Data Mining Project

Team #10

1. Title

뉴스 데이터를 통한 인기 키워드 추출과 뉴스 추천 시스템

2. Abstract

1) Problem/Motivation/Objective

포털사이트 실시간 검색어 기능이 삭제되어 트렌드를 파악하는 것이 어려워졌다. 또한, 최근 뉴스 제공 사이트의 무분별한 추천, 광고 등으로 인해 사용자가 제대로 된 컨텐츠 기반, 히스토리 기반 뉴스를 찾아보기 어렵다.

2) A statement of the problem and objectives

위의 문제를 해결하고자 최근 뉴스 데이터를 분석해 실시간 인기 키워드를 제공한다. 또한, 컨텐츠와 유저 히스토리 기반 뉴스 추천이 모두 가능한 시스템을 구현하여 사용자의 불편함을 덜어주고자 한다.

3) Methods or Approach you (will) use

실시간 인기 키워드는 최근 기사들에 TF-IDF 알고리즘을 적용해 기사 제목에 많이 나온 단어를 분석해 도출한다. 사전에 수집한 대량의 기사를 K-means Clustering으로 군집화하고 TF-IDF 알고리즘을 적용해 content-based filtering을 구현한다. 기존 뉴스 데이터를 바탕으로 특정한 뉴스와 비슷한 뉴스를 도출한다. 유저 기반 추천은 여러 유저의 Rating 정보와 대량의 뉴스 데이터에 SVD를 이용한 collaborative filtering을 적용하여 구현한다. 특정한 유저에 대해 다른 유저들의 Rating에 기반한 추천 뉴스를 도출한다.

4) Summary results

위와 같은 구현을 통해 실시간 인기 키워드 및 컨텐츠 기반, 유저 기반 뉴스 추천 서비스를 제공한다.

5) Conclusions and comments

결과적으로, 사용자별 맞춤형 뉴스 기사 추천 및 실시간 키워드 대체가 가능할 것으로 예상된다.

3. Introduction

1) Background & Related work

Content-Based Filtering

사용자가 특정 아이템을 선호하는 경우, 그 아이템과 비슷한 컨텐츠를 가진 다른 아이템을 추천해주는 방식이다. 본 프로젝트에서는 Clustering, TF-IDF와 cosine similarity를 활용해 아이템 분석을 수행하고 뉴스 제목과 내용으로 유사도를 판단해 사용자가 선택한 기사에 대해 가장 유사도가 높은 뉴스들을 추천한다.

Collaborative Filtering

많은 사용자들로부터 얻은 선호정보에 따라 특정 사용자의 선호정보를 자동적으로 예측하게 해주는 방식이다. 본 프로젝트에서는 뉴스 기사에 대해 view, scrap에 대해 선호정보를 기록하고 SVD 알고리즘을 통해 user-based collaborative filtering으로 뉴스를 추천한다.

KoNLPY

KoNLPY란 자연어처리(NLP)에서 형태소 단위 토큰화 시 한국어 데이터의 전처리를 위한 파이썬 패키지이다. 본 프로젝트에서는 KoNLPY의 Okt(구 Twitter) 클래스를 사용해 기사의 제목과 내용에 대한 형태소 단위 토큰화를 수행해 데이터 전처리 과정을 거친다.

Stopword

불용어란 언어를 분석할 때 자주 등장하지만 분석에 큰 의미를 갖지 않는 단어나 조사 등의 단어를 말한다. 영어와 달리 한국어는 불용어 처리에 대한 라이브러리를 지원하지 않으므로 원하는 단어를 추가해 리스트로 만들어 제거하여 사용해야 한다.

4. Team Information

사이버보안학과 201720606 이상일

- User-based 추천 알고리즘 구현, 뉴스 기사 수집, 발표

사이버보안학과 201720550 이주현

- Content-based 추천 알고리즘 구현, 실시간 뉴스 수집, 프로토타입 제작

소프트웨어학과 201823776 김한성

- 인기 Keyword 추출, K-means Clustering 및 Cluster 별 핵심어 추출 알고리즘 구현

5. Methods

Description of Data

1) 추천을 위한 뉴스 데이터

BeautifulSoup 라이브러리를 통해 네이버 뉴스의 IT과학, 정치, 세계, 생활문화, 사회, 경제 카테고리에서 추천을 위한 약 5만개의 뉴스를 수집한다. 수집한 뉴스는 제목, 내용, 날짜, URL을 포함하며 다음 그림과 같다. 이 뉴스 데이터를 통해 K-means clustering 및 content-based, user-based 추천을 한다.

6	2021.11.01. 오후	IT과학	'가상 공장에서 물품 제조'...	KBS 대전 앵커 쌍방향 소통이 ㄱ	https://news.naver.com/main/read.naver?mode=LS
7	2021.11.01. 오후	IT과학	"점심 장사 망쳤는데..." KT	KT '먹통사태' 보상안 발표 경향	https://news.naver.com/main/read.naver?mode=LS
8	2021.11.01. 오후	IT과학	영상전장사업 가속 페달 밟	벤츠 ADAS 카메라 이어 르노 인	https://news.naver.com/main/read.naver?mode=LS
9	2021.11.01. 오후	IT과학	이큐웨어 IT항균 브랜드 'EC	스포츠경향 IT 항균 전문 기업 이	https://news.naver.com/main/read.naver?mode=LS
10	2021.11.01. 오후	IT과학	오락가락 택시요금 미리 정	우버·티맵 합친 '우티' 서비스 사	https://news.naver.com/main/read.naver?mode=LS
11	2021.11.01. 오후	IT과학	코인뉴스 "비트코인 상승은	피터 틸 페이팔 공동창업자 "코인	https://news.naver.com/main/read.naver?mode=LS
12	2021.11.01. 오후	IT과학	현장속으로 로봇과 함께 시	KBS 창원 앵커 그동안 참아 왔던	https://news.naver.com/main/read.naver?mode=LS
13	2021.11.01. 오후	IT과학	'가상 공장에서 물품 제조'...	KBS 대전 앵커 쌍방향 소통이 ㄱ	https://news.naver.com/main/read.naver?mode=LS

2) 인기 Keyword를 위한 뉴스 데이터

Selenium 라이브러리를 통해 특정 날짜에 대한 인기 Keyword를 도출한다. 수집하는 데이터는 뉴스의 제목이며 다음 그림과 같다. sklearn TF-IDF Vectorizer를 사용해 단어의 중요도를 계산해 상위 20개의 인기 Keyword를 도출한다.

114	112	윤석열, '적폐청산 칼잡이'서 '조국 수사'로 야당 대선후보 되다
115	113	윤석열 "후회 되는 실언이 한두개겠나" [일문일답]
116	114	이준석 "尹 축하...홍카콜라·유치타·대장동강사 꿈도 실현"
117	115	'국힘 경선'에 이재명 테마주도 들썩 [3분 국내주식]
118	116	서울고검, '조국 수사팀' 감찰에... 한동훈 "치졸한 보복"
119	117	살인에 시신 100여구 능욕까지...英병원 34년 미제 풀려
120	118	'예산안 삭감' 두고 갈등 깊어지는 서울시·시의회

3) 유저 히스토리 데이터

유저 히스토리 데이터는 userId, articleId, scrap, view, weight 컬럼으로 이루어져 있다. 뉴스는 영화 등과 다르게 Rating 정보가 없으므로 이를 대체하기 위해 조회 및 스크랩에 대해 임의로 가중치를 부여해 Rating 개념으로 사용했다. 유저 히스토리 데이터는 수집이 어렵기 때문에 임의로 유저 및 해당 유저가 본 뉴스와 가중치를 설정했다. 먼저, 유저들의 Rating 정보를 가져온다. 기사 조회는 1회당 1씩 weight를 증가시키고, 스크랩은 weight를 2만큼 증가시킨다. 유저들의 Rating 정보는 다음 그림과 같은 형식으로 저장되어 있다. articleId는 기사의 번호이고 weight는 조회와 스크랩 가중치를 더해 Rating의 개념으로 사용할 값이다.

userId	articleId	scrap	view	weight
1	232	2	1	3
1	124	0	1	1
1	92	2	3	5
1	405	2	1	3
1	851	0	1	1
2	551	0	1	1
2	398	0	1	1
2	556	0	1	1
2	252	2	1	3
2	306	2	2	4
3	169	0	1	1
3	511	0	2	2
3	209	0	1	1
3	83	0	1	1
3	752	2	1	3
4	87	2	1	3

4) K-means clustering 결과 데이터

K-means clustering을 진행하기 위해 수집한 뉴스 데이터에서 제목(title)과 내용(content)을 추출한다. Clustering 작업 전에 기사의 내용을 전처리해 content_cleaned에 저장하고 content_cleaned를 기준으로 뉴스 데이터의 k-means clustering을 진행하였다. clustering을 마친 후 각각의 뉴스들은 클러스터의 개수인 100에 맞춰 1부터 100까지의 labels와 함께 저장된다. 같은 labels에 속하는 뉴스들은 상대적으로 높은 유사도를 보이며 같은 cluster에 포함된다. clustering 결과 데이터는 다음 그림과 같다.

	title	content	content_cleand	labels
0	中 니오 전	10월 3667대 불과... 주문량은 역디	대 불과 주문량은 역대 최고 기록 경신 중	19
1	페북도 링	점유율 1위 애플워치...메타와의 디	점유율 위 애플워치 메타와의 대결구도	16
2	넥슨 '던전	넥슨은 1일 자회사 네오플이 개발	넥슨은 자회사 네오플이 개발한 D 액션	26
3	KT "개인·7	앵커 통신마비 사태로 전국적 혼란	앵커 통신마비 사태로 전국적 혼란을 초래	11
4	'가상 공장	KBS 대전 앵커 쌍방향 소통이 가능	KBS 대전 앵커 쌍방향 소통이 가능한 가상	0
5	"점심 장사	KT '먹통사태' 보상안 발표 경향신	KT 먹통사태 보상안 발표 경향신문 YHA	11
6	영상전장시	벤츠 ADAS 카메라 이어 르노 인포	벤츠 ADAS 카메라 이어 르노 인포테인먼	19

Methods Contents

Architecture and Environment

Windows 10, Google Colab default setting with Python 3

Methods Used

1) K-means Clustering

뉴스데이터의 clustering 작업 전, 여러 문장을 토큰화하고 어근을 추출하기 위해 sklearn에서 제공하는 TfidfVectorizer를 이용하여 TF-IDF 벡터화를 수행한다. TF-IDF 벡터들은 normalize 함수를 통해 L2 정규화가 진행되고 미리 정한 cluster 개수로 k-means clustering이 진행된다. 이후 sklearn에서 k-means clustering과 함께 제공하는 labels_를 이용하여 클러스터에 관한 정보를 저장한다.

또한 clustering 모델을 바탕으로 cluster 내부의 핵심 단어를 추출했다. get_cluster_details 함수는 cluster에 사용된 모델과 데이터를 통해 center를 기준으로 정렬한 상위 n개의 feature 단어를 추출한다. get_cluster_details 함수를 바탕으로 상위 10개의 feature를 저장하는 함수인 save_cluster_details를 통해 csv 파일로 만들고, 이 데이터는 뉴스 추천을 위해 사용된다.

2) TF-IDF Based Similarity

1)번에서 cluster 별 기사와 cluster 별 feature 들에 대한 결과를 토대로 유저가 선택한 기사에 대해 get_similar_clusters 함수를 통해 가장 유사한 cluster 들을 찾고 get_recommend_news 함수를 통해 그 cluster들 안의 뉴스 기사들에 대해서 유사도가 높은 상위 5개의 기사를 추천한다. get_similar_clusters 함수는 cluster 별 feature들에 대해 유저가 선택한 기사의 내용인 target과 비교 후 가장 유사한 3개의 cluster를 반환한다. get_recommend_news 함수는 3개의 cluster들에 포함되어 있는 기사들을 rawdata에 저장한다. 이 후 rawdata를 TfidfVectorizer를 사용해 Counter Vector 화 하고 target에 대해 토큰화를 수행한다. 유사도 판단은 target을 KoNLPY Okt(구 Twitter)로 토큰화 후 DTM(Document-Term Matrix)를 생성, rawdata의 Counter Vector와 비교해 Score가 가장 높은, 즉 유사도가 가장 높은 5개의 기사를 추천한다.

3) SVD를 이용한 collaborative filtering

유저의 기존 Rating을 기반으로 뉴스를 추천해주는 시스템을 구현한다. 위에서 언급한 유저 히스토리 데이터 테이블을 SVD를 수행하기 위해 numpy matrix로 변경 후, scipy 라이브러리를 사용해 SVD를 적용한다. SVD를 실행하기 전 평균을 빼고, SVD 실행 후 다시 평균을 더한다. 이는 가중치를 대체적으로 높게 주는 유저와 가중치를 대체적으로 낮게 주는 유저에 대한 정규화를 위함이다.

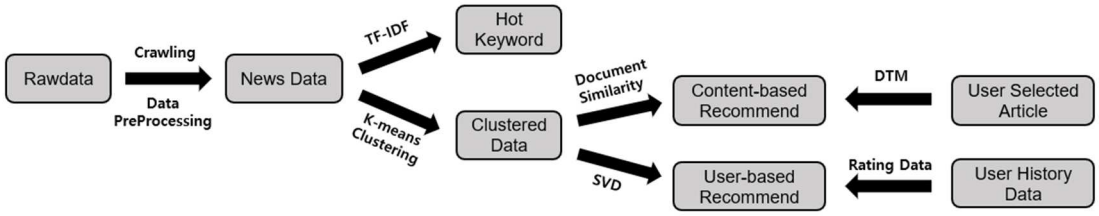
이 과정을 통해 만들어진 prediction과 유저의 기존 기록인 rating_data를 통해 추천뉴스를 도출하는 함수를 구현했다. 추천뉴스 도출 함수는 rating_data에서 특정 유저에 대한 데이터를 뽑아 뉴스데이터와 합친다. 이후 뉴스데이터에서 유저가 본 뉴스를 제외한 데이터를 추출하고, 유저의 Rating이 높은 순으로 정렬된 데이터와 위에서 만든 prediction을 합친다.

추천뉴스 도출함수를 호출하면 뉴스데이터(news_data)와 유저 Rating(rating_data)를 기반으로 특정 유저에게 정해진 수의 뉴스를 추천해준다. 추천뉴스 도출함수 호출 후, 해당 유저가 이미 Rating을 완료한 뉴스 및 추천 뉴스를 모두 출력할 수 있다.

DM Pipeline

뉴스 기사의 수집에서부터 클러스터링, 뉴스 추천까지의 Data Mining Pipeline은 다음 그림과 같다.

뉴스 기사의 크롤링 및 전처리 과정을 거쳐 추천을 위한 기사를 수집한다. 수집된 뉴스 데이터는 인기 키워드의 추출과 클러스터링에 사용된다. K-means clustering의 결과 데이터는 컨텐츠 기반 및 유저 기반 추천에 사용된다. 컨텐츠 기반 추천은 클러스터링 결과와 유저가 선택한 기사에 대해 TF-IDF 기반 문서 유사도 측정을 통해 수행하고 유저 기반 추천은 클러스터링 결과와 유저 히스토리 데이터에 대해 SVD 알고리즘을 적용해 수행한다.



Evaluation Criterion

K-means Clustering는 clustering 후 center에서 군집의 데이터 간의 거리를 합산한 값인 inertia value를 통한 정량적 평가가 가능하다. Hot Keyword는 결과로 나온 20개의 키워드 중에 불용어의 유무나 트렌드를 잘 반영하고 있는지에 대한 정성적인 평가가 가능하다. Content-based와 User-based 추천 또한 유저가 선택한 기사 또는 유저 히스토리 정보를 바탕으로 얼마나 유사한 기사를 추천했는지에 대한 정성적인 평가가 가능하다. 따라서 정성적인 평가만 가능한 경우 여러 입력 값에 대한 결과로 구현한 추천 시스템의 정확도를 평가해야 한다.

6. Results

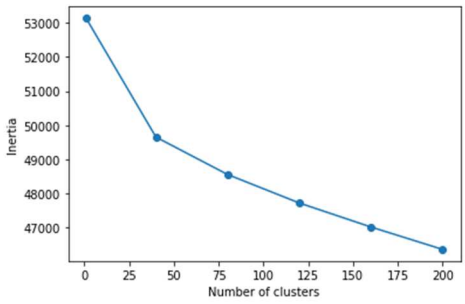
1) Popular keywords 결과

다음 그림은 TF-IDF를 통해 추출한 2021년의 12월 7일 인기 keyword에 대한 결과이다. 20개의 단어들은 트렌드를 잘 반영하고 있지만, 단어의 어근으로 추출되었기 때문에 '최다'와 같이 해당 단어로만 의미 파악이 어려운 단어들이 존재한다. 따라서 추가로 불용어에 대한 전처리 과정이나 토큰화 과정에서의 개선이 필요한 것으로 판단된다.

	단어	빈도
6978	오미크론	25.294947
7430	윤석열	23.951808
7637	이재명	23.351120
10651	확진	17.087247
9536	코로나	15.977355
5404	서울	14.870845
9261	최다	14.823323
3117	대통령	14.318194
6788	역대	13.836339
4464	백신	13.173180
92	12월	13.054496
7518	의혹	12.834472
7448	윤우진	12.736699
9537	코로나19	12.197964
9275	최대	11.665542
7020	오후	11.531843
8313	정부	11.402238
4654	보이콧	11.092690
7060	올림픽	10.895499
5466	선대위	10.883135

2) K-means clustering 결과

k-means clustering의 적합한 k를 찾기 위해, clustering이 된 후에 각 center에서 군집의 데이터 간의 거리를 합산한 값인 inertia value를 이용했다. k의 값이 높을수록 inertia는 작아져 높은 응집도를 보이지만 클러스터 내에 인위적인 경계가 생기거나 성능이 떨어진다는 단점이 있다. 다음 그림은 k가 1부터 200일 때까지의 inertia를 부분적으로 계산하고 그래프로 나타낸 것이다. 이 그래프를 통해 적당한 knee point로 추정되는 100을 k값으로 설정했다.



14 ['중국', '중국의', '코로나', '상하이', '베이징', '중국에서', '검사를', '중국이', '중국은', '지난', '미국', '지난달', '것으로', '따르면', '제조업', '야후는', '홍콩', '부동산', '최근', '인터넷']
15 ['대통령은', '문재인', 'cop', '대통령이', '메탄', '영국', '한반도', '정상회의', '감축', '한국은', '청가리', '온실가스', '당사국총회', '대통령과', '글래스고', '프란치스코', '바이든', '청와대', '협력을', '산림']
16 ['대구', '수성못', '김현태', '홍준표', '드리는', '특별', '의원이', '경북', '기자회견을', '공정식', '이상화', '국민의힘', '시비', '대선', '지역구인', '국민께', '대권주자인', '오후', '상화동산에서', '승리를']
17 ['경찰은', '혐의를', '혐의로', '불법', '경찰', '것으로', '경찰에', '위반', '받고', 'a씨는', '경찰이', '공수처는', 'a씨를', '수사', '손준성', '검찰에', '명을', '지난달', '검찰은', '남성이']
18 ['kt', '인터넷', '장애', 'kt는', '서창석', '광화문', '보상', '네트워크', '무선', '서울', '오전', '서비스', '전무가', '유무선', '재발방지대책', '보상안', '네트워크혁신tf', 'kt본사에서', '관련', '설명회에서']
19 ['로마', '이탈리아', '누볼라', '최재구', '현지시간', '문재인', '한글학교를', '대통령이', '컨벤션센터에서', '김정숙', '여사가', '양자회담장에서', '연합뉴스', '기자', '정상회의에', '방문', '컨벤션센터', '메르켈', '앙겔라', '총리와']
20 ['기시다', '일본', '자민당', '중의원', '총리가', '후미오', '총리는', '자민당이', '도쿄', '석을', '총선', '의석', '선거', '과반', '단독', '아베', '총리', '자민당은', '의석을', '선거에서']

위 그림은 k-means clustering 후 cluster 별 핵심어들을 추출한 내용의 일부이다. 비슷한 분야의 단어들끼리 clustering 된 것을 확인할 수 있다.

3) Content-Based Filtering 결과

유저가 선택한 게시물인 target에 대한 content-based 추천 결과는 다음 그림과 같다. target의 핵심 Keyword는 “라이엇게임즈”, “리그오브레전드”, “아케인”, “넷플릭스” 등이다. 이에 대한 결과로 나온 Keyword는 “넷플릭스”, “라이엇게임즈”, “리그오브레전드”, “오징어게임”, “OTT” 등으로 유사한 것을 확인

할 수 있다.

```
target =
(
    "Riot 아케인 이벤트 대표 이미지 라이엇게임즈 제공 지난 달간 리그오브레전드 LoL 유니버스 게임의 월간 사용자 MAU 억 기록했다 라이엇게임즈는 내용을 밝히면서 역대 최고 기록이라고 설명했다
    "지난 년간 리그오브레전드 유니버스 게임을 즐긴 세계 플레이어는 억 명에 달했다 니콜로 러렌트 라이엇게임즈 CEO는 때보다 게임을 즐기는 이들이 많아진 애니메이션 아케인을 통해 세계 각지
    "플레이어와 손잡고 게임을 글로벌 엔터테인먼트 산업의 중심으로 자리 잡게 하고자 한다 고 했다 아케인 은 리그오브레전드 지식재산권 기반 첫 번째 장편 애니메이션이다 오는 롤드컵 이후
    "넷플릭스에서 세계 동시 공개된다 첫 화는 트위치에서도 동시 중계로 감상할 라이엇게임즈는 아케인 출시를 기념해 발로란트 와일드 리프트 전략적 팀 전투 자사 모든 게임에 적용하는 Riot
    "아케인 이벤트 한다 게임별 이벤트는 애니메이션 주제와 내용을 테마로 신규 콘텐츠 등에 초점을 맞췄다 라이엇게임즈는 롤드컵 개막식 날인 오는 오전 시 분 아케인 애니메이션을 처음
    "글로벌 프리미어 행사 미국 로스앤젤레스에 위치한 라이엇게임즈 본사에서 한다 사라 슈츠 라이엇게임즈 익스피리언스 부문 책임자는 우리는 리그오브레전드 지식재산권의 기원인 게임을 아케인과
    "관련된 경험의 출발점으로 삼았다 고 했다"
)
```

```
2.2627631872109917 : 경향신문 지난 7일 공개된 넷플릭스 애니메이션 시리즈 아케인 . 이 작품은 인기 게임 '리그 오브 레전드' 의 세계관을 기반으로 만들어졌다. 넷플릭스 제공 한국의 넷플릭스 오리지널 드
2.198997643876838 : 라이엇게임즈 인기 게임 LoL 기반 애니메니션 서울경제 한국의 넷플릭스 오리지널 드라마 시리즈 '오징어게임' 이 46일만에 TV쇼 부문 1위 왕좌 자리를 내줬다. 새롭게 왕좌 자리에 오른
2.0194884196770126 : 크래프톤 영화 그라운드 제로 이어 16일 배그 세계관 네이버웹툰 연재 컴투스 서머너즈 워 日 애니와 협업 LoL 넷플릭스 통해 애니 아케인 공개 네오위즈 브라운더스트 웹소설 연재 크래프
1.864491784177706 : 성공한 게임 원작 드라마 · 웹툰 · 소설 · 만화책으로 무한 확장 46일간 전 세계 인터넷 동영상 서비스 OTT 시장을 휩쓸었던 넷플릭스 오리지널 드라마 시리즈 '오징어 게임' 이 8일 현지 시
1.830967966167316 : 넷플릭스 드라마 '오징어게임' 한 장면. 넷플릭스 제공 넷플릭스 최대 흥행작 '오징어 게임' 이 46일 만에 넷플릭스 TV쇼 부문에서 1위 자리를 내줬다. 8일 현지시간 미국 온라인 동영상
```

4) SVD를 이용한 collaborative filtering의 결과

SVD를 이용한 collaborative filtering을 적용해 유저 10에 대한 추천 기사 5개를 도출했다. 다음 그림을 참고하면 우선 유저 10이 보거나 스크랩한 뉴스들의 cluster를 확인하면 2, 3, 4, 20, 28 등과 같은 cluster에 속한 뉴스들이다.

```
already_rated, predictions = recommend_news(df_svd_preds, 10, news_data, rating_data, 5)
already_rated.head(10)
```

	userId	articleId	scrap	view	weight	title	content	content_cleand	labels
0	10	1	2	3	5	폐북도 링 위 오른다...애플에 스마트워치 도전장	점유율 1위 애플워치...메타와의 대결구도 관심 삼성전자 갤럭시워치도 메타와 유력 경쟁...	점유율 위 애플워치 메타와의 대결구도 관심 삼성전자 갤럭시워치도 메타와 유력 경쟁...	2
1	10	2	2	2	4	넥슨 '던전앤파이터 모바일' 내년 1분기 국내 출시	넥슨은 1일 자회사 네오플이 개발한 2D 액션 역할수행게임 RPG '던전앤파이터 모...	넥슨은 일 자회사 네오플이 개발한 D 액션 역할수행게임 RPG 던전앤파이터 모...	20
2	10	3	2	1	3	KT "개인·기업 천 원 소상공인 7천 원 감면"...고객 분통	앵커 통신마비 사태로 전국적 혼란을 초래했던 KT가 u200b보상 방안을 내놨습니다...	앵커 통신마비 사태로 전국적 혼란을 초래했던 KT가 u b보상 방안을 내놨습니다...	28
6	10	7	2	1	3	이큐웨어 IT항균 브랜드 'EQ MEDIC' 론칭	스포츠경향 IT 항균 전문 기업 이큐웨어가 항균 원료를 원자재에 배합하는 기술을 바...	스포츠경향 IT 항균 전문 기업 이큐웨어가 항균 원료를 원자재에 배합하는 기술을 바...	3
11	10	12	2	1	3	"중국판 넷플릭스 투자했는데..." '지리산' 혹평에 곤혹	드라마 지리산 중 일부. 드라마 지리산 방송 캡처 헤럴드경제 박지영 기자 "중국판 ...	드라마 지리산 중 일부 드라마 지리산 방송 캡처 헤럴드경제 박지영 기자 중국판 ...	20
12	10	12	2	1	3	"중국판 넷플릭스 투자했는데..." '지리산' 혹평에 곤혹	드라마 지리산 중 일부. 드라마 지리산 방송 캡처 헤럴드경제 박지영 기자 "중국판 ...	드라마 지리산 중 일부 드라마 지리산 방송 캡처 헤럴드경제 박지영 기자 중국판 ...	20
13	10	14	2	1	3	KT 보상시간 10배라지만... 소상인 7000원·개인 1000원 불만	3500만 회선·최대 400억대 규모 내달 청구요금서 자동감면 방침 라우팅 오류확산...	만 회선 최대 억대 규모 내달 청구요금서 자동감면 방침 라우팅 오류확산...	28
14	10	15	2	1	3	네이버 새로운 검색 경험 제공 '에어서치' 선보여	"검색 전반에 AI 기술 녹여" 네이버 서치 Search CIC 김상범 책임리더. ...	검색 전반에 AI 기술 녹여 네이버 서치 Search CIC 김상범 책임리더 ...	4
7	10	8	0	2	2	오락가락 택시요금 미리 정해두고 타세요	우버·티맵 합친 '우티' 서비스 사전 확정요금제에 합승도 허용 1일 온라인으로 열린...	우버 티맵 합친 우티 서비스 사전 확정요금제에 합승도 허용 일 온라인으로 열린...	4
8	10	9	0	2	2	코인뉴스 "비트코인 상승은 심각한 인플레 때문"...지금이 사야할 때	피터 틸 페이팔 공동창업자 "코인 많이 매수 못한 것 후회한다" 비트코인 7334만...	피터 틸 페이팔 공동창업자 코인 많이 매수 못한 것 후회한다 비트코인 만...	2

다른 유저들의 데이터를 바탕으로 유저 10에게 추천된 5개의 뉴스들의 cluster는 각각 2, 4, 28로 위그림에서 확인한 유저10의 히스토리 뉴스 cluster들에 포함된다. 추천이 잘 되는지 확인하기 위해 유저 10의 히스토리는 모두 동일한 카테고리의 뉴스로 설정했고, 해당 카테고리에 속하고 같은 cluster에 속하는 뉴스들이 잘 추천된 것을 확인할 수 있다.

```
[39] predictions.head(5)
```

	articleId	title	content	content_cleand	labels
772	786	KT 89분 망 장애 10배 수준 15시간 보상...소상공인엔 10일	추가 단말 알뜰폰 재판매 인터넷도 해당 KT가 지난 25일 발생한 전국적 유무선 통...	추가 단말 알뜰폰 재판매 인터넷도 해당 KT가 지난 일 발생한 전국적 유무선 통...	28
512	526	KT 장애 요금감면으로 일괄보상...중복회선 포함 400억원	협력업체 구상권 청구도 검토 재발방지대책 발표하는 KT 서창석 네트워크혁신TF 전무...	협력업체 구상권 청구도 검토 재발방지대책 발표하는 KT 서창석 네트워크혁신TF 전무...	28
110	124	노웅래 의원 3일 가상자산 과세 현안 점검 토론회 개최	가상자산 과세 유예 및 가상자산 투자자 보호책 마련 논의 선 先 보호 후 後 과세해...	가상자산 과세 유예 및 가상자산 투자자 보호책 마련 논의 선 보호 후 과세해...	2
837	851	현대오토에버 ESG 평가 통합 A등급 획득	디지털데일리 이상일기자 현대오토에버 대표이사 서정식 는 한국기업지배구조원 KCGS ...	디지털데일리 이상일기자 현대오토에버 대표이사 서정식 는 한국기업지배구조원 KCGS ...	4
306	320	이슈분석 SK텔레콤 AI 기반 디지털 인프라 회사로 재탄생... SKB와 원팀 강조...	유영상 대표 고객 기술 서비스 3대 키워드 실질적 원팀 기반 운영...시너지 극대화...	유영상 대표 고객 기술 서비스 대 키워드 실질적 원팀 기반 운영 시너지 극대화...	4

7. Conclusion

“BeautifulSoup”과 “Selenium” 라이브러리를 사용하여 네이버 뉴스페이지에서 뉴스 기사들을 추출했다. 추출된 뉴스 데이터는 목적에 맞게 컬럼을 나누어 인기 키워드 검출과 추천 시스템에 사용됐다.

인기 키워드를 추출하기 위해서 사용된 알고리즘은 TF-IDF로 불용어를 제외하며 문서 내 핵심 단어들의 빈도를 측정했다. 빈도가 높은 순으로 20개의 단어를 검출하여 확인해본 결과, 단어들은 트렌드를 잘 반영하지만 어절 단위로 끊겨 단어 자체만으로 해석이 어려운 부분이 있었다. 띄어쓰기의 내용도 포함한 단어의 의미를 학습하여 토큰화하는 모델을 사용한다면 이러한 문제를 해결할 수 있을 것으로 보인다.

Content-based와 User-based 추천 시스템의 구현을 위해 유사한 내용의 뉴스들을 군집으로 묶는 군집화 과정을 진행했다. 이 때 사용한 방법은 k-means clustering으로 적절한 군집 수 k를 알아내기 위해 inertia 그래프를 확인했다. 뉴스 데이터의 수가 많아 그래프는 다소 불완전한 knee point를 보였지만 그래프와 추천 시스템의 결과를 고려하여 k 값을 설정할 수 있었다.

Content-based 추천 시스템 구현 시 유저가 선택한 기사에 대해 군집화 된 뉴스 기사들 중 cluster 별 feature와 비교해 가장 유사한 cluster를 찾은 후 그 cluster들 내의 기사들에 대해서만 유사도 측정을 진행했다. 여러 번의 다른 기사에 대해 높은 유사도의 결과를 보여주었다. 하지만 Clustering의 결과에 의존적일 수 있다는 점, 그리고 cluster의 feature에 포함된 단어들이 포함되지 않는다면 가장 유사한 cluster를 찾는 것이 정확하지 않을 수 있다는 점이 한계점으로 작용할 수 있다.

User-based 추천 시스템 구현 시 유저 히스토리 정보와 뉴스에 대한 Rating 정보가 필요했는데, 뉴스에는 Rating 정보가 없어 이를 가중치 정보로 대체했다. 또한 유저 히스토리 정보를 수집할 수 없어 임의로 유저 히스토리 데이터를 생성해서 구현한 시스템을 평가했다. 이와 같은 한계점이 있지만, 설계한 시스템은 적절한 추천을 하고 있다고 평가했다. 특정 카테고리의 뉴스만 봤다고 가정한 임의의 유저에 대해 user-based 추천을 수행한 결과, 유저 히스토리에 있던 뉴스들과 카테고리 및 cluster가 동일한 뉴스가 추천되었다.

8. References

[1] 공부하는시몬쯔, "추천시스템 Collaborative Filtering(CF) python 기반 [3]," 5 2020. [Online]. Available: https://simonezz.tistory.com/25?category=852348. [Accessed 05 12 2021].

[2] 조영훈, "[NLP] 문서 군집화(Clustering)와 문서간 유사도(similarity) 측정하기," 앎의 공간, 08 2020. [Online]. Available: https://techblog-history-younghunjo1.tistory.com/114. [Accessed 27 11 2021].

[3] rachitgupta1997, "User-Based Collaborative Filtering," 16 7 2020. [Online]. Available: https://www.geeksforgeeks.org/user-based-collaborative-filtering/. [Accessed 25 11 2021].

[4] yoonicon, "Scikit-learn의 CountVectorizer를 이용한 "관련 게시물 찾기"," 6 7 2017. [Online]. Available: theyoonicon.com/scikit-learn을-이용한-군집화관련-게시물-찾기/. [Accessed 20 11 2021].

[5] Wikipedia, "Document-term matrix," 4 9 2021. [Online]. Available: https://en.wikipedia.org/wiki/Document-term_matrix. [Accessed 05 12 2021].

[6] lumyjuwon, "KoreaNewsCrawler," 2 1 2021. [Online]. Available: https://github.com/lumyjuwon/KoreaNewsCrawler. [Accessed 30 11 2021].