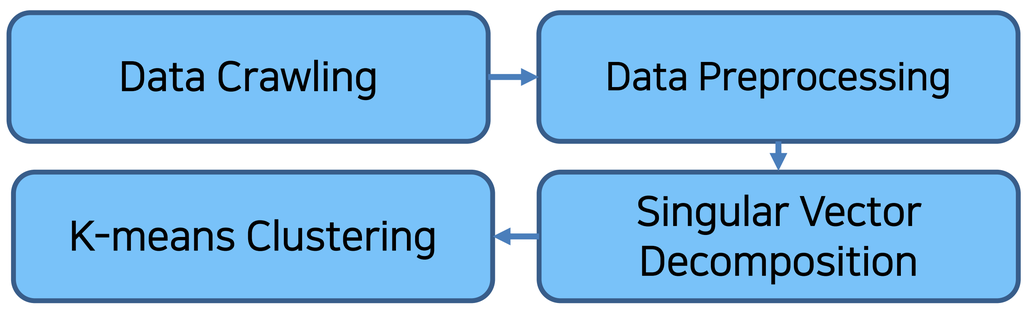
Youtube News Keyword Analysis Using Data Mining

팀 명 박정규 팀 원 박정규 지도교수 이슬 멘 토

개발 동기 및 목적

우리는 최근 TV와 포털사이트 이외에도 유튜브를 이용하여 뉴스를 접한다. KBS, SBS, MBC 등 방송사 유튜브 채널에 시간당 5~6개의 영상이 업로드 되고 있고 하루에 약 100개의 영상이 업로드 되고있다. 시간이 흐름에 따라 지속적으로 들어오는 Stream Data 를 분석하면 뉴스에서 집중적으로 조명하는 특정 키워드를 알 수 있고 사람들은 어떤 단어를 주로 댓글에 사용하며 어떤 주제에 관심이 많은지 알 수 있을 것이다.

개발 내용



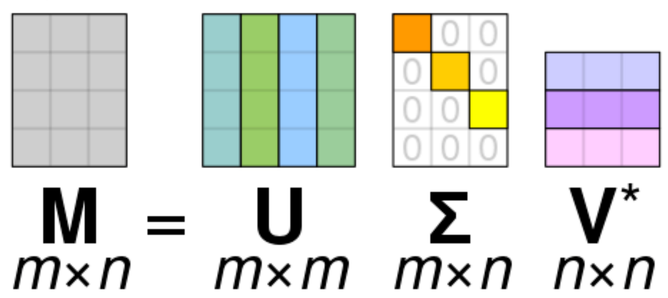
키워드 분석을 위해 가장 먼저 Python Selenium과 BeautifulSoup를 이용하여 데이터 크롤링을 진행했다. 크롤링 데이터는 일반적으로 label 되어있는 데이터와는 다르게 Data Preprocessing 과정이 매우 중요하다. 영상 제목과 댓글에는 키워드 분석에 필요 없거나 노이즈를 발생시키는 불용 어구(특수문자, 유니코드, 이모티콘 등) 단어들이 존재하므로 정규 표현식을 사용해 모두 제거해 준다. 또한 크롤링 중 인터넷 오류 또는 Colab 오류로 인해 크롤링이 되지 않아 row 값이 비어 있는 데이터도 존재하므로 이 또한 제거해 준다. 전처리된 데이터들을 토큰 화해 주기 위해 한국어 정보처리 라이브러리인 KoNLPy를 사용해 문장에서 명사 키워드만 추출해 데이터 분석에 사용한다.

토큰화된 단어 데이터들을 Word2Vec 라이브러리를 이용해 Hyper Parameter를 분석에 적합하게 설정하여 단어를 벡터로 수치화해준다. DM Algorithm인 SVD(Singular Vector Decomposition)를 사용해 벡터화된 단어를 벡터 행렬 **U**, **Sigma**, **Vt** 로 분리 해주고 **Sigma**의 Singular Value값이 급격히 감소하는 지점을 찾아 해당 지점 이후의 값들은 모두 삭제하고 행렬의 곱셈으로 원본 행렬로 복원한다.

비슷한 단어의 묶음을 보기 위해 K-means Clustering 사용했고 SSE(Sum Squared Error)를 이용한 elbow method를 사용해 최적의 K값을 찾아 Clustering을 진행했다.

주요기술

1. Data Crawling
   - Python Selenium, BeautifulSoup, Pandas를 사용했다. Selenium으로 크롬을 자동제어해 유튜브 영상 클릭 후 스크롤을 가장 아래로 내린다. 그 후 해당 페이지의 html 소스를 BeautifulSoup를 사용해 영상 제목과 댓글을 가지고 있는 html tag를 parsing후 Pandas를 사용해 엑셀로 저장한다.
2. Data Preprocessing
   - Python Re 모듈의 정규표현식을 사용해 불용어 단어들을 제거해준다. (특수문자, 유니코드, 이모티콘 등)
3. Text Vectorization
   - Word2Vec 라이브러리의 skip-gram을 사용해 중심 단어에서 주변 단어를 기반으로 단어를 벡터화 해준다.
4. Data Mining Algorithm
   - SVD(Singular Vector Decomposition) – 특잇값 분해



특잇값 분해(SVD)를 사용하여 M행렬을 특정 구조로 분해 후 고윳값 행렬에서 급격히 값이 감소하는 지점 이후의 값들은 제거 하는 차원축소 과정을 지난 후 행렬의 곱셈으로 원래 행렬로 복원한다.

- K-means Clustering

K-means Clustering을 사용하여 연관성이 높은 단어를 군집화 해서 분석 결과를 확인하기 쉽게 했다.

결과 및 분석

```
Cluster 0 ⇒ ['김종대', '지은', '과학', '독재정', '무기징역', '명신', '애국자', '조심', '지옥', '전과']
Cluster 1 ⇒ ['진짜', '우리', '이제', '인간', '하나', '정부', '얼마나', '집회', '이상', '다시']
Cluster 2 ⇒ ['날씨', '용산', '바이든', '공작', '가을', '정은', '강등', '상품', '해체', '캠핑']
Cluster 3 ⇒ ['국민', '수사', '이재명', '윤석열', '김건희', '민주당', '대통령', '국힘', '대선', '대표']
Cluster 4 ⇒ ['특검', '범인', '거부', '자가', '대장동']
Cluster 5 ⇒ ['혼란', '노고', '아무나', '엠비씨', '김순호', '중단', '코시', '좌우', '리가', '소름']
Cluster 6 ⇒ ['고스톱', '수갑', '요안나', '재능', '지진', '미세먼지', '윤성렬', '이형', '심해', '물증']
Cluster 7 ⇒ ['검찰', '나라', '지금', '대한민국', '검사', '구속', '조작', '증거', '응원', '모두']
Cluster 8 ⇒ ['가맹', '점주', '본사']
Cluster 9 ⇒ ['사람', '생각', '처벌', '정말', '문제', '경찰', '그냥', '저런', '우리나라', '세상']


Cluster 0 ⇒ ['문화', '데스크', '글로벌', '위반', '부산', '제빵', '결정', '신규', '노영민', '정치']
Cluster 1 ⇒ ['광장', '가을', '처리', '원도', '열사', '버스', '병원', '불발', '언제', '사망자']
Cluster 2 ⇒ ['검찰', '뉴스', '민주', '브리핑', '외전', '세계', '자막', '시각', '경제', '불법']
Cluster 3 ⇒ ['보기', '신고', '거짓', '확대', '해양', '동전', '대우조선', '기계', '정지', '절반']
Cluster 4 ⇒ ['이재명', '김용', '수사', '구속', '특검', '서욱', '피격']
```

영상 제목과 댓글의 특잇값 분해를 적용 후 k-means Clustering을 적용한 각 클러스터의 데이터 내용이다. 주로 정치와 관련한 키워드가 많은 것을 볼 수 있었으며 10월 26일 기준 당시 이슈와 관련된 키워드도 볼 수 있다. 실제로 사람들이 어떤 단어를 주로 뉴스 영상 댓글에 사용하고 있는지 알 수 있고, 주로 어떤 단어 종류를 사용하고 있는지 알 수 있다.