

NLP 키워드 추출 기반  
**경제 트렌드 뉴스레터**

---

6조 윤희재 임홍주 조은정

SUBSCRIBE



# 목차

PART 1.

## 주제 선정 배경

PART 2.

## NLP 프로세스

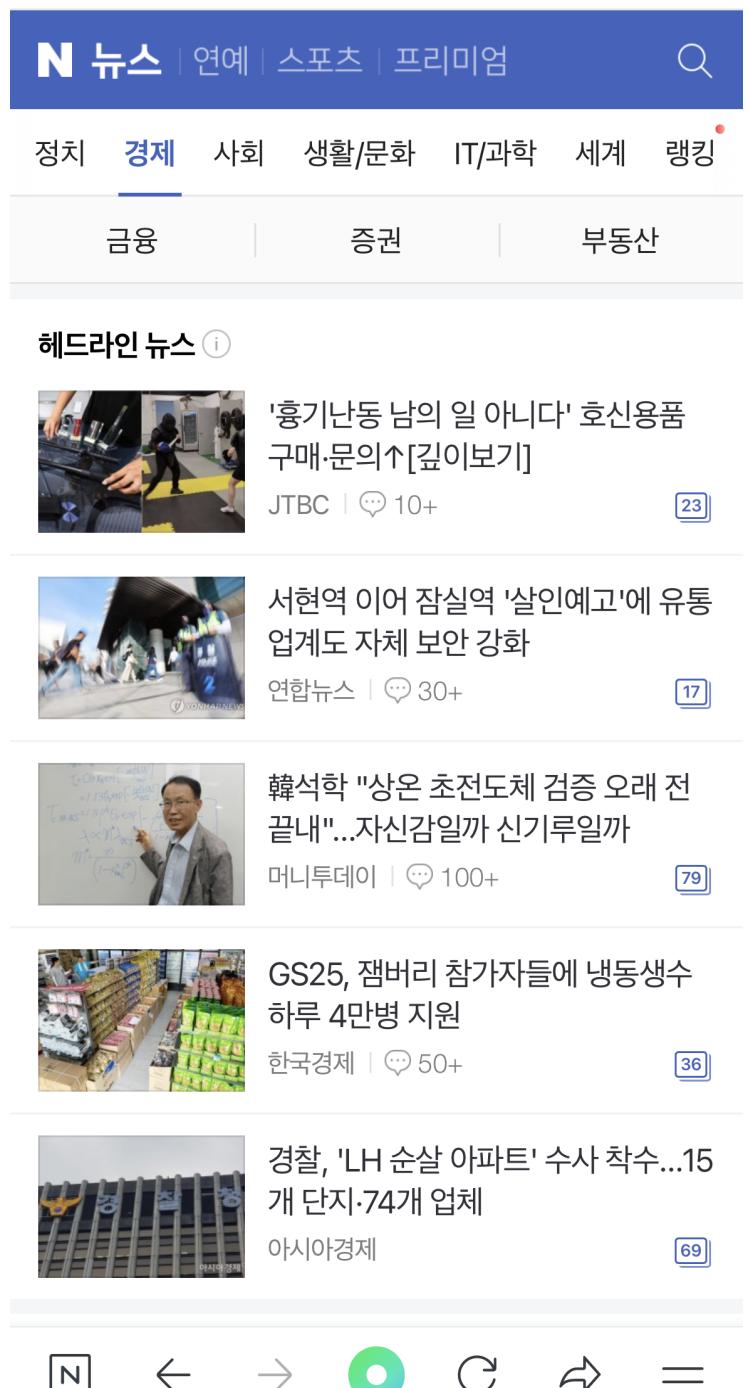
PART 3.

## 프로젝트 결과

- 배경 및 목적
- 뉴스 수집 및 전처리
- 기사별 Top-K개 키워드 선정
  - 1차: 빈도수 기반
  - 2차: 코사인 유사도 거리 기반
- 대주제 탐색
- 주제별 기사 요약
- 결과물
- 의의 및 한계점

# [Part 1] 주제 선정 배경

## 배경



하루 평균 경제 기사만 약 1500개  
어떤 뉴스를 읽어야 할까? 오늘 이슈 뉴스는 무엇일까?

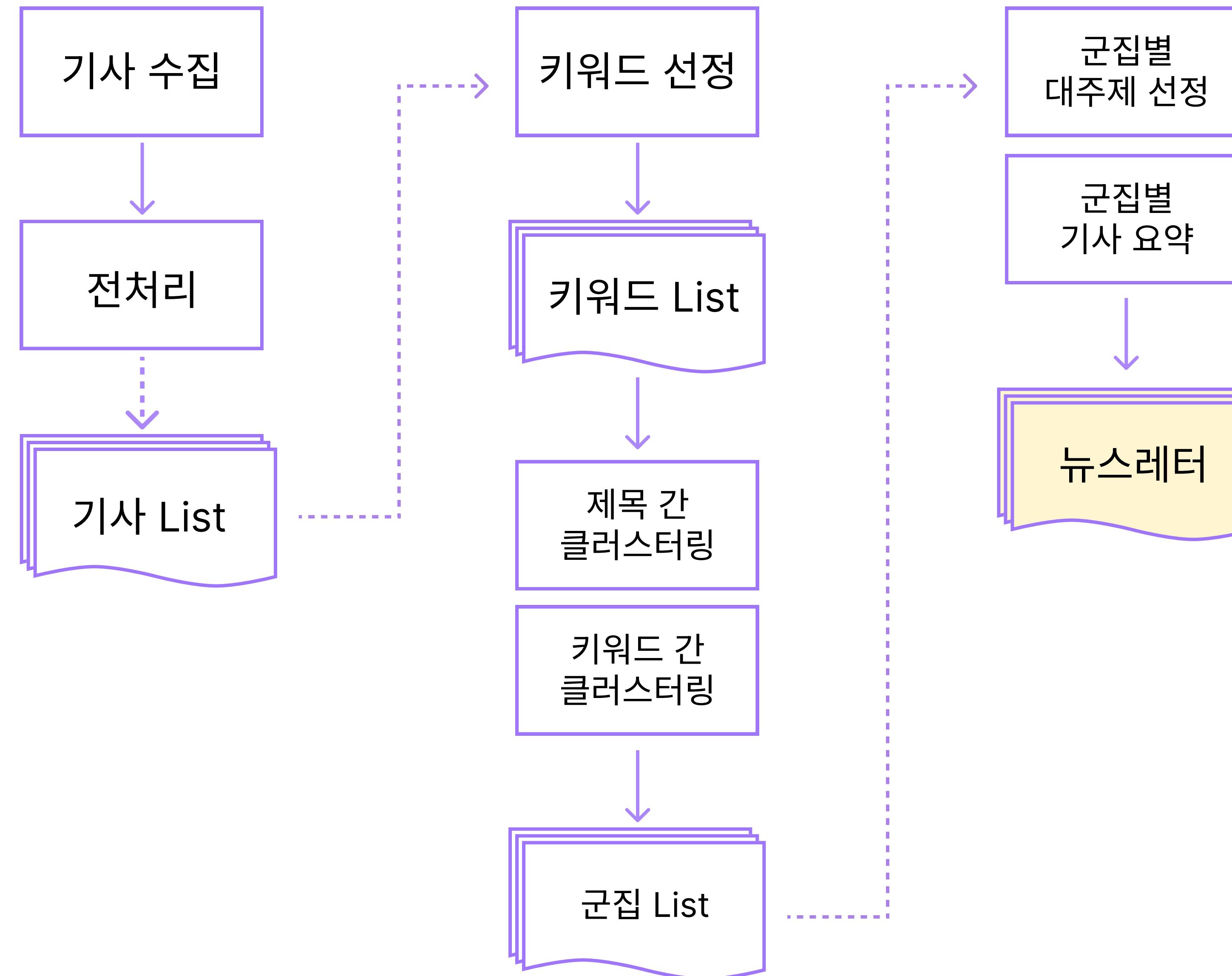
## 목적

A screenshot of the NEW NEEK homepage. The main header reads 'NEW NEEK'. Below it is a large cartoon illustration of a hedgehog holding a newspaper with a speech bubble that says 'WAKE UP!!'. To the right of the hedgehog, the text reads '“이러다 오늘도 가장 유식하겠는데”' (Like this, today will also be the most delicious). Below this, there is a 'MONEYLETTER' section with the text '어서와, 뉴닉은 처음이지?' (Hello, this is my first time, New Nick). To the right, there is a sidebar titled '돈 되는 돈 이야기' (Stories of Money) which includes sections for 'MONEYLOG' and 'SERIAL'. The 'MONEYLOG' section shows a person holding money with the text '돈 번 곳에 세금 내시오!' (Pay taxes at the place where you make money!). The 'SERIAL' section shows a hamster in a wheel with the text '돈 뒤의 사람 이야기' (Stories of people behind money).

경제 기사를 읽고자 하는 사람을 대상으로 **매일 짧은 시간 내에**  
**당일 경제 트렌드를 쉽게 파악할 수 있도록** 핫 키워드별 뉴스레터 제공

## [Part 2] NLP 프로세스

### 분석 개요



# 1. 뉴스 수집 및 전처리

## a. 크롤링

### 크롤링 데이터

- 뉴스레터 하루치 분량  
: 07.13(목) 06시부터 07.14(금) 06시까지

### 뉴스 크롤링 기준

- 실시간 최신순 뉴스 (헤드라인 뉴스 X)
- 페이지당 기사 20개 크롤링
- 모든 기사의 제목, 본문, 날짜, 언론사 정보 수집

### 크롤링 코드

```
# 본문 - 사진 캡션 제거
captions_list = content_paragraphs.findAll(name="span", attrs={"class":"end_photo_org"})
for caption in captions_list:
    caption.decompose()

# 본문 - 문장 구분
...
1. 문장 부호가 없어 구분되지 않는 문장을 사이에 공백 추가하여
요약문 한 줄 마지막 단어와 다음 줄 첫 단어가 이어지지 않도록
& 소제목 마지막 단어와 본문 첫 단어가 이어지지 않도록
2. strip()으로 본문 시작과 끝에 있는 줄바꿈과 공백을 제거
...
content_corpus = content_paragraphs.getText(separator=" ").strip() # 텍스트만
content_corpus = content_corpus.strip()

# 날짜
date_html = soup.find(name="span",
                       attrs={"class":"media_end_head_info_datestamp_time_ARTICLE_DATE_TIME"}) # 최초 입력 기준, 수정 시각 무시
date = date_html.attrs["data-date-time"]

# 언론사
press_logo = soup.find(name="img", attrs={"class":"media_end_head_top_logo_img"})
press = press_logo.attrs["alt"]

# 모든 정보를 하나의 데이터 프레임에 저장합니다.
row = [date, row_title, content_corpus, press, page_url]
series = pd.Series(row, index=df.columns)
df = df.append(series, ignore_index=True)
```

### 결과

	title	content	date	press
0	한 집에 한 대가 안돼, 차댈 곳 없어 '뱅뱅'.. "임대 살아서?"	1가구당 고작 0.79대 주차.. 비용 때문에 분양 0.68대~1.28대↔임대 0.2...	2023-07-15 16:22:01	JIBS
1	[데일리안 오늘뉴스 종합]尹대통령, 우크라이나 깜짝 방문...젤렌스키와 정상회담...	[데일리안 = 정진주 기자] ▲尹대통령, 우크라이나 깜짝 방문...젤렌스키와 정...	2023-07-15 17:53:01	데일리안
2	"최선 차선 모두 잘못됐네"...붕괴 사고로 머리 아픈 조합원들	아이파크 붕괴사고 후 GS건설로 시공사 변경했더니 조합 "집값 하락·단지 안전..."	2023-07-15 13:11:01	매일경제
3	"한 집에 한 대도 못 대"...임대아파트 주차난 어쩌나	가구당 평균 주차가능대수 0.79대에 불과 결국 비용 문제...주차공간 늘리면 마진 줄...	2023-07-15 14:11:01	아시아경제
4	용진이형도 맛 본 새우깡 동생 '먹태깡'... 중고거래서 대박난 까닭	먹태깡 품귀현상, 중고마켓에서 2~3배 웃돈 거래까지 1회당 4봉지 제한...식품업계...	2023-07-15 14:18:01	이코노미스트
5	"미국은 이미 중국에 졌다"...美포드 회장 작심한 듯 '전기차 미래' 솔아냈다 [...]	미포드 회장, 인터뷰에서 작심발언 '중국이 미래차 시장 재편할 것' 전망 비야디,...	2023-07-15 13:37:01	매일경제
6	웃돈만 9억인데 사라진 아파트...한남3구역 현금청산 날벼락[부동산 360]	한남3구역 현금청산자 216명...20명 중 1명 끝 일부는 소송전에도 나서 투기과열지...	2023-07-15 17:42:01	헤럴드경제
7	'영끌' 오픈런 명품소비 1위 한국...이부진도 불평도 루이비통家와 각별한 관계로?	CNBC "작년 한국인 명품 구매액 세계 1위" 루이비통 셋째 며느리 구이엇 이부진...	2023-07-15 18:11:01	서울경제
8	오바마의 첫 알바가 이 곳에서? 무더위 속 그가 얄은 교훈 [추동훈의 흥부전]	[브랜드로 남은 창업자들-09] 배스킨 라빈스 #미국의 첫 흑인 대통령, 버락 오바...	2023-07-15 18:26:00	매일경제
9	어떻게 빨려들어갈지 모르는 민감정보 '블랙홀' 챗GPT 규제책은	[인공지능의 두 얼굴(17)] 개발 및 이용 과정에서 이용자 정보 대량 학습·유출...	2023-07-15 18:06:01	미디어오늘

## 1. 뉴스 수집 및 전처리

## b. 전처리

# 1. 특수문자 종류와 빈도수 확인

- 한글로 변환 해석(% → 퍼센트)
  - 단위 치환: 특수문자를 알파벳으로 통일 (km)
  - 한자 치환

## 2. 일반 텍스트 전처리

- 바이라인(기자 이메일, 언론사 URL) 제거
  - 한글, 숫자, 알파벳, 일부 특수문자(., & ~), 공백만 유지
  - 경제 기사답게 퍼센트(%, %)가 가장 많이 등장 -> 단위 통일

### 3. 불필요한 텍스트 제거

- [단독] 제거

4. 결과: 새로운 열 content\_p, title\_p, len (content\_p 본문 길이) 생성

전처리 코드

```

'''기사 형식'''
# 바이라인: 이메일, url 제거
import re
pattern_mail = re.compile(r'[\w\w.]+@\w\w?[\w\w]+\.\w\w.+\w')
# 공백 포함: mbcjebo @ mbc.co.kr

text = re.sub(pattern_mail, '', text)
pattern_url = re.compile(r'^(?:https?:\/\/)?[-_0-9a-z]+(?:[\.\.\w\w] [-_0-9a-z]+)*', flags=re.IGNORECASE)
text = re.sub(pattern_url, '', text)
pattern_call = re.compile(r'\w\w{2,3}-\w\w{3,4}-\w\w{4}')
text = re.sub(pattern_call, '', text)

'''특수문자'''
# 한자를 한글로 치환 (兆, 조)
text = text.replace('比', '대비')
text = hanja.translate(text, 'substitution')
text = text.replace('年', '년')
text = text.replace('季', '이') # 이창용 한은 총재

# 단위 치환
text = text.replace('km', 'km')
text = text.replace('kg', 'kg')
text = text.replace('ml', 'ml')
text = text.replace('mg', 'mg')
text = text.replace('ℓ', 'l')
text = text.replace('cm', 'cm')
text = text.replace('pH', 'pH')
text = text.replace('$', '달러')
text = text.replace('m', 'm')
text = text.replace('g', 'g')
text = text.replace('kW', 'kW')
text = text.replace('Hz', 'Hz') # 헤르츠도 있음 - 통의어 사전 필요함
text = text.replace('℃', '섭씨')
text = text.replace('kal', '칼로리')
text = text.replace('%', '%')
text = text.replace('%', '퍼센트')
text = text.replace('...', '...') # 추후 띄어쓰기 위함

# 특수문자 제거
# 예외: .(1.2%), &(B&S 홀딩스, S&P), ~(30~40)
pattern_special = r'[^가-힣0-9a-zA-Z\w,\w\w\w\w\w\w] ' # 한글, 숫자, 영어, 공백, 기타 문자만 유지
text = re.sub(pattern_special, ' ', text)

```

## 결과

	date	title	content	press	link	title_p	content_p	len
259	2023-07-13 09:18:03	파라다이스시티 부티크 호텔 '아트파라디소' 3년 만에 재개장	(서울=연합뉴스) 차민지 기자 = 파라다이스시티는 부티크 호텔 '아트파라디소'를 새...	연합뉴스	<a href="https://n.news.naver.com/mnews/article/001/001...">https://n.news.naver.com/mnews/article/001/001...</a>	파라다이스시티 부티크 호텔 아트파라디소 3년 만에 재개장	서울 연합뉴스 차민지 기자 파라다이스시티는 부티크 호텔 아트파라디소를 새로 단장 하...	263
268	2023-07-13 09:20:59	나노엔텍, 에이플러스그룹 품에 안긴 나노엔텍이 강세다. 13일 오전 9시19분 나노엔텍은 전...	에이플러스그룹 품에 안긴 나노엔텍이 강세다. 13일 오전 9시19분 나노엔텍은 전...	머니투데이	<a href="https://n.news.naver.com/mnews/article/008/000...">https://n.news.naver.com/mnews/article/008/000...</a>	나노엔텍, 에이플러스그룹 시너지 기대감..↑	에이플러스그룹 품에 안긴 나노엔텍이 강세다. 13일 오전 9시19분 나노엔텍은 전일...	291
297	2023-07-13 09:29:10	이마트24, '라떼는쑥파르페' 출시..."할매니얼 저격"	컵 탑 파르페...투 플러스 원 행사 [서울=뉴시스] 심동준 기자 = 이마트24는 자...	뉴시스	<a href="https://n.news.naver.com/mnews/article/003/001...">https://n.news.naver.com/mnews/article/003/001...</a>	이마트24, 라떼는쑥파르페 출시 할매니얼 저격	컵 탑 파르페 투 플러스 원 행사 서울 뉴시스 심동준 기자 이마트24는 자체 브랜...	294
520	2023-07-13 10:36:41	서대문구, 홍제1구역 '서대문푸르지오센트럴파크' 준공 인가	(서울=뉴스1) 김도엽 기자 = 서울 서대문구는 '서대문푸르지오센트럴파크' 아파트가...	뉴스1	<a href="https://n.news.naver.com/mnews/article/421/000...">https://n.news.naver.com/mnews/article/421/000...</a>	서대문구, 홍제1구역 서대문푸르지오센트럴파크 준공 인가	서울 뉴스1 김도엽 기자 서울 서대문구는 서대문푸르지오센트럴파크 아파트가 들어선 홍...	305
359	2023-07-13 09:51:35	최태원 "전 세계 블록화로 기업 살아남기 어려워.. 정부·민간 함께 뛰어야"	최태원 대한상공회의소 회장이 최근 경제 상황에 대해 "미·중 갈등으로 전 세계 블록...	MBC	<a href="https://n.news.naver.com/mnews/article/214/000...">https://n.news.naver.com/mnews/article/214/000...</a>	최태원 전 세계 블록화로 기업 살아남기 어려워 정부·민간 함께 뛰어야	최태원 대한상공회의소 회장이 최근 경제 상황에 대해 미·중 갈등으로 전 세계 블록화...	346

## 2. 기사별 키워드 선정

### a. 목표

#### 목표

명사/명사구만을 남겨  
문서를 키워드 단위로 요약



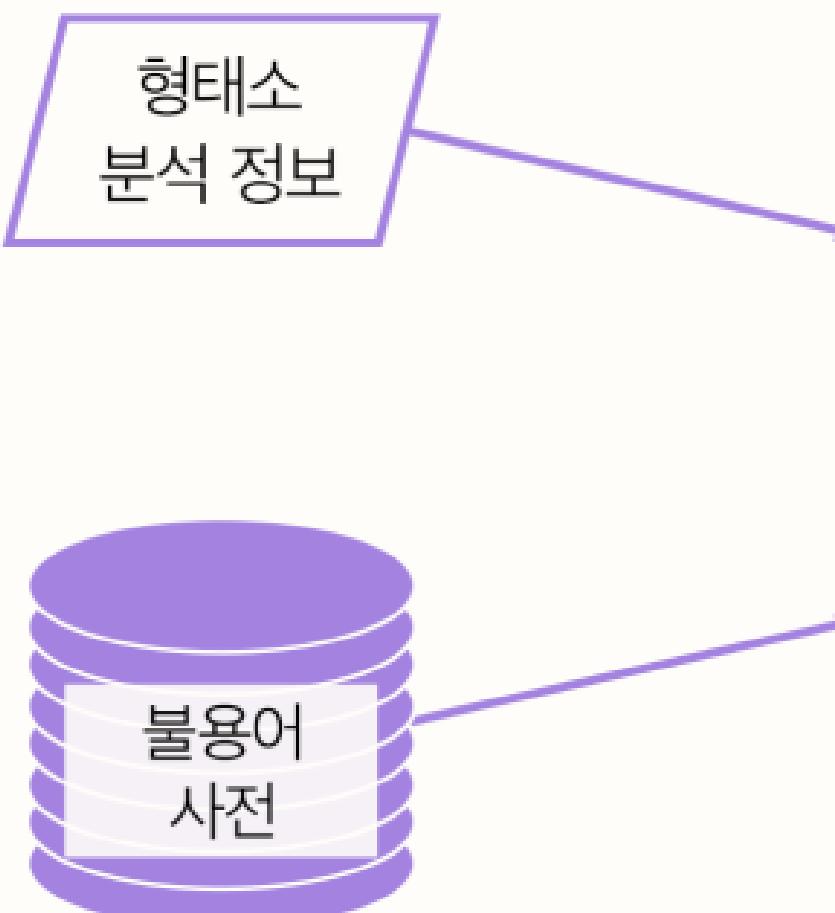
빠르고 간단한  
정보 전달

#### 방법

1. 빈도수 기반
2. 코사인 유사도

## 키워드 후보군 추출 예시

...  
이번에 공개하는 오픈 베타 서비스는 아직은 초기 버전의 서비스이다. 'A' 서비스내 캐릭터가 고객과 교감하는 기간을 통해 성장하며 진화할 예정이다. 이를 위해 해당 오픈 베타 서비스 기간 동안 서비스를 자주 이용하고, 우수한 제언을 하는 참여자들에게 혜택을 제공하는 프로모션을 진행할 예정이며, 이러한 사용자 피드백을 다른 어떤 서비스보다도 빠르고 적극적으로 반영하여 개선해 나갈 것이다.



## 2. 기사별 키워드 선정

### b. 빈도수 기반 선정: TF - IDF

#### TF - IDF

문서 내에서 특정 단어의 중요도를 계산하는 방법

TF : 하나의 문서 내에서 특정 단어가 나타나는 빈도

IDF : 여러 문서에서 특정 단어가 얼마나 공통적으로 등장하는지를 나타내는 값

#### TF \* IDF :

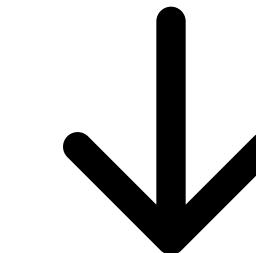
한 문서에 많이 등장하면서 다른 문서에서 적게 등장할수록 중요한 단어로 인식

#### TF - IDF 키워드 예시

['듯하다', '소비자물가는', '안으로는',  
'금리인상을', ... , ]

#### 형태소 분석을 먼저 진행하지 않은 이유

1. 형태소 분석 시 기업명을 잘 잡지 못함
2. 명사 + 명사 붙어 있는 경우 서로 다른 단어로 인식  
(EX. 임대보증금 > 임대, 보증금)
3. 띄어쓰기 되어 있을 경우 하나의 단어로 잡지 못함



#### 주요 단어의 누락

```
# TF-IDF를 사용하여 문서 벡터화 진행
tfidf_matrix = tfidfv.fit_transform(df1['content_p'])

for i in range(10):
    doc_tfidf = tfidf_matrix[i].toarray().flatten() # 벡터화시킨 것을 다시 문자로
    sorted_indices = doc_tfidf.argsort()[-1] # tf-idf값이 높을수록 중요한 키워드인데 높은 순으로 정렬
    top_keywords = [tfidfv.get_feature_names_out()[idx] for idx in sorted_indices[:15]] # 상위 10개 단어만 추출
    # 숫자나 영어 단어만 나올 경우 제거하고 다음 키워드를 올려서 쓸거기 때문에 넉넉히 뽑음
    # 키워드를 저장할 빈 리스트 생성
    keywords = []
    keywords.append(top_keywords)
    # 각 행에 알맞은 키워드 부여
    df1[['TF-keywords']][i] = keywords
print(df1[['TF-keywords']][1])

<ipython-input-7-bc3a2566a001>:12: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user\_guide/indexing.html#returning-a-view-versus-a-copy
df1[['TF-keywords']][i] = keywords
[['제로', '열량은', '말티률은', '발암', '설탕', '아스파탐', '다이어트에', '설탕이', '열량', '진로', '혼란스러워하고', '도수를', '설탕의', '식품회사들이', '제로슈거']]
```

## 2. 기사별 키워드 선정

### b. 빈도수 기반 선정 : N-gram

#### 연어(Collocation)

두 단어가 연속적으로 쓰여 뜻을 가짐

2음절 = BigramAssocMeasures

3음절 = TrigramAssocMeasures

한 [귀 로] 들고 한 귀 로 훌린 다

한 귀 [로 들고] 한 귀 로 훌린 다

한 귀 로 [들고 한] 귀 로 훌린 다

음절의 단위를 형태소 분석 결과로 나누어서 정확도를 높임

2음절 = 2 단어 연어, 3음절 = 3 단어 연어

```
measures = collocations.BigramAssocMeasures()
print(type(measures))
doc = selectsample(20, content=True)

# bigram
# 품사 부착
print('#nA. Collocations among tagged words:')
tagged_words = okt.pos(doc) # 품사 부착
# tagged_words = Kkma().pos(doc) # 품사 부착
finder = collocations.BigramCollocationFinder.from_words(tagged_words)
# finder.apply_freq_filter(3) # only bigrams that appear 3+ times
pprint(finder.nbest(measures.pmi, 10)) # top 10 n-grams with highest PMI

# 알고리즘: pmi
print('#nB. Collocations among words:')
words = [w for w, t in tagged_words] # 단어+품사 중 단어만 선택
ignored_words = [u'되다 되었다 됐다 하다 하였다 했다 이다 이었다 였다 있다 없다'] # 불용어
finder = collocations.BigramCollocationFinder.from_words(words)
finder.apply_word_filter(lambda w: len(w) < 2 or w in ignored_words) # 제외조건: 길이가 1이거나 ignored_words에 포함되면 제외
finder.apply_freq_filter(3) # only bigrams that appear 3+ times
pprint(finder.nbest(measures.pmi, 10)) # top 10 n-grams with highest PMI

# 품사만
print('#nC. Collocations among tags:')
tags = [t for w, t in tagged_words] # 품사만 선택
finder = collocations.BigramCollocationFinder.from_words(tags)
pprint(finder.nbest(measures.pmi, 5))

<class 'nltk.metrics.association.BigramAssocMeasures'>

A. Collocations among tagged words:
[('100억', 'Number'), ('달러', 'Noun'),
 ('11일', 'Number'), ('증가', 'Noun'),
 ('13일', 'Number'), ('오전', 'Noun'),
 ('1만', 'Number'), ('5천', 'Number'),
 ('20억달러', 'Number'), ('약', 'Noun'),
 ('2조', 'Number'), ('6천억', 'Number'),
 ('3일', 'Number'), ('부터', 'Foreign'),
 ('5', 'Number'), ('배', 'Noun'),
 ('66만', 'Number'), ('9천', 'Number'),
 ('9시', 'Number'), ('30분', 'Number')]
```

## 2. 기사별 키워드 선정

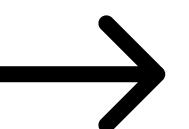
### c. 형태소 분석 및 불용어 처리

#### 형태소 분석

비타민은      언제나      즐거워  
    ↖↗                  ↓                  ↓  
    Noun Josa      Adverb      Adjective

#### “형태소 분석기 = Okt”

['듯하다',  
'소비자물가는',  
'안으로는',  
'금리인상을',  
'마켓 pro',  
'여겨진다',  
'고용과',  
'가운데',  
'방향성을',  
'입장이',  
'25bp',  
'인해',  
'비교적',  
'물가에',  
'따라']



['듯',  
'소비자 물가',  
'',  
'금리 인상',  
'마켓 pro',  
'',  
'고용',  
'가운데',  
'방향성',  
'입장',  
'25 bp',  
'',  
'비교 적',  
'물가',  
'']

#### 불용어 처리

##### 기존 불용어 사전

+

##### 불용어 직접 추가 (EX. A씨, 덕분에, 인해 등등)

```
# Okt 형태소 분석기 초기화
okt = Okta()

# 불용어 사전 파일을 읽어와서 불용어 사전 구성
# 배포되어 있는 불용어 사전에 개인작업으로 추가 불용어들 추가
with open('/content/drive/MyDrive/Colab Notebooks/new_stopwords.txt', 'r', encoding='utf-8') as f:
    korean_stopwords = set(f.read().splitlines())

# 형태소 분석 후 명사와 숫자, 알파벳, 접미사만 남기는 함수 정의
def filter_words(text):
    morphemes = okt.pos(text)
    filtered_words = [word for word, pos in morphemes if pos in ['Noun', 'Number', 'Alpha', 'Suffix'] and word not in korean_stopwords]
    return ' '.join(filtered_words)

# 빈 리스트 할당
df1['new_key'] = [[] for _ in range(len(df1['TF-keywords']))]

for i in range(len(df1['TF-keywords'])):
    for j in range(len(df1['TF-keywords'][i])):
        filtered_words = filter_words(df1['TF-keywords'][i][j])
        df1['new_key'][i].append(filtered_words)

print(df1['new_key'])
```

## 2. 기사별 키워드 선정

### d. 키워드와 기사 간 코사인 유사도 계산해 비교

```
for i in range(len(df)):
    # 제목, 본문, 키워드 후보 데이터
    title = df['title_p'][i]
    body = df['content_p'][i]
    keyword_candidates = df['merged_col'][i]

    # TF-IDF 벡터화 객체 생성
    tfidf_vectorizer = TfidfVectorizer()

    # 제목과 본문을 따로 TF-IDF 벡터화
    title_vector = tfidf_vectorizer.fit_transform([title])
    body_vector = tfidf_vectorizer.transform([body])

    # 키워드 후보들을 TF-IDF 벡터화
    keyword_candidate_vectors = tfidf_vectorizer.transform(keyword_candidates)

    # 제목과 본문의 TF-IDF 벡터를 가중합(Weighted Sum) 및 정규화(Normalization)하여 의미 벡터 생성
    alpha = 0.7 # 제목 벡터 가중치 (0과 1 사이의 값을 선택)
    beta = 0.3 # 본문 벡터 가중치 (0과 1 사이의 값을 선택)

    title_vector = title_vector.toarray()
    body_vector = body_vector.toarray()
    title_body_vector = alpha * title_vector + beta * body_vector

    # 정규화(Normalization)하여 의미 벡터를 단위 벡터로 변환
    title_body_vector = normalize(title_body_vector)[0]

    # 키워드 후보 벡터와의 의미 유사도를 측정하여 Top-K애의 키워드 선정
    k = 3 # Top-K 개수를 선택해주세요

def cosine_similarity(vec1, vec2):
    dot_product = np.dot(vec1, vec2)
    norm_vec1 = np.linalg.norm(vec1)
    norm_vec2 = np.linalg.norm(vec2)
    return dot_product / (norm_vec1 * norm_vec2)

similarity_scores = [cosine_similarity(title_body_vector, candidate.toarray()[0]) for candidate in keyword_candidate_vectors]
top_k_indices = np.argsort(similarity_scores)[-k:]
selected_keywords = [keyword_candidates[i] for i in top_k_indices]

print("선택된 키워드:")
for keyword in selected_keywords:
    print(keyword)
```

## 지금까지의 키워드

문맥 고려 X      단어 의미 고려 X

키워드 벡터

제목+본문 벡터

키워드들을 벡터화시켜줌

1. 제목과 본문 각각 벡터화
2. 제목 : 본문 = 7 : 3 가중치 부여 후 합산
3. 합산된 벡터값 정규화

두 벡터간의 코사인 유사도가 높은 순으로 10개 선택

키워드와 본문간의 연관성 향상

### 3. 대주제 탐색

#### a. DBSCAN으로 주제 군집화

##### 군집화란?

- 비슷한 특성을 가지는 데이터들을 함께 그룹화하는 비지도 학습 (Unsupervised Learning) 기법

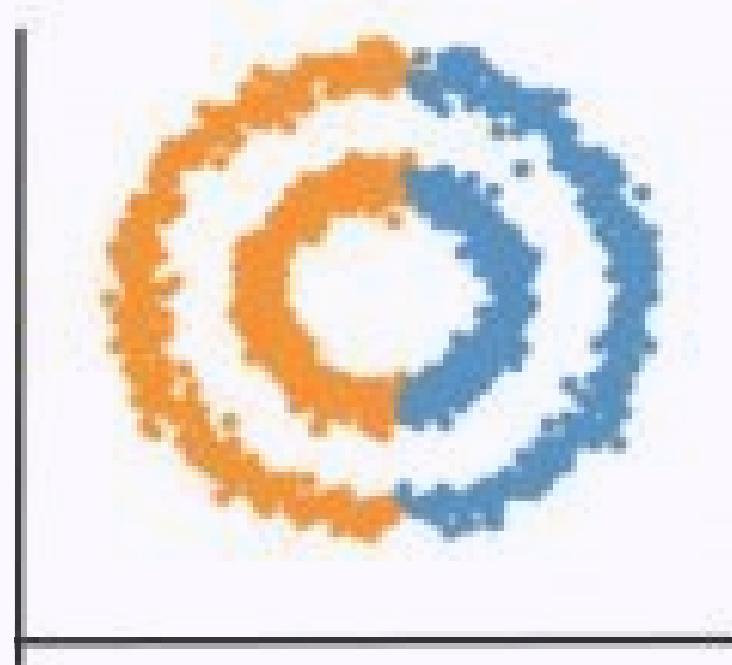
##### 군집화를 하는 목적

- 매일 다양한 주제에 관한 기사가 새로 발행됨
- 크롤링된 데이터는 어떠한 분류도 없이 존재
- 다양한 주제속에서 이슈가 되는 주제를 선정하고자 함

### 3. 대주제 탐색

#### a. DBSCAN으로 주제 군집화

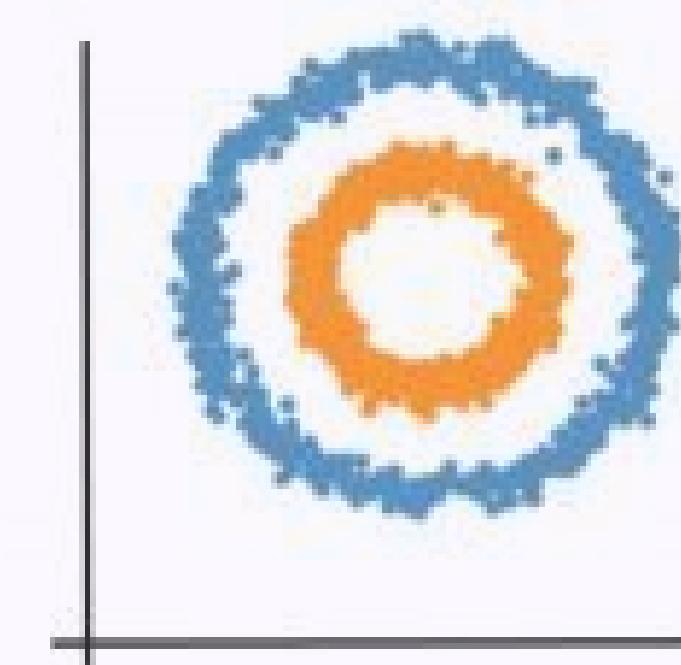
##### K-means Clustering (거리 기반 클러스터링)



K-MEANS

군집 설정 필요 O  
이상치 특정 불가능  
초기 군집 존재 O

##### Dbscan (밀도 기반 클러스터링)



DBSCAN

군집 설정 필요 X  
이상치 특정 가능  
초기 군집 존재 X

#### Dbscan 구성

##### 1. 핵심 포인트(Core-point)

주어진 반경(eps) 내에 최소 데이터 포인트 수(min\_samples)  
이상의 이웃포인트를 가지는 데이터 포인트

##### 2. 노이즈 포인트(Noise-point)

주어진 반경(eps) 내에 최소 데이터 포인트 수(min\_samples)  
이상의 이웃점을 가지지 않는 데이터 포인트

##### 3. 경계 포인트(Border-point)

주어진 반경(eps) 내에 최소 데이터 포인트 수(min\_samples)  
미만의 이웃점을 가지지만, 핵심 포인트의 이웃 포인트인  
데이터 포인트를 경계 포인트로 정의합니다.

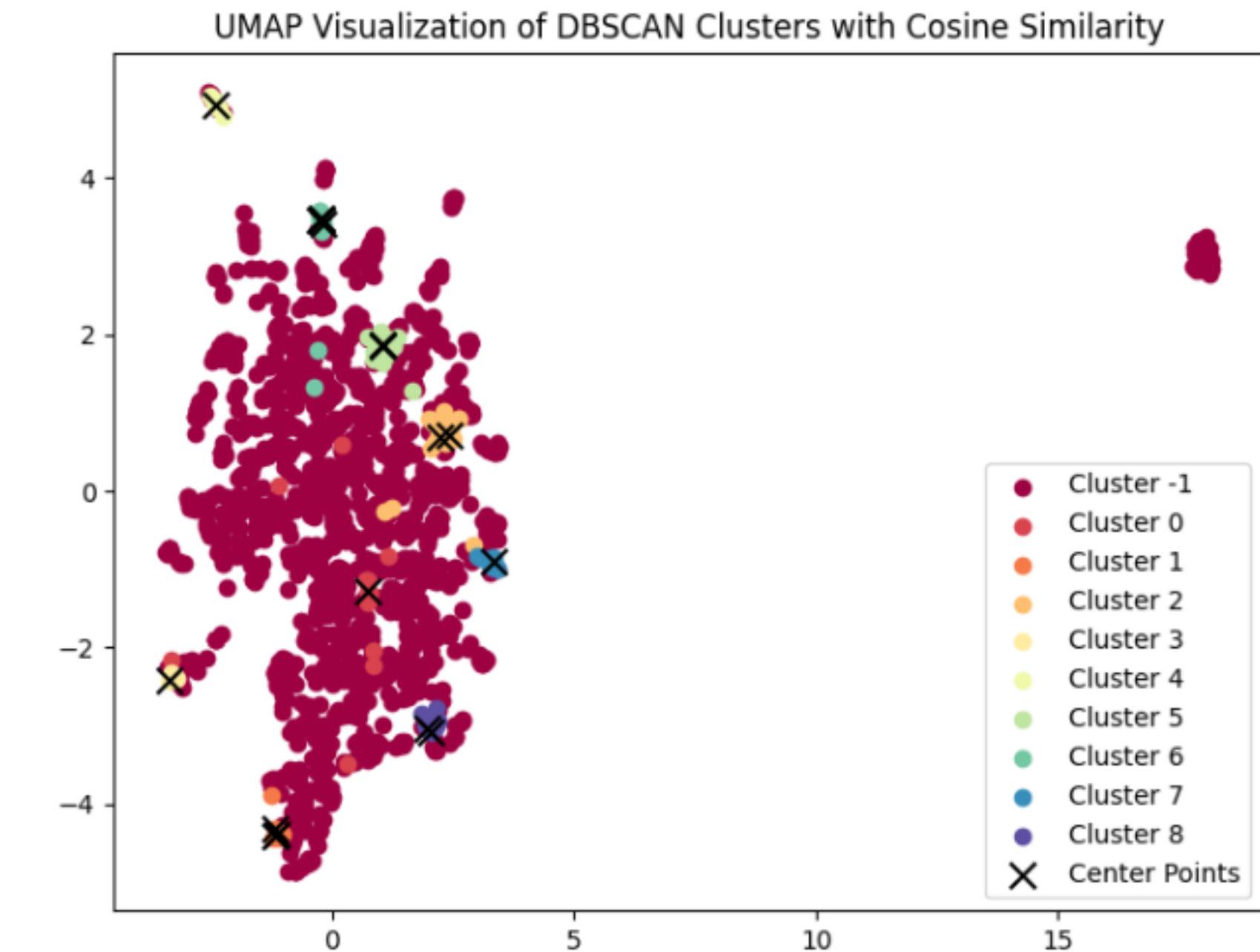
### 3. 대주제 탐색

#### a. DBSCAN으로 주제 군집화

## 기사 “제목”으로 DBSCAN

- \*  $\text{eps} = 0.6$ ,  $\text{min\_samples} = 12$
- \*  $\text{eps}$  = 두 데이터 포인트가 이웃으로  
간주되기 위한 최대 거리
- \*  $\text{min\_samples}$  = 하나의 클러스터를  
형성하는 최소 데이터 포인트 수

$\text{eps}$ 가 낮을수록 군집이 밀접하게 형성되고  
 $\text{min\_samples}$ 가 높을수록 큰 군집이 형성된다.



```
model.core_sample_indices_
array([ 114,  363,  376,  453,  469,  639,  645,  759,  950, 1100, 1177,
       1313, 1347, 1356, 1363, 1440, 1494])
```

-1의 경우 군집에 속하지 못한 데이터들을 모아둔 군집임 ← 군집의 개수 = 7개

### 3. 대주제 탐색

#### a. DBSCAN으로 주제 군집화

## 기사별 “키워드”로 DBSCAN

\* **eps = 0.7, min\_samples = 14**

대부분의 군집이

제목으로 군집화한 결과와 비슷했지만

유일하게 군집 8번만 다른 주제

#### 군집의 적절성 판단

1. 군집내의 키워드들 중 많이 반복된 것이 있는지
2. 군집 내의 기사들이 일관성 있는지

```
import matplotlib.pyplot as plt
import umap.umap_
from sklearn.cluster import DBSCAN

# DBSCAN Clustering
from sklearn.cluster import DBSCAN

# DBSCAN 모델의 파라미터 조정
model = DBSCAN(eps=0.6, min_samples=12, metric="cosine")
result = model.fit_predict(vector)
df['result'] = result

# UMAP을 사용하여 데이터를 2차원으로 변환
umap_model = umap.UMAP(n_neighbors=15, min_dist=0.1, metric='cosine')
umap_vector = umap_model.fit_transform(vector)

# 중심 포인트 가져오기
center_points = umap_vector[model.core_sample_indices_]

# 시각화
plt.figure(figsize=(8, 6))

unique_labels = np.unique(result)
colors = plt.cm.Spectral(np.linspace(0, 1, len(unique_labels)))

for i, label in enumerate(unique_labels):
    cluster_points = umap_vector[result == label]
    plt.scatter(cluster_points[:, 0], cluster_points[:, 1], color=colors[i], label=f'Cluster {label}')

# 중심 포인트 표시
plt.scatter(center_points[:, 0], center_points[:, 1], color='black', marker='x', s=100, label='Center Points')

plt.title('UMAP Visualization of DBSCAN Clusters with Cosine Similarity')
plt.legend()
plt.show()
```

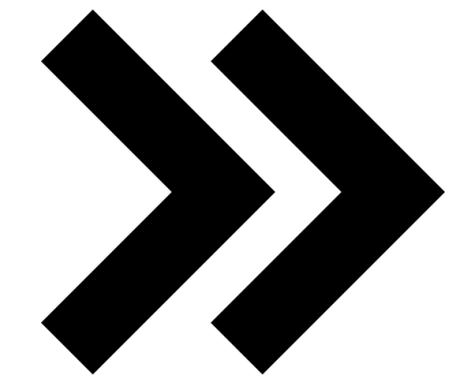
### 3. 대주제 탐색

#### a. DBSCAN으로 주제 군집화

8(7+1)개의 최종 군집 결정

#### 군집 예시

cluster num : 1  
한은 금통위 긴급 예측 기준금리 4연속 동결 유력 정철진 경제평론가 머니 클래스  
한은 기준금리 4연속 동결 2퍼센트p 금리차 부담 속 경기 집중 상보  
한은, 3.5퍼센트 기준금리 4연속 동결 한 미 금리차 2퍼센트p 임박 상보  
금통위, 금융 불안에 기준금리 4연속 동결 상보  
새마을금고도 불안한데 한은, 기준금리 3.5퍼센트 4연속 동결 상보  
금리차 2퍼센트P 임박 한은, 미국과 다른 길 택했다  
기준금리 4연속 동결 경기 회복 지켜본다 종합  
한은, 기준금리 4연속 동결 미국과 초유의 2퍼센트P 차이 눈앞  
한은 기준금리 4연속 동결, 미달 한미금리차 2퍼센트p 확대 전망  
한은 기준금리 4연속 동결에도 꿈틀대는 대출 금리  
한은, 기준금리 3.5퍼센트 4연속 동결 미와 사상최대 2퍼센트p 차 벌어지나  
한은, 기준금리 4연속 동결 한 미 금리차 역대 최대 눈앞 뉴스 투데이  
기준금리 4연속 동결 한은, 3.5퍼센트 유지



동일한 주제의 기사가 군집화

### 3. 대주제 탐색

#### b. KeyBERT로 대주제 생성

- **목표:** 군집 개수만큼 총 8개의 대주제와 기사 요약문 생성
- **대주제 생성:** KeyBERT

#### KeyBERT

- KeyBERT: BERT를 이용한 키워드 추출 방식
- 임베딩 방식: paraphrase-multilingual-MiniLM-L12-v2 모델  
(50개 이상의 언어로 학습된 다국어 SBERT)
- 입력값: 군집 내 기사 제목을 합친 하나의 텍스트
- KeyBERT가 계산한 중요도 기준 3개 키워드 채택

```
1 ## title로 한단어~두단어 키워드 출력
2
3 for i in range(1,10):
4     df = df[df['cluster_num']==i]['title']
5     df = ' '.join(df)
6     morphemes = okt.pos(df)
7     filtered_words = [word for word, pos in morphemes
8                        if pos in ['Noun', 'Number', 'Alpha', 'Suffix']]
9     filtered_words = ' '.join(filtered_words)
10    kw_model = KeyBERT('paraphrase-multilingual-MiniLM-L12-v2')
11    keywords = kw_model.extract_keywords(filtered_words,
12                                         keyphrase_ngram_range=(1,2),
13                                         use_maxsum=True,
14                                         nr_candidates = 20,
15                                         top_n = 3,
16                                         use_mmr = True,
17                                         diversity=0.7)
18    print(keywords)
19    result = []
20    result.append(keywords)
```

```
[('금리 역대', 0.7091), ('미국 다른', 0.4164), ('회복', 0.1713)]
[('반도체 교실', 0.6628), ('대한민국 저력', 0.4274), ('회복 1년', 0.2534)]
[('회장 2030년', 0.711), ('매출 성장', 0.3287), ('열풍', 0.13)]
[('당첨 음모론', 0.6343), ('조작 정부', 0.4512), ('2등 600', 0.28)]
[('총장 ai', 0.5314), ('cfo lounge', 0.2488), ('교통사고', 0.0678)]
[('출시 삼성', 0.6766), ('vs lg', 0.393), ('tv 경쟁', 0.318)]
[('최저임금 10620원', 0.7455), ('선거 보조', 0.2004), ('폭풍', 0.081)]
[('경제 불안', 0.64), ('adp 대비', 0.2572), ('발목 이착용', 0.0392)]
```

## 4. 주제별 기사 요약

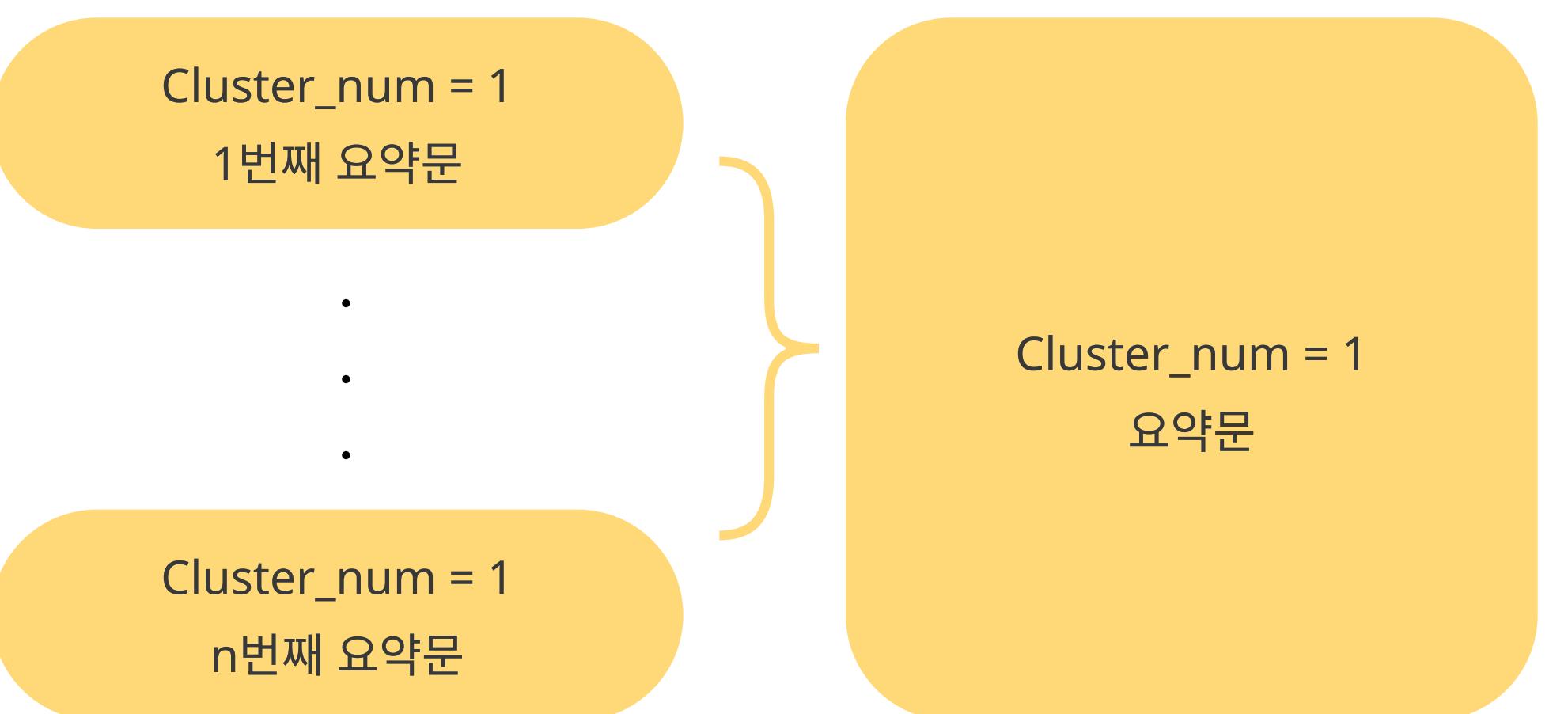
### a. KoBART로 기사 요약문 생성

- **목표:** 군집 개수만큼 총 8개의 대주제와 기사 요약문 생성
- **기사 요약문 생성:** KoBART

#### KoBART

- KoBART: 한국어 BART
- 각 군집에 대한 요약문 생성: 군집 내 기사 요약 후 합쳐서 다시 요약

\*BART: 입력 텍스트 일부에 노이즈를 추가하여 이를 다시 원문으로 복구하는 autoencoder의 형태로 학습



### 예시) 1번 클러스터

- 1번 클러스터에 해당하는 기사 요약문 하나로 합치기

```
1 dff = df[df['cluster_num']==1]
2 dfff = ' '.join(dff['summary'])
3 dfff
```

"J은행 금융통화위원회가 잠시 후 오전 9시, 회의를 열고 현재 3.50%인 기준금리를 조정할지 말지 결정하는데, 이번에도 동결하면 4번 연속 동결인데, 이렇게 될 것 같다는 전망이 우세하며, 이는 이미 역대 최대로 떨어진 한·미 금리 차이가 변수고, 정월진 경제평론가와 함께 오늘도 맥 한 번 짊어보겠습니다. 한일 금융통화위원회는 13일 오전 통화정책방향 결정회의에서 기준금리를 현행 연 3.50%로 유지해 지난 2월, 4월, 5월에 이은 4연속 동결했으며 한미 금리 격차가 이달 말 2%포인트(p)까지 벌어질 것으로 전망되나 2%대로 낮아진 물가 오름세와 아직은 확신할 수 없는 하반기 경기 회복을 고려한 결정으로 풀이된다. 한국은행 금융통화위원회가 13일 통화정책방향 결정회의를 열고 기준금리를 3.50%로 동결한 가운데 올 2월과 4월, 5월에 이어 이달까지 4회 연속 금리 동결로 한은의 금리 인상 사이클이 사실상 마무리되었다는 '인상 풍물론'이 확실시 되고 있으며 미 연방준비제도(연준·Fed)는 오는 25~26일(현지 시각) 열리는 연방공개시장위원회(FOMC) 회의에서 금리 인상을 재개할 것이 우력시되고 시카고상품거래소(CME) 페드워치에 따르면 이날 오전 9시 기준 연방기금금리(FFR) 선물 시장에서 연준이 이달 베이비스텝(금리 13일 한국은행은 7월 금융통화위원회를 열고 새마을금고 사태 등에 따른 금융시장 불안을 이유로 기준금리를 4연속 동결했지만 '매파적(긴축 선호)' 스탠스는 유지할 것으로 보이며 다만 4연속 동결에도 '매파적' 스탠스는 유지할 것으로 보인다. 한국은행 금융통화위원회가 지난 2월, 4월, 5월에 이어 13일부터 기준금리를 연 3.5%로 동결했는데 물가 경로가 한은 예상대로 흘러가는 가운데 물가가 오름세와 함께 금리를 올려 경기에 찬물을 끼얹을 이유가 없다고 판단한 것으로 해석되며 새마을금고 부실 논란도 고려한 것으로 분석된다. 한은행이 13일 기준금리를 4연속 동결하면서 한미 금리차가 더 커질 것을 우려하며 외국인 자금 유출과 원화 가치 하락 압력이 커질 것으로 보아, 한국 경제는 수출과 내수 회복 지연으로 정부나 한은이 기대하는 하반기 경기 반등이 불투명한 상태이다. 한국마을 금고 사태의 영향으로 물가는 잡히는 듯 하지만 여전히 우려가 있고 경기 회복은 길 길이 멀어, 한국은행이 기준금리를 4연속 동결했으나, 새마을 금고 사태도 금융 시장 불안으로 안심할 수 없어, 한은은 '매파적(긴축 선호)' 스탠스를 유지할 것으로 보인다. 13일 한국은행 금융통화위원회가 기준금리를 현 3.50%로 동결하여 한·미 기준금리 역전 폭은 이달 말 미 연방준비제도(Fed)가 예상대로 금리를 인상하면 최대 기록을 또다시 경신하면서 사상 최초로 2%포인트대에 진입할 가능성이 높아졌으며, 미 Fed가 이달 말 예상대로 정책금리(기준금리)를 0.25%포인트 더 올리면 한·미 금리 차가 사상 최초인 2%포인트로 떨어지고, 외국인 자금 유출과 원화 가치 하락 압력이 더욱 커질 것으로 우려된다. 한국은행이 13일 지난 2, 4, 5월에 이어 4회 연속으로 기준금리를 3.50%로 동결했는데 미국이 오는 25~26일 연방공개시장위원회(FOMC)에서 정책금리를 0.25% 인상할 가능성이 커 이달 한미금리차가 2.00%로 확대될 전망이다. 13일 기준금리를 네 차례 연속 동결했지만 지난달 28일 변동형 금리가 연 4.21~6.12%에 분포하던 것과 비교해 상단금리가 더 뛰었고 같은 기간 연 4.00~5.81%에서 4.06~6.00%를 나타내며 상단금리가 6%대로 올라서는 등 시중은행의 대출금리는 오히려 상승세를 그리고 있다. 인플레이션이 우려한 둔화세를 이어가고, 경기 침체 우려가 커짐에 따라 한국은행이 13일 기준금리를 연 3.5%로 동결했고 이달 말 미국의 기준금리 인상 가능성은 높기 때문에 한미 금리 역전 폭은 2%포인트로 확대될 것으로 예측된다. 한국은 목표 수준보다 높은 물가 추이와 경기 위축 우려 등을 고려해 4연속 금리 동결로 한국과 미국의 기준금리 격차는 1.75%포인트(미 상단 기준)를 유지, 미국 연방준비제도(Fed·연준)가 이달 말 예상대로 정책금리(기준금리) 0.25%포인트 인상에 나설 경우, 한·미 금리차는 2.00%포인트까지 벌어져 역대 최대치를 경신할 전망이다. 한국은행이 13일 기준금리를 연 3.50%로 동결하며 올 2월, 4월, 5월에 이어 4연속 기준금리 유지 결정을 내렸고, 미국 연방준비제도가 이달 추가 금리 인상에 나설 것으로 예상되는 가운데 한미 간 금리 차이가 처음으로 2%포인트까지 벌어질 가능성성이 높아졌다."

- 전체 텍스트로 요약문 생성하기

```
12 summary_text_ids = model.generate(
13     input_ids = inputs_ids,
14     bos_token_id = model.config.bos_token_id,
15     eos_token_id = model.config.eos_token_id,
16     length_penalty = 2.0,
17     max_length = 200,
18     min_length = 32,
19     num_beams = 4)
20
21 tokenizer.decode(summary_text_ids[0], skip_special_tokens=True)
```

'한은행이 13일 기준금리를 4연속 동결하면서 한미 금리차가 더 커질 것을 우려하며 외국인 자금 유출과 원화 가치 하락 압력이 커질 것으로 보아, 한국 경제는 수출과 내수 회복 지연으로 정부나 한은이 기대하는 하반기 경기 반등이 불투명한 상태이며 한미 금리차가 2%포인트 까지 벌어질 것으로 전망되나 2%대로 낮아진 물가 오름세와 아직은 확신할 수 없는 하반기 경기 회복을 고려한 결정으로 풀이된다.'

## 4. 주제별 기사 요약

### a. KoBART로 기사 요약문 생성

Cluster #

Kobart 요약문

Cluster 1

한은행이 13일 기준금리를 4연속 동결하면서 한미 금리차가 더 커질 것을 우려하며 외국인 자금 유출과 원화 가치 하락 압력이 커질 것으로 보아, 한국 경제는 수출과 내수 회복 지연으로 정부나 한은이 기대하는 하반기 경기 반등이 불투명한 상태이며 한미 금리차가 2%포인트 까지 벌어질 것으로 전망되나 2%대로 낮아진 물가 오름세와 아직은 확신할 수 없는 하반기 경기 회복을 고려한 결정으로 풀이된다.

Cluster 3

신임 2주년을 맞은 신동원 농심 회장이 최근 임직원에게 보낸 메세지 메시지에서 2025년 미국 제3공장을 착공하고 2030년까지라면 시장 1위를 달성하겠다는 목표를 제시하며 신 회장의 현장 경영으로 세계 100여 개국으로 수출하는 글로벌 식품기업으로 성장했다고 강조하며 오는 2030년 까지 미국 시장에서 지금의 3배 수준인 연 매출 15억 달러를 달성하고라면 시장 1위에 오르겠다며 오는 2030년까지 매출 15억 달러와 함께 미국라면 시장 1위 역전을 이뤄내겠다는 포부를 밝혔다.

Cluster 9

이창용 한국은행 총재는 13일 기자간담회에서 우리나라가 명목 국내총생산(GDP) 기준 우리나라 순위가 13위로 하락한 것과 관련해 "단기환율은 언제든 바뀔 수 있지만 중장기적으로 우리나라의 저출산·고령화 등 여러 구조조정을 미뤄서 경쟁력이 둔화되고 성장률이 낮아져 경제(규모)순위가 떨어지는 게 더 큰 문제(가 될 수 있다)"라고 우려했고, 소비자물가상승률이 지난달 2.7%에서 연말에는 3% 내외로 움직일 것이라고 예측한다며 향후 추가 금리 인상 가능성은 언급하고 물가 목표인 2%대로 충분히 수렴한다는 확신이 들 때 인하를 논의할 것이라고 밝혔다.

\*점수 높은 순

Keybert 생성 키워드

#금리 역대 #미국 다른 #회복

#회장 2030년 #매출 성장 #열풍

#경제 불안 #gdp 대비 #발목 이창용

DBSCAN 중심점 기사  
제목과 키워드

한은 기준금리 4연속 동결...'2%p 금리차' 부담 속 경기 집중(상보)  
#3.50퍼센트 #금리 #연속동결 #안정 #회복 #물가 #한미

신동원 농심 회장 취임 2주년..."2030년 美 라면시장 1위 목표"  
#1위 #농심 #라면 #미국 #시장 #성과 #미국시장 #15억

이창용 한은 총재 "기준금리 3.75% 인상 열어둬야...가계부채, 금리로 대응할수도"  
#가계대출 #GDP #명목GDP #13위로 #GDP가계부채비율 #환율 #구조개혁 #중장기적

: 같은 '한은' 키워드지만 논지가 달라  
다른 군집으로 분류

## 4. 주제별 기사 요약

### a. KoBART로 기사 요약문 생성

**Cluster #**

**Cluster 4**

**Cluster 7**

**Cluster 8**

**Kobart 요약문**

13일 기획재정부 복권위원회에 따르면 한국정보통신기술협회(TTA)는 로또복권에 대한 조작이 불가능하다는 것을 확인하고 최근 다수 당첨 역시 확률·통계적으로 충분히 발생 가능한 범위에 있다는 판단인데 이는 위·변조 행위를 방지하기 위한 다양한 장치가 마련돼 있어 조작이 불가능하고 신뢰성을 저해할 만한 위험 요소가 없다고 말했고 기획정부통신기술협회 검증 결과에 따르면 현 복권시스템과 추첨과정에는 내외부에서 시도할 수 있는 위·변조가 방지하기 위한 다양한 장치가 마련돼 있어 조작이 불가능하며, 로또복권의 신뢰성을 저해할 만한 위험요소가 없음을 확인했다"고 했다.

13일 TV 시그니처 올레드 M은 세계 최초로 4K·120Hz 무선 전송 기술을 더해 전원을 제외한 모든 선(線)을 없앤 무선 올레드 TV로 한국을 시작으로 북미, 유럽 등 글로벌 주요 시장에 순차 출시, TV전자와 LG전자가 나란히 TV 시장이 침체를 겪는 가운데 프리미엄 TV로 승부수를 띠 우며 나란히 90형대 프리미엄 TV 신제품을 선보였는데 삼성전자는 거거익선 트렌드에 따라 98형 네오(Neo) QLED 8K 신모델을 국내 출시하고, LG전자는 세계 최초로 무선 올레드(OLED·유기발광다이오드) TV인 'LG 시그니처 올레드 M'을 출시했다.

최저임금위원회는 13일 정부세종청사에서 제13차 전원회의를 열고 논의할 예정으로 최저 임금이 1만원을 넘어설 수 있을지가 관건인 가운데 공익위원안을 통한 표결 가능성도 나오고 있어 오늘밤 늦게 결정될 것으로 예상되며, 다만 이번에는 노동계에서 정부 개입 최소화를 요구했고, 이에 공익위원 측에서는 최대한 개입을 자제하는 분위기가 강한것으로 전해졌다.

**Keybert 생성 키워드**

#당첨 음모론 #조작 정부 #2등

#출시 삼성 #vs lg #경쟁

#최저임금 10620원 #선거 보조 #폭풍

**DBSCAN 중심점 기사 제목과 키워드**

"로또 조작 불가능...통계적으로 1·2등 다수 당첨 가능"  
#추첨과정 #접근제어 #복권 #추첨 #당첨번호 #연구소 #통계적 #접근

삼성·LG, 4000만원대 TV 동시출격 '프리미엄 대격돌'  
#신제품 #네오 #예약 #퀀텀 #초대형 #무선 #8k #적용 #시장 #시청

노동계 "1만1040원" 경영계 "9755원" 최저임금 5차 수정안  
#노사 #중재 #공익위원 #최저임금위원장 #최대한 #제시 #최저임금위 #115원 #심의촉진구간

## 4. 주제별 기사 요약

### a. KoBART로 기사 요약문 생성

Cluster #

Cluster 2

Cluster 6

Kobart 요약문

최태원 대한상공회의소 회장은 12일 제주 해비치호텔&리조트에서 열린 제46회 대한상의 제주포럼 계기에 진행한 기자간담회에서 "반도체 업다운 사이클이 빨라진 데 이어 진폭 자체가 커지는 문제점에 봉착하고 있다"고 우려하며 업사이클로 올라가는 흐름으로 회복 시점으로는 6개월 1년 뒤라고 내다봤고 유튜브 채널과 강남구청 인터넷 수능방송 사이트에 온라인 반도체 특강 시리즈를 올려 누적 조회수 57만 뷰를 기록한 SK하이닉스는 올해도 전국 17개 고등학교를 방문해 찾아가는 반도체 교실'을 진행한다고 13일 밝혔는데 이번에 두 번째 프로그램인 오프라인 특강을 통해 미래 반도체 산업을 이끌어갈 고등학생들에게 새로운 경험을 제공한다.

이광형 카이스트 총장은 13일 제주 해비치호텔에서 열린 대한상의 제주포럼에서, 10년 후, 50년 후를 내다보지 않고 기술만 개발해선 안된다고 말하며 AI 시대에 자손이 떳떳하게 살려면 한국형 AI를 만들어야 하고 만약 독도가 일본에 의해 침략당한다고 했을 때 혼란이 있을 수 있도록 AI는 앞으로 국가를 지키는 기반이라고 거듭 한국형 AI 당위성을 설파했다.

Keybert 생성 키워드

#반도체 교실 #대한민국 저력 #회복 1년

#총장 ai #cfo lounge #교통사고

DBSCAN 중심점 기사 제목과 키워드

삼성전자, 美브로드컴에 시가총액 잡혀...반도체 4위로  
#연중 #브로드컴 #순위 #반도체시가총액순위 #비메모리 #메모리  
#팹리스 #3673억달러 #주가 #반도체시가총액

"잃어버린 30년 되찾겠다"…日, 반도체 재건에 사활  
#공급망 #반도체기업 #일본정부 #라피더스 #개발 #반도체산업  
#tsmc #투자 #장비

"자동차가 갓나와 교통사고 났을 때..." 빌게이츠의 'AI 낙관론'  
#변화 #등장 #딥페이크 #허위정보 #현실 #블로그 #허위 #컴퓨터 #게이츠

탈통신 SKT...신무기는 'AI·UAM'  
#데이터 #SK텔레콤 #돌파구 #가입자 #비교  
AI 챗봇 뒤 팩트검증 노동자들의 그림자..."과로·낮은 급여 시달려"  
#바드 #구글 #해고 #AI챗봇 #답변 #외주 #업무  
척 보면 알아요, 맛있는 거봉 쑥쑥 골라내는 AI  
#거봉 #비파괴 #산도 #자동 #선별 #롯데마트

: 포괄적인 분야 혹은 키워드만 일치



한입에 떠먹여드려요!

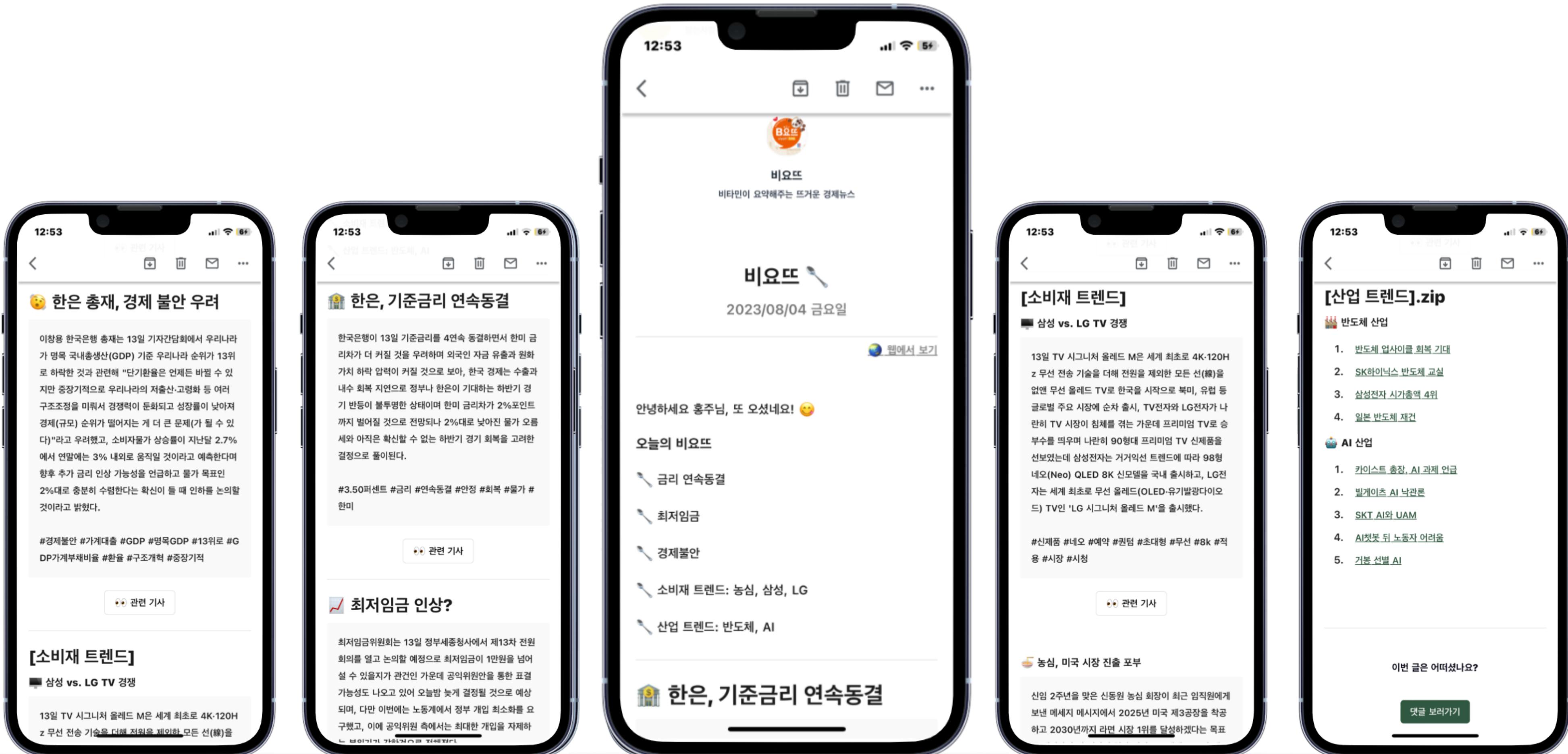
비  
요  
뜨

타민이  
약해주는  
거운 경제뉴스



# [Part 3] 프로젝트 결과

## a. 결과물



## [Part 3] 프로젝트 결과

### b. 의의 및 한계점

#### 의의

- 바쁜 현대인들에게 쉽고 금방 읽을 수 있는 경제 뉴스레터 제공
- 정보 습득 시간 단축 -> 중복되는 기사 대신에 핵심 주제에 대한 내용을 빠르게 접함

#### 한계점

- 전처리 시, 한국어 데이터 처리 보완 필요
  - 보다 정확한 키워드 추출을 위해 개체명 유무 및 이형동의어와 대용어 사전 활용하고자 하였으나 데이터 부족으로 활용하지 못한 부분이 아쉬움
  - 추출한 키워드 간 의미의 유사성 고려 못 함

NLP 키워드 추출 기반  
**경제 트렌드 뉴스레터**

6조 윤희재 임홍주 조은정

SUBSCRIBE



# End Of Document