텍스트 연관성(상관계수) 분석/네트워크그래프제작

TECHNOLOGY & PROGREMMER

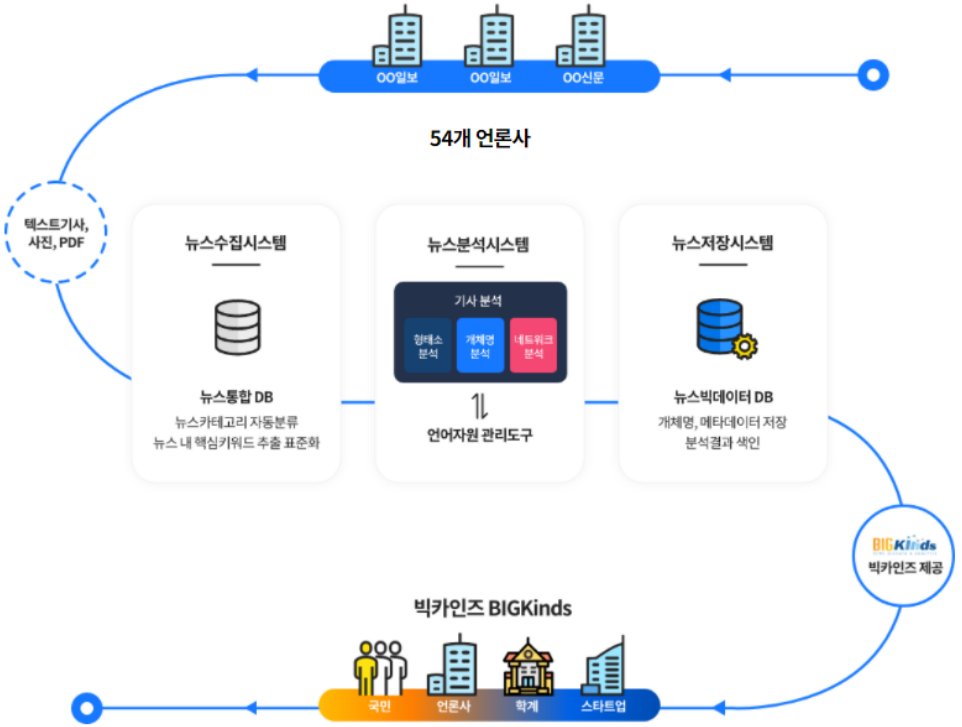
목차

Part1. 뉴스빅데이터 분석 서비스, BIGKinds 활용

1. 소개 https://www.bigkinds.or.kr/

한국언론진흥재단의 무료 플랫폼으로
뉴스수집시스템, 분석시스템, 저장시스템 등으로
구성돼 있으며,
저장된 뉴스 분석 정보는 국민, 언론사, 학계,
스타트업 등이 활용할 수 있는
뉴스빅데이터 분석서비스 '빅카인즈(BIGKinds)'로
제공됨

비정형 신문기사 → 정형화된 토큰 으로 제공함



Part1. 뉴스빅데이터 분석 서비스, BIGKinds 활용

2. 활용논문 보기

출처: (논문제목)빅카인즈로 분석한 COVID-19 전후의 건강지향적 베이커리 트렌드 분석

초록·키워드 목차

오류제보하기 >

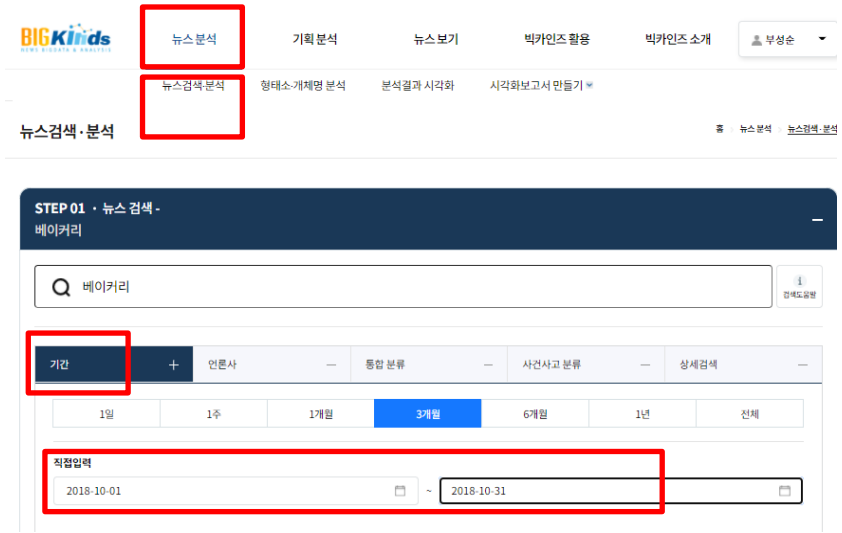
본 연구는 빅카인즈(Bigkinds)를 활용하여, 최근 5년(2016~2020)간 뉴스 빅데이터의 연도별 베이커리 성분 변화에 대한 분석을 진행하였다. 베이커리 관련 트렌드는 성분을 중심으로 변화하고 있었고 COVID-19(Corona Virus Disease-19)발생 전인 2016~2019년에는 베이커리 제품에 대한 맛, 재료, 영양, 성분 및 신선도와 관련된 키워드가 많이 나타났고 2020년부터 언택트(Untact)문화가 자리잡으면서 '구독서비스', 'HMR(Home Meal Repl acement)'와 관련된 키워드가 새롭게 등장하였다. 본 연구에서는 뉴스 빅데이터를 분석하고, 1인가구 증가와 더불어 간편식을 추구하는 현상에서 사람들이 웰니스(Wellness)를 위해 베이커리 소비패턴이 어떻게 변화하는지 조사하였다. 연구 결과, COVID-19전에는 비건, 채식주의 등의 친환경적인 소비패턴으로 사람들이 건강 증진을 위한 것으로 나타났고 COVID-19이후, 폴리페놀, 면역력 증진 등 예방 차원에서의 소비패턴을 보였다. 또한, 도출된 키워드들에 대하여 전문가 인터뷰를 통해 결과를 재검증했다. 본 연구를 통해 기존 빅데이터 분석 툴(Tool)이나 프로그램(Program)을 설치하지 않아도 다양한 분야에서 빅카인즈를 통해 용이하고 신속하게 대용량의 빅데이터를 분석하고 인사이트를 얻을 수 있다는 점에서 실무적 시사점을 도출하였다.

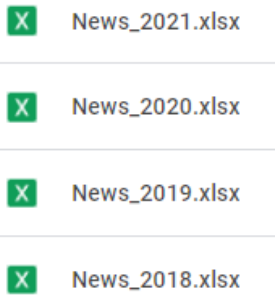
Part1. 뉴스빅데이터 분석 서비스, BIGKinds 활용

3. 빅카인즈 데이터셋 다운로드 및 파이썬 실행

A https://www.bigkinds.or.kr/ 회원가입후 로그인

B 뉴스분석-뉴스검색분석-검색어 입력과 기간설정하여 엑셀 자료 다운로드





Part1. 뉴스빅데이터 분석 서비스, BIGKinds 활용

3. 빅카인즈 데이터셋 다운로드 및 파이썬 실행

```
1 import pandas as pd
2 df = pd.read_excel('./data/News_2018.xlsx')
3 df.head()
```

```
/usr/local/lib/python3.7/dist-packages/openpyxl/styles/stylesheet.py:226: UserWarning: Workbook contains no default style, apply openpyxl's default
  warn("Workbook contains no default style, apply openpyxl's default")
```

	뉴스 식별자	일자	언론사	기고자	제목	통합 분류 1	통합 분류 2	통합 분류 3	사건/사고 분류 1	사건/사고 분류 2	사건/사고 분류 3	인물	위치	기관	키워드	특성추출(가중치순 상위 50개)	본문
0	1.300201e+06	20181130	강원일보	허남윤	`2030 여성 인기' 출시 6개월만에 50만병 판매	사회>여성	NaN	NaN	NaN	NaN	NaN	NaN	서울,횡성	코엑스,국순당,유산균 막걸리	인기,여성,출시,50만,판매,국순당,유산균,막걸리,전통주,기업,횡성,국순당,여성,디...	유산균,국순당,서울디저트쇼,여성층,마리,전시회,서울,코엑스,1병,전통주,50만,횡성	전통주 기업인 횡성 국순당이 2030 여성들을 잡기 위해 `디저트 마케팅'을

	18만	1병	20만	250여개	40만	50만	5천	brewing	hmr	가정간	...	출시	카페노무	커피나무	코엑스	타깃	판매	평가	홈페이지	홍보관	횡성
0	0	1	0	0	0	1	0	0	0	0	...	1	0	0	1	1	1	1	0	0	1
1	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	1	1	...	0	0	0	0	0	0	0	0	0	0
3	1	0	1	0	1	0	0	0	0	0	...	0	0	0	0	0	0	0	1	0	0
4	0	0	0	1	0	0	1	1	0	0	...	0	1	1	0	0	0	0	0	1	0

5 rows × 111 columns

Part2. 단어간 상관도 분석의 이해

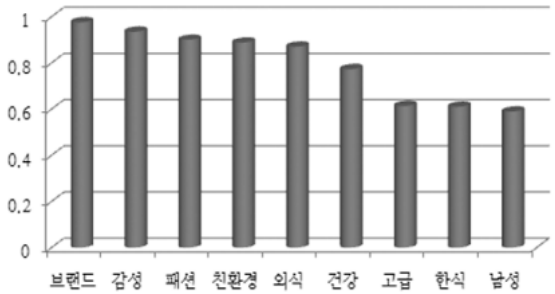
1. 논문 보기

신문 텍스트로 살펴본 문화 소비 현상의 트렌드

(A Trend Analysis of Cultural Consumption Based on Newspaper Texts)

요 약 본 논문은 물결 21 코퍼스, 즉 동아, 조선, 중앙, 한겨레 신문의 2000-2009년까지의 약 4억어절의 신문 자료에서 나타나는 문화 소비 현상의 트렌드에 대한 분석이다. 구체적으로, 명사 '트렌드'와의 공기어(공기 명사) 중에서 10년 동안 꾸준히 증가하는 단어들(일반 명사, 고유 명사)을 살펴보고 이것들의 속성에 따라서 명사를 분류하여 공기어의 증감도를 살펴본다. t-점수(t-score)를 이용하여 공기어를 추출하고 만-켄달 분석을 통하여 증감도를 분석하여 매년 공기하여 나타는 정도가 증가하는 단어를 대상으로 연구한다. 또한 이러한 명사들이 주제어와 변수 사이에 상관관계가 있는지 분석하여 10년 동안 트렌드에 대해서 주제별로 분석한다. 결과적으로, 신문에 나타나는 단어를 통해 사회적, 문화적 트렌드를 관찰할 수 있다.

키워드 : 트렌드, [물결 21] 코퍼스, 신문, t-점수, 공기어, 만-켄달 분석, 증감도 분석, 상관관계 분석



'트렌드' 단어의 상관도

3. 트렌드와 공기어

다음 표는 '트렌드'의 공기어를 증감도 분석에 의해 증가 추세로 나타난 단어들(0.5이상)을 추출하여 [13-15]에 따라 주제별로 분류한 것이다.

표 1 '트렌드'의 공기어 중 10년 동안 증가추세 단어(괄호 속 숫자는 tau value)

패션(0.91)	스타일(0.86)	브랜드(0.86), 드라마(0.85), 디자이너(0.82), 명품(0.72), 캐주얼(0.72),
	뷰티(0.69)	피부(0.74), 화장품(0.6), 메이크업(0.58)
	컬러(0.64)	검정(0.51)
	소재(0.64)	고급(0.69), 고급화(0.61)
	아이템(0.85)	옷(0.81), 구두(0.72), 의류(0.63), 가방(0.61), 핸드백(0.6), 시계(0.51)
	남성(0.64)	
웰빙(0.63)	외식(0.86), 한식(0.6), 카페(0.61)	
여행(0.59)	친구(0.86), 혼자(0.82), 자전거(0.68), 관광(0.55), 자동차(0.51)	
생활(0.61)	친환경(0.86), 인테리어(0.82), 장식(0.55)	
감성(0.87)	문화(0.77)	
소비자(0.86)	고객(0.95), 소비(0.77), 시장(0.73), 쇼핑(0.58)	

Part2. 단어간 상관도 분석의 이해

2. 파이썬을 이용한 상관도 분석

txt=['파이썬 차트 파이썬 머신러닝',
'차트 파이썬 R 차트',
'R 분석 시각화']

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \times \sum(y_i - \bar{y})^2}}$$

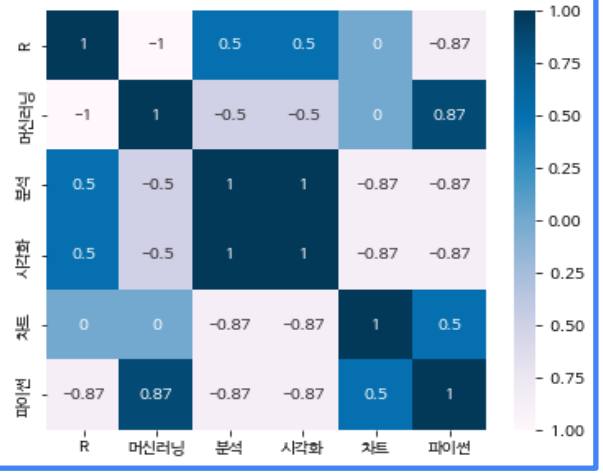
데이터프레임구조 df.corr()

	머신러닝	분석	시각화	차트	파이썬
0	1	0	0	1	2
1	0	0	0	2	1
2	0	1	1	0	0

넘파이 구조 np.corrcoef

```
[[1 0 0 1 2]
 [0 0 0 2 1]
 [0 1 1 0 0]]
```

히트맵 시각화

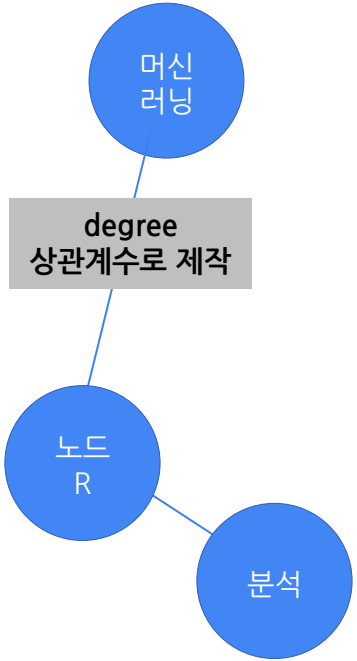


Part2. 단어간 상관도 분석의 이해

3. 상관계수를 이용한 네트워크 그래프 작성

	R	머신러닝	분석	시각화	차트	파이썬
R	1.000000	-1.000000	0.500000	0.500000	0.000000	-0.866025
머신러닝	-1.000000	1.000000	-0.500000	-0.500000	0.000000	0.866025
분석	0.500000	-0.500000	1.000000	1.000000	-0.866025	-0.866025
시각화	0.500000	-0.500000	1.000000	1.000000	-0.866025	-0.866025
차트	0.000000	0.000000	-0.866025	-0.866025	1.000000	0.500000
파이썬	-0.866025	0.866025	-0.866025	-0.866025	0.500000	1.000000

[('R', '머신러닝', -1.0),
('R', '분석', 0.4999999999999999),
('R', '시각화', 0.4999999999999999),
('R', '차트', 0.0),
('R', '파이썬', -0.8660254037844385),
('머신러닝', '분석', -0.5),
('머신러닝', '시각화', -0.5),
('머신러닝', '차트', 0.0),
('머신러닝', '파이썬',
0.8660254037844387),
('분석', '시각화', 0.9999999999999998),
('분석', '차트', -0.8660254037844385),
('분석', '파이썬', -0.8660254037844385),
('시각화', '차트', -0.8660254037844385),
('시각화', '파이썬', -0.8660254037844385),
('차트', '파이썬', 0.5)]



Part3. 파이썬 프로그램 제작

1. MyTf.py 모듈로 DTM 제작

2. 넘파이에서 제공하는 corr함수 적용

3. 직접 corr 함수 제작

4. corr값을 이용하여 시각화 및 네트워크 그래프 제작

5. 빅카인즈 제공 신문기사 자료로 상관도 및 연관분석(미션)