



향수 추천 챗봇

2023 KUBIG NLP 분반
챗봇 1팀 박종혁 이영노 하예은

목차



01

주제 선정
동기



02

데이터
전처리



03

임베딩



04

추천 과정
및 결과

01 주제 선정 동기

- *So I said, “CHANEL N°5!”*



01 주제 선정 동기

향미자 - 향수에 미친 자 | 2023.02.14.

클린 워코튼 향수 비누향 역대급 호불호!



그래서 향수를 고를 때도 코튼향 비누향 향수 들을 선호하는데요, 클린 워코튼은 그런 제 취향에 정말 딱 맞는... 향 자체가 크게 변한 건 아니었지만 약간의 달달함과 화사함이 더해져 전체적으로 풍성해진 듯한 느낌?...

뷰티 크리에이터 ✓브이로그 | 인플루언서 | 2023.01.02.

20대 30대 남자향수 추천, 여름 겨울 지속력좋은 비누향 향수



바이레도 입문기 남자 비누향 향수 추천 안녕하세요. 오늘은 저의 니치향수 입문기 입니다. 평소에 저렴한... 남자향수 추천 해준 친구가 거의 매니아 수준으로 여러 종류를 사용 중인데요. 이 제품이, 야간 꽃향같은 비누향...

상구앤상추와 함께 행복하자 | 2023.01.14.

바이레도 블랑쉬 비누향 대표 향수

깨끗한 비누 느낌과 함께 화이트 머스크 향이 맑고 시원하게 느껴지다 보니 중성적인 매력이 있어서 더 마음에 들어요. 저처럼 비누향 향수를 좋아하거나 머리 아픈...



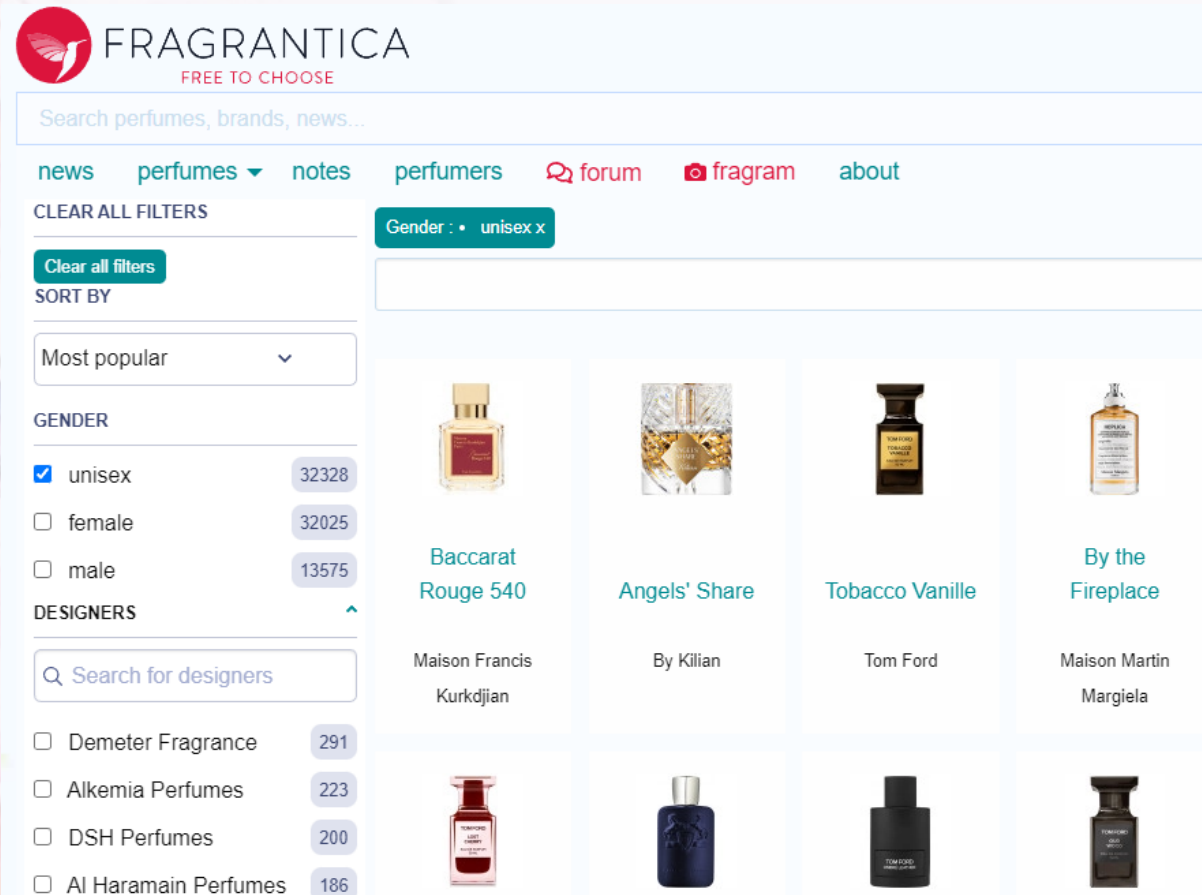
여러 향수에 대한 종합적인 평가? 🤖
자기 취향을 모를 때 향수 추천? 🤖



자유롭게 내 취향을 입력하여
취향에 맞는 향수를 추천하고
해당 향수의 리뷰 키워드를 보여주는 챗봇!

02 데이터 전처리

1. 데이터 확보



FRAGRANTICA
FREE TO CHOOSE

Search perfumes, brands, news...

news perfumes notes perfumers forum fragram about

CLEAR ALL FILTERS

Gender : unisex x

Clear all filters

SORT BY

Most popular

GENDER

- ☒ unisex 32328
- ☐ female 32025
- ☐ male 13575

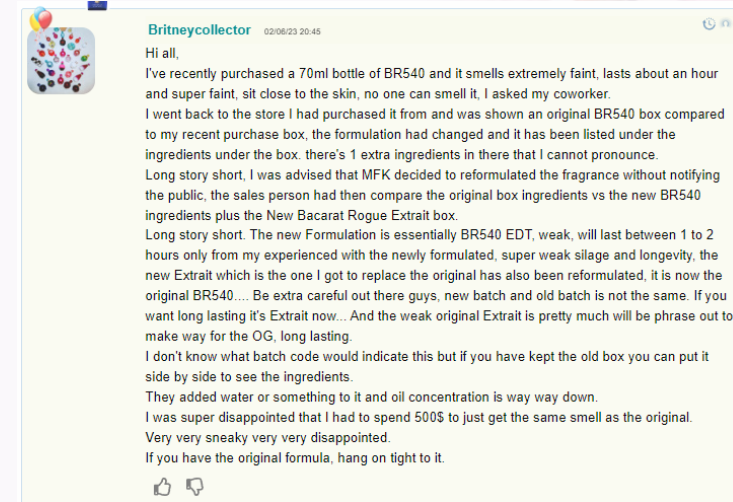
DESIGNERS

Search for designers

- ☐ Demeter Fragrance 291
- ☐ Alkemia Perfumes 223
- ☐ DSH Perfumes 200
- ☐ Al Haramain Perfumes 186

Perfume Name	Brand
Baccarat Rouge 540	Maison Francis Kurkdjian
Angels' Share	By Kilian
Tobacco Vanille	Tom Ford
By the Fireplace	Maison Martin Margiela

FRAGRANTICA에서
unisex 향수 144개
각각 200개 리뷰 크롤링



Britneycollector 02/06/23 20:45

Hi all,

I've recently purchased a 70ml bottle of BR540 and it smells extremely faint, lasts about an hour and super faint, sit close to the skin, no one can smell it, I asked my coworker.

I went back to the store I had purchased it from and was shown an original BR540 box compared to my recent purchase box, the formulation had changed and it has been listed under the ingredients under the box. there's 1 extra ingredients in there that I cannot pronounce.

Long story short, I was advised that MFK decided to reformulated the fragrance without notifying the public, the sales person had then compare the original box ingredients vs the new BR540 ingredients plus the New Baccarat Rouge Extrait box.

Long story short. The new Formulation is essentially BR540 EDT, weak, will last between 1 to 2 hours only from my experienced with the newly formulated, super weak silage and longevity, the new Extrait which is the one I got to replace the original has also been reformulated, it is now the original BR540.... Be extra careful out there guys, new batch and old batch is not the same. If you want long lasting it's Extrait now... And the weak original Extrait is pretty much will be phrase out to make way for the OG, long lasting.

I don't know what batch code would indicate this but if you have kept the old box you can put it side by side to see the ingredients.

They added water or something to it and oil concentration is way way down.

I was super disappointed that I had to spend 500\$ to just get the same smell as the original.

Very very sneaky very very disappointed.

If you have the original formula, hang on tight to it.

👍 👎

리뷰 예시

02 데이터 전처리

2. 텍스트 전처리

- 1) 정규표현식으로 숫자, 특수문자 제거
- 2) nltk 불용어 제거
- 3) wordnet lemmatizer로 표제어 추출
- 4) 200개 → 1개의 리뷰로 이어 붙이기

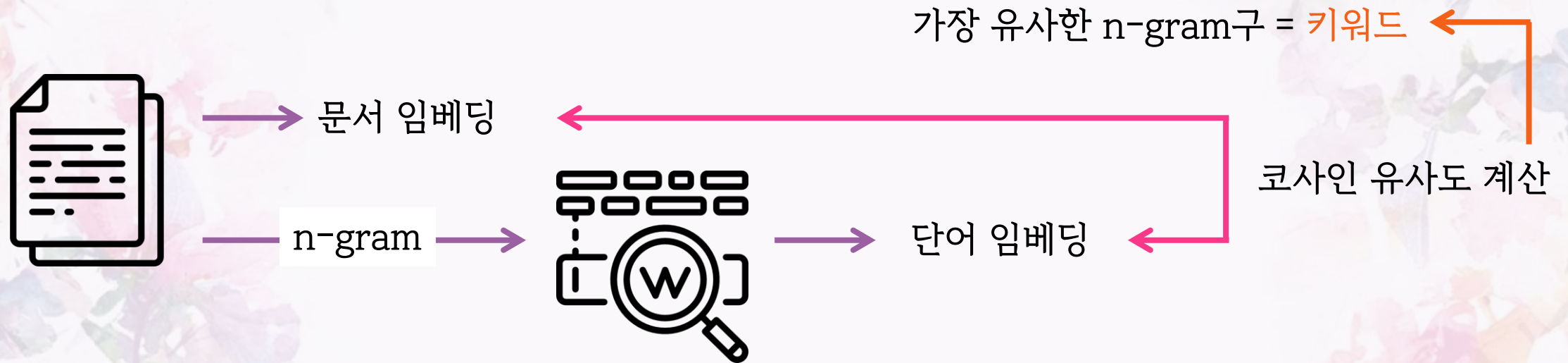
'skin get strawberry jelly candies. nostalgic pretty know smelt though popular still really like decide feel this. love fresh yet warm sweet something singe nostril forget hype for get cloud comparison forget tiktok forget world second admit perfume truly unique masterpiece. nose like smell period. smell like nearly every flight attendant since . lovely chic somewhat saturated marketplace. ew price even ew. honestly longest time thought anyone smelled like wearing lalique soleil finally able try figure getting hand tester. nearly identical soleil bit crisp citrusy base tone basically complete dupe. anyone like want smell similar tenth price lalique soleil scent botox filled designer clothed basic rotten souled woman driving range rovers. keep far away hahahah bought social medium right . month get olfactory fatigue thinking overhyping one month closet killing everyone street strong nice bossy every occasion every day . nice nose trip dentist room saffron plantation really old place old t...'

❁ Lemmatization(표제어 추출) vs Stemming(어간 추출)

example	candy	candies	smell	smelt	comforting	comfort
Lemmatization	candy	candy	smell	smelt	comforting	comfort
Stemming	candi	candi	smell	smelt	comfort	comfort

02 데이터 전처리

3. 키워드 추출 - 1) KeyBERT



- ❖ MMR(Maximal Marginal Relevance) : 중복 최소화 & 다양성 극대화
use_mmr = TRUE 옵션으로 사용, diversity 인자로 정도 조절 가능

02 데이터 전처리

3. 키워드 추출 - 1) KeyBERT

	keyword	weight
0	vanilla	0.3475
1	butteriness	0.2829
2	tea	0.2805
3	coconutiness	0.2500
4	cupcake	0.2481

	keyword	weight
0	grapefruit	0.3437
1	tea	0.3074
2	perfume	0.3068
3	liquorice	0.2828
4	gingerbread	0.2769
5	aftershave	0.2622
6	vampire	0.2366
7	syrupy	0.2326

	keyword	weight
0	vanilla	0.3907
1	sugarsweet	0.3876
2	strawberry	0.3731
3	creamier	0.3694
4	strawberries	0.3654
5	lemonade	0.3630
6	hazelnut	0.3504
7	popcorn	0.2789
8	tea	0.2769
9	britney	0.2220
10	intoxicating	0.2120
11	gastronomy	0.2078

	keyword	weight
0	cinnamony	0.4199
1	buttercream	0.3648
2	vanilla	0.3600
3	apples	0.3464
4	fruity	0.3284
5	cognac	0.3187
6	tea	0.2764
7	marshmallow	0.2351
8	starbucks	0.2309
9	toothpaste	0.2225
10	prettiest	0.2163
11	intoxicated	0.2009

mmr 사용 & 상위 20개 추출
유사도(weight)가 0.2 이상인 키워드 선정

but strawberries, apples 등
표제어 추출이 제대로 안 된 키워드 존재

02 데이터 전처리

3. 키워드 추출 - 2) 개체명 인식(Named Entity Recognition)

❁ 개체명 인식 : 이름을 가진 객체를 인식하는 것

예시) 영노[인명]는 정경관[지명]에서 종혁이[인명]와 6시[시간]에 만나기로 약속하였다.

1) 향에 대한 단어 200개 태깅 작업

☒ FRAGRANCE NEW TAG EDIT TAGS

Orpheon is quite linear on my skin. It 's a powdery FRAGRANCE , slightly sweet FRAGRANCE white floral FRAGRANCE fragrance that is clean FRAGRANCE and fresh FRAGRANCE but not new. It can get boring after a few weeks of consistent wear (though I always switch it up every day / every few hours sometimes) . I can not imagine having this as your ONLY fragrance. I really want the bottle just for the bottle itself and not the juice , so I might grab this one day as a collectible. It's a like , not a love , and I heavily recommend getting it with a discount ! It's clean and nice and pleasant and all those adjectives , but it did n't wow me .

2) Spacy로 키워드 추출

- label 된 단어들을 정답으로 인식 → 단어 간 관계를 파악하여 문서를 가장 잘 설명하는 키워드 추출

02 데이터 전처리

3. 키워드 추출 - 3) 키워드 병합

❁ 추출된 키워드에 대해 표제어 추출 한 번 더 진행

❁ keyBERT 키워드의 개수가 최소 2개 이상으로 일정하지 않은 문제 발생

1) keyBERT 키워드가 5개 이상일 경우

- keyBERT 키워드 5개 + NER 키워드 5개

2) keyBERT 키워드가 5개 미만일 경우

- keyBERT 키워드 2개 + NER 키워드 8개

```
0      [vanilla, sweet, sugarsweet, feminine, strawbe...
1      [cinnamony, sweet, buttercream, cognac, vanill...
2      [hazelnut_kitty, tobacco, vanilla, vanilla, ho...
3      [vanilla, vanilla, tea, sweet, marshmallows, s...
4      [perfume, cherry, cinnamon, sweet, rose, cherr...
...
139     [lipsticky, incense, intoxicating, dark, wealt...
140     [perfumey, cedar, vanilla, woody, cinnamon, sw...
141     [grapefruit, woody, perfume, spicy, vanilla, u...
142     [caramely, sweet, vanilla, vanilla, lipstick, ...
143     [vanilla, vanilla, butteriness, sweet, tea, su...
Name: keyword, Length: 144, dtype: object
```


03 임베딩

1. 각 향수 리뷰 임베딩

각 향수별 리뷰를 하나의 텍스트로 합치고 ELMo로 임베딩

❖ ELMo(Embeddings from Language Model) : 문맥을 반영한 워드 임베딩

	text	token	word_embedding_elmo	index
0	kin get strawberry jelly candies nostalgic pre...	kin	[-1.3006852865219116, 0.3651913106441498, -1.2...	0
1	kin get strawberry jelly candies nostalgic pre...	get	[0.0042253658175468445, -0.4184453785419464, -...	0
2	kin get strawberry jelly candies nostalgic pre...	strawberry	[-0.2695613503456116, 0.13173973560333252, 0.5...	0
3	kin get strawberry jelly candies nostalgic pre...	jelly	[-0.3758509159088135, -0.17479586601257324, 1....	0
4	kin get strawberry jelly candies nostalgic pre...	candies	[-0.7830489873886108, 0.8652899265289307, 0.01...	0

예) Baccarat Rouge 540에 대한 리뷰 임베딩

size : 1 x 512

03 임베딩

2. 키워드 임베딩

리뷰 임베딩 결과에서 token이 키워드인 임베딩 값 추출

	text	token	word_embedding_elmo	index
3890	kin get strawberry jelly candies nostalgic pre...	sugarsweet	[-0.6196644902229309, 0.9715945720672607, 0.35...	0

	keyword	embeddings
0	jasmine	[-0.280725359916687, -0.6509876251220703, 0.36...
1	vanilla	[-0.31130504608154297, -0.5989245772361755, 0....
2	butteriness	[-1.5927531719207764, -0.1343236267566681, -0....
3	sweet	[-0.2673688530921936, -0.4698314964771271, 0.1...
4	tea	[-0.4590742588043213, -0.7632981538772583, 0.5...
5	summer	[0.5416969060897827, 0.6820259094238281, -0.08...
6	coconutiness	[-0.8889951109886169, -0.13306169211864471, 0....
7	creamy	[-0.6275959610939026, -0.04324675351381302, 0....
8	cupcake	[-0.27284932136535645, -0.13251814246177673, 0...

향수 별로 임베딩 → 같은 키워드라도 다른 값으로 임베딩
문맥을 반영한 임베딩!

04 추천 과정 및 결과

- 1) 입력값 전처리
- 2) keybert로 입력값 키워드 추출(*불용어 : scent, fragrance 등 ‘향’ 자체를 나타내는 단어)
- 3) 입력값 전체에 대한 임베딩
 - 입력값과 리뷰 임베딩 벡터의 차원이 다르므로 padding으로 차원 크기를 맞춤
- 4) 입력값 키워드에 대한 임베딩 벡터 추출
- 5) 각 향수의 키워드 임베딩 벡터들간의 코사인 유사도 계산
- 6) 코사인 유사도가 가장 높은 향수 출력

“Is there something fresh and sharp with cool vibes?
Would be nice if it is more like masculine perfume.”



	name	cosine_similarity
46	Chocolate Greedy	0.247541
137	La Capitale	0.232257
141	Halfeti	0.222774
47	Erba Pura	0.222262
16	Ani	0.218438

**감사합니
다.**

