캡스톤 디자인(2) 최종 발표

Web Crawling, NLP를 활용한
Word Cloud 기반 기사 통합 요약 시스템

목차

프로젝트 소개

프로젝트 개발 동기

프로젝트 개발 동기

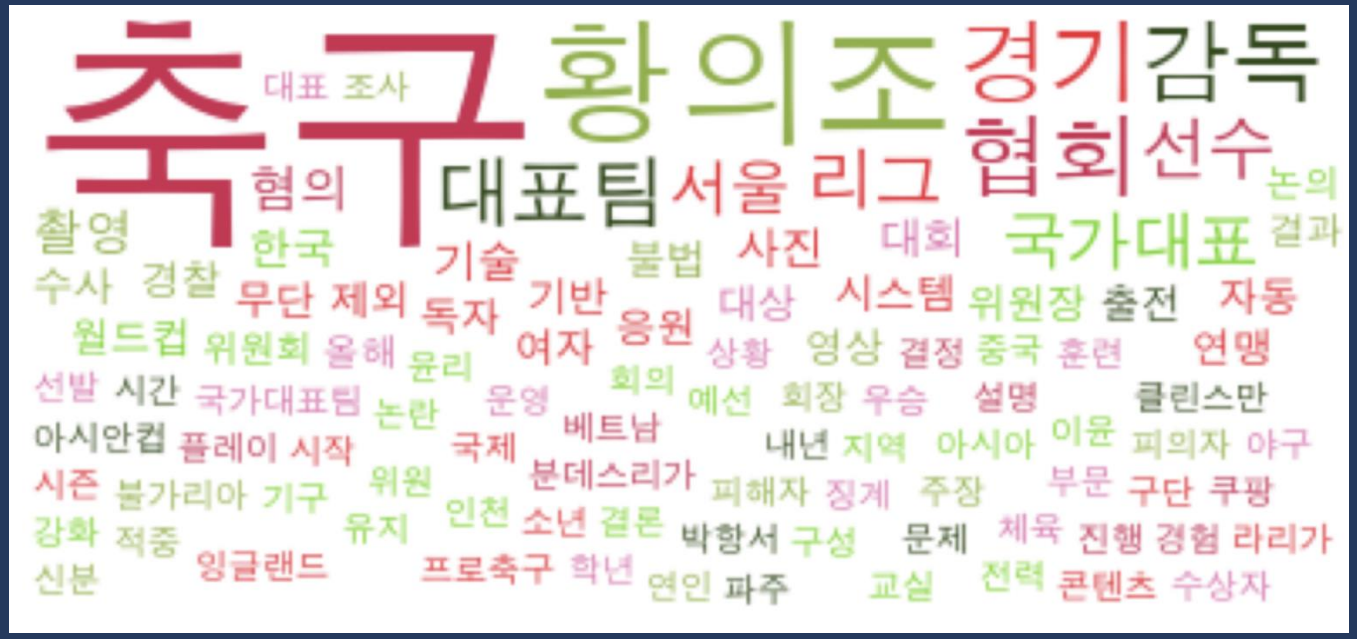

N 뉴스 | 연예 | 스포츠 | 날씨 | 프리미엄
로그인
언론사별 정치 경제 사회 생활/문화 IT/과학 세계 랭킹 신문보기 오피니언 TV 팩트체크 알고리
11.30(목)
언론사편집 기자 연재 구독설정 >
조선일보 11월 30일 23:19 + 구독
"뼈 구조 보니 내 딸" 가수 오유진 스토킹한 60대 재판행
14세 트로트 가수 오유진과 가족을 지속적으로 스토킹한 60대 남성...
한국경제TV 11월 30일 22:49 + 구독
엑스포 어쩌지...늑장 공사에 결국
2025년 오사카·간사이 만국박람회(이하 오사카 엑스포) 개막 500...
비즈워치 11월 30일 21:18 + 구독
방송뉴스
LIVE YTN 뉴스나이트 방송보기 >
LIVE 24시간 보도전문채널 연합뉴스TV 방송보기 >
LIVE KBS N 방송보
언론사별 심층기획
더팩트 사진이 더 팩트다! 더팩트 11월 30일 17:35 + 구독
이종석 헌법재판소장 임명동의안 '국회 본...
국회의장실 앞 연좌 농성하는 국민의힘 [TF...
인사혁 국교사
축구
통합검색
축구
축구 국가대표 평가전
축구 일정
축구 국가대표 일정
축구중계
축구 국가대표
축구화
축구 아시안게임
축구공
축구경기
지역지

프로젝트 개요

프로젝트 개요

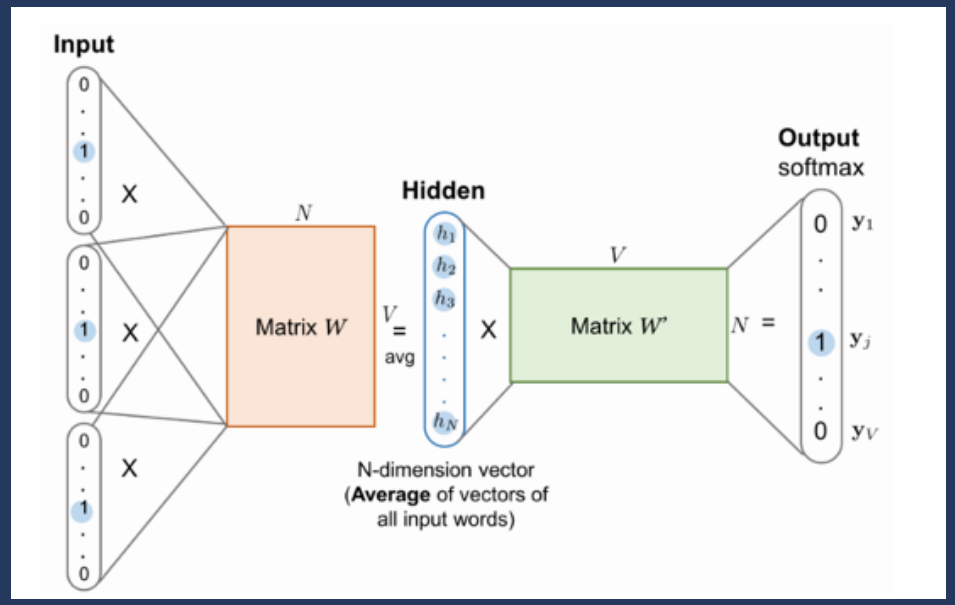

축구
황의조
협회
감독
국가대표
경기
선수
리그
대표팀
서울
한국
혐의
불법
사진
위원장
대회
기술
출전
올해
제외
경찰
무단
연맹
기반
결과
시스템
자동
대상
독자
윤리
월드컵
영상
응원
결정
조사
논의
수사
훈련
회의
클린스만
위원회
대표
시간

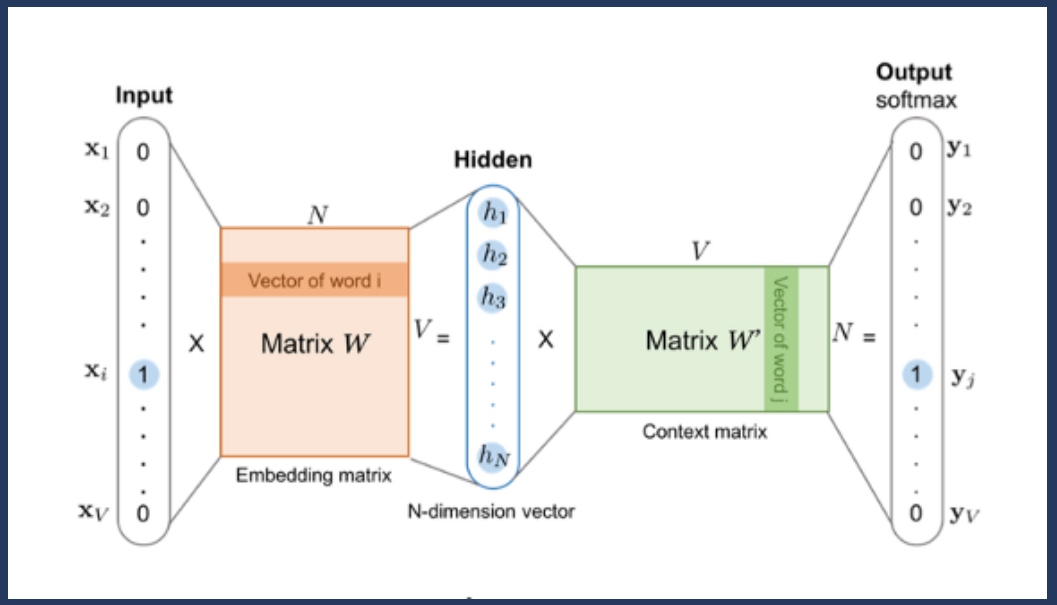
프로젝트 개요


축구
황의조
경기
감독
대표 조사
협회
선수
혐의
대표팀
서울
리그
논의
촬영
한국
불법
사진
대회
국가대표
결과
기술
수사
경찰
무단
제외
독자
기반
대상
시스템
위원장
출전
자동
응원
원드컵
위원회
올해
여자
상황
영상
결정
중국
훈련
연맹
윤리
회의
선발
시간
국가대표팀
논란
운영
예선
회장
우승
설명
클린스만
베트남
아시안컵
플레이
시작
국제
내년
지역
아시아
이윤
피의자
야구
분데스리가
시즌
불가리아
기구
위원
피해자
징계
주장
부문
구단
쿠팡
인천
소년
결론
유지
박항서
구성
문제
체육
진행
경험
라리가
강화
적중
잉글랜드
프로축구
학년
연인
파주
교실
전력
콘텐츠
수상자
신분

프로젝트 배경 : KoNLPy

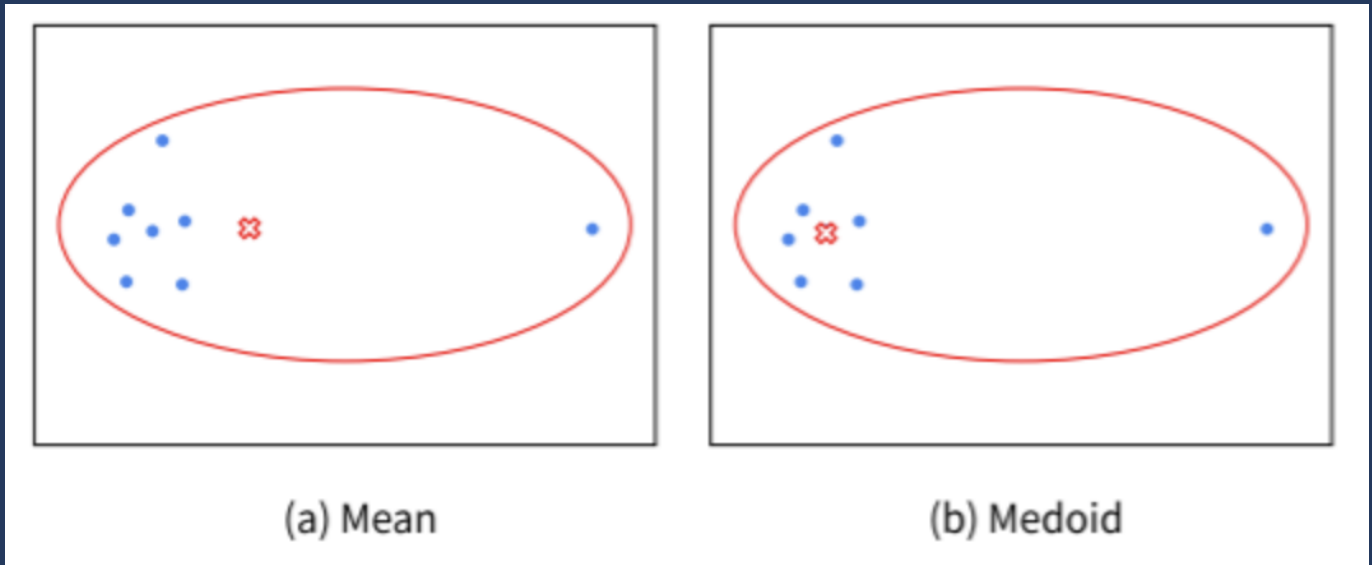
프로젝트 배경 : Word2Vec



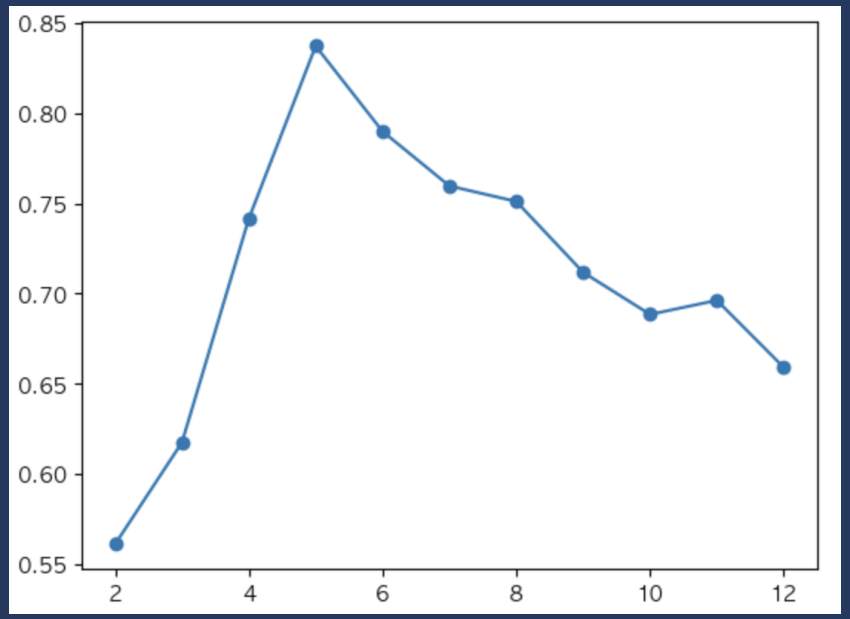


프로젝트 배경 : t-SNE

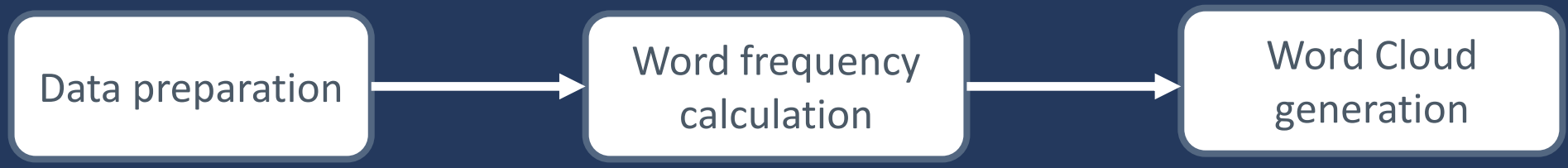
프로젝트 배경 : K-medoids



프로젝트 배경 : K-medoid – Silhouette method

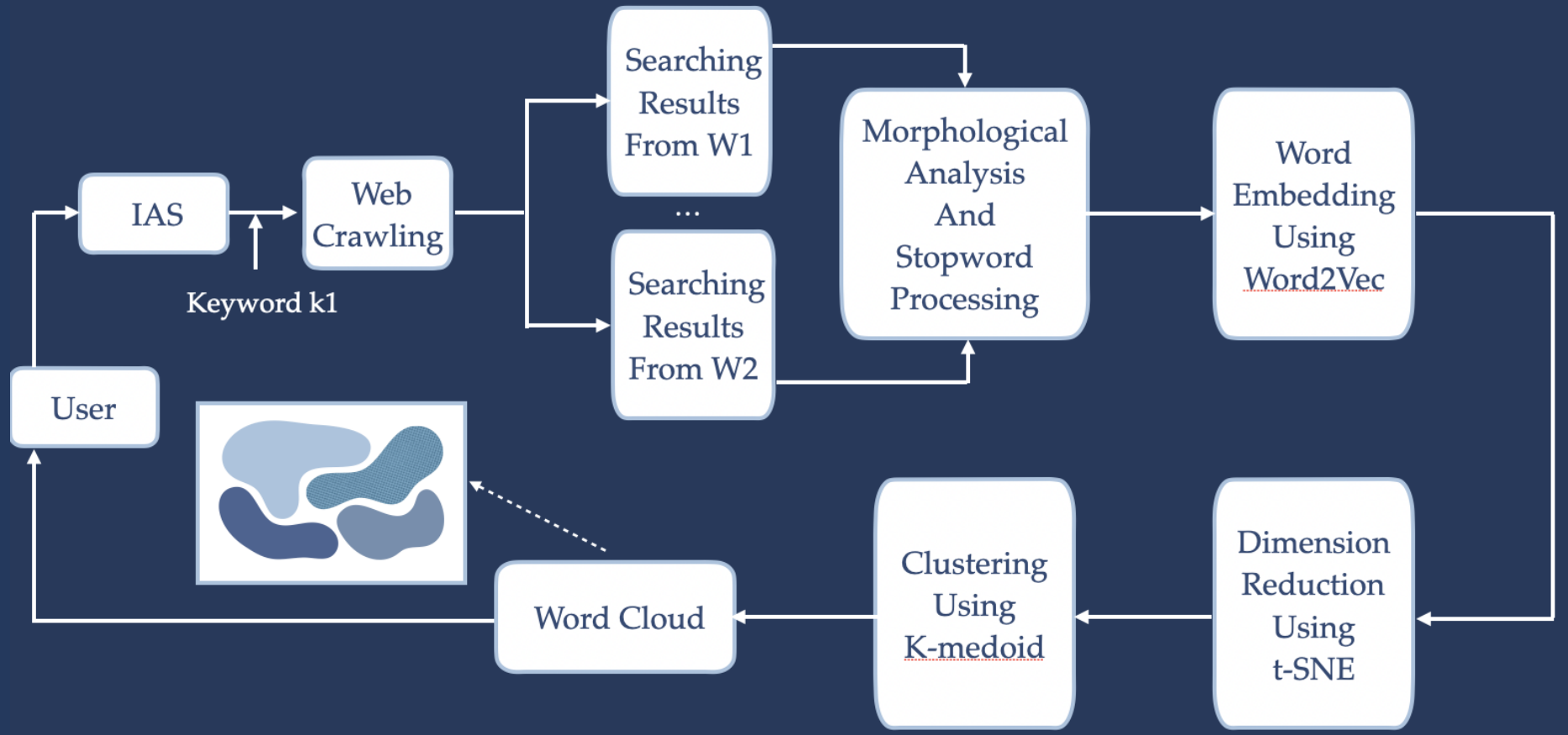


프로젝트 배경 : Word Cloud



설계 및 구현

Software Architecture



Dataset: Web Crawling

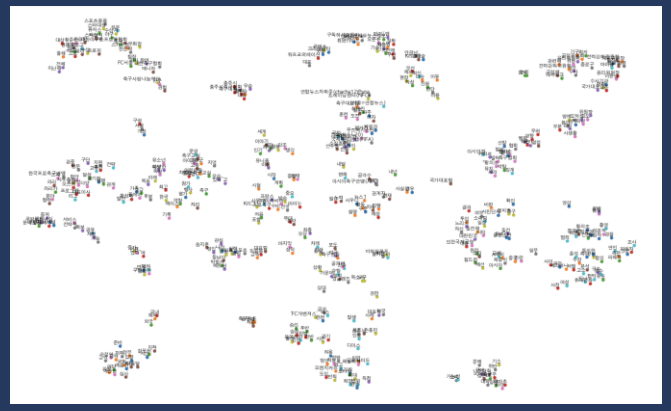
	date	link	content
0	2023.11.28.	https://n.news.naver.com/mnews/article/001/001...	[축구 국가대표 황의조[연합뉴스 자료사진](서울=연합뉴스) 안홍석 기자 = 축구 국...
1	2023.11.28.	https://n.news.naver.com/mnews/article/366/000...	[대한민국 축구 국가대표팀 황의조 /뉴스1 축구 국가대표 황의조(노리치...
2	2023.11.28.	https://n.news.naver.com/mnews/article/052/000...	[©연합뉴스불법 촬영 혐의로 경찰 수사를 받는 축구 국가대표 황의조(31·노리치시티...
3	2023.11.28.	https://n.news.naver.com/mnews/article/020/000...	[대한민국 축구 국가대표팀 황의조가 19일 오전 2026 FIFA 북중미 월드컵 아...
4	2023.11.28.	https://sports.news.naver.com/news.nhn?oid=421...	[(서울=뉴스1) 김성진 기자 = 이윤남 대한축구협회 윤리위원장이 28일 오후 서울...

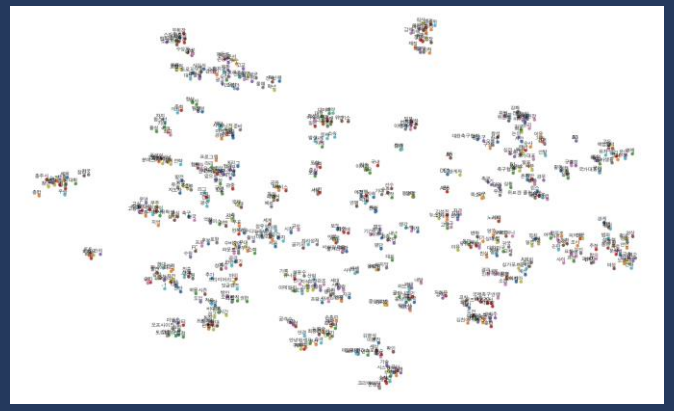
Preprocessing : 형태소 분석 및 불용어 처리

Preprocessing : Word2Vec

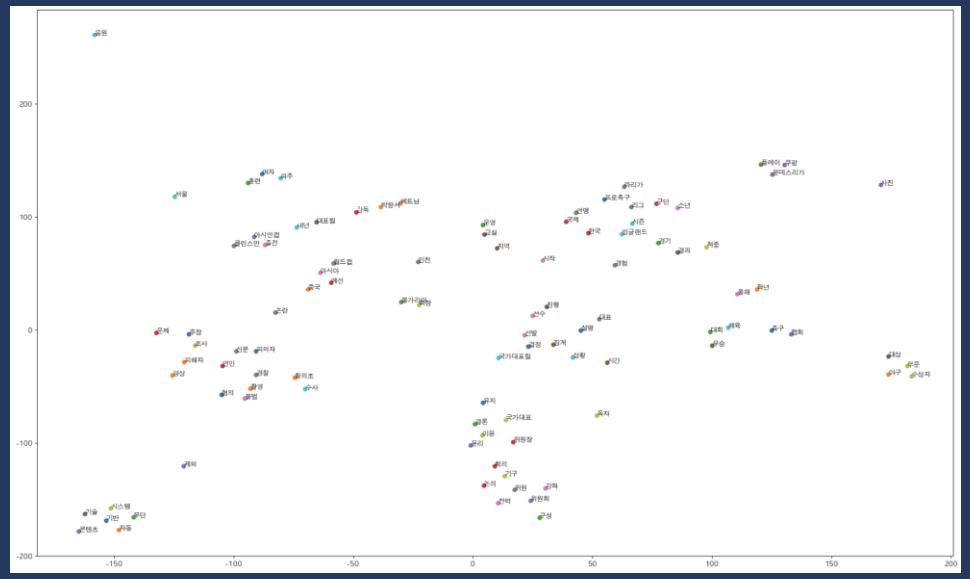
```
model_kkma = word2vec.Word2Vec(sentences=preprocess_kkma, vector_size=300, window=10, min_count=5, sg=1)
model_okt = word2vec.Word2Vec(sentences=preprocess_okt, vector_size=300, window=10, min_count=5, sg=1)
model_hannanum = word2vec.Word2Vec(sentences=preprocess_hannanum, vector_size=300, window=10, min_count=5, sg=1)
model_komoran = word2vec.Word2Vec(sentences=preprocess_komoran, vector_size=300, window=10, min_count=5, sg=1)
```

Preprocessing : t-SNE 기반 차원 축소





Preprocessing : t-SNE 기반 차원 축소 및 빈도 수 포함



Preprocessing : t-SNE 기반 차원 축소 및 빈도 수 포함

```
print(labels_okt[30])
```

여자

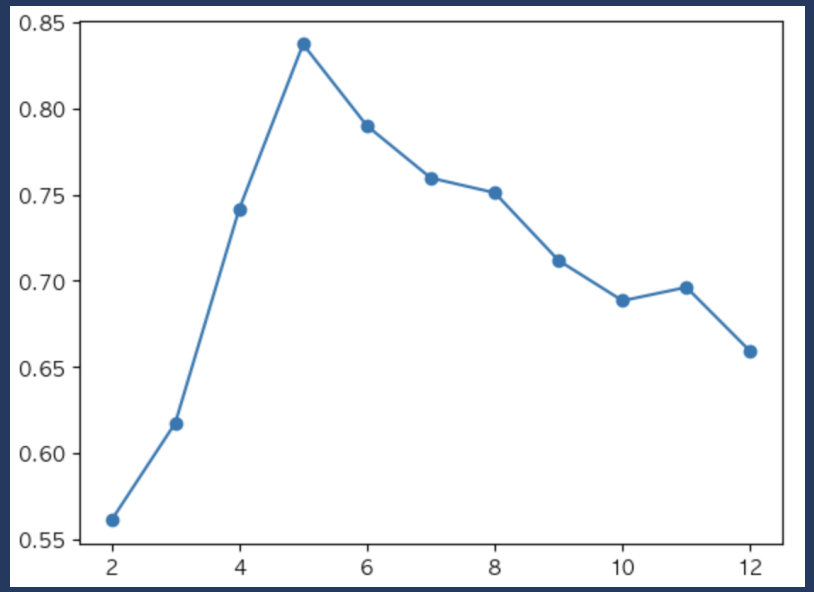
```
print(labels_okt[-1])
```

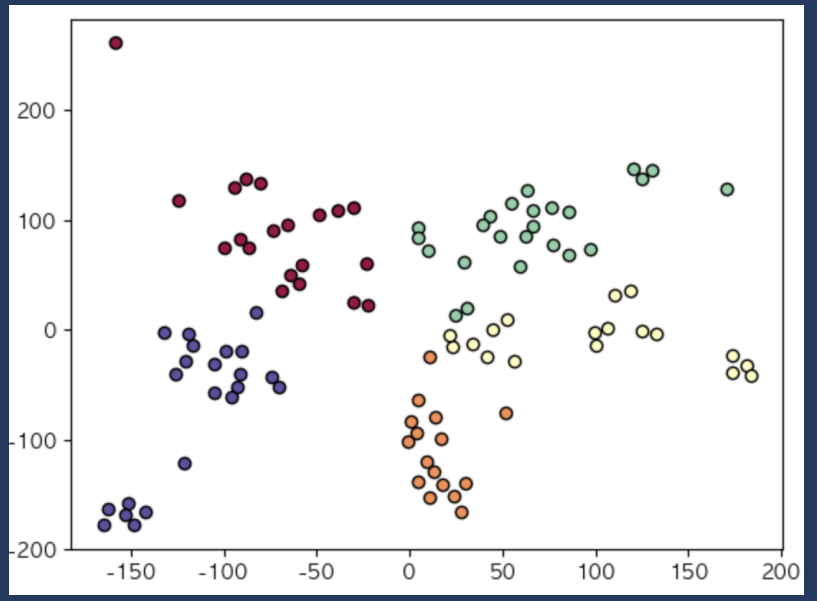
파주

```
print(cos_sim(values_okt[30], values_okt[-1]))
```

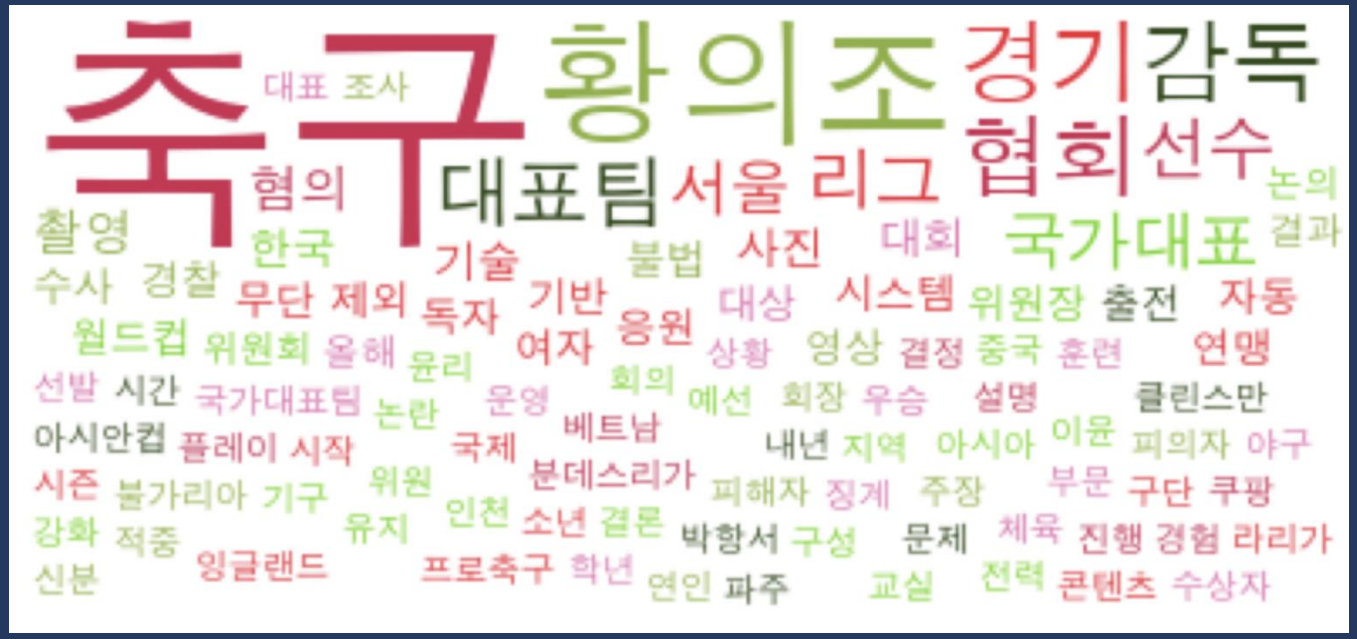
```
0.99949986
```

Preprocessing : K-medoids



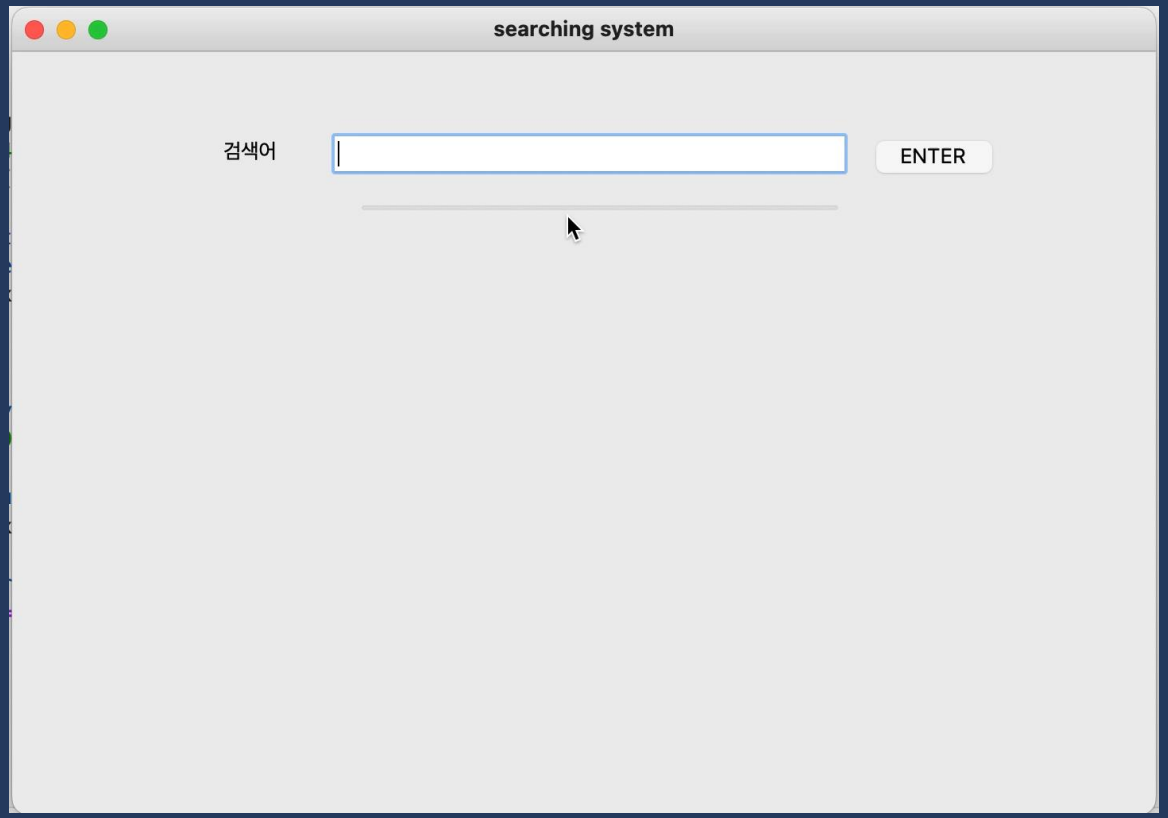


Word Cloud : 클러스터링 기반 색상 및 재배치



실행 결과

시연 결과



향후 확장 가능성

프로젝트 의미와 향후 확장 가능성

- 정보 습득의 효율 상승
- Word Cloud의 시각화 기능 향상
- Sematic Cloud의 대중화
- 이후 병렬화 도입하여 사용 가능성 향상 예정

참고 문헌

참고 문헌

- ‘Natural Language Processing with Transformers; Building Language Applications with Hugging Face’, Lewis Tunstall, Leandro von Werra & Thomas Wolf, O’Reilly
- ‘Word Cloud Explorer: Text Analytics Based on Word Clouds’, Heimerl, Florian. 2014 47th Hawaii International Conference on System Sciences ISBN
- ‘Semantic Word Cloud Generation Based on Word Embeddings’, Jin Xu, Yubo Tao, Hai Lin
- ‘Context Preserving Dynamic Word Cloud Visualization’, Weiwei Cui, Yingcai Wu, Shixia Liu, Furu Wei, Michelle X.Zhou