2022년 데이터 분석 청년인재

한류 데이터 사전 구축 및 분석 프로세스 수립

한국문화정보원 문화빅데이터부

조성주, 한지수

목차

01
분석 프로세스
수립 배경

분석 프로세스 수립 배경

'한류 외신기사 데이터 분석 사업' 소개

분석 내용

- 한국문화정보원과 해외문화홍보원 업무 협약 체결을 통한 주간 한류 외신데이터 분석 추진
- 해외문화홍보원이 운영 중인 세계 28개국 33개 한국문화원의 언론 모니터링을 바탕으로 주간 평균 120~130건 내외 한류 기사에 대한 키워드 분석

분석 활용 현황

- 분석 데이터를 개방하여 문체부 소속기관(88명), KOCIS(76명), 대통령실, 문화원에서 활용
- 문화 빅데이터 플랫폼을 통한 주간 보고서 및 데이터 개방을 통한 대국민 활용

문화 · 한류 주간 외신동향

2023. 2. 2.(5호) KOCIS(해문홍) 외신분석팀

1 금주 보도 현황

○ 보도 비중 (해문홍 자체 외신기사DB 기준)

❶ 1.26 ~ 2.2(8일 간) 문화·한류 기사는 21개국 93개 매체에서 98건 보도

- 문화 기사는 同기간 보고된 전체 외신 보도(33개국 235개 매체 645건)의 15.1%

❷ (주요 이슈) ▲英매체, 런던 최고의 전시 중 하나로 '한류! 코리안 웨이브'* 선정 (*문체부 해외 박물관의 한국실 지원사업 일환) ▲K팝 인기, 한국어 열풍 주도

- 관광·여행, 서울에서 가볼 만한 카페, 2023년 한국 최고의 벚꽃 명소 등 소개

- 주목할 만한 보도, <뉴욕타임스> 메타버스는 재미있을까? 한국에 물어보라

❸ (권역별 보도량) 아시아(44건) > 영·유럽(33건) > 북미(18건) > 중동(3건) 순

15.1%

• 보도 건수 상위 12국(건)
▸ 미국(18), 인도(10)
▸ 태국(10), 튀르키예(10)
▸ 벨기에(8), 필리핀(7)
▸ 영국(5), 일본(4)
▸ 호주(3), 카자흐(3)
▸ 독일(3), UAE(3)

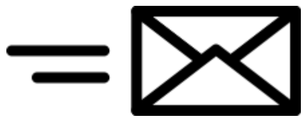
○ 문화·한류 관련 키워드 분석 (데이터분석 지원 : 한국문화정보원)

K팝	172	작품	58
중국	158	설날	55
드라마	111	공연	54
팬	94	BTS	52
인기	76	음식	50
영화	76	태국	49
블랙핑크	74	한류	48
미국	69	브랜드	46
새해	68	축제	43
행사	64	감독	41

분석 프로세스 수립 배경

'한류 외신기사 데이터 분석 사업'의 키워드 분석 업무를 진행

기간 : 2023.01.01 ~ 2023.02.16



> 2023-2호 문화한류 주간외신 동향 보고서 로우데이터를 송부드립니다.
> 지난주보다는 기사량이 늘었습니다.
> 23개국 87개 매체 101건 보도됐습니다.

해외문화홍보원에서 로우데이터 전달

워드클라우드 시각화 작업 진행

문화 한류 주간 외신동향 보고서 발간

분석 프로세스 수립 배경

매주 새로 생기는 외신기사 데이터에 대한 시각화를 진행하기 때문에

매주 새로운 불용어가 등장하고, 매주 새로운 트렌드 키워드가 계속해서 생겨남

따라서 데이터를 정제하는 과정이 **단순 반복 작업**으로 **비효율적**이라는 문제점 인식

효율적인 데이터 정제와 유의미한 키워드 도출을 위해

한류 데이터 사전을 구축하고 분석 프로세스를 수립하고자 하였음

02
분석 프로세스

분석 프로세스

1. 형태소 분석
2. 한류 데이터 사전
3. TF-IDF 모델
4. 워드클라우드 생성

1. 형태소 분석

기존에 사용하던 형태소 분석기 Okt는 **고유명사를 제대로 추출하지 못하는 문제점** 존재

한류 외신기사 키워드를 Okt 분석기로 적용해 본 결과, 다음과 같음

ex) 한국관광공사

형태소 분석기	형태소 분석 결과		
Okt	한국	관광	공사

-> '한국', '관광', '공사'는 개별적으로 빈도수가 추출되기 때문에 '한국관광공사'의 빈도수가 추출되지 않는 문제점 존재

ex) 친절한 금자씨

형태소 분석기	형태소 분석 결과		
Okt	친절한	금	자씨

-> 고유명사를 제대로 추출하지 못해 '자씨'만 보고 원본 명사를 파악하는 데 어려움 존재

1. 형태소 분석

Okt 문제점을 인식 후, 형태소 분석기를 변경하고자 한국관광공사' 키워드를 하나의 예시로 형태소 분석기 비교

형태소 분석기	형태소 분석 결과		
Okt	한국	관광	공사
Kkma	한국	관광	공사
Pecab	한국	관광공사	
Komoran	**한국관광공사**		

- 유일하게 Komoran만 '한국관광공사'를 하나의 명사로 추출
- '한국관광공사' 외에 한류 외신에 등장한 키워드를 형태소 분석기로 비교한 결과,
  명사 추출 성능이 가장 좋은 Komoran으로 형태소 분석기 변경

2. 한류 데이터 사전

- 형태소 분석기는 신어, 전문 용어 등 새롭게 등장하는 키워드를 명사로 추출하지 못하는 문제점 존재
- 따라서 형태소 분석기 수준에서 추출하지 못하는 단어를 대상으로 한류 데이터 사전 구축

  ex) 블랙핑크, 오징어 게임
- 기존에는 이러한 문제를 해결하기 위해 임의로 단어를 합치는 작업을 추가적으로 진행

	word	count
21	핑크	39
26	블랙	36
34	결심	30
61	게임	17
71	오징어	14

한류 데이터 사전 등록

이러한 정제 작업에 많은 시간이 소요되어 업무의 효율성을 높이기 위하여 한류 데이터 사전 구축

2. 한류 데이터 사전

한류 데이터 사전 구축에 관한 객관적인 기준 필요

기준 예시 1. 한류 데이터 사전에 정의할 단어인지

ex) 인기 드라마, 히트 영화, 고급 음식

기준 예시 2. 어디까지를 하나의 단어로 보아야 하는지

ex) 할리우드 영화, 고전 할리우드 영화

2. 한류 데이터 사전

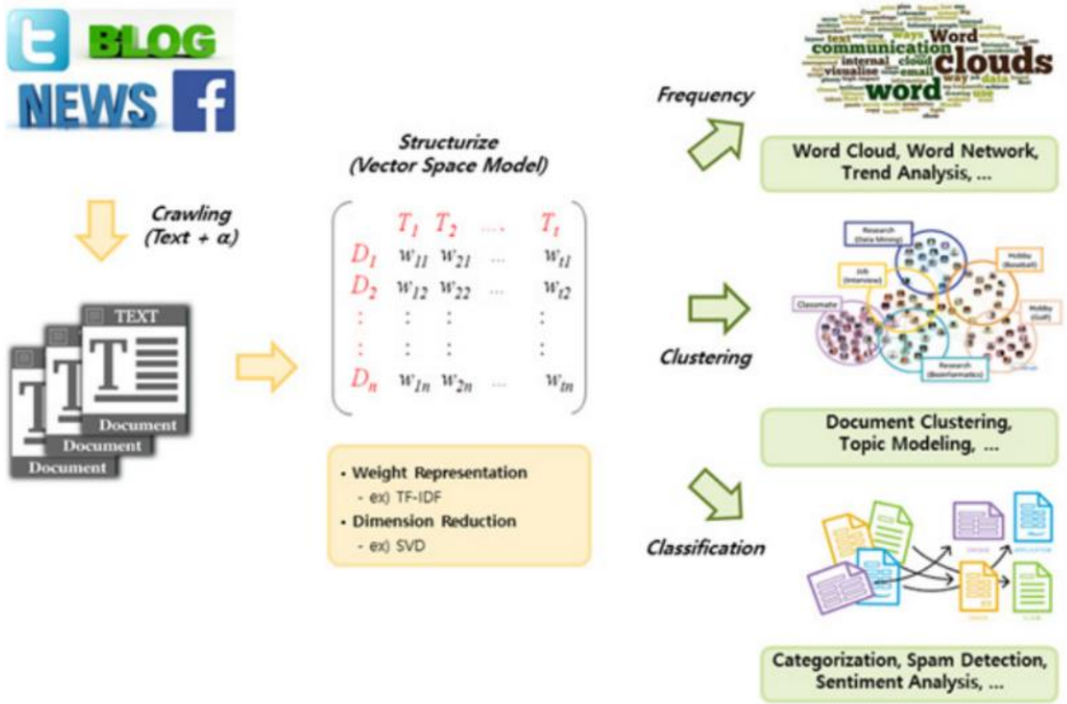
- '표준국어대사전'에 실린 어휘뿐만 아니라
  신어, 전문 용어 등 다양한 어휘가 수록되어 있는
  '우리말샘' 사전을 기준으로 사전 구축
  ex) 한류 관광, 한류 열풍, 추천 영상, 외국어영화상 등

- '우리말샘' 사전이란?
  - 국립국어원에서 만든 개방형 한국어 사전

3. TF-IDF 모델

- TF-IDF*란?
  - ○ 단어의 빈도와 역 문서 빈도를 사용하여 각 단어들마다 중요한 정도를 가중치로 주는 방법

    → 문서 내의 단어들에 각각 부여한 중요도를 나타내는 숫자값



[출처: 김남규, 텍스트 분석 기술 및 활용 동향, 2017]

참조 * : 공공빅데이터 표준분석모델 매뉴얼 (민원분야)

3. TF-IDF 모델

- 기존 프로세스에서 '것', '이', '수', '그' 등의 불용어가 빈도수 상위권에 존재하는 문제 발생
- 불용어 처리 과정은 많은 시간이 소요되는 비효율적인 단순 반복 작업
- 이를 개선하여 **불용어 처리 과정의 효율성을 높이기 위해** TF-IDF 적용
- TF-IDF를 적용한 결과 상위권에 유의미한 단어들이 존재하는 것을 확인할 수 있음

기존 프로세스 (**빈도수 사용**)

개선 프로세스 (**TF-IDF 사용**)

3. TF-IDF 모델

TF-IDF 모델 코드

- 한글, 영어 형태소 분석을 동시에 하는 모듈이 존재하지 않기 때문에 언어에 맞는 형태소 분석 후,

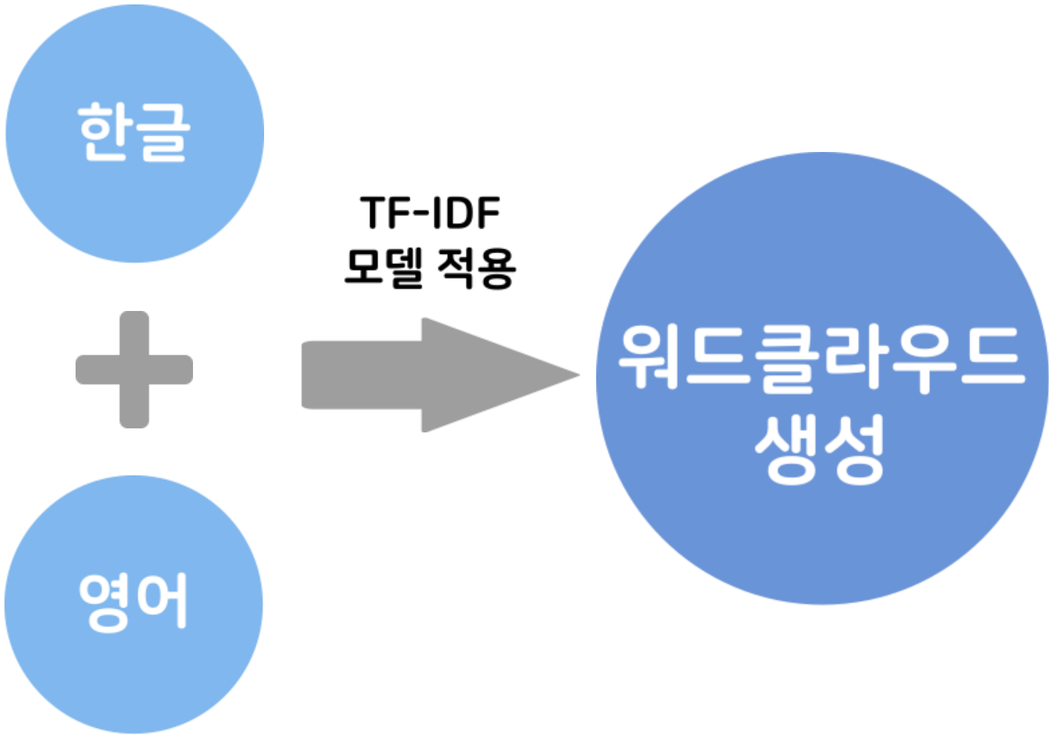
  하나의 리스트로 합쳐 TF-IDF 적용

```
total = []
for i in range(len(df)):
    temp = []
    temp.append(kor_total[i])
    temp.append(eng_total[i])
    total.append(' '.join(temp))
```

```
#TfidfVectorizer

from sklearn.feature_extraction.text import TfidfVectorizer
vectorizer = TfidfVectorizer()
tfidf = vectorizer.fit_transform(total)
```

```
tfidf_weights = [(word, tfidf.getcol(idx).sum())
                 for word, idx in vectorizer.vocabulary_.items()]
tfidf_weights[0:10]
```



03
프로세스 수립 결과

프로세스 수립 결과

1월	불용어 개수	한류 데이터 사전 (176개)
1주차	978개 -> 96개 (▼90%)	블랙핑크, 코로나19, 더 글로리 등 (65개)
2주차	823개 -> 92개 (▼89%)	박찬욱 감독, 헤어질 결심, 한류 열풍 등 (39개)
3주차	686개 -> 104개 (▼85%)	오징어 게임, 한복 디자이너 등 (43개)
4주차	639개 -> 129개 (▼80%)	디올, 소프트파워 등 (29개)

프로세스 수립 결과, **불용어 개수 80% 이상 감소**하여 데이터 정제 과정 효율화

한류 데이터 사전에 블랙핑크, 헤어질 결심, 오징어 게임 등이 등록되어 **한류 트렌드 반영**

1월 1주차부터 4주차까지 총 **176개의 키워드**를 한류 데이터 사전에 등록 완료

프로세스 수립 결과

한류 데이터 사전에 등록된 단어가 워드클라우드 상에 생성된 것을 확인할 수 있음

1월 1주차

1월 2주차

1월 3주차

1월 4주차

04
프로세스 효과 및 적용 확장방안

프로세스 효과

1) 업무 효율성 제고

- 기존 프로세스에서 한계점을 발견 후, 이를 개선하고자 함
- TF-IDF를 활용하여 불용어 처리 시간 기존 85분에서 15분으로 단축 (▼82%)
- 프로세스 소요 시간을 단축시켜 업무 효율성 제고

2) 한류 데이터 사전 활용

- 한류 도메인을 바탕으로 한 전문 용어, 신어까지 워드클라우드에 생성 가능
- 한류 데이터 사전 구축 완료 후, 데이터를 개방하여 활용 가능하도록 함

프로세스 적용 확장방안

사용자 사전 구축 시, 원문 데이터의 문맥 파악 과정에서
상당 시간 소요 & 인력 부족 한계점 발생

Transformer* 기반의 딥러닝 모델인 BERT를 사용하여
프로세스 적용 확장 가능

*Transformer 모델 : 문장 속 단어와 같은 순차 데이터 내의 관계를 추적해 맥락과 의미를 학습하는 신경망

참고문헌

- 정한영, "트랜스포머 모델, 다양한 산업에서 인공지능 혁신!...'트랜스포머 AI' 시대로" *인공지능 신문*, 2022년 4월 6일, https://url.kr/tr8w1e
- *공공 빅데이터 표준분석모델 매뉴얼: 민원분야* (행정자치부·한국정보화진흥원, 2017), 310.
- 유원준·안상준, "04-04 TF-IDF(Term Frequency-Inverse Document Frequency)" 위키독스, 2022년 11월 14일, https://wikidocs.net/31698
- "우리말샘 소개" 우리말샘, https://opendict.korean.go.kr/service/helpList
- 김남규·이동훈·최호창, *"텍스트 분석 기술 및 활용 동향" 한국통신학회논문지 V.42 NO.2(2017)* 471-492

THANK YOU