소셜 데이터를 활용한 텍스트 분석
: 인스타그램 기반 검색 키워드 관련 트렌드 제공

이승건, 모준우, 온유나
경희대학교 산업경영공학과
leegun9371@naver.com, liarcrown@naver.com, ohnyn823@gmail.com

Text Analysis Using Social Data
: Provision of trend related to Instagram-Based Search Keyword

Lee Seung Geon, Mo Joon Woo, Ohn Yoo Na
Kyung Hee University Industrial & Management Systems Engineering

요 약

특정 인사이트를 얻기 위해 다양한 웹에서 정보를 크롤링하는 방법이 많이 연구되고 있다. 또한 SNS 사용이 활성화됨에 따라 방대한 규모의 소셜 데이터가 자발적으로 눈 상에서 생성되고 있다. 본 논문에서는 Selenium 라이브러리를 이용하여 연령별 이용수가 가장 많은 인스타그램의 소셜 데이터를 통해 현재 사용자들의 관심사와 needs를 분석한다. 이를 통해 주 소비자층인 20, 30대가 관심을 갖는 특정 키워드와 관련된 다른 검색어들의 빈도수를 확인하고, 그 결과로 분석된 트렌드를 시각화하여 파악하였다. 제안된 분석 결과를 통해 SNS에서 정보를 얻는 사용자들이 특정 키워드에 대한 트렌드를 파악하고 이를 개인적 의사 결정 과정 또는 마케팅 전략에 활용할 수 있다.

1. 서 론

코로나19가 장기화함에 따라, 소셜미디어 시장이 상승 추세로 진입했다. 시장조사업체 DMC미디어의 '2021 소셜 미디어시장 및 현황 분석 보고서'에 따르면 2021년 1월 기준 우리나라의 소셜 미디어 이용률은 89.3%로 세계 평균(53.6%)보다 약 1.7배 높다.

그림 1. 국가별 소셜미디어 이용률

또한 국내 인스타그램과 트위터, 틱톡 이용자는 전년 대비 증가하는 추세를 보인다. 그중 연령별로 가장 많이 사용하는 소셜미디어는 10대, 20대, 30대 모두 인스타그램으로 나타났다.

그림 2. 연령별 가장 많이 이용하는 소셜미디어 top 3[1]

이렇듯 최근 주 소비자층의 SNS 사용이 증가함에 따라 특정 토픽에 대한 개인의 소셜 데이터 또한 증가하고 있다.[2] 본 논문에서는 주 소비자층이 가장 많은 인스타그램 소셜 데이터를 크롤링하고 대량의 데이터로부터 특정 키워드에 대한 새로운 패턴 또는 동향 발생을 살펴본다.[3]

인스타그램에서 소셜 데이터는 사용자, 장소 데이터, 검색어, 해시태그 등의 형태로 다양하게 나타난다. 주요 소비자층이 인스타그램을 사용한다는 가정하에, 게시글 데이터와 해시태그 데이터를 구별하여 데이터 별 가중

[1] DMC MEDIA, 2021 소셜 미디어시장 및 현황 분석 보고서, 2021

[2] 이채완, 오정민, 이상호 , "사람들은 왜 관광경험을 SNS 에 전시하는가?" , International Journal of Tourism and Hospitality Research Volume 34 , p21~32 (2020.3)

[3] 윤현주, 신재영, 인스타그램 게시물 데이터를 활용한 건강기능식품 브랜드 분석 및 평가, 한국컴퓨터정보학회 하계학술대회 논문집, 29, 2, 2021

치를 부여하고 빈도수에 대해 정렬하는 등 다양한 알고리즘을 통해 특정 키워드에 대한 트렌드를 시각화한다. 본 논문의 구성은 다음과 같다. 2 장에서는 분석에 활용한 방법 및 알고리즘에 대한 설명, 3 장에서는 언급한 방법을 통해 분석한 결과 값을 해석한다. 4 장에서는 프로젝트를 진행하면서 느꼈던 소감에 대해 서술하고 결론을 맺는다.

2. 방 법

2.1. TfidfVectorizer

TfidfVectorizer 는 불용어처럼 모든 문서에 자주 등장하는 용어는 덜 중요하며, 특정 문서에 자주 등장하는 용어가 해당 문서에서 더 중요하다고 여기는 워드 임베딩 방식이다. 가중치를 부여함으로서 분석 목적에 맞는 효율을 느끼고 데이터들의 대표성을 확보할 수 있다. [4] 모든 게시글 단어를 document data matrix 로 변환한 뒤, TfidfVectorizer 를 통해 각 단어에 가중치, 즉 TF-IDF 값을 부여한다.

2.2. Word2Vec

Word2Vec 워드 임베딩은 [그림 3]과 같이, 단 하나의 은닉층을 이용하는 neural network 방식이다.

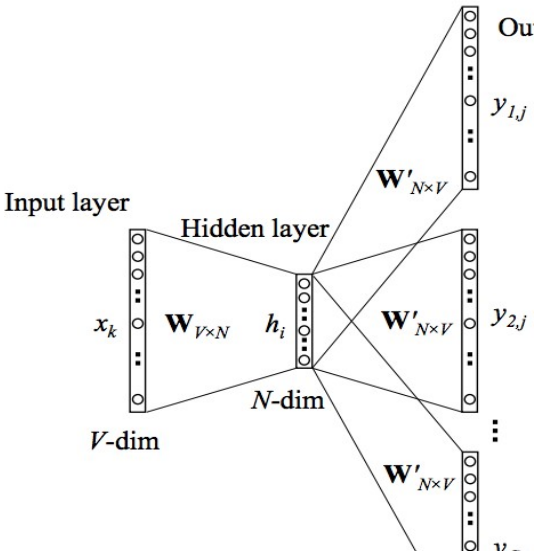

그림 3. Word2Vec의 구조

CBOW 와 skip-gram 중 skip-gram 을 선택하였는데, 이는 중심 단어로 주변 단어를 학습하여 예측하는 방식이다. 모든 단어를 중심 단어로 삼아, 각 단어의 윈도우 크기(학습할 단어 개수)만큼 주변 단어를 가져와 학습하는 과정을 진행한다. 학습을 통해 중심 단어와 주변 단어 벡터의 내적이 cosine similarity 가 되도록 벡터를 벡터 공간에 임베딩할 수 있다.

[4] 김규성, 황영은, 박진우 "패널조사에서 가중치 부여 방법 및 효과에 관한 연구"

임베딩 후 사용자 입력 키워드와 모든 단어들 간 cosine similarity 계산 및 평균값을 도출하고, 평균값을 넘는 similarity 값을 갖는 단어들만 추출한다. 이 과정을 통해 검색 키워드의 트렌드와 무관한 단어들을 제거할 수 있다.

2.3. Topic Modeling

토픽 모델링이란 자연어처리 과정에서 문서 집합의 잠재적인 주제를 발견하기 위한 통계적 모델이다. 텍스트 마이닝 기법을 통해 데이터의 숨겨진 의미를 발견할 수 있다.

본 프로젝트에서는 토픽모델링 기법 중 하나인 잠재 디리클레 할당, LDA 를 사용하여 토픽모델링을 진행하였다. LDA 란 주어진 문서에 대해 어떤 주제들이 확률적으로 존재하는지에 대한 내용을 다룬다. 사용자가 특정 토픽 수를 정하면, LDA 는 특정 토픽에 특정 단어가 토픽을 뽑아낸다는 건, 특정 선택한 토픽 문서에서 앞서 나타난 단어들의 출현 확률분포를 바탕으로 문서에 사용할 단어를 고르는 과정이다.

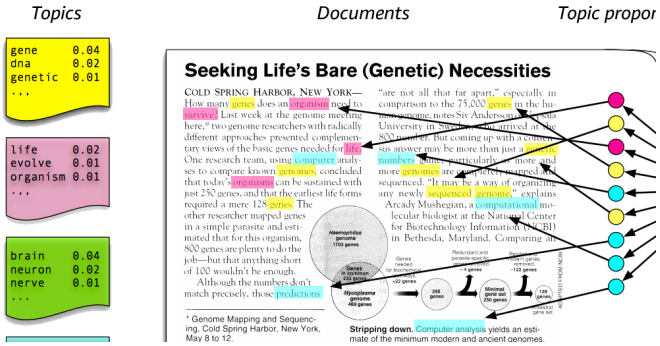

그림 4. LDA의 개략적인 도식

인스타그램의 게시글 데이터를 크롤링하면, Mallet 모듈을 통해 LDA 객체를 만든다. 각 객체를 크롤링한 데이터를 이용해 학습시키고, 그렇게 학습된 모델은 각 단어의 확률분포에 대한 결과를 보여준다.

마지막으로 wordcloud 와 pyLDAvis 를 이용하여 시각화를 진행한다. wordcloud 에서 큰 단어들은 빈도수가 높은 단어들을 의미한다. Intertopic Distance map 에서, 좌측의 원들은 각 토픽을 의미하고 각 토픽들의 상이성 유무는 원들의 거리로 판단한다. 두개의 원이 겹치면 두 토픽은 유사한 토픽이라는 의미이다.

3. 결 과

3.1. Word Embedding - TfidfVectorizer

'네이마르'를 검색 키워드로 설정 시 크롤링된 인스타그램 게시글 데이터를 TfidfVectorizer 로 임베딩한 뒤, 각 단어의 TF-IDF 가중치 값을 계산했다. 내림차순 정렬 후 10 개의 순위만 나타내었다.

TF-IDF 값 순위

1 : 브라질 0.59581
2 : 선수 0.34859
3 : 경기 0.3397
4 : 축구 0.29346
5 : 직관 0.26678
6 : 대한민국 0.17785
7 : 한국 0.15117
8 : 손흥민 0.1494
9 : 이마루 0.09604
10 : 진짜 0.08359

그림 5. TF-IDF 값에 따른 내림차순 순위

3.2. Word Embedding - Word2vec

'네이마르'에 대해 크롤링된 인스타그램 게시글 데이터를 Word2Vec 으로 임베딩 한 후, 검색 키워드에 대한 cosine similarity 가 평균 이상인 단어만 추출한다. 이들의 wordcloud 를 그려 검색 키워드와 유사한 동시에, 자주 등장한 게시글 단어가 무엇인지 사용자가 직관적으로 파악하게 된다.

그림 6. Word2Vec과 cosine similarity를 통한 wordcloud

3.3. Topic Modeling

상단과 동일한 키워드로 크롤링된 인스타그램 게시글 데이터의 각 토픽별 빈도수를 바탕으로 그려진 그림이다. 키워드는 최근 가장 핫한 키워드인 네이마르를 지정했다. 임의로 정해진 토픽 4 가지에서 다음과 같은 wordcloud 가 나타났다. 각 토픽 순서대로, 축구/경기장/나라/선수를 임의로 선정했다.

그림 7. Topic Modeling을 통한 wordcloud

위 결과는 최근 있었던 한국 vs 브라질 경기 관련 내용이 적절하게, 각 wordcloud 에 토픽별로 분류되어있음을 보여준다.

3.4. Hashtag

상단과 동일한 키워드로 크롤링한 해시태그 데이터에 대해 전처리를 실시했다. 의미 없는 광고성 태그 혹은 이모지 등 불용어를 제거하고, 의미 있는 단어만 추출하여 리스트에 저장했다.

[그림 8]은 키워드 '네이마르'에 대한 해시태그 크롤링 데이터의 단어 빈도수를 바탕으로 제작한 wordcloud 이다.

그림 8. 단순 해시태그 단어 빈도수 wordcloud

앞서 게시글 데이터로 제작한 wordcloud 와 비슷한 모양의 wordcloud가 나타난다. 이를 통해 사용자들이 게시글과 같은 맥락의 해시태그를 사용하고 있음을 확인할 수 있다.

3.5. 좋아요 수에 따른 인기 게시글 순위

상단과 동일한 키워드로 크롤링된 데이터 재사용성에 대해 논의한 결과, 세가지 주요 기능 이외의 부가적인 기능 또한 구현하였다.

좋아요 수 데이터를 전처리를 통해 정수 데이터로 변환, sorting 을 통해 크롤링 된 데이터 중 좋아요 수에 대한 순위를 정렬하고 이를 보여준다.

	content	date	like	place	
134	끝까지 훈훈하네요..! #네이마르 #브라질 #축구 #브라질전 #축구대표팀 #축덕...	2022-06-06	998		[네이마르, 브라질, 축구, 브라질전, #]
423	호감도 정점찍은경기 직 후 네이마르 인터뷰.한국와서 호감도 오지게 얻고 감ㅋㅋㅋㅋㅋ...	2022-06-04	935		[축구, 축구선수, 브라질, 브라질대
229	'날강두' 사태 겪은 사람이 건 미친 조건 ㅋㅋㅋㅋㅋ	2022-06-05	893		
200	2022.06.02 ⚽⚽⚽ 🎬 NO.10 NEYMAR JR (네이마르...	2022-06-05	820	Seoul World Cup Stadium	
59	토트넘 새 유니폼 놀리는 에브라:1990년부터 매년 똑같음 ㅋㅋㅋㅋㅋㅋ	2022-06-07	783		
231	네이마르 '45분 이상 출전' 조항 추가 의무 조건"만약 네이마르가 한국에서 뛰...	2022-06-05	743		[네이마르, 네이마르45분이상출전, 네이마르
750	서울 경기보다 더 신났던 브라질전 ⚽⚽축구장에서 이렇게 많은 사진을 남겨보는 것...	2022-06-03	715	서울월드컵경기장	
751	오늘 하루 행복해질 수 있는 마법 @lululala.lime———————————...	2022-06-03	674		[축구, 친성경기, 한국, 브라질,

그림 9. 좋아요 수에 따른 인기 게시글 순위 결과

3.6 인기있는 장소 순위 (핫플레이스)

상단과 동일한 키워드로 크롤링된 데이터의 재사용성 관점에서, 수집된 장소 데이터를 사용해서 인기있는 장소를 파악하는 기능을 부가적으로 구현했다. 사용자가 특정 키워드를 입력하면, 해당 키워드에 관한 핫플레이스를 언급된 횟수 기준으로 정렬하였다.

[그림 10]은 정렬된 장소 데이터를 그래프를 통해 시각화한 결과이다. 최근 트렌드였던 한국 vs 브라질 전이 열린 서울 경기장에 관한 데이터가 높은 순위를 차지하고 있음을 확인할 수 있다.

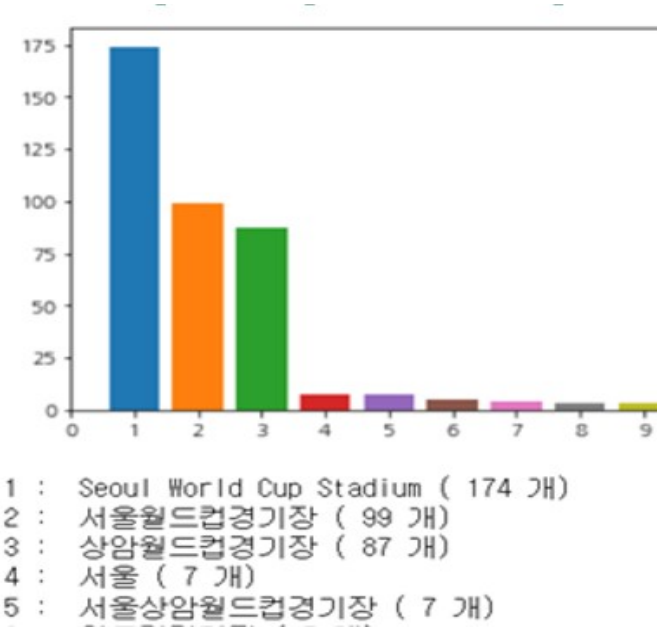

그림 10. 키워드에 관한 핫플레이스 순위

4. 소 감

조원들과 머리를 맞대 데이터를 일상속에서 유용하게 사용할 수 있는 주제에 대해 생각했었다. 공학도는 항상 이론적인 부분을 실생활에 적용해서 의미있는 결과물을 내야 한다고 생각한다. 최근 가장 유망한 자원이라고 평가되는 데이터에 대해 다양한 이론을 적용해보고 그 이론이 분석 주제에 맞게 활용되는 점이 뿌듯했고 더 동기부여가 되는 느낌이었다. 매주 꾸준하게 특정 주제에 관해 고민하고 교수님의 피드백을 반영해가는 과정이 의미있었다. 생각만 했던 주제에 대한 결과물이 나와, 생각을 현실로 만들 수 있는 데이터 분석이라는 활동에 더욱 흥미를 느꼈다. 또한 이론만으로 배웠던 분석 알고리즘이나 데이터 마이닝에 대한 전반적인 과정을 느낄 수 있어서 좋았고 이론에선 배울 수 없던 것들을 실제로 겪어보고 부딪히면서 배울 수 있는 점이 좋았던 것 같다.

프로젝트 중 크롤링된 데이터의 재사용성에 관한 논의가 있어 DB를 이용해 보다 효율적인 데이터 활용을 기대했으나 기간 및 기존 코드와의 호환성 문제로 재고하였다. 또한 데이터 전처리 과정에서 연관성이 떨어지는 과정에서 K-MEANS 알고리즘을 활용할 예정이었으나, 유사도 값이 분석 목적에 더욱 적절하다고 판단하여 K-MEANS 알고리즘을 사용하지 않았다. 위와 같이 데이터 분석 미흡으로 인해 진행하지 못한 과정들이 많이 있었다. 그렇기에 앞으로 기회가 된다면 부족한 부분을 후속 연구로 진행하여 더욱 발전된 결과를 도출하고 싶다.