**K-평균 군집을 이용한 일일 뉴스 헤드라인 요약**

대건고등학교 20928 정준영

목차

[I. 탐구 동기 2](#_Toc184591489)

[II. 비지도학습과 K-평균 3](#_Toc184591490)

[III. 일일 뉴스 요약 모델 개발 5](#_Toc184591491)

[1. 데이터 수집 5](#_Toc184591492)

[2. 코드 5](#_Toc184591493)

[3. 모델 5](#_Toc184591494)

[4. 결과 6](#_Toc184591495)

[IV. 결론 8](#_Toc184591496)

[I. 출처 9](#_Toc184591497)

# 탐구 동기

머신러닝의 학습법인 지도학습, 비지도 학습, 강화학습에 대해 공부하며 여러 프로젝트를 진행하던 중 비지도 학습에 대해 심층적으로 다뤄보고자 ‘[세바스찬 라시카, 바히드 미자리리. 박해선 옮김. 길벗 출판사. 『머신 러닝 교과서 with 파이썬, 사이킷런, 텐서플로』.](https://www.yes24.com/Product/Goods/98809693)’를 읽어보며 비지도학습 알고리즘 중 K-평균 알고리즘에 대해 알게되었고 이를 실습하는 동시에 실생활에 도움이 되는 모델을 만들고 싶어 최근 여러 사건들이 발생하며 이를 알려주는 뉴스를 매일 요약하여 그날 그날 어떠한 일들이 발생하였는지를 대략적으로 알 수 있게 도와주는 프로그램을 만들고자 ‘일일 뉴스 요약’ 모델을 만들어 보게 되었다. 아래의 탐구 내용들은 위에 언급한 책을 기반으로 추가적인 내용을 인터넷에서 검색한 내용을 토대로 작성한 것이다.

# 비지도학습과 K-평균

비지도학습이란 입력 데이터에 레이블이 없이 주어진 데이터만을 가지고 해당 데이터가 어떻게 구성되었는지를 알아내는 방법이다. 여기서는 세부분야 중 하나인 군집 분석에 대해 자세히 알아볼것이다. 군집분석은 같은 군집(클러스터)안의 아이템이 다른 클러스터의 아이템보다 더 비슷해지도록 데이터에 있는 자연스러운 그룹을 찾는것이다. 레이블이 지정되지 않은 데이터가 주어지면 그 데이터들의 유사점 또는 차이점을 찾아 서로 비슷한 클러스터 끼리 묶어 그룹들을 형성하게 된다. 군집화 알고리즘은 분류되지 않은 데이터들을 스스로 분석한 일정한 구조적 패턴에 기반하여 그룹으로 처리하는데 사용된다.

K-평균(K-means) 군집화는 데이터의 요소들을 k개의 클러스터로 묶는 알고리즘이다. 데이터간의 유사도를 측정하여 가장 유사한 데이터끼리 묶어 k개의 클러스터를 형성하게 된다. 책에 서술된 K-평균의 주요 알고리즘은 다음과 같다.

1. 샘플 포인터(데이터)에서 랜덤하게 k개의 센트로이드(평균)를 초기 클러스터 중심으로 선택
2. 각 샘플을 가장 가까운 센트로이드에 할당
3. 할당된 샘플들의 중심으로 센트로이드를 이동
4. 클러스터 할당이 변하지 않거나, 허용오차 또는 최대 반복 횟수에 도달할 때 까지 2와 3 반복

이때 2번에서 ‘가장 가까운’의 의미는 유클리디안 거리의 제곱이 최소가 되는 센트로이드란 뜻이다. 유클리디안 거리는 m-차원 공간에 있는 두 포인트 x, y에 대해

라 정의한다. 이 식의 의미는 샘플포인트 x와 y의 j번째 인덱스(차원) 사이의 거리의 제곱을 j=1부터 m까지 합한 값이다. 이 값이 작을수록 해당 데이터가 특정 클러스터 중심과 근접하다는 의미이다. 제곱을 사용하는 의미는 미분하여 최적화하기 편하게 하기 위함이다.

본 탐구에서는 책에서와 동일하게 사이킷런을 통해 k-평균을 구현하였다. 사이킷런의 KMeans 클래스를 임포트 한 후 사용 할 수 있다. 해당 클래스의 형태는 다음과 같다.

KMeans(n\_clusters=k, init="k-means++", n\_init=20, max\_iter=300, tol=1e-04, random\_state=42)

각 하이퍼파라미터(사용자가 직접 수정 가능한 파라미터)들은

n\_clusters: 클러스터의 개수

init: 초기화 방법

n\_init: 초기 중심위치 시도 횟수. 이 중 가장 우수한 결과를 최종적으로 채텍함

max\_iter: 최대 반복 횟수

tol: 허용 오차

random\_state: 랜덤 시드값

이다. 여기서 초기화 방법이 k-means++인데, 이는 초기 센트로이드가 서로 멀리 떨어지도록 위치시켜 기존(랜덤)보다 더 우수한 결과를 만들수 있도록 해준다.

클러스터의 개수는 기본적으로 직접 지정해 줘야한다. 이는 곧 k-평균의 단점이 된다. 이를 해결하기 위한 방법으로는 엘보우 방법과 실루엣 그래프 방법이 있다. 엘보우 방법은 k-평균 결과물의 SSE(왜곡)을 이용하여 최적의 클러스터 개수를 찾는 방법이다. 왜곡이 낮을수록 주로 좋은 성능을 가지며 k가 증가할수록 왜곡은 낮아진다. 이때 기울기가 빠르게 감소하는 k값을 찾으면 그 k값이 최적의 클러스터 k값이라 할 수 있다. 실루엣 방법은 실루엣 계수(-1~1 사이 값으로, 샘플들이 얼마나 조밀하게 모여있는지를 보여줌)를 분석하는 방법으로 실루엣 그래프를 그려서 실루엣 계수가 0에서 멀수록 군집이 잘 되었음을 보여준다. 본 탐구에서는 엘보우 방법만을 사용해 보았다.

# 일일 뉴스 요약 모델 개발

이제 앞에서 조사한 내용들을 토대로 모델을 직접 코딩을 통하여 개발하였다.

## 데이터 수집

데이터는 [네이버 뉴스](https://news.naver.com/)의 언론사별 랭킹 뉴스에서 [많이 본 뉴스](https://news.naver.com/main/ranking/popularDay.naver)와 [댓글 많은 뉴스](https://news.naver.com/main/ranking/popularMemo.naver)를 크롤링하여 사용하였다. 데이터는 날짜별로 json 형식으로 저장하였다. 뉴스데이터의 날짜 범위는 2024-12-01 ~ 2024-12-07 이다.

## 코드

데이터 수집, 모델 코드와 결과 자료들은 [깃허브](https://github.com/jjy0809/dailyNewsKMeans)에 업로드 해두었다.

## 모델

앞서 말했듯이 사이킷런을 이용하여 파이썬으로 구현하였다. k값을 10, 20, 30으로 각각 설정해보았으나 동일한 결과가 출력되었다. 추가로 엘보우 그래프도 출력하였지만 앞서 공부한 바와는 달리 눈에띄는 기울기 하락 지점은 보이지 않았다. 이 외 하이퍼파라미터 설정은 책을 따라했다.

주요 헤드라인 추출 방법은 우선 k값 만큼 클러스터를 군집화한 다음 군집의 크기가 큰 순으로 정렬하여 상위 5개의 군집에서 중심에 가장 가까운 헤드라인을 각 1개씩 총 5개를 추출하도록 하였다.

## 결과

생각보다 준수한 성능을 보여줬던 것 같다. 최근 자주 들리는 이슈 키워드로는 ‘윤석열, 이재명, 민주당, 국민의힘, 국회, 김건희, 명태균, 탄핵, 계엄’와 같은 정치적 사건들을 중심으로 ‘동덕여대 공학 전환 폭력 시위, 정우성-문가비 혼외자 출산’ 등의 기타 사건들이 있었다. 결과 중 특이한 사항으로는 비상 계엄이 선포된 2024-12-03의 결과의 경우

"20241203": [

"당신은 90세에 죽습니다 사망일 알려주는 시계 나왔다",

"속보 윤 대통령 비상계엄 선포",

"속보 윤 대통령 비상 계엄 선포",

"속보 한동훈 비상계엄 선포 잘못된 것 국민과 함께 막겠다",

"오세훈 이어 홍준표도 명태균 강혜경 고소 똑같은 여론조작 사기꾼"

],

이며, 계엄에 관한 헤드라인이 3개나 존재한다. 본래 이론대로라면 ‘계엄’이라는 하나의 클러스터안에 위 뉴스(샘플)이 모두 들어가거나 한두개의 클러스터 정도로만 분리되어야 마땅하나 클러스터 3개 이상으로 분화되었으며 심지어 그 중 두 개의 헤드라인은 100% 동일하다. 군집화가 잘 이루어 지지 않은것인지는 잘 모르겠으나 해당 날짜에 ‘계엄’과 관련된 뉴스가 매우 많아 k값 20이상에서 클러스터를 분류할 때 그만큼의 k값으로 불류할 만큼 뉴스의 다양성이 부족했던것으로 보인다. 아마 위 문제는 k값을 감소시키면 해결될것으로 보인다. 이외 김건희 특검법과 윤석열 탄핵소추 의결일인 2024-12-07의 결과의 경우

"20241207": [

"지금 이럴 때가 아니야 개미들은 계좌부터 팠다",

"국민의힘 안철수 김예지 김상욱 윤 대통령 탄핵안 표결 참여",

"담화나오자 일사불란 한동훈 윤 대통령 구해주고 침묵",

"국민의힘 10시간여 의총 끝 尹 탄핵 반대 당론 유지",

"속보 국민의힘 대구시당 앞에 모인 시민들 국민의힘 해체하라 의원들 나가라"

]

이며 해당 날짜의 실제 주요 이슈들에 맞게 뉴스 헤드라인이 잘 추출된것으로 보인다. 이외 모든 결과들은 깃허브에 있다.

# 결론

머신러닝의 학습법 중 비지도 학습, 그 중에서도 군집화 모델, 그 안에서 K-평균에 대해 중점적으로 다루어 보았다. 이후 K-평균을 이용하여 ‘일일 뉴스 요약’ 모델을 만들어 실제로 사용해 보았다. 결과는 꽤 괜찮게 나왔다고 생각하며, 추후 데이터를 더 넓은 범위에서 더 많이 수집하여 더 정교하며 포괄적인 요약 모델을 만들고 싶다.

# 출처

<https://ko.wikipedia.org/wiki/%EB%B9%84%EC%A7%80%EB%8F%84_%ED%95%99%EC%8A%B5>

<https://www.ibm.com/kr-ko/topics/unsupervised-learning>

<https://ko.wikipedia.org/wiki/K-%ED%8F%89%EA%B7%A0_%EC%95%8C%EA%B3%A0%EB%A6%AC%EC%A6%98>

<https://datascienceschool.net/03%20machine%20learning/16.02%20K-%ED%8F%89%EA%B7%A0%20%EA%B5%B0%EC%A7%91%ED%99%94.html>

그 외 모든 내용은 앞서 언급한 책을 참고함.