_examples(val_examples)
    ce_evaluator(cross_model)
    # 0.8650250798639563
```

```
0.7242223818224929
```

Application



Application: Demo

https://colab.research.google.com/drive/1oUCJ_aEKqFQhlj8k2T4ImHLI36DeLQsK?usp=sharing

Further Research

1. Mitigating Overfitting(오버피팅 완화)
2. Web Serving
   - -Construct VectorDB
   - -Serve on web via FastAPI or own website

   https://olavvn.github.io/pour_monsieur_web/
3. Cutting edge coding
   - -object oriented clear coding
   - -Improving RAG with filtering UI

References

Kim, Jooyoung, Kangrok Oh, and Beom-Seok Oh. 2024. "An NLP-Based Perfume Note Estimation Based on Descriptive Sentences" Applied Sciences 14, no. 20: 9293. https://doi.org/10.3390/app14209293

Brian K. Lee, Emily J. Mayhew, Benjamin Sanchez-Lengeling, Jennifer N. Wei, Wesley W. Qian, Kelsie Little, Matthew Andres, Britney B. Nguyen, Theresa Moloy, Jane K. Parker, Richard C. Gerkin, Joel D. Mainland, Alexander B. Wiltschko. 2022. "A Principal Odor Map Unifies Diverse Tasks in Human Olfactory Perception"bioRxiv 2022.09.01.504602; doi: https://doi.org/10.1101/2022.09.01.504602

Huh, Jungjoon, 2024 "LLM을 활용한 실전 AI 애플리케이션 개발",책만