NLP & Information Retrieval Class

Team Project

SoftNLP | 유승욱, 김민주, 정훈석

Tabloid Discriminator

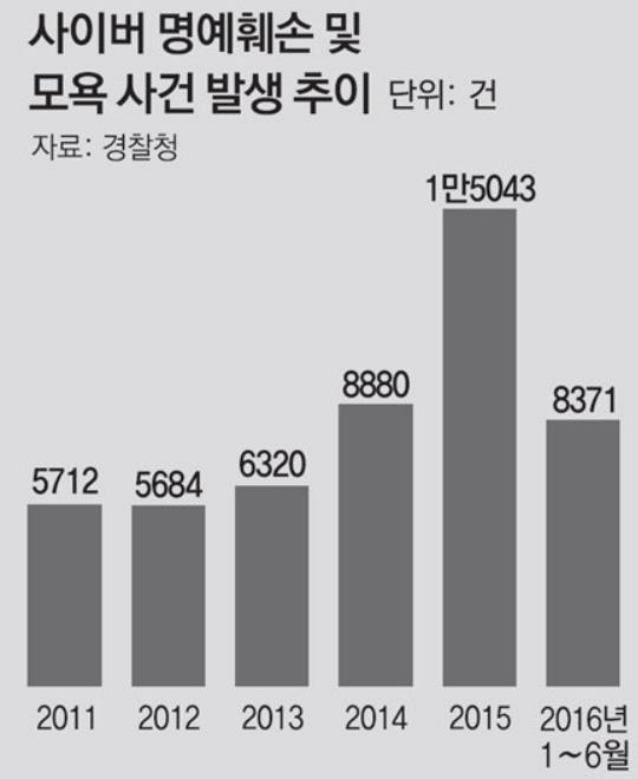
Contents

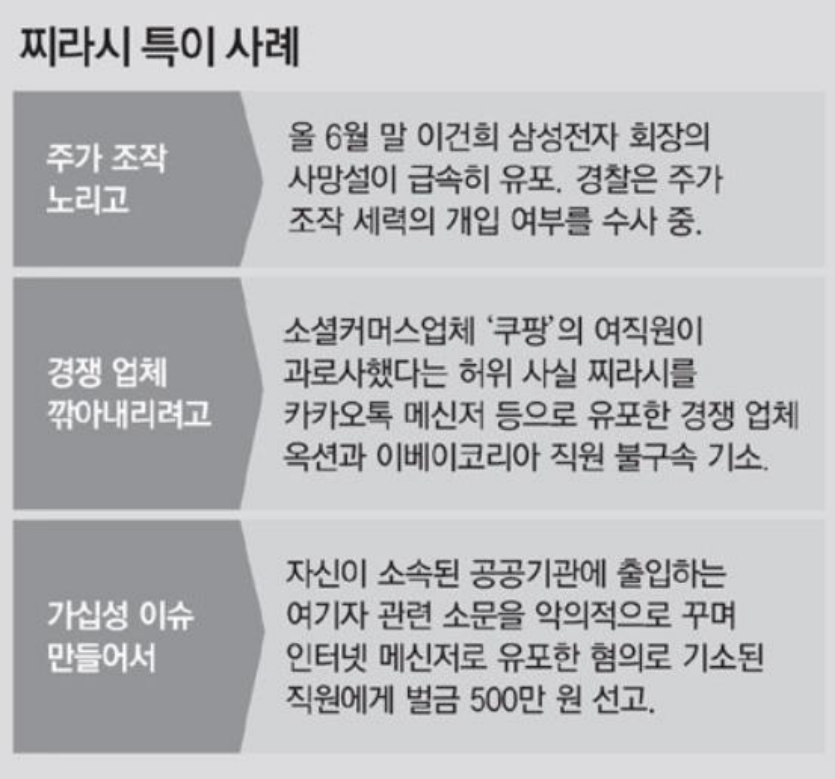
1. Topic Overview
2. Process: Data
3. Process: Retrieval
4. Process: Summary
5. Team Roles

1. Topic Overview

친구가 보낸 '카톡 찌라시' 퍼나르면… 나도 모르게 "범법자"

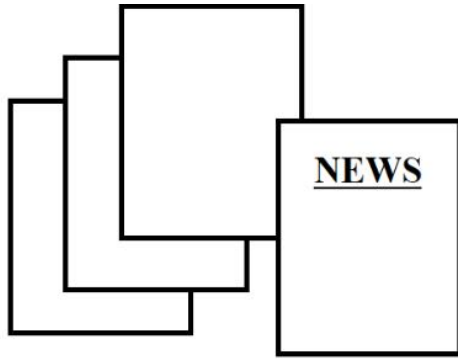
If you spread the Kakaotalk tabloid that your friend sent…
You can be a criminal without your knowledge



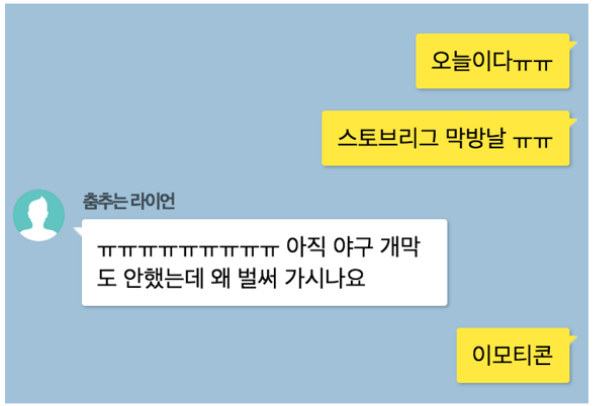


1. Topic Overview

Latest Online News



Crawling News Texts



Kakaotalk Conversation Image

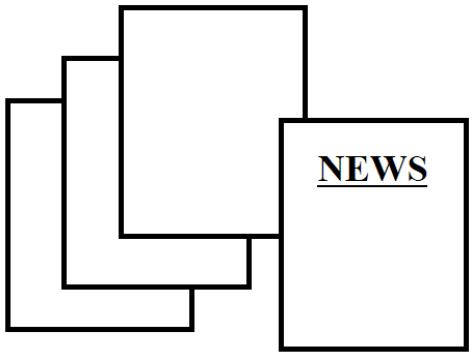
"오늘이다ㅠㅠ"
"스토브리그 막방날ㅠㅠ"
"ㅠㅠㅠㅠㅠㅠㅠㅠㅠㅠ 아직 야구 개막도 안했는데 왜 벌써 가시나요"
"이모티콘"

Extracting Conversation Texts

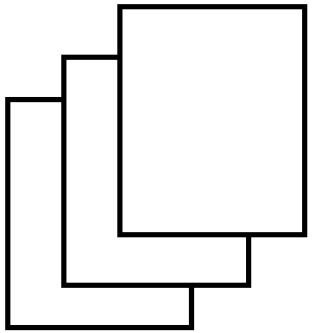
1. Topic Overview

"오늘이다ㅠㅠ"
"스토브리그 막방날ㅠㅠ"
"ㅠㅠㅠㅠㅠㅠㅠㅠㅠㅠ 아직 야구 개막도 안했는데 왜 벌써 가시나요"
"이모티콘"

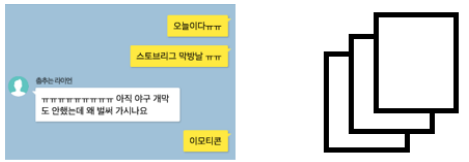
Extracting Conversation Texts



Crawling News Texts



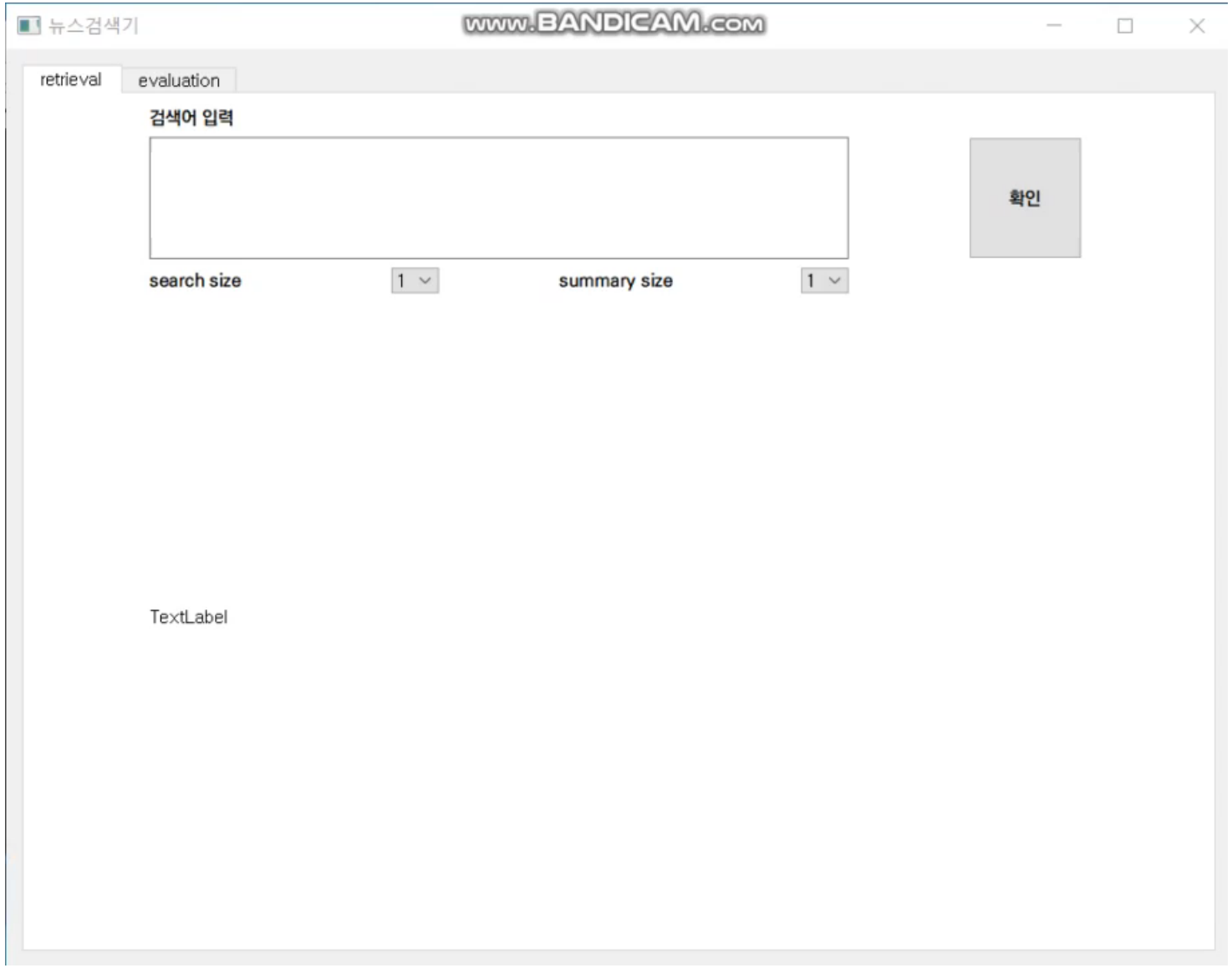
Ranking Similar News



News Title: ~~~~~~~~~~
News Date: ~~~~~
News Content: ~~~~~~~~~~~~~~~~~~~~~~~~~~~~
Summarization: ~~~~~~~~~~~~~~~~~~~~

Similar News Information,
Similar News Summarization

1. Topic Overview



2. Process: Data

2. Process: Data

Select some Daum 뉴스 categories that may lead to tabloid rumors

society/affair, society/others, society/labor,
politics/administration, politics/dipdefen, economic/consumer

2. Process: Data

Crawling 6 categories during 2022 May 1st to 31st by Selenium

	제목	주소	본문	날짜/카테고리
0	훔친 차 몰다 사고 낸 고교생..의식 잃은 동승자 버리고 도주	https://v.daum.net/v/20220501223102728	파주에서 무면허 상태로 훔친 차량을 몰다 사고를 낸 고교생이 의식 잃은 동승자를 버...	20220501/society/affair
1	훔친차 몰다 사고낸 뒤 동승자 두고 달아난 고교생 덜미	https://v.daum.net/v/20220501220916537	경찰 마크. 경향신문 자료사진 훔친 차량을 몰다 사고를 낸 고교생이 의식을 잃은 동...	20220501/society/affair
2	고교생, 훔친 차량 몰다 사고..'의식불명' 동승자 버리고 도주	https://v.daum.net/v/20220501220127479	훔친 차를 몰다 사고를 낸 고교생이 의식을 잃은 동승자를 아무런 조처 없이 차량 밖...	20220501/society/affair
3	고교생, 무면허로 훔친차 몰다 사고 후 도주..동승자 '의식불명'	https://v.daum.net/v/20220501215259356	훔친 차를 몰다 사고를 낸 고교생이 의식을 잃은 동승자를 아무런 조처 없이 차량 밖...	20220501/society/affair
4	익산 태양광 발전시설 ESS서 화재..2억원 재산 피해	https://v.daum.net/v/20220501215023318	1일 오후 2시 50분께 전북 익산시 망성면 어량리의 한 태양광 발전설비 에너지저장...	20220501/society/affair
...	...	...	...	...
69403	W컨셉, 리조트룩·아웃도어룩 최대 83% 할인	https://v.daum.net/v/20220531060009417	W컨셉은 오는 6월 112일 여름 휴가를 콘셉트로 기획전을 실시하고 관련 상품을 최...	20220531/economic/consumer
69404	SSG닷컴, 6월 한 달간 '스마일클럽' 고객 대상 프로모션 실시	https://v.daum.net/v/20220531060001392	뉴스1 SSG닷컴은 6월 한 달간 스마일클럽 가입 고객 대상 특별 프로모션을 열...	20220531/economic/consumer
69405	업사이클링 제품 300여개 한눈에	https://v.daum.net/v/20220531030621223	30일 경기 성남시 현대백화점 판교점에서 업사이클링 제품을 한데 모은 하우 투 리그...	20220531/economic/consumer
69406	중년층 이탈에.. TV홈쇼핑 업체들 'TV 탈출 작전'	https://v.daum.net/v/20220531030218046	TV홈쇼핑 업체들의 탈 TV가 가속화하고 있다. TV 방송의 영향력이 감소하고, 주...	20220531/economic/consumer
69407	횡성 국순당-칠성사이다 협업 '칠성막사' 내일 출시	https://v.daum.net/v/20220531001042395	70년 전통 막걸리와 사이다가 만나면 어떤 맛을 낼까. 횡성 향토기업 국순당이 칠성...	20220531/economic/consumer

69408 rows × 4 columns

Text Preprocessing and Tokenization

Removing special characters and Korean stopwords

Tokenization reflecting the characteristics of Korean and morpheme analysis

	제목	주소	본문	날짜/카테고리
0	홈친 차 몰다 사고 낸 고교생..의식 잃은 동승자 버리고 도주	https://v.daum.net/v/20220501223102728	파주에서 무면허 상태로 홈친 차량을 몰다 사고를 낸 고교생이 의식 잃은 동승자를 버...	20220501/society/affair
1	홈친차 몰다 사고낸 뒤 동승자 두고 달아난 고교생 덜미	https://v.daum.net/v/20220501220916537	경찰 마크. 경향신문 자료사진 홈친 차량을 몰다 사고를 낸 고교생이 의식을 잃은 동...	20220501/society/affair
2	고교생, 홈친 차량 몰다 사고..'의식불명' 동승자 버리고 도주	https://v.daum.net/v/20220501220127479	홈친 차를 몰다 사고를 낸 고교생이 의식을 잃은 동승자를 아무런 조처 없이 차량 밖...	20220501/society/affair
3	고교생, 무면허로 홈친차 몰다 사고 후 도주..동승자 '의식불명'	https://v.daum.net/v/20220501215259356	홈친 차를 몰다 사고를 낸 고교생이 의식을 잃은 동승자를 아무런 조처 없이 차량 밖...	20220501/society/affair
4	익산 태양광 발전시설 ESS서 화재..2억원 재산 피해	https://v.daum.net/v/20220501215023318	1일 오후 2시 50분께 전북 익산시 망성면 어량리의 한 태양광 발전설비 에너지저장...	20220501/society/affair
...	...	...	...	...
69403	W컨셉, 리조트룩·아웃도어룩 최대 83% 할인	https://v.daum.net/v/20220531060009417	W컨셉은 오는 6월 112일 여름 휴가를 콘셉트로 기획전을 실시하고 관련 상품을 최...	20220531/economic/consumer
69404	SSG닷컴, 6월 한 달간 '스마일클럽' 고객 대상 프로모션 실시	https://v.daum.net/v/20220531060001392	뉴스1 SSG닷컴은 6월 한 달간 스마일클럽 가입 고객 대상 특별 프로모션을 열...	20220531/economic/consumer
69405	업사이클링 제품 300여개 한눈에	https://v.daum.net/v/20220531030621223	30일 경기 성남시 현대백화점 판교점에서 업사이클링 제품을 한데 모은 하우 투 리그...	20220531/economic/consumer
69406	중년층 이탈에.. TV홈쇼핑 업체들 'TV 탈출 작전'	https://v.daum.net/v/20220531030218046	TV홈쇼핑 업체들의 탈 TV가 가속화하고 있다. TV 방송의 영향력이 감소하고, 주...	20220531/economic/consumer
69407	횡성 국순당-칠성사이다 협업 '칠성막사' 내일 출시	https://v.daum.net/v/20220531001042395	70년 전통 막걸리와 사이다가 만나면 어떤 맛을 낼까. 횡성 향토기업 국순당이 칠성...	20220531/economic/consumer

69408 rows × 4 columns

	제목	주소	본문	날짜/카테고리	제목토큰	본문토큰
0	홈친 차 몰다 사고 낸 고교생..의식 잃은 동승자 버리고 도주	https://v.daum.net/v/20220501223102728	파주에서 무면허 상태로 홈친 차량을 몰다 사고를 낸 고교생이 의식 잃은 동승자를 버...	20220501/society/affair	['잃다', '도주', '몰다', '동승자', '내다', '홈치다', '버리다', ...	['몰다', '확인태다', '위반', '주유소', '조사', '고교생', '파주시'...
1	홈친차 몰다 사고낸 뒤 동승자 두고 달아난 고교생 덜미	https://v.daum.net/v/20220501220916537	경찰 마크. 경향신문 자료사진 홈친 차량을 몰다 사고를 낸 고교생이 의식을 잃은 동...	20220501/society/affair	['고교생', '덜미', '몰다', '동승자', '사고내다', '홈치다자', '달아...	['몰다', '위반', '주유소', '조사', '고교생', '파주시', '투숙객',...
2	고교생, 홈친 차량 몰다 사고..'의식불명' 동승자 버리고 도주	https://v.daum.net/v/20220501220127479	홈친 차를 몰다 사고를 낸 고교생이 의식을 잃은 동승자를 아무런 조처 없이 차량 밖...	20220501/society/affair	['고교생', "'사고'의식불명'", '도주', '몰다', '동승자', '차량', '...	['몰다', '위반', '주유소', '조사', '고교생', '파주시', '투숙객',...
3	고교생, 무면허로 홈친차 몰다 사고 후 도주..동승자 '의식불명'	https://v.daum.net/v/20220501215259356	홈친 차를 몰다 사고를 낸 고교생이 의식을 잃은 동승자를 아무런 조처 없이 차량 밖...	20220501/society/affair	['고교생', '무면허', '몰다', '홈치다자', '도주동승자', "'의식불명'"...	['타고', '않다', '미숙', '발견', '경기', '고양', '출구', '주차...
4	익산 태양광 발전시설 ESS서 화재..2억원 재산 피해	https://v.daum.net/v/20220501215023318	1일 오후 2시 50분께 전북 익산시 망성면 어량리의 한 태양광 발전설비 에너지저장...	20220501/society/affair	['피해', '재산', '발전시설', '익산', '태양광']	['오후', '현재', '망성면', '놀길', '경찰', '내다', '꺼지다', '...
...	...	...	...	...	...	...
69170	W컨셉, 리조트룩·아웃도어룩 최대 83% 할인	https://v.daum.net/v/20220531060009417	W컨셉은 오는 6월 112일 여름 휴가를 콘셉트로 기획전을 실시하고 관련 상품을 최...	20220531/economic/consumer	['할인', '최대']	['콘셉트', '브랜드', '휴양지의', '대주', '데이즈데이즈', '헤어', '...
69171	SSG닷컴, 6월 한 달간 '스마일클럽' 고객 대상 프로모션 실시	https://v.daum.net/v/20220531060001392	뉴스1 SSG닷컴은 6월 한 달간 스마일클럽 가입 고객 대상 특별 프로모션을 열...	20220531/economic/consumer	["'스마일클럽'", '대상', '하다', '실시', '프로모션', '고객']	['프리미엄', '결제하다', '등록되다', '연다', '브랜드', '월요일', '...
69172	업사이클링 제품 300여개 한눈에	https://v.daum.net/v/20220531030621223	30일 경기 성남시 현대백화점 판교점에서 업사이클링 제품을 한데 모은 하우 투 리그...	20220531/economic/consumer	['업사이클링', '한눈', '제품']	['현사', '경기', '현대백화점', '한국환경산업협회', '브랜드', '제품',...
69173	중년층 이탈에.. TV홈쇼핑 업체들 'TV 탈출 작전'	https://v.daum.net/v/20220531030218046	TV홈쇼핑 업체들의 탈 TV가 가속화하고 있다. TV 방송의 영향력이 감소하고, 주...	20220531/economic/consumer	['이탈', '탈출', '중년층', "작전'"]	['설립하다', '않다', '진출하다', '중장년층', '따르다', '브랜드', '...
69174	횡성 국순당-칠성사이다 협업 '칠성막사' 내일 출시	https://v.daum.net/v/20220531001042395	70년 전통 막걸리와 사이다가 만나면 어떤 맛을 낼까. 횡성 향토기업 국순당이 칠성...	20220531/economic/consumer	["'칠성막사'", '내일', '횡성', '협업', '국순당', '칠성사', '출시']	["'전통', '맡다', '만나다', '제품', '구현하다', '느끼다', '막거리'...

69175 rows × 6 columns

after text preprocessing and tokenization

3. Process: Retrieval

3. Process: Retrieval

Implement news retrieval system by scikit-learn tf-idf

User query
tf-idf vector

compare cosine similarities

	제목	주소	본문	날짜/카테고리
0	홈친 차 몰다 사고 낸 고교생, 의식 잃은 동승자 버리고 도주	https://v.daum.net/v/20220501223102728	파주에서 무면허 상태로 훔친 차량을 몰다 사고를 낸 고교생이 의식 잃은 동승자를 버...	20220501/society/affair
1	훔친차 몰다 사고낸 뒤 동승자 두고 달아난 고교생 멀미	https://v.daum.net/v/20220501220916537	경찰 마크. 경향신문 자료사진 훔친 차량을 몰다 사고를 낸 고교생이 의식을 잃은 동...	20220501/society/affair
2	고교생, 훔친 차량 몰다 사고..'의식불명' 동승자 버리고 도주	https://v.daum.net/v/20220501220127479	훔친 차를 몰다 사고를 낸 고교생이 의식을 잃은 동승자를 아무런 조처 없이 차량 밖...	20220501/society/affair
3	고교생, 무면허로 훔친차 몰다 사고 후 도주..동승자 '의식불명'	https://v.daum.net/v/20220501215259356	훔친 차를 몰다 사고를 낸 고교생이 의식을 잃은 동승자를 아무런 조처 없이 차량 밖...	20220501/society/affair
4	익산 태양광 발전시설 ESS서 화재..2억원 재산 피해	https://v.daum.net/v/20220501215023318	1일 오후 2시 50분께 전북 익산시 망성면 어량리의 한 태양광 발전설비 에너지저장...	20220501/society/affair
...	...	...	...	...
69403	W컨셉, 리조트룩 아웃도어룩 최대 83% 할인	https://v.daum.net/v/20220531060009417	W컨셉은 오는 6월 112일 여름 휴가를 콘셉트로 기획전을 실시하고 관련 상품을 최...	20220531/economic/consumer
69404	SSG닷컴, 6월 한 달간 '스마일클럽' 고객 대상 프로모션 실시	https://v.daum.net/v/20220531060001392	뉴스1 SSG닷컴은 6월 한 달간 스마일클럽 가입 고객 대상 특별 프로모션을 열...	20220531/economic/consumer
69405	업사이클링 제품 300여개 한눈에	https://v.daum.net/v/20220531030621223	30일 경기 성남시 현대백화점 판교점에서 업사이클링 제품을 한데 모은 하우 투 리그...	20220531/economic/consumer
69406	중년층 이탈에.. TV홈쇼핑 업체들 'TV 탈출 작전'	https://v.daum.net/v/20220531030218046	TV홈쇼핑 업체들의 탈 TV가 가속화하고 있다. TV 방송의 영향력이 감소하고, 주...	20220531/economic/consumer
69407	횡성 국순당-칠성사이다 협업 '칠성막사' 내일 출시	https://v.daum.net/v/20220531001042395	70년 전통 막걸리와 사이다가 만나면 어떤 맛을 낼까. 횡성 향토기업 국순당이 칠성...	20220531/economic/consumer

69408 rows × 4 columns

Crawled news tf-idf vector

3. Process: Retrieval

Remove similar news from retrieved results

Define a 'similar threshold',
If cos similarity of news pairs (in the retrieved results) is greater than 'similar threshold',
remove one news of that pairs

	제목	링크	본문	날짜/카테고리	제목토큰	본문토큰
0	훔친 차 몰다 사고 낸 고교생..의식 잃은 동승자 버리고 도주	https://v.daum.net/v/20220501223102728	파주에서 무면허 상태로 훔친 차량을 몰다 사고를 낸 고교생이 의식 잃은 동승자를 버...	20220501/society/affair	['도주', '몰다', '사고', '잃다', '고교생', '의식', '내다', '동승자',...	['오전', '주차돼다', '잃다', '파주', '크게', '인근', '위반', '...
1	훔친차 몰다 사고낸 뒤 동승자 두고 달아난 고교생 덜미	https://v.daum.net/v/20220501220916537	경찰 마크. 경향신문 자료 사진 훔친 차량을 몰다 사고를 낸 고교생이 의식을 잃은 동...	20220501/society/affair	['덜미', '사고내다', '고교생', '몰다', '두다', '훔치다차', '동승자...	['오전', '고양지역', '주차돼다', '잃다', '크게', '인근', '알려지다...
2	고교생, 훔친 차량 몰다 사고..'의식불명' 동승자 버리고 도주	https://v.daum.net/v/20220501220127479	훔친 차를 몰다 사고를 낸 고교생이 의식을 잃은 동승자를 아무런 조처 없이 차량 밖...	20220501/society/affair	['도주', "사고'의식불명'", '고교생', '몰다', '차량', '동승자', '...	['오전', '주차돼다', '잃다', '크게', '인근', '말하다', '위반', ...
3	고교생, 무면허로 훔친차 몰다 사고 후 도주..동승자 '의식불명'	https://v.daum.net/v/20220501215259356	훔친 차를 몰다 사고를 낸 고교생이 의식을 잃은 동승자를 아무런 조처 없이 차량 밖...	20220501/society/affair	['무면허', '고교생', '몰다', '도주동승자', '사고', '훔치다차', '"...	['택시', '오전', '없이', '화중로', '병원', '숨다', '있다', '잃...
4	익산 태양광 발전시설 ESS서 화재..2억원 재산 피해	https://v.daum.net/v/20220501215023318	전북 익산시 망성면 어량리의 한 태양광 발전설비 에너지저장...	20220501/society/affair	['익산', '피해', '발전시설', '재산', '태양광']	['잡다', '진압', '망성면', '불길', '전기', '배터리', '정확하다',...

remove when near-similar news appeared like this

Retrieval Examples

Input query: '검찰청장 수사'

default

	링크	본문
1	https://v.daum.net/v/20220502161620767	김지용 대검찰청 형사부장이 지난 4월 29일 서울 대검찰청에서 검수완박 수정안 관련...
2	https://v.daum.net/v/20220502104017060	황운하 전 울산경찰청장이 2일 오전 서울 서초구 서울중앙지방법원에서 열린 청와대 하...
3	https://v.daum.net/v/20220502104021065	황운하 전 울산경찰청장이 2일 오전 서울 서초구 서울중앙지방법원에서 열린 청와대 하...
4	https://v.daum.net/v/20220520164645528	이정수 서울중앙지검장이 20일 오후 서울 서초구 서울중앙지검에서 열린 이임식에 참석...
5	https://v.daum.net/v/20220520164617511	이정수 서울중앙지검장이 20일 오후 서울 서초구 서울중앙지검에서 열린 이임식에서 이...

when similar threshold = 0.9

	링크	본문
1	https://v.daum.net/v/20220502161620767	김지용 대검찰청 형사부장이 지난 4월 29일 서울 대검찰청에서 검수완박 수정안 관련...
2	https://v.daum.net/v/20220502104017060	황운하 전 울산경찰청장이 2일 오전 서울 서초구 서울중앙지방법원에서 열린 청와대 하...
3	https://v.daum.net/v/20220520164645528	이정수 서울중앙지검장이 20일 오후 서울 서초구 서울중앙지검에서 열린 이임식에 참석...
4	https://v.daum.net/v/20220520164657532	이정수 서울중앙지검장이 20일 오후 서울 서초구 서울중앙지검에서 열린 이임식에서 국...
5	https://v.daum.net/v/20220520164610505	이정수 서울중앙지검장이 20일 오후 서울 서초구 서울중앙지검에서 열린 이임식에서 생...

when similar threshold = 0.8

	링크	본문
1	https://v.daum.net/v/20220502161620767	김지용 대검찰청 형사부장이 지난 4월 29일 서울 대검찰청에서 검수완박 수정안 관련...
2	https://v.daum.net/v/20220502104017060	황운하 전 울산경찰청장이 2일 오전 서울 서초구 서울중앙지방법원에서 열린 청와대 하...
3	https://v.daum.net/v/20220520164645528	이정수 서울중앙지검장이 20일 오후 서울 서초구 서울중앙지검에서 열린 이임식에 참석...
4	https://v.daum.net/v/20220502161720826	경찰이 더불어민주당 대선 후보였던 이재명 전 경기지사의 성남FC 후원금 의혹 수사를...
5	https://v.daum.net/v/20220513100554611	이성윤 서울고등검찰청 검사장이 13일 오전 서울 서초구 중앙지방법원에서 김학의 전 ...

Evaluation for retrieved news

To compare retrieve method

	Our news retrieval system	Daum news retrieval system
News posting date	2022 May	2022 May
News category	Only 6 categories	All categories
Retrieve method	Comparing cos similarity between tf-idf query and news	unknown

3. Process: Retrieval

Evaluation for retrieved news

Crawling searched Daum news when given an input query

searched news
from our retrieval system

compare cosine similarities

searched news
from daum retrieval system

4. Process: Summary

Summarization for retrieved news

본문: 훔친 차를 몰다 사고를 낸 고교생이 의식을 잃은 동승자를 방치한 채 도망쳤다가 경찰에 붙잡혔다. 1일 경기 파주경찰서는 절도와 도로교통법 위반 등 혐의로 고교생 A군을 붙잡아 조사 중이라고 밝혔다. A군은 이날 오전 2시쯤 고양시 덕양구 화중로에 주차돼있던 SM5 승용차를 훔친 혐의를 받고 있다. A군은 차 안에 키가 있는 것을 찾아 고교생 B양을 태운 뒤 (중략)

요약: 훔친 차를 몰다 사고를 낸 고교생이 의식을 잃은 동승자를 방치한 채 도망쳤다가 경찰에 붙잡혔다. 이 사고로 조수석에 타고 있던 B양이 머리를 크게 다쳐 의식을 잃었으나 A군은 구호조치를 취하지 않고 B양을 인근 모텔 주차장에 옮긴 뒤 인근 야산으로 도망쳤다

본문: 법원 절도 범죄로 수 차례 처벌을 받고도 가석방된 지 9개월 만에 택시기사들을 상대로 돈을 훔친 20대가 실형을 선고받았다. 춘천지법 형사3단독 차영욱 판사는 상습야간주거침입절도와 사기 혐의로 기소된 A씨에게 징역 1년 6개월을 선고했다고 1일 밝혔다. A씨는 지난해 11월 초 춘천의 한 택시 조수석에 탄 후 입고있던 패딩 점퍼를 벗어 조수석 앞과 미터기를 (중략)

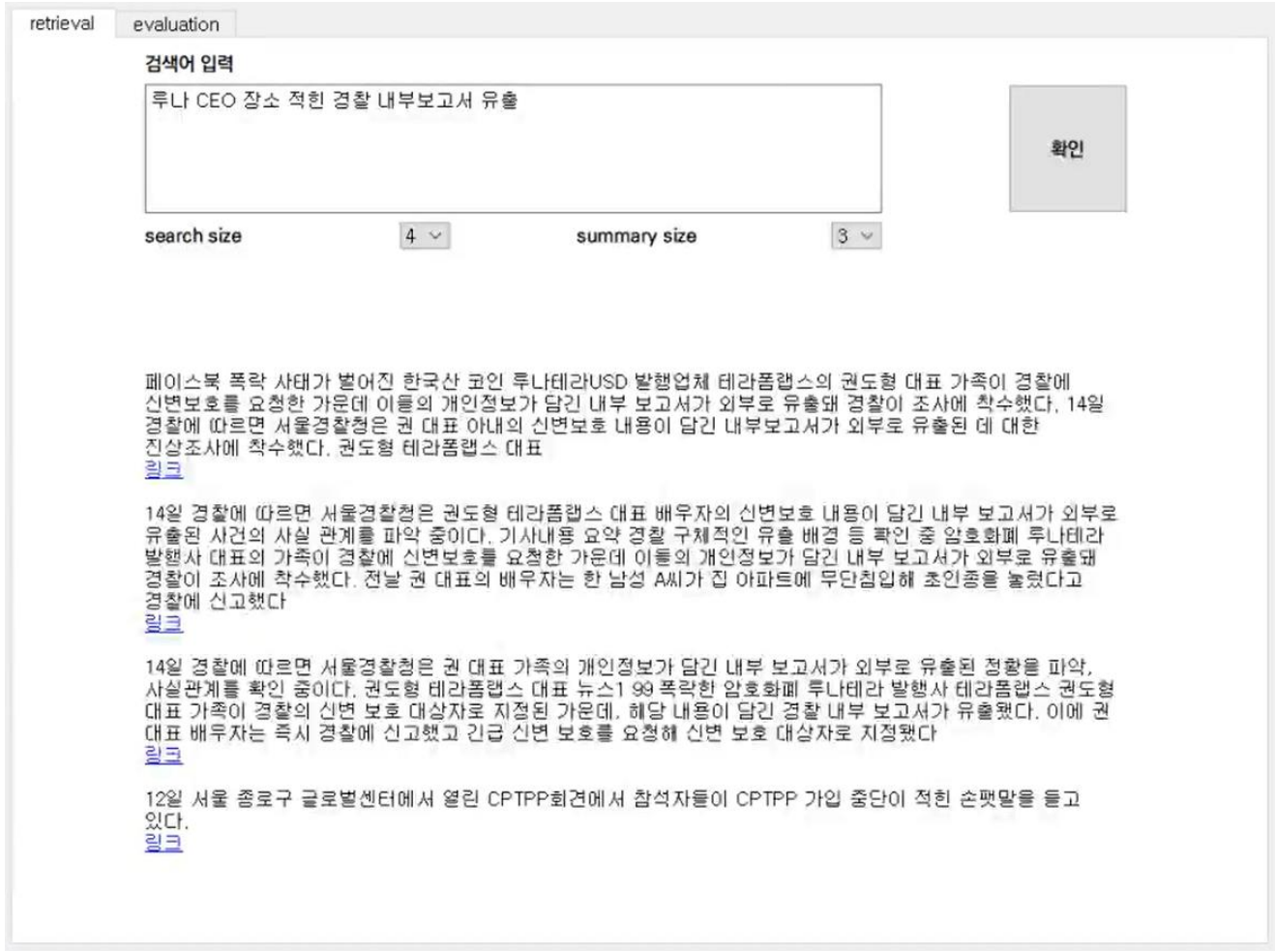
요약: 지난 2020년 상습야간건조물침입절도죄로 징역 10개월을 선고받은 A씨는 지난해 2월 가석방된 이후 9개월 만에 또다시 범행을 저질렀다. 법원 절도 범죄로 수 차례 처벌을 받고도 가석방된 지 9개월 만에 택시기사들을 상대로 돈을 훔친 20대가 실형을 선고받았다

본문: 614억원의 회삿돈을 횡령한 혐의로 구속영장이 청구된 우리은행 직원 A씨와 동생이 영장실질심사를 받기 위해 서울 서초구 서울중앙지법으로 들어가고 있다. 뉴시스 우리은행 회삿돈을 빼돌린 혐의를 받는 직원의 동생도 공범으로 함께 구속됐다. 서울중앙지법 허정인 판사는 1일 우리은행에서 거액을 빼돌린 A씨의 동생 B씨의 구속 전 피의자 심문을 진행한 뒤 증거인멸과 (중략)

요약: 뉴시스 우리은행 회삿돈을 빼돌린 혐의를 받는 직원의 동생도 공범으로 함께 구속됐다. 614억원의 회삿돈을 횡령한 혐의로 구속영장이 청구된 우리은행 직원 A씨와 동생이 영장실질심사를 받기 위해 서울 서초구 서울중앙지법으로 들어가고 있다

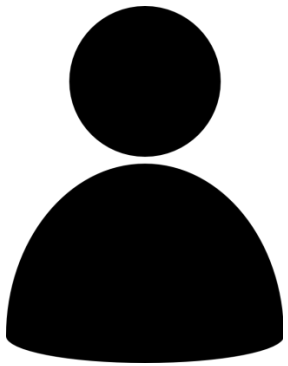
4. Process: Summary

Service visualization through GUI with



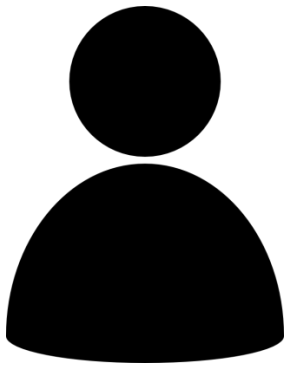
5. Team Roles

5. Team Roles



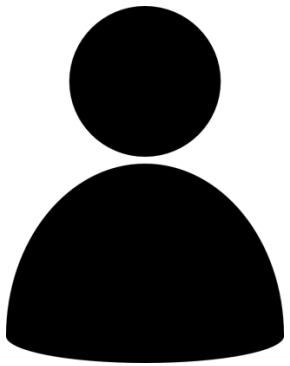
유승욱, Seunguk Yu
(Team Leader)

- Crawling for data gathering & evaluation
- Text preprocessing and tokenization
- Remove similar news from searched results



김민주, Minju Kim

- Retrieval news through tf-idf
- Evaluation for retrieved news



정훈석, Hunseok Jeong

- Summarization for retrieved news
- Service visualization through GUI

NLP & Information Retrieval Class

Team Project

SoftNLP | 유승욱, 김민주, 정훈석

Thanks!