


ARTIFICIAL INTELLIGENCE PROJECT

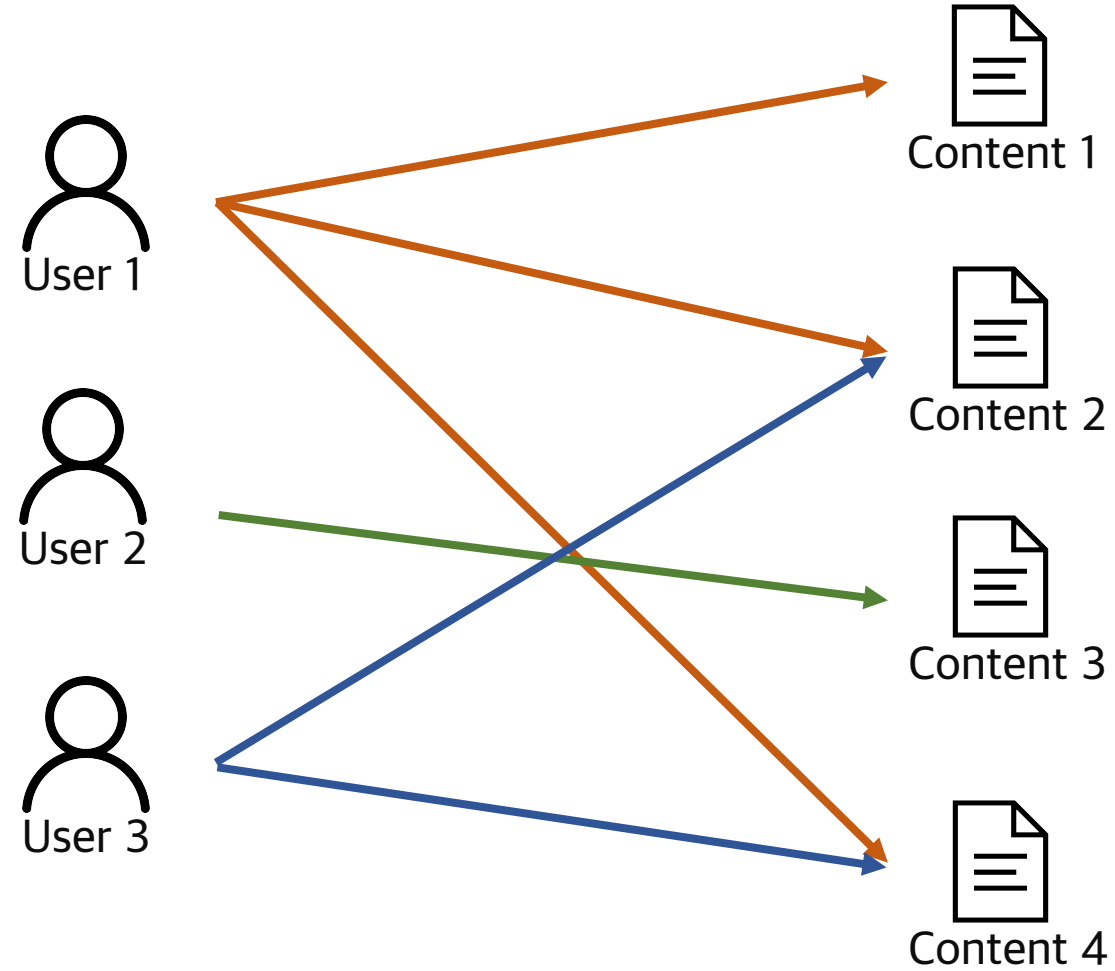
Measure Text Similarities and Clustering for News Recommendation

B211044 김연태
B211095 박현호
B211192 임종완

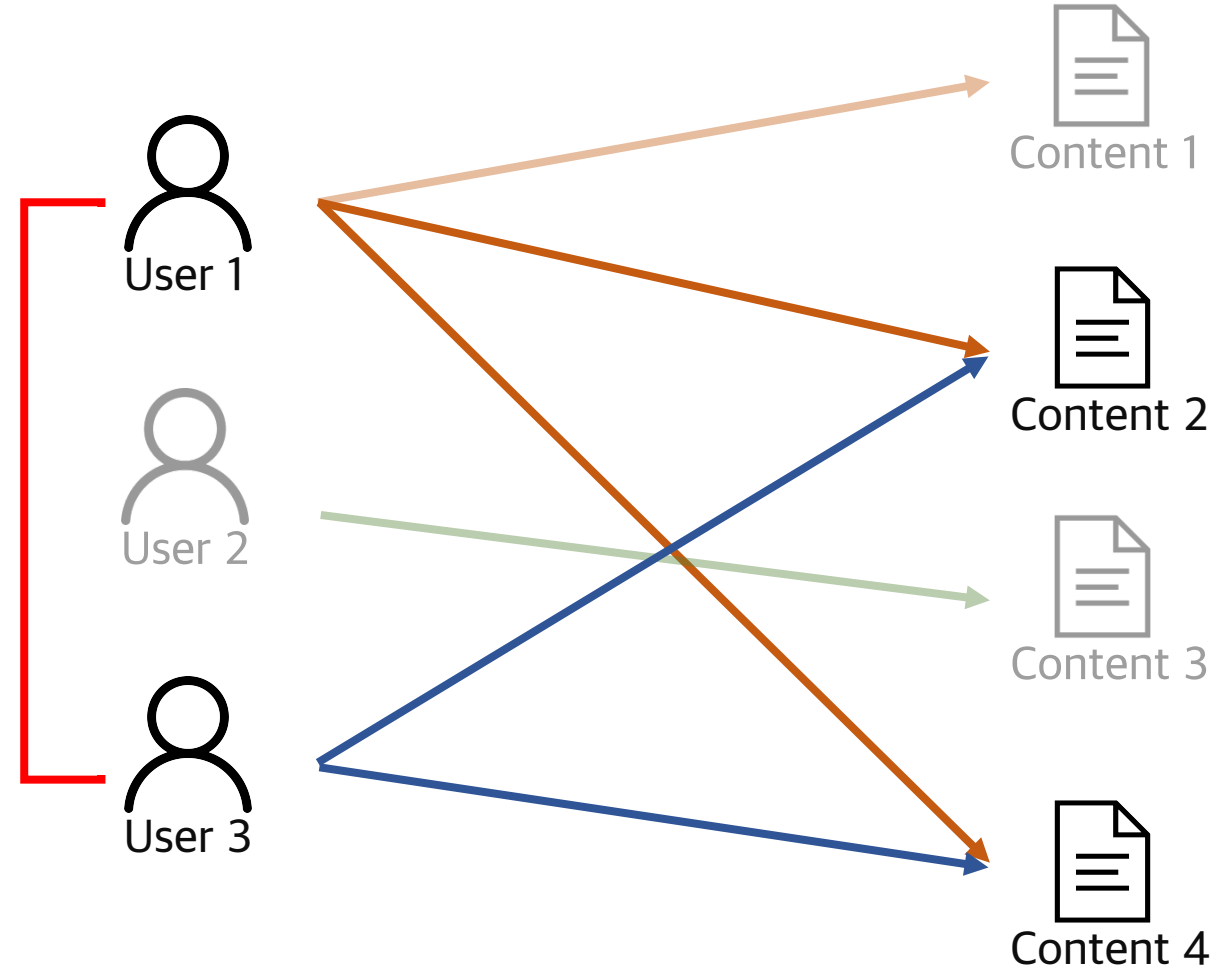


Background
System Design
Progress
Improvement
Personal Role

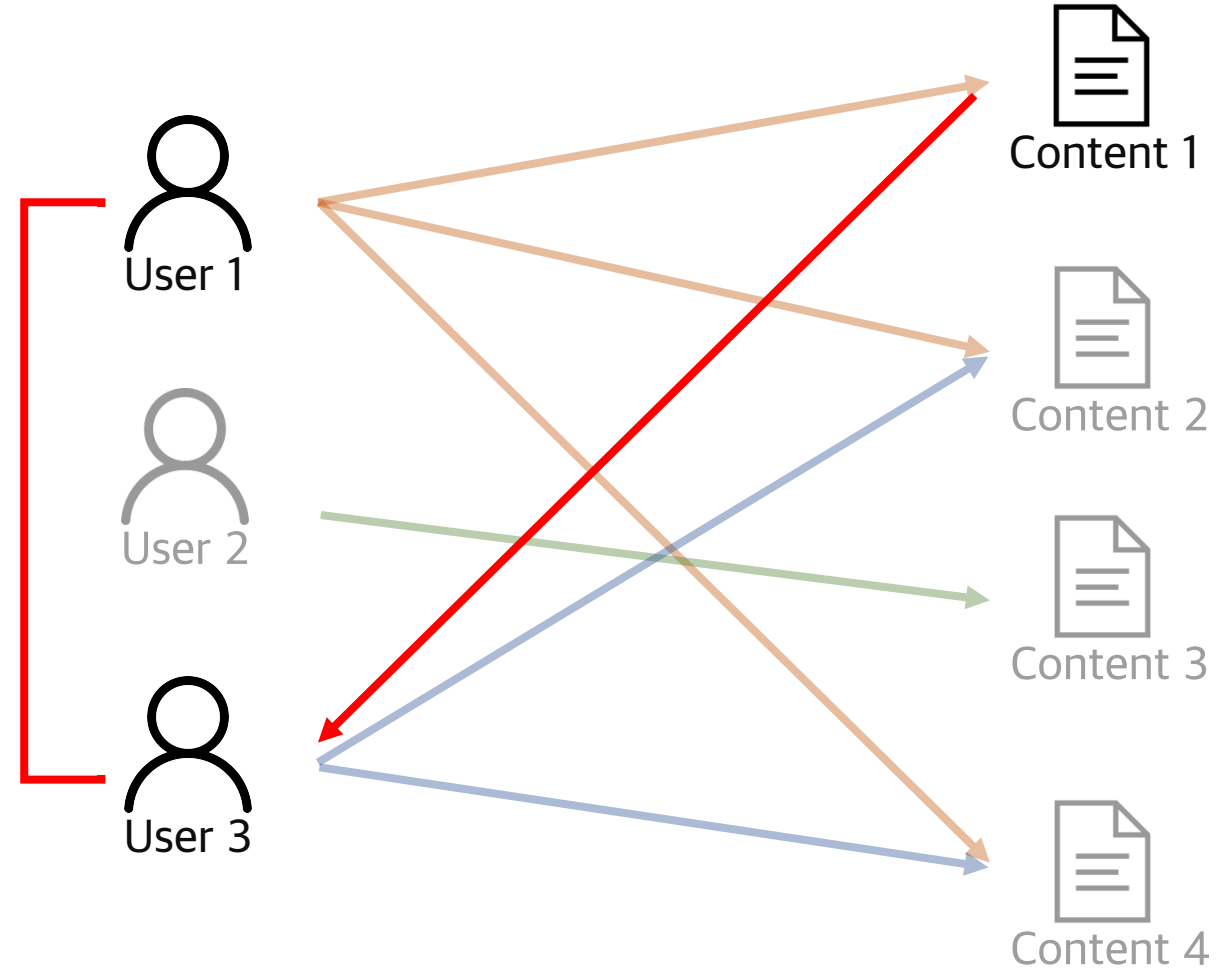
User-based Collaborative Filtering



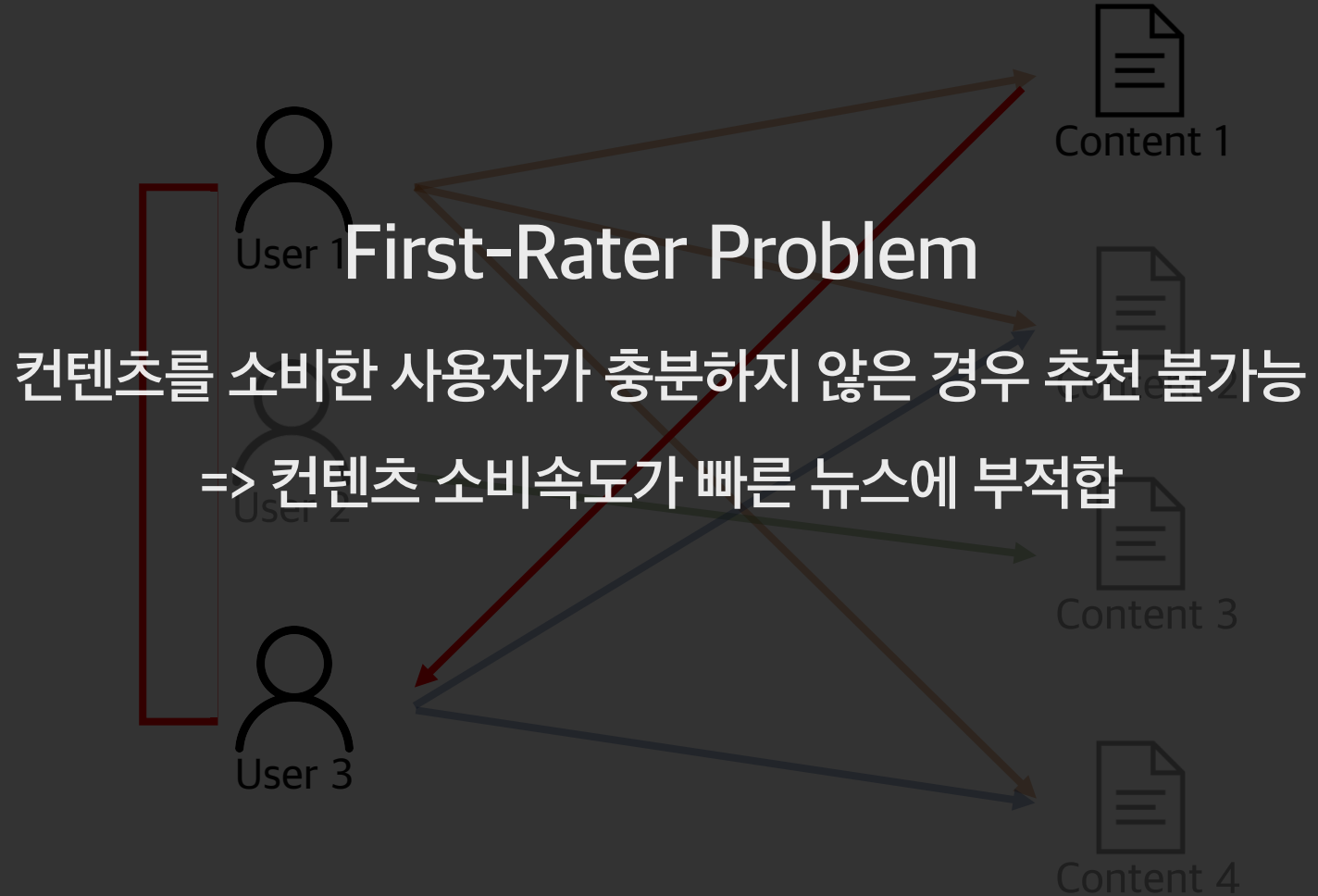
User-based Collaborative Filtering



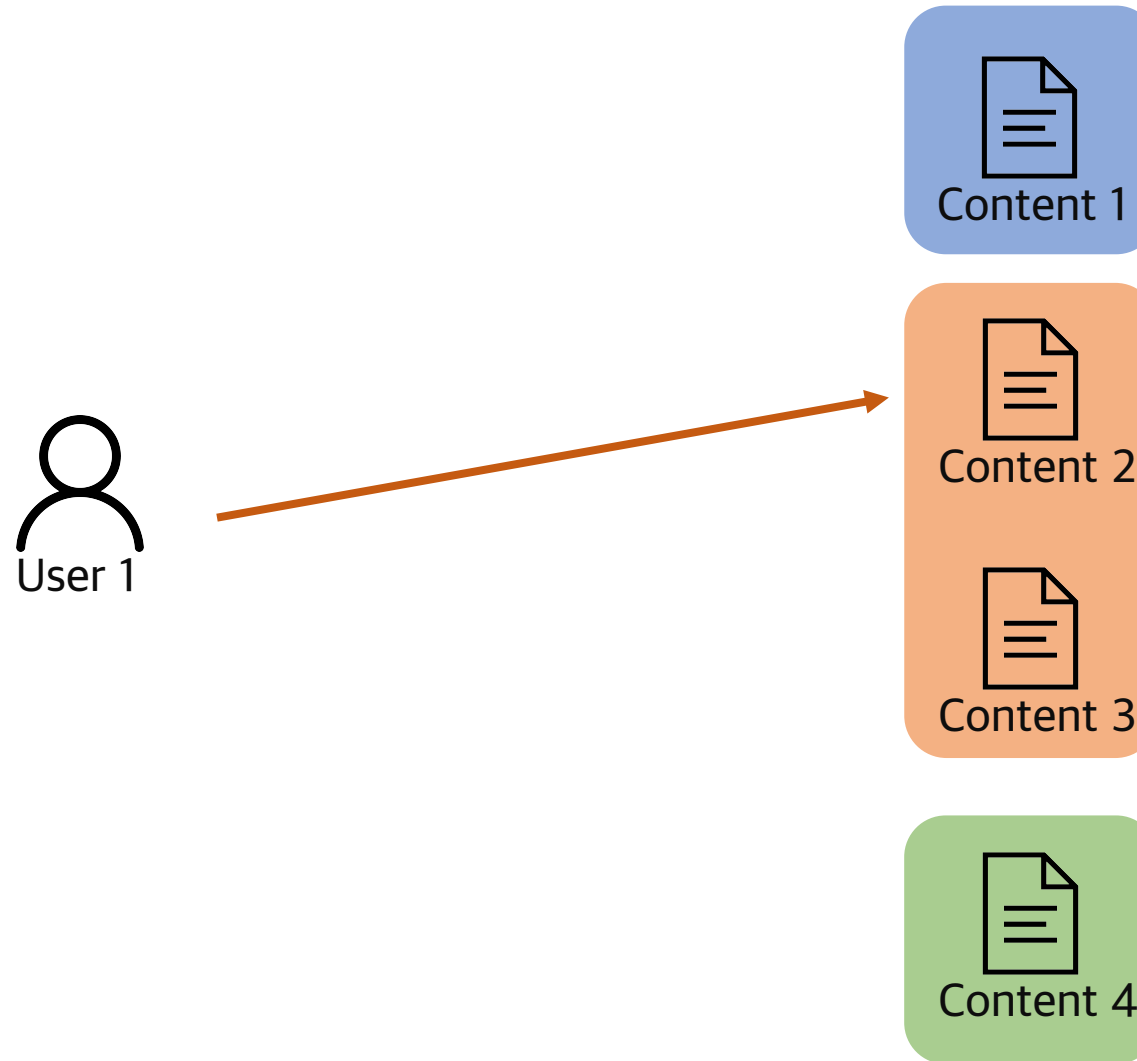
User-based Collaborative Filtering



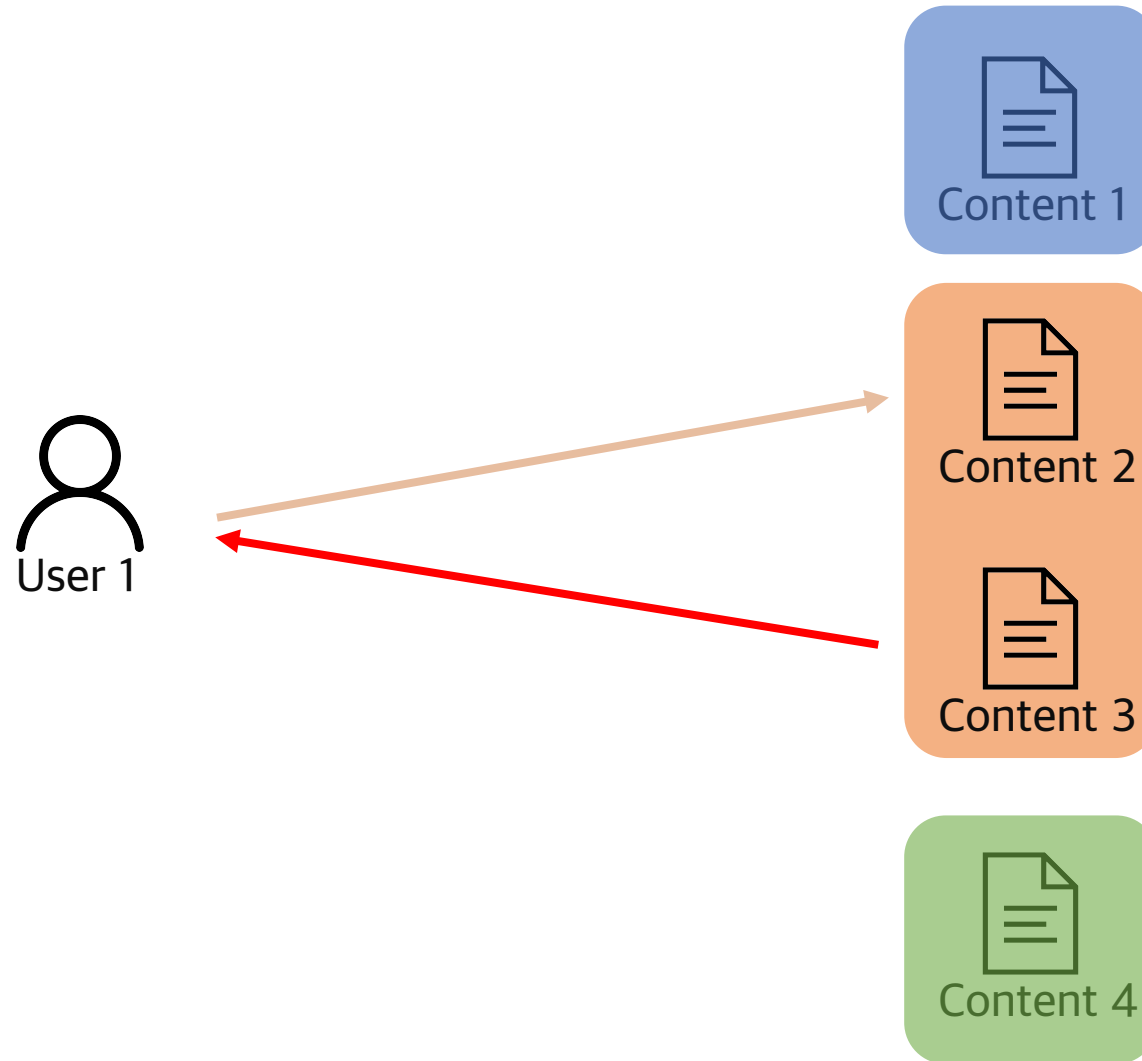
User-based Collaborative Filtering



Contents Based Filtering



Contents Based Filtering

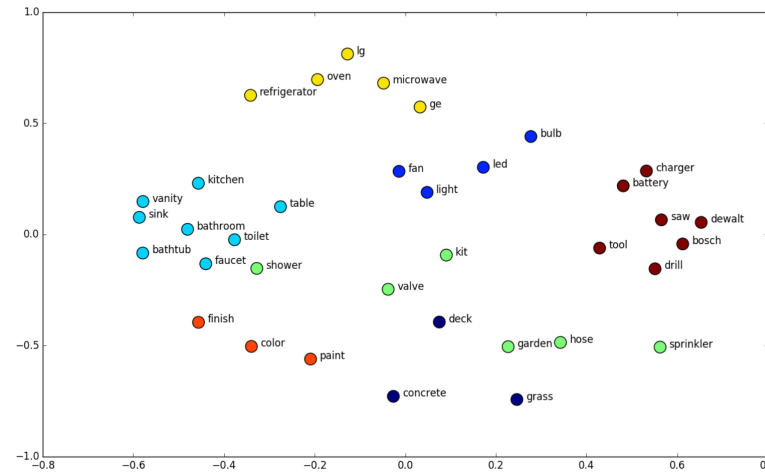


Measuring Similarities



Word Embedding

- Gensim
- word2vec
- konlpy

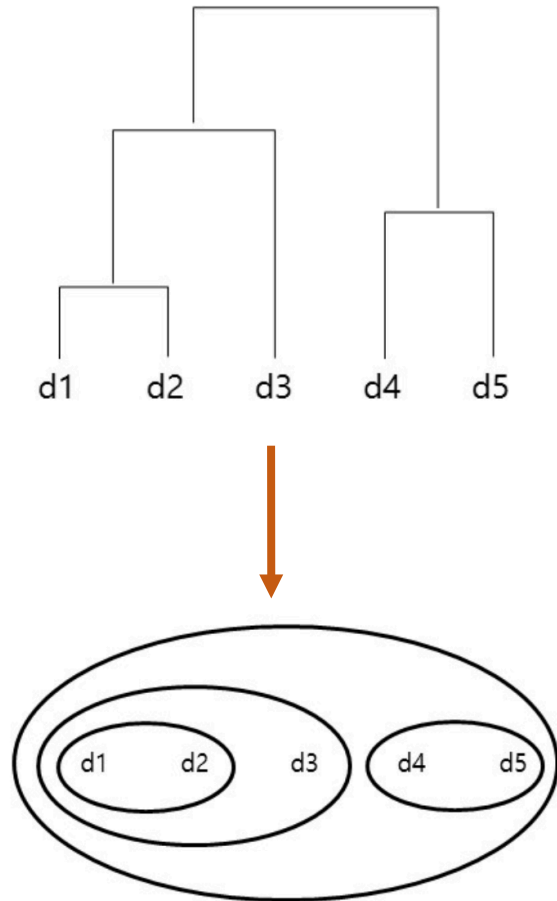


Calculate Distance

- Euclidean distance
- weighted value

Distance
Value

Clustering



- **Complete Linkage**
- **Average Linkage**
- **Single Linkage**

| 크롤링 모듈 개발

- 정치, 경제, 사회, 스포츠, 연예 분야 뉴스 크롤링
- 분야별 20개/일 * 15일 크롤링 완료

| 정치, 스포츠, 연예 분야 뉴스 20개씩 유사도 측정

- Word2Vec 사용
- Skip-gram 적용
- 1000회 반복학습
- 100차원 Vectorization
- 1회라도 출연하는 단어에 대해 학습
- 가중치로 거리의 세제곱 사용

Measuring Similarities

	정치	스포츠	연예
정치	3610	6047	5588
스포츠	6047	5136	5700
연예	5588	5700	4212

<6월 4일 정치, 스포츠, 연예 뉴스 유사도 평균>

Measuring Similarities

	정치	스포츠	연예
정치	3610	6047	5588
스포츠	6047	5136	5700
연예	5588	5700	4212

<6월 4일 정치, 스포츠, 연예 뉴스 유사도 평균>

Nearest Distance News

< [여론조사①] 재보선 12곳 중 후보 낸 11곳 '민주당 우세' >

오늘(4일) 8시 뉴스는 다음 주 수요일 지방선거를 앞두고 민심을 엿볼 수 있는 여론조사 결과로 시작하겠습니다. 이번 선거에서는 지역 일꾼과 함께 전국 12곳의 국회의원도 다시 뽑게 됩니다. SBS가 KBS, MBC와 함께 국회의원 재보궐 선거 여론조사를 했는데 민주당이 후보를 낸 11곳에서 모두 우세를 보였습니다. 먼저 수도권과

1249

1250

863

< 6.13 재보궐 선거...민주당 11곳 앞서 >

6.13 지방선거에서는 국회의원 재보선도 치러지죠. 모두 12곳, '미니 총선급'입니다. MBC 등 방송 3사가 공동여론조사를 해봤더니 11곳에서 민주당 우세로 나타났습니다. 정시내 기자입니다....

< 민주당, 재보궐 12곳 중 11곳서 선두... >

민주당, 송파을·PK 지역에서도 우위... 호남서도 압도적 한국, '보수텃밭' 김천서도 무소속에 밀려...1위 전무. 더불어민주당이 6.13 재보궐 선거 지역 12곳 중 후보를 낸 11곳 모두에서 1위.....

< [여론조사] 민주당, 재보선 12곳 中 11곳 1위 >

KBS 등 방송 3사가 국회의원 재보궐선거가 치러지는 12개 선거구에서 여론조사를 실시했습니다. 1곳을 제외한 11곳에서 더불어민주당이 1위로 나타났습니다. 정연욱 기자입니다.....

Longest Distance News

< [여론조사①] 재보선 12곳 중 후보 낸 11곳 '민주당 우세' >

오늘(4일) 8시 뉴스는 다음 주 수요일 지방선거를 앞두고 민심을 엿볼 수 있는 여론조사 결과로 시작하겠습니다. 이번 선거에서는 지역 일꾼과 함께 전국 12곳의 국회의원도 다시 뽑게 됩니다. SBS가 KBS, MBC와 함께 국회의원 재보궐 선거 여론조사를 했는데 민주당이 후보를 낸 11곳에서 모두 우세를 보였습니다. 먼저 수도권과

9205

7740

10635

< 손흥민의 이적시장 가치는 '1130억'>

유럽 축구 선수들에 대한 경제적 가치를 주기적으로 산정 및 발표하고 있는 CIES(국제스포츠연구센터)가 2018년 6월을 기준으로 한 새로운 선수 이적시장 가치 TOP 100....

< NC, "김경문 감독 퇴진은 구단 결정... >

계약기간인 2019년까지 잔여 연봉 지급, 복지 제공도 (서울=뉴스1) 정명의 기자 = 김경문 감독이 7년만에 NC 다이노스 지휘봉을 내려놓았다. 구단의 결정이었고, 김 감독은 이에 따랐다. NC는 ...

< '꽃보다 할배' 프리퀄로 먼저 본다... >

'꽃보다 할배' 리턴즈 첫 방송 이전, 프리퀄 형태의 하이라이트 방송으로 먼저 만난다. 4일 마이데일리 취재 결과, 케이블채널 tvN '꽃보다 할배 리턴즈' 첫 방송에 앞서....

Compare News in Sports category

< ‘월드컵 우승 도전’ 호날두 “전 세계가 보고 있다, 뛰러 갈 시간” >

크리스티아누 호날두가 월드컵을 앞두고 비장한 각오를 보였다. 발롱도르, 챔피언스리그, 유로까지 석권한 호날두의 목표는 월드컵 우승이다. 2018 국제축구연맹(FIFA) 러시아 월드컵 개막이 초읽기에 들어갔다. 14일(한국시간) 개막전을 시작으로 월드컵 트로피를 향한 32개 팀의 혈투가 벌어진다. 월드컵 출전국은 FIFA에 최종 명단 23인을 제출해 대회 막판 준비에 여념없다....

3567

< 신태용 감독 "세트피스 철저히 숨긴다... 본선에서 보라" >

“세트피스는 볼리비아전에도 볼 수 없을 것이다.” 신태용 감독은 23명의 최종엔트리가 가려진 뒤 진행되는 오스트리아 레오강 훈련에서 러시아 월드컵 16강의 토대를 닦겠다고 다짐했다....

5505

< NC, 김경문 감독 경질 ‘썩썩한 뒷맛’>

NC 김경문 감독(사진)이 전격 경질됐다. 지난 3일 마산 삼성전을 앞두고 황순현 NC 대표이사가 유영준 단장에게 감독대행을 맡으라고 통보했고 김경문 감독에게는 구단 고문 자리를 권유했다.....

Word Embedding

1) Word Vector 전처리

'늘어나는': 1830,
'늘어나고': 1830,
'늘어난': 1831,
'늘어서': 1832,
'늘었다': 1833,
'늘었을': 1834,



- 어간을 중심으로 단어군 생성
- 단어군 내 평균을 통해 임베딩 벡터 재조정

<같은 어간을 사용하지만 별개로 학습되는 단어>

Word Embedding

2) 음절기반 단어 임베딩

- 전처리를 하는 과정은 형태소 분석기의 성능에 결과가 바뀜
- 어근, 어간에 대한 학습이 불가능
- 한국어는 영어와 달리 음절 하나하나에 의미가 있는 한자어가 다수 존재



- 어절이 아닌 음절을 기준으로 단어 분리
- 같은 음절에 대해 위치와 앞뒤 조합 구분 필요
- CNN을 사용하여 음절기반 단어 임베딩

김연태

- 크롤링 모듈 개발
- 워드 임베딩 모델 개발

박현호

- 워드 임베딩 모델 개발
- 유사도 측정 모델 개발

임종완

- 클러스터링 모델 개발