데이터 분석 보고서

다국어 검색어 관련 분석

1. 개요
   -. 현 상황 진단
   -. 선행 연구 탐색 및 연구가설 설정

2. 분석 결과
   -. 분석 프로시져
   -. 각 언어별 기초통계량 확인
   -. 검색어 품질 관련 이슈 우려 사항

3. 결론

개요

개요

1. 현 상황 진단

번호	유저ID	검색어	번역어	검색일자
646	162414	by love j love	Bit by Bit 비트바이비트	2022-11-08
…				
649	150615	플라워 스커트	플라워	2022-11-08
650	167554	무스탕	무스탕	2022-11-08
651	167187	背心	나시	2022-11-08

〈2022년 11월 8일의 검색어 / 번역어 로그〉

해외 유저의 이용 편의 증진을 **위해**
Papago 번역 기능을 **기존 검색어 시스템**에 추가 도입

2022년 11월 8일 이후로 **검색어 및 번역어**와 관련된
히스토리(로그)를 **적재**하기 시작

검색어의 **현 상황**을 **포괄적**으로 **살펴보고**, **검색어 시스템 도입 관련** 서비스 품질에 **유저**가 **만족**하는지 여부 등
검색어와 관련된 **인사이트**를 **도출**하여 서비스 개선에 활용

본문

1. 선행 연구 검토 및 연구 가설 설정

1) 선행 연구 탐색

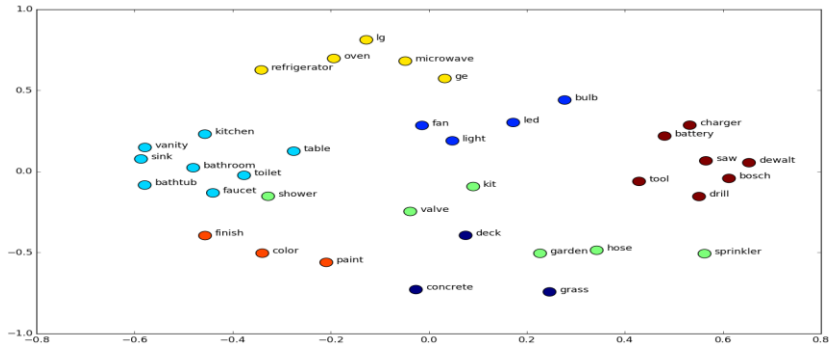
-. Word Embedding

① 문맥을 기반으로 비슷한 문맥의 문장에 있는 단어들을 비슷한 위치로 이동시키는 딥러닝 학습 방법론

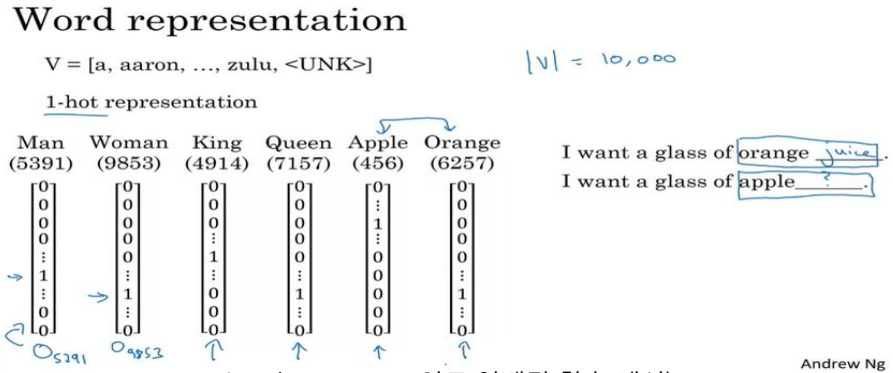
② 예를 들어, “I want a glass of orange”와 “I want a glass of apple”이라는 문장이 있다면, 두 문장은 비슷한 문맥에 ‘orange’와 ‘apple’이라는 단어만 다르기 때문에 학습이 반복될수록 apple과 orange는 비슷한 위치에 점차 접근하게 됨

③ 학습을 수천 ~ 수십만개의 문장에 대해 진행하면 모델은 거리상 가까운 단어들은 가깝게, 먼 단어들은 점차 떨어뜨리는 방향으로 학습을 수행

④ 이번 분석엔 Word - Embedding 기법 중 **Sentence-BERT**와 **FastText**를 이용



〈2차원 워드 임베딩〉



〈apple - orange 워드 임베딩 학습 예시〉

1. 선행 연구 검토 및 연구 가설 설정

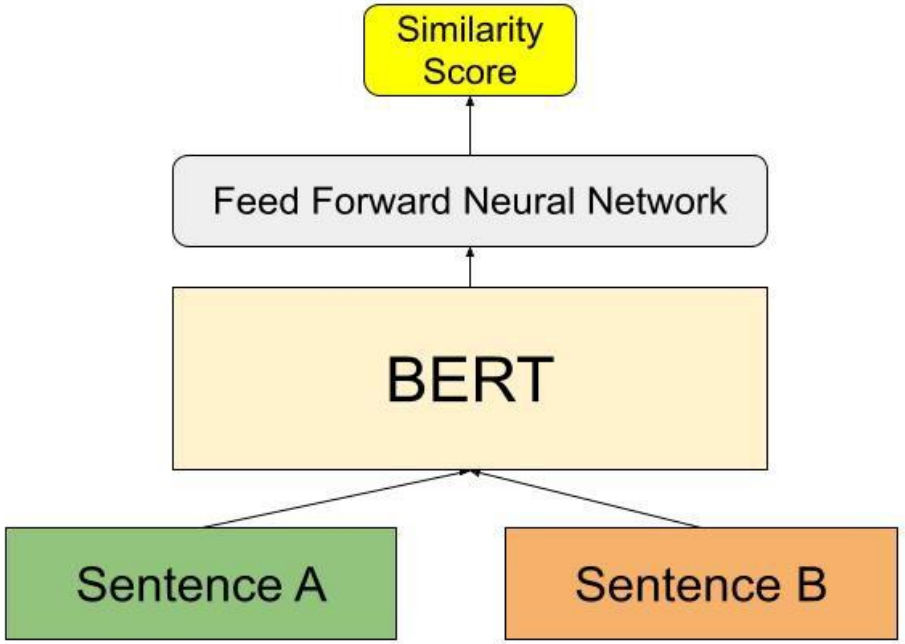
1) 선행 연구 탐색

-. Sentence BERT

① Word Embedding을 이용하여 문장간 유사도를 학습한 Pre-Trained 모델

② A문장(ex. 영어)과 B문장(ex. 한국어)을 짝지어 '유사' - '중립' - '반대' 짝을 딥러닝 모델에 투입하여 학습을 수행

③ 유사한 문장은 유사한 문장끼리, 반대 문장은 반대 문장끼리 서로 가까워지거나 혹은 멀어지며 각 문장간 '문맥 ' 의 유사도를 학습



〈Sentence-Bert 모델 구조도〉

1. 선행 연구 검토 및 연구 가설 설정

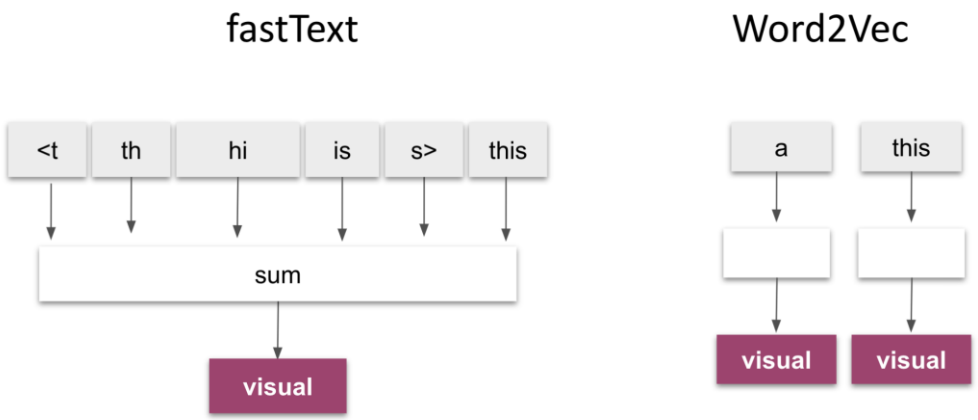
1) 선행 연구 탐색

-. FastText

① Word Embedding 방법론 중 가장 진보된 방법론

② 페이스북에서 개발하여 공개

③ 기존의 Word Embedding 방법론을 차용한 모델이 '전혀 처음 보는 단어'엔 대응할 수 없다는 단점을 보완하기 위해, 단어를 가장 작은 어근 수준까지 쪼개 학습을 수행



1. 선행 연구 검토 및 연구 가설 설정

2) 가설 설정

-. 가설 1 : 파파고 번역 기능 도입 이후 각 언어권별 유저들은 자국어 검색어 사용 빈도를 늘렸을 것이다

-. 가설 2 : 파파고 번역 기능은 검색 품질 향상을 위해 현재 검색어 번역을 올바르게 수행하고 있을 것이다

2. 데이터 분석 프로시져

1) 데이터 테이블 정의

1. 검색어 히스토리

번호	컬럼명	설명
1	**userId**	검색 수행 유저의 ID
2	**Originword**	원본 검색어
3	**Transword**	파파고가 번역한 번역된 검색어
4	**createdAt**	검색 수행 일자

-. 수집 기간 : 2022-11-08 ~ 2022-11-22
-. 건수 : 221,554 건

2. 바이어 데이터

번호	컬럼명	설명
1	**buyerId**	바이어 ID
2	**countryCodePlace**	사업장 소재지 기준 국가
3	**LastLoginAt**	마지막 접속일자

-. 수집 기간 : 2016-01-01 ~ 2022-11-22
-. 건수 : 54,042 건

3. 셀러 데이터

번호	컬럼명	설명
1	**sellerId**	셀러 ID
2	**BrandID**	브랜드 ID
3	**BrandName**	브랜드명(한글, 영어)

-. 수집 기간 : 2016-01-01 ~ 2022-11-22
-. 건수 : 17,018 건

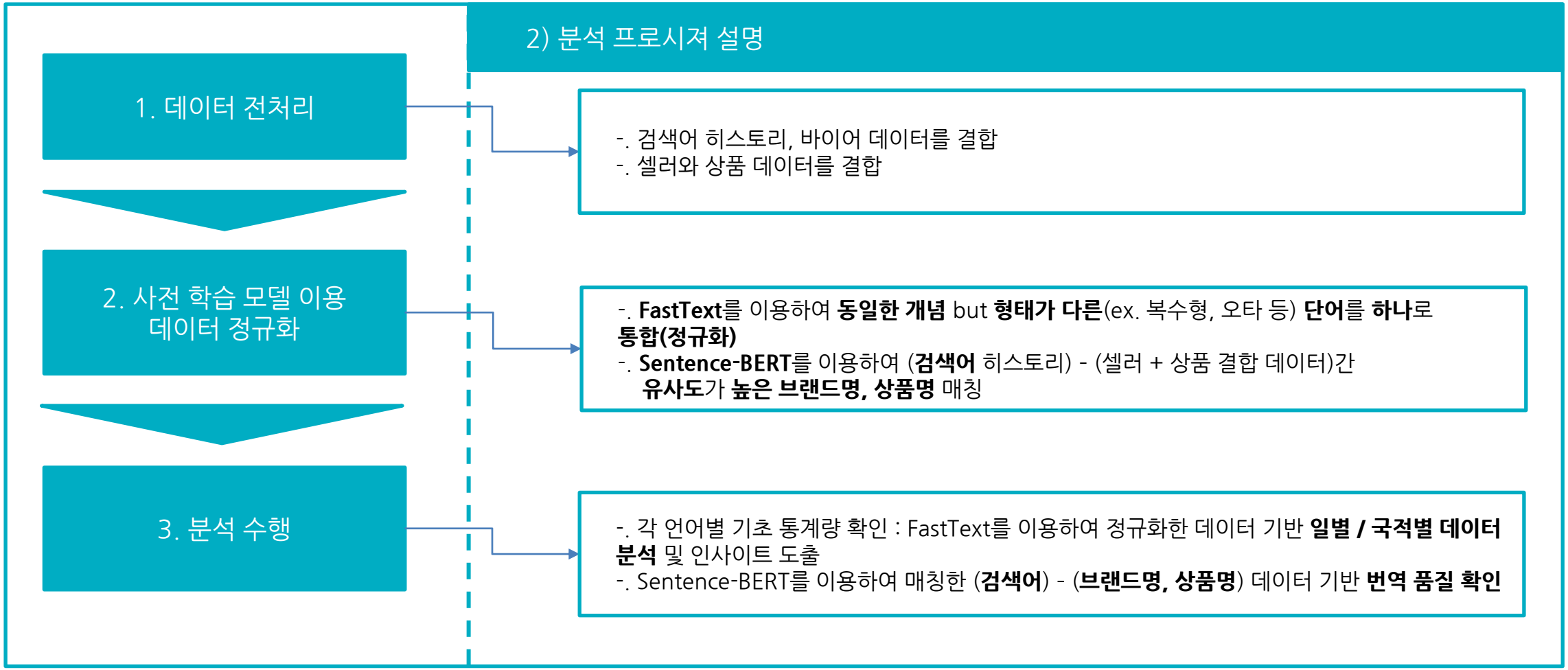
4. 상품 데이터

번호	컬럼명	설명
1	**sellerId**	셀러 ID
2	**BrandID**	브랜드 ID
3	**BrandName**	브랜드명(한글, 영어)

-. 수집 기간 : 2022-09-01 ~ 2022-11-22
-. 건수 : 1,024,763건

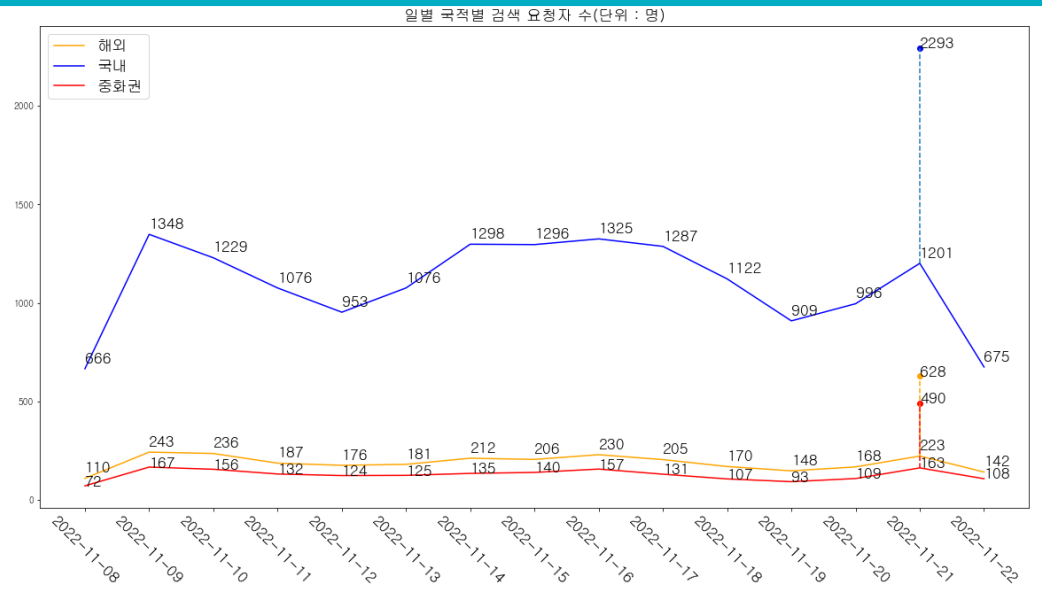
본문

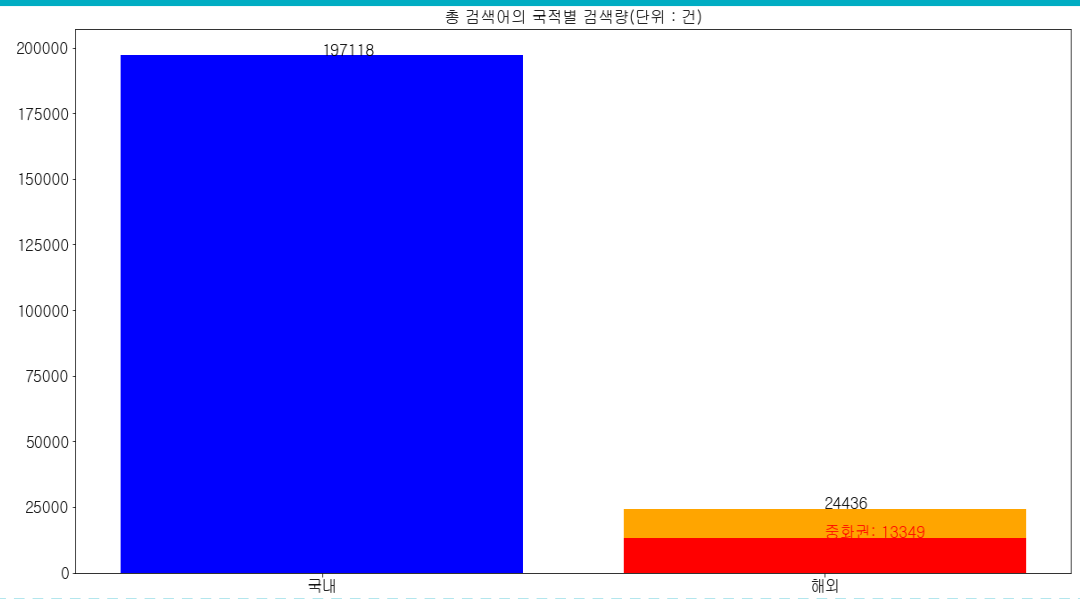
2. 데이터 분석 프로시져



3. 데이터 분석 결과

1) 각 언어별 기초 통계량 확인





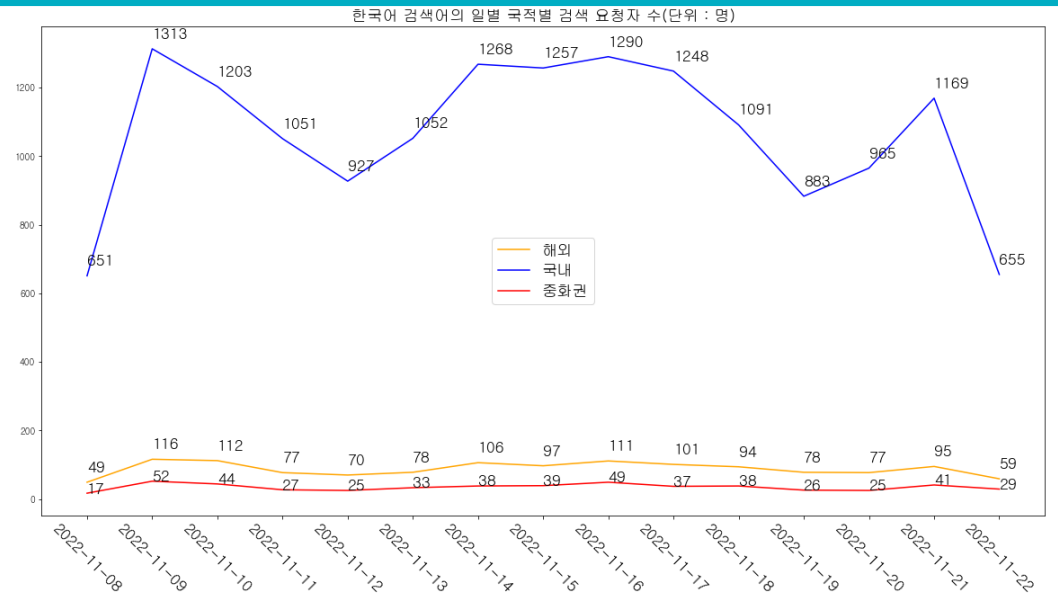
-. 모든 검색어의 검색자수 / 검색량 통계

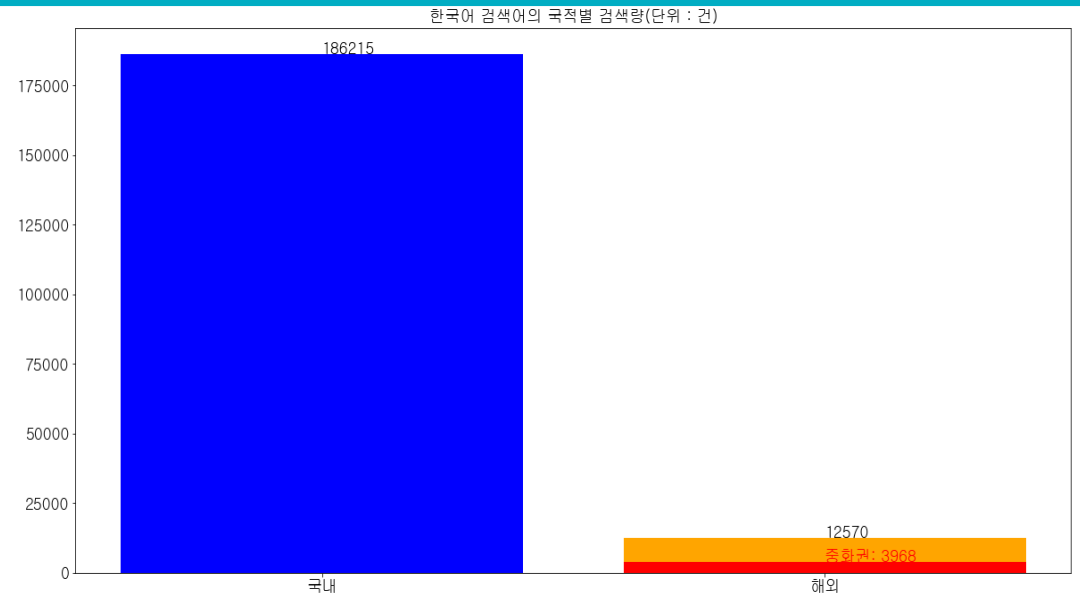
① **국내 유저**와 **해외 유저**의 **검색자(명)**의 비중은 약 **9:1 수준**으로, 국내 유저가 압도적이나 **해외 유저의 비중**도 **적지 않음**

② 명 수가 아닌 건수로 보면 총 **221,554건** 중 **국내 유저**의 검색건수 **19만 7118건**, **해외 유저**의 검색건수 **2만 4436건**으로 약 **9:1 비중**임

3. 데이터 분석 결과

1) 각 언어별 기초 통계량 확인





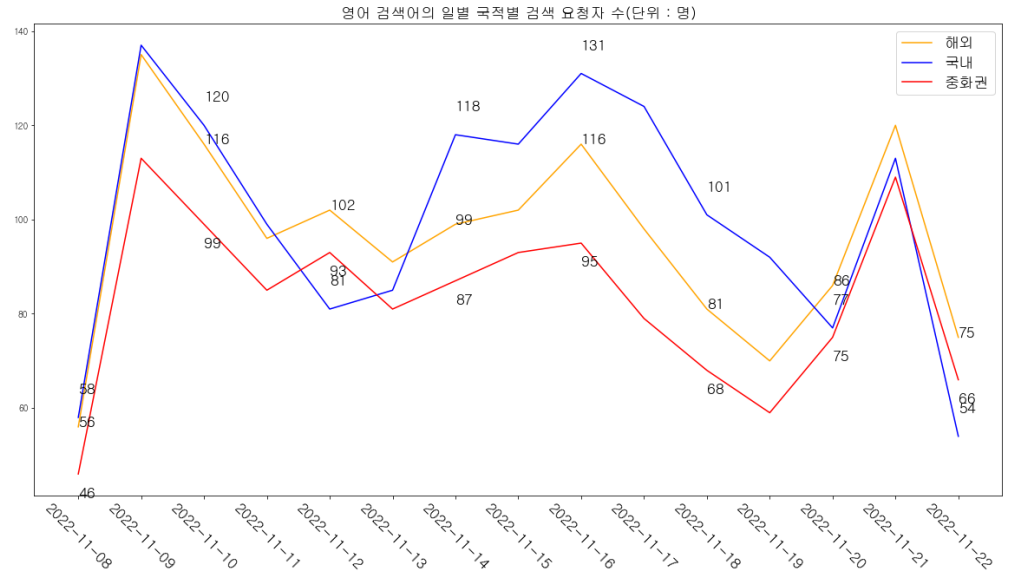
-. 한국어 검색어의 검색자수 / 검색량 통계

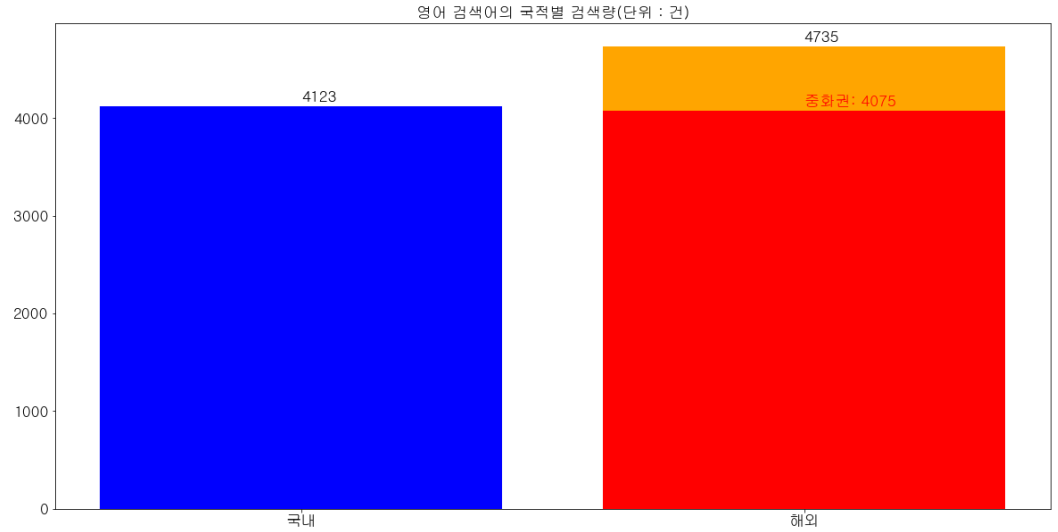
① **한국어 검색어**만 따로 뽑아 확인한 결과 주로 **국내 유저**가 **한국어**로 **검색**한 경우 **多**, But **해외 유저**도 **한국어**로 **검색**한 경우 **적지 않음**

② **중화권** 기준으로 **일별 평균 3~40명**의 검색자수와 **총 3,968건**의 검색이 발생

3. 데이터 분석 결과

1) 각 언어별 기초 통계량 확인





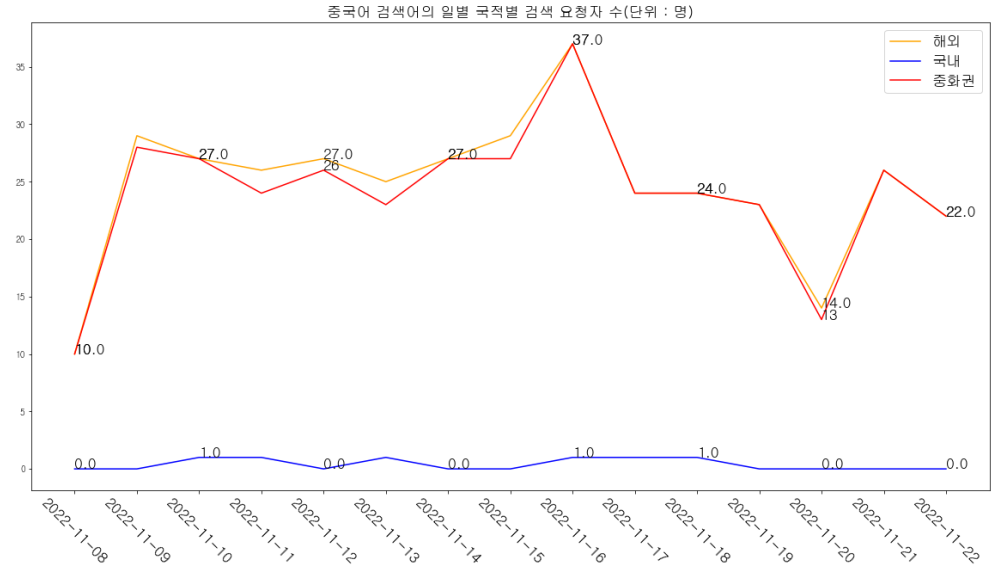
-. 영어 검색어의 검색자수 / 검색량 통계

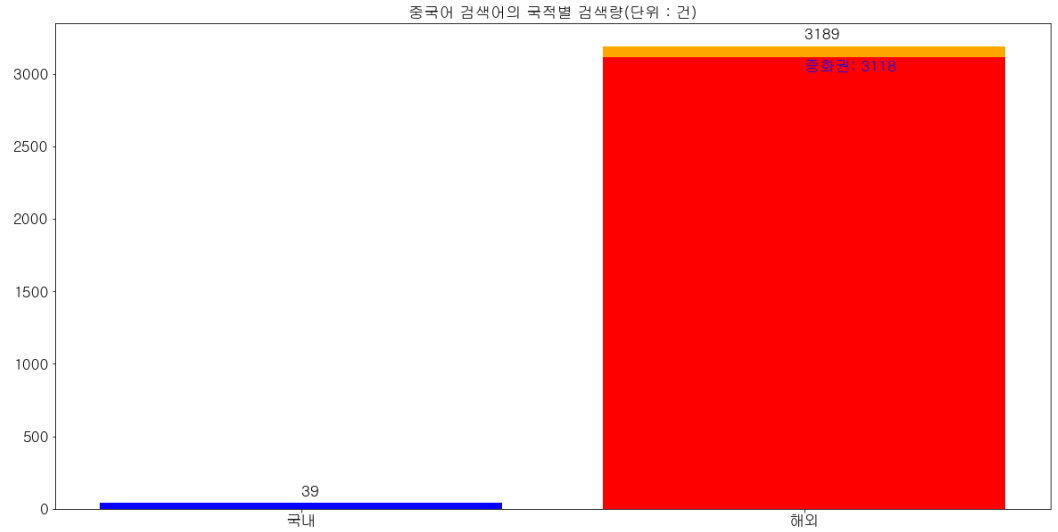
① **영어 검색어**의 경우 **국내 / 해외 유저**의 **검색자수(명)**가 **엇비슷**한 수준을 기록

② **국내 유저**가 **4,123건**의 **검색어**를 생산 할 동안 **해외 유저**의 경우 **4,735건**의 검색량을 생산했으며, 그 중 **중화권** 유저가 **4,075건**이었음

3. 데이터 분석 결과

1) 각 언어별 기초 통계량 확인



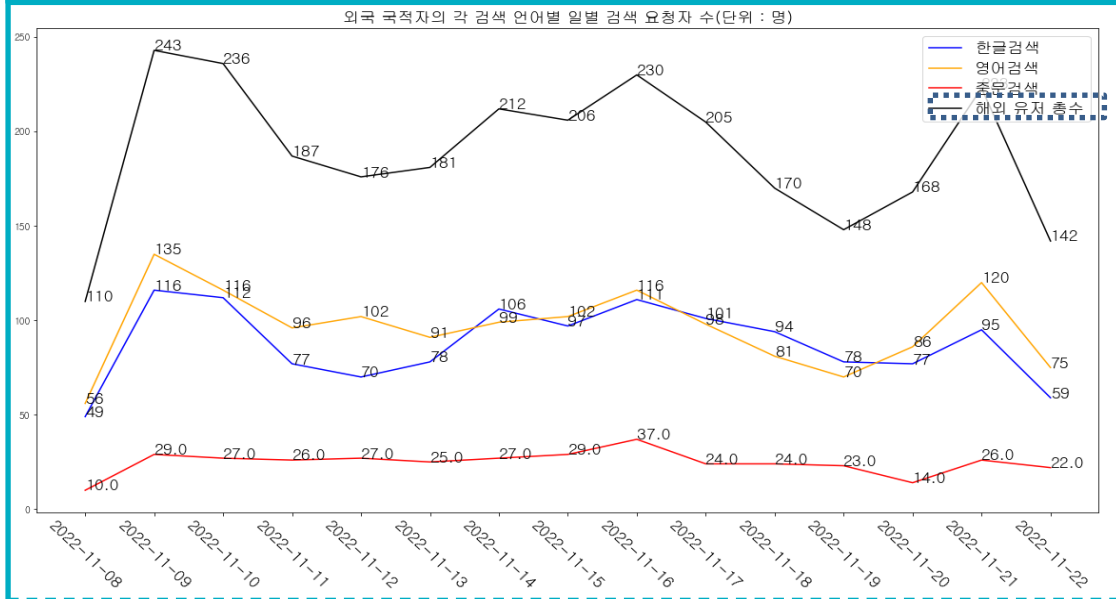


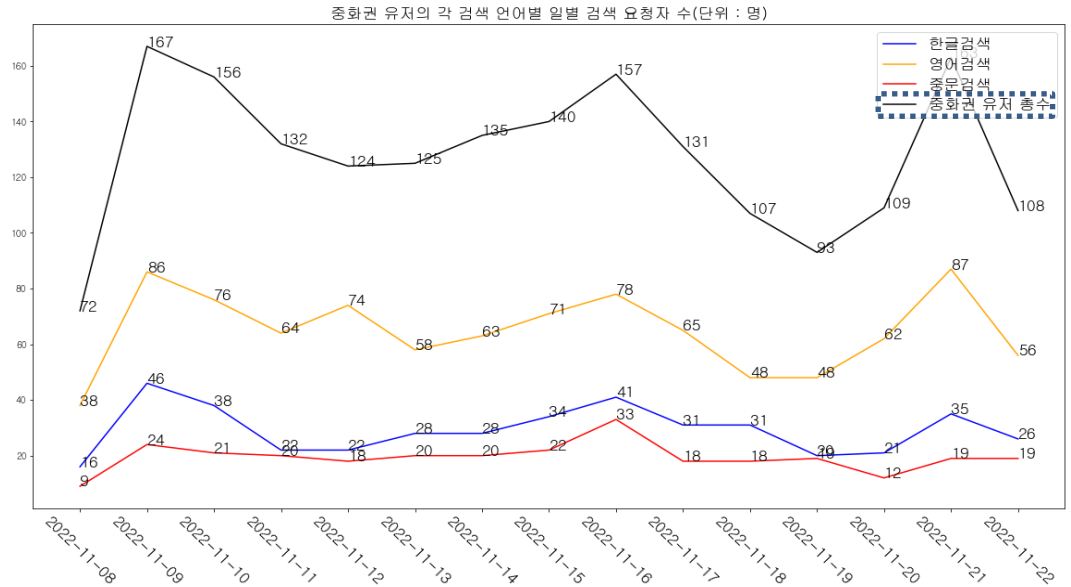
-. 중국어 검색어의 검색자수 / 검색량 통계

① 중국어 검색의 경우 **국내 유저**가 **검색**한 경우는 **거의 없었고**, **절대 다수**가 **중화권** 유저에 의해 생산

3. 데이터 분석 결과

1) 각 언어별 기초 통계량 확인





-. 해외 유저 한정 한글 / 영어 / 중문 검색 현황 비교

① **해외 유저(중화권 포함)**로 한정 지어서 각 **언어별 검색자수(명)**를 비교하면 **영어 / 한국어 비중 비슷**. 해외임에도 불구하고 한국어 高
② **중화권 유저**의 경우 **중국어 검색어**를 그대로 입력하기보단 **영어**로 **입력**하거나, **한국어**로 **번역**해서 **입력**하는 비중이 높은 것으로 추정 가능

3. 데이터 분석 결과

1) 각 언어별 기초 통계량 확인

-. 한국어 검색어 Top 20

순위	검색어	빈도
1	신상	2062
2	니트	2052
3	패딩	1716
4	세일	1471
5	무스탕	1176
6	원피스	1162
7	트위드	1117
8	패딩조끼	935
9	맨투맨	931
10	나이키	848
11	코트	788
12	기모	728
13	가디건	710
14	잠옷	705
15	숏패딩	689
16	양말	676
17	크리스마스	637
18	골프	624
19	조거팬츠	563
20	핸드메이드코트	525

-. 영어 검색어 Top 20

순위	검색어	빈도
1	mtm	80
2	scarf	43
3	sock , socks ※	29
4	tweed	25
5	shoes	22
6	set	19
7	bag	18
8	monroe	18
9	leather pants	18
10	t	18
11	apm	16
12	egg	16
13	tweed jacket	16
14	bra top	15
15	sale	15
16	aosta	14
17	g	14
18	brooch	14
19	party	13
20	lace	12

-. 중국어 검색어 Top 20

순위	검색어	빈도
1	帽	87
2	圍巾	82
3	羊毛外套	73
4	短裙	72
5	帽子	61
6	套裝	54
7	襪子	52
8	連衣裙	49
9	外套	45
10	背心	40
11	針織	40
12	牛仔外套	39
13	長裙	39
14	鞋	34
15	毛衣	32
16	起毛	31
17	短褲	30
18	衛衣	29
19	長款洋裝	28
20	馬甲背心	27

※ FastText 모델이 동일 단어로 판단하여 병합

3. 데이터 분석 결과

1) 각 언어별 기초 통계량 확인

-. 국내 유저 검색어 Top 20

순위	검색어	빈도
1	신상	1757
2	니트	1689
3	패딩	1625
4	세일	1300
5	무스탕	1108
6	트위드	1045
7	패딩조끼	858
8	맨투맨	829
9	나이키	814
10	원피스	761
11	코트	738
12	기모	720
13	잠옷	688
14	가디건	642
15	양말	628
16	숏패딩	620
17	골프	611
18	크리스마스	594
19	조거팬츠	534
20	핸드메이드코트	513

-. 해외 유저 검색어 Top 20

순위	검색어	빈도
1	원피스	401
2	니트	363
3	신상	305
4	부츠	276
5	스웨이드	208
6	세일	171
7	아우터	123
8	가을신상	114
9	울	112
10	맨투맨	102
11	패딩	91
12	帽	87
13	mtm	86
14	圍巾	82
15	패딩조끼	77
16	羊毛外套	72
17	트위드	72
18	短裙	72
19	숏패딩	69
20	가디건	68

-. 중화권 유저 검색어 Top 20

순위	검색어	빈도
1	원피스	313
2	부츠	241
3	니트	201
4	아우터	122
5	가을신상	111
6	울	110
7	신상	105
8	帽	87
9	圍巾	82
10	mtm	80
11	短裙	72
12	羊毛外套	72
13	帽子	61
14	가디건	55
15	套裝	54
16	襪子	52
17	맨투맨	51
18	連衣裙	49
19	힐	45
20	外套	45

3. 데이터 분석 결과

1) 각 언어별 기초 통계량 확인

-. 국내 유저의 영어 검색어 Top 20

순위	검색어	빈도
0	ops	135
1	jk	66
2	apm	58
3	set	56
4	xxl	39
5	sale	38
6	usa	35
7	Vess Square Padding Vest Jacket	35
8	pxg	34
9	child sweatshirt	34
10	ct	31
11	miu	23
12	mtm	23
13	laine	22
14	denim jacket	22
15	ny	20
16	cd	20
17	nike	18
18	silhouette	18
19	sa	17
20	ugg	17

-. 중화권 유저의 영어 검색어 Top 20

순위	검색어	빈도
0	mtm	80
1	scarf	43
2	sock , socks ※	29
3	tweed	25
4	shoes	22
5	set	19
6	bag	18
7	monroe	18
8	leather pants	18
9	t	18
10	apm	16
11	egg	16
12	tweed jacket	16
13	bra top	15
14	sale	15
15	aosta	14
16	g	14
17	brooch	14
18	party	13
19	lace	12
20	blackfriday	12

※ FastText 모델이 동일 단어로 판단하여 병합

3. 데이터 분석 결과

2) 검색어 품질 관련 이슈 우려 사안

-. 고유명사를 동사로 번역한 사례

① 브랜드명, 상품명 등의 고유명사를 동사형태로 번역한 사례 다수 발견

② 이는 기본적으로 '문장 대 문장' 번역을 상정하고 학습된 파파고의 한계로 추정

③ 예를 들어, playd는 영어 형태 'played'를 번역한 단어인 '재생됨 ' 으로 번역하나, 브랜드명 'play'가 이미 존재. 또 雪紡은 시폰(비단, 나일론)의 재질을 의미하나, 이를 알 수 없는 동사어로 번역함

검색어	번역어	모델 매칭 결과	유사도
playd	재생됨	Play(브랜드)	0.91
Joa	싸고좋아	조아(브랜드)	0.90
Goodgirl	잘했어	Goodboys(브랜드)	0.80
Arrow	화살을 쏘다	Arrows(브랜드)	0.95
Let your beautiful	당신의 아름다움을 맡기다	-	-
Bottleseoul	영혼을 병에 담다	Bottle(브랜드)	0.92
Flatyou	너를 때려눕혀라	Flat, flatyou(브랜드)	0.95, 0.89
Bowen official	공손히 인사하다	-	-
Angkor	웅크리다	Gonggam, 앙꼬(브랜드)	0.85, 0.83
Mienn	광산을 하다	Mignon, 미엔느(브랜드)	0.91, 0.85
雪紡	양복의 가죽이 누더분하다	스노우 앙벙거지, 스노우 누빔	0.86 0.85
深藍色上衣	야구공이 너덜너덜하다	블루종셔츠 블루문원피스	0.82, 0.81

3. 데이터 분석 결과

2) 검색어 품질 관련 이슈 우려 사안

-. 고유명사를 형용사로 번역한 사례

① 브랜드명, 상품명 등의 고유명사를 그대로의 형용사형으로 번역한 사례 다수 발견

② Double을 '두배로'로, And More를 '그리고 더보기 ' 로 번역하여 적용 중

검색어	번역어	모델 매칭 결과	유사도
viviene	살림살이의	Vivienne(브랜드)	0.93
Glowy	광채가 나는	딜라이트(브랜드)	0.90
Berkeley	냉정하게	Bervery(브랜드)	0.89
DOUBLE	두 배로	DOUBLE.H(브랜드)	0.89
Winsome	애교 있는	Wisse(브랜드)	0.88
And More	그리고 더 보기	AND MORE(브랜드)	0.85
CHRISTIAN	기독교의	크리스티(상품)	0.86
Blackondeday	어느 날 까맣게	Blackpig(브랜드)	0.90
Herohero	영웅적인	Hero(브랜드)	0.90

3. 데이터 분석 결과

2) 검색어 품질 관련 이슈 우려 사안

-. 고유명사를 의역, 직역, 오역한 사례

① 브랜드명, 상품명 등의 고유명사를 의역하거나, 직역하거나, 혹은 전혀 엉뚱한 단어로 오역한 경우가 다수 발견

② Recipe는 브랜드 D'recipe를 의미하는 것으로 추정되나, 이를 '요리법 ' 으로 번역하여 적용

③ Corner는 브랜드 '코너 ' 를 의미하는것으로 보이나, 이를 '코너링 ' 으로 의역하여 적용

④ 加绒卫裤 는 양털로 이루어진 펑퍼짐한 바지를 의미하나, 이를 '캔버스 가방 ' 으로 오역하여 적용

검색어	번역어	모델 매칭 결과	유사도
Recipe	요리법	D'recipe(브랜드)	0.91
Goldtree	골드	Gold tree(브랜드)	0.99
Corner	코너링	코너(브랜드)	0.94
Atcorner	티셔츠	Thecornershop (브랜드)	0.83
Jay	젠체하는 사람	Jy, jaydoc(브랜드)	0.93, 0.91
卫裤	미니 원피스	속바지(상품)	0.86
加绒卫裤	캔버스 가방	기모 청바지, 쿤스배기바지(상품)	0.83 0.83
女装恤衫	여성 셔츠	여리여리 셔츠 원피스, 드레스핫찌셔츠	0.83 0.83
魚尾裙	롱 양장	첼시 원피스	0.80
紅色	니트 집업 코트	레드하트y(상품)	0.82
鯊魚	야구	오징어셋트	0.86
雙面	송편 캔버스화 쿠키	양면 뽀글골덴벙	0.89

결론

결론

LinkShops

1. 결론 요약

검색어 관련

-. 해외, 특히 중화권 유저의 경우 중국어 자체를 입력하기보단 한국어, 영어로 번역하여 입력하는 경우가 다수
-. 검색어 번역을 신뢰하지 않거나, 혹은 일차 시도 후 원하는 결과를 얻지 못했음을 추론 가능

검색 결과 품질 관련

-. 고유어(브랜드, 상품명)를 직역하거나 다른 품사(형용사, 동사 등)로 번역하여 오히려 검색 품질 저하를 초래하는 사례 다수 발견

결론

2. 연구의 한계 향후 시사점

분석의 한계

-. 분석 방법론 : 한국어 단어 - 타 언어 단어 교차 Pre-trained Model(FastText)의 부재로 인하여 차선책으로 문장 기반 모델(S-BERT)을 사용하였으나, 문장 기반이라는 한계점이 존재하여 유사도를 엄밀하게 도출 못함

-. 분석 서버 성능 : 저장 용량 한계로 영어 혹은 다국어가 섞인 Pre-trained Model로만 분석 수행

향후 분석 진행 방향

-. 검색어 유형 분류 : 유저가 입력한 검색어를 카테고리 / 브랜드 / 상품명 등 사전에 정한 기준에 따라 분류하여 비율 도출

참고문헌

참고문헌

- Piotr Bojanowski et.al, “**Enriching Word Vectors with Subword Information**”, arXiv(2016)
- Nils Reimers and Iryna Gurevych, “**Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks**”, arXiv(2019)