NLP를 이용한 카페 추천 알고리즘

목다현[1], 변규린[2], 추현승[12*]
[1]성균관대학교 전자전기컴퓨터공학과
[2]성균관대학교 수퍼인텔리전스학과
{dahyun1025, byungyurin21, choo}@skku.edu

Cafe recommendation algorithm using NLP

Dahyun Mok[1], Gyurin Byun[2], Hyunseung Choo[12*]
[1]Dept. of Electrical and Computer Engineering, Sungkyunkwan University
[2]Dept. of Superintelligence Engineering, Sungkyunkwan University

요 약

본 논문은 맞춤형 카페 추천 서비스를 제안한다. 대중적인 포털 사이트의 카페 정보와 사용자 리뷰를 크롤링 하여 지역별, 키워드별 카페 데이터를 수집한다. 사용자가 원하는 지역과 임의의 키워드를 기준으로 데이터셋 내의 키워드와 비교하여 가장 유사한 키워드를 추출한다. spaCy 라이브러리의 사전 학습된 모델 중 similarity method를 사용하여 추출된 키워드를 바탕으로 해당하는 카페를 추천한다. 이를 통해 사용자는 불필요한 정보를 걸러내고 쉽게 원하는 정보를 얻을 수 있다.

1. 서론

우리 사회는 정보화 시대를 맞이하면서 정보의 양이 기하급수적으로 증가하고 있다. 우리는 원하는 정보를 찾기 위해 많은 시간과 노력을 투자해야 한다. 본 논문에서는 키워드를 기반으로 spaCy 라이브러리를 이용하여 구축된 필터 알고리즘을 사용한다. 불필요한 정보를 걸러내고, 사용자가 원하는 양질의 정보만을 제공하고자 한다.

키워드를 정형화하기 위해 많이 이용하는 포털사이트의 정보와 이용자 리뷰를 크롤링하여 전처리 과정을 거친다. 이때, 키워드는 이분법적인 단어가 아닌 묘사하는 키워드로 다양하게 세분화된다. 사용자가 원하는 지역과 조건을 입력하면 해당 조건에 맞는 카페를 추천해준다.

2. 관련 연구

최근 대규모 언어 모델의 등장으로 자연어 처리 분야는 빠르게 발전하고 있다. 검색 엔진, 챗봇, 기계 번역, 텍스트 분류, 음성 인식, 정보 추출 등 다양한 분야에서 활용되고 있으며, 이를 위해 NLTK(Natural Language Toolkit), spaCy, gensim, Konlpy 등의 라이브러리를 이용하여 많은 연구가 진행되고 있다[1]. spaCy 라이브러리는 높은 처리속도와 효율적인 메모리 관리, 다양한 언어와 기능을 제공하는 측면에서 우수하다[2]. 이러한 이유로 본 연구에서는 spaCy를 기반으로 알고리즘을 구현한다.

3. 실험

3.1 데이터셋 생성

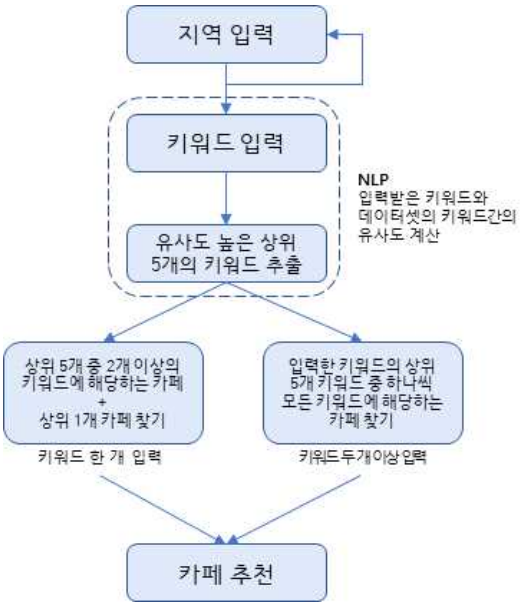
한국 농수산식품유통공사의 식품산업통계정보시스템에 따르면, 2020년 기준 커피전문점의 점포 수가 108,466개로 집계된다[3]. 일반적인 고객이 원하는 조건을 만족하는 카페를 찾으려면 많은 시간과 노력이 필요하다. 따라서 복잡한 과정을 간소화하기 위해 네이버지도와 카카오맵에 카페를 검색 후, 세부 정보를 파이썬으로 크롤링하여, 카페에서 제공하는 편의시설 정보와 분위기, 이용자가 직접 작성한 리뷰를 수집하였다. 추천 정확도를 높일 수 있도록 유사어와 함께 이분법적인 키워드가 아닌 형용사 형식의 키워드로 정형화하였다. 이를 바탕으로 '대화하기 좋은', '조용한', '힙한' 등의 형식으로 카페를 설명할 수 있는 키워드를 정의하였다.

이를 기반으로 유동인구가 많은 서울의 13개 지역 각각의 키워드에 해당하는 카페 데이터셋을 생성하였다. 13개 지역은 '서울 카페 매장수 추이'에서 중상위권에 해당하는 구의 주요 동네[4]와 포털 및 인스타그램에서 서울 핫플레이스를 검색했을 때 나오는 빈도수가 높은 동네로 선정하였다. 또한, 행정구역이 아닌 일반적으로 사용되는 지명으로 정의한

후, 지번 주소에서 동을 추출하여 데이터셋에 포함 여부를 결정하였다.

3.2 spaCy를 이용한 카페 추천 알고리즘

spaCy 자연어 처리 라이브러리 기반 pretrained 된 'ko_core_news_md' 언어 모델을 추천 알고리즘으로 사용한다. 세부 전체적인 알고리즘 동작 순서는 [그림 1]과 같다. 사용자는 먼저 원하는 지역을 입력하고, 찾고자 하는 카페의 조건들을 한 개 또는 여러 개 입력한다. 입력된 지역명과 키워드를 기반으로 유사도를 측정하기 위해, 각 키워드의 임베딩 벡터를 계산한 후 코사인 유사도를 이용한다. 코사인 유사도는 벡터 간의 각도를 이용하여 두 벡터가 얼마나 비슷한 방향을 가지는지를 측정하는 지표이다. 벡터 간의 코사인 유사도 값은 -1과 1 사이의 값으로, 1에 가까울수록 두 벡터가 유사함을 의미한다. 이를 바탕으로 입력값과 유사한 상위 5개 키워드를 출력한다. 해당 알고리즘을 이용하여 효율적이고 정확하게 사용자의 입력에 따라 유사 키워드와 카페 정보를 찾아 제공한다.



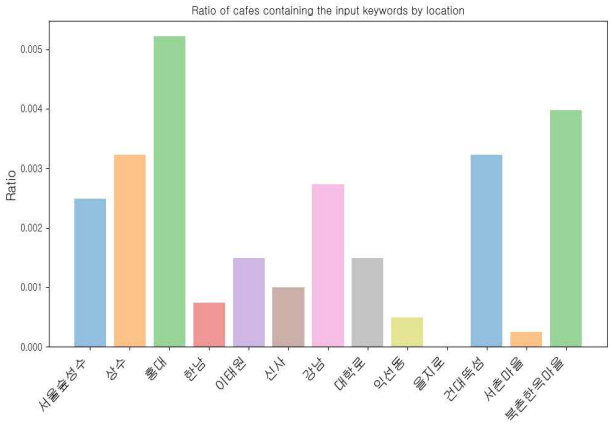
[그림 1] 알고리즘 동작도

4. 실험 결과

매칭 정확도를 높이기 위해 상황을 두 가지로 나누어 실험하였다. 입력 키워드가 하나인 경우는 유사도 상위 1개 키워드의 카페 또는 상위 5개 키워드 중 두 개 이상 포함되는 카페를 출력한다. 원하는 키워드가 2개 이상인 경우, 각 키워드에 대한 상위 5개 키워드를 찾은 후, 각 유사 키워드 중 하나 이상 모두 해당하는 카페를 추천한다. 신사에서 데이트를 하기로 한 고객이 남자친구와 함께 디저트가 맛있는 색다른 감성 카페에 가고 싶어한다. '남자친구', '새로운', '감성카페', '케이크'라는 4개의 키워드를 입력했을 때 출력되는 상위 1개 키워드와 그 유사도는 [표 1]과 같다. 키워드를 완전히 동일하게 입력했을 때의 유사도는 1이 나오고, 찾은 유사 키워드에 대해서는 0.5 ~ 0.7 정도의 유사도를 보인다. [그림 3]은 앞서 예시를 든 키워드를 기반으로 13개 지역에 대해 전체 카페 수 대비 모든 키워드에 해당하는 카페 수의 비율을 나타낸다. 이를 통해 각 지역에서 원하는 조건에 맞는 카페 후보군을 좁혀줌으로서 번거로움을 줄이고 불필요한 시간 소모를 없앨 수 있다.

입력 키워드	찾은 키워드	유사도
남자친구	데이트	0.52
새로운	이색적인	0.65
감성카페	감성카페	1.00
케이크	디저트카페	0.56

[표 1] 입력된 키워드와 검색결과 키워드 간 유사도



[그림 2] 지역 별 검색 카페의 비율

5. 결론 및 향후 연구 계획

본 연구는 키워드 기반 서울의 카페 추천 알고리즘을 제안한다. 맞춤형 추천이라는 큐레이션 역할을 하여 불필요한 추천을 줄이고, 각 사용자에게 맞는 카페를 소개함으로써 개인의 니즈가 충족되도록 도와주고 선택의 폭을 줄여준다.

본 논문에서는 키워드 기반으로 서울 지역의 카페를 추천했지만, 추후에는 지역을 확장하고 사용자들이 많이 검색한 키워드를 바탕으로 추천하도록 한다. 또한, 다른 라이브러리를 함께 이용하여 정확도를 높일 계획이다.

Acknowledgement

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 ICT명품인재양성(IITP-2023-2020-0-01821), 4단계 BK21 사업과, 글로벌핵심인재양성지원사업(RS-2022-00155415)(기여율:50%), 인공지능 혁신 허브(No.2021-0-02068), 지원을 받아 수행된 연구 결과임.

참고문헌

[1] Parvathi et al., "Identifying relevant text from text document using deep learning." 2018 International Conference on Circuits and Systems in Digital Enterprise Technology (ICCSDET). IEEE, 2018.

[2] Yu et al ., "spaCy 를 이용한 TV 드라마 대본의 주인공 및 배경 분석." 한국정보과학회 학술발표논문집 (2020): 1379-1381.

[3]https://www.atfis.or.kr/home/pdf/view.do?path=/board/202302/0dfd28e7-07af-4116-a9a6-c181e87d620e.pdf

[4]https://post.naver.com/viewer/postView.nhn?volumeNo=26158801&memberNo=8963019.