NLP& Information Retrieval Class

Team Project

SoftNLP

유승욱,김민주,정훈석

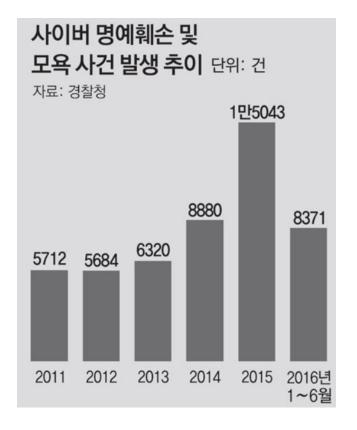
Tabloid Discriminator

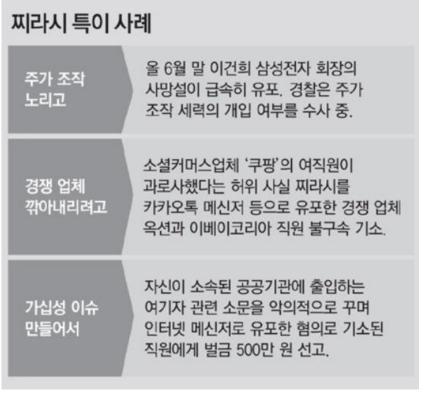
Contents

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

친구가 보낸 '카톡 찌라시' 퍼나르면… 나도 모르게 "범법자"

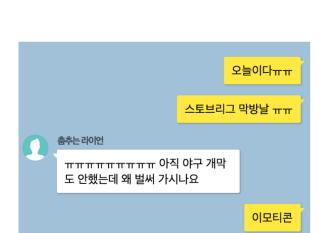
If you spread the Kakaotalk tabloid that your friend sent... You can be a criminal without your knowledge



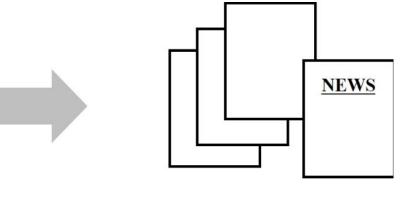




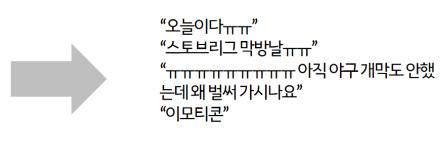
Latest Online News



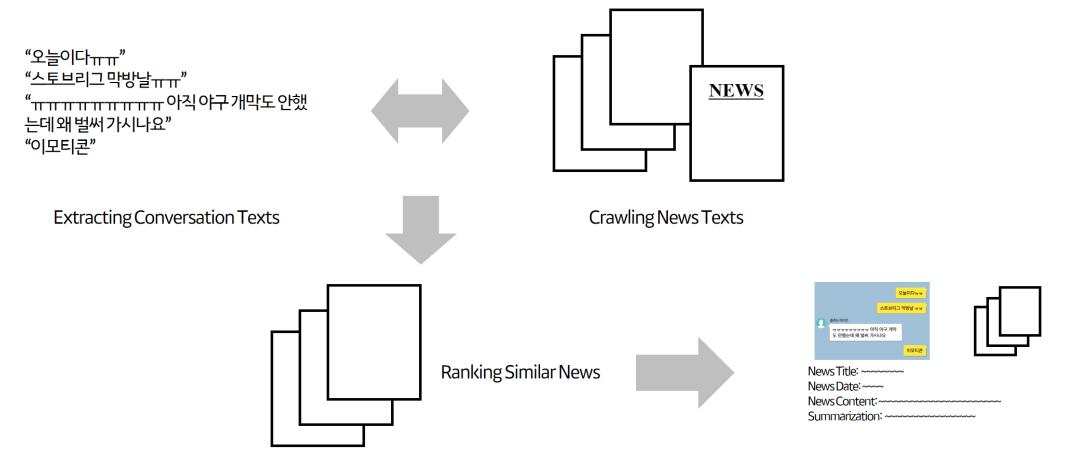
Kakaotalk Conversation Image



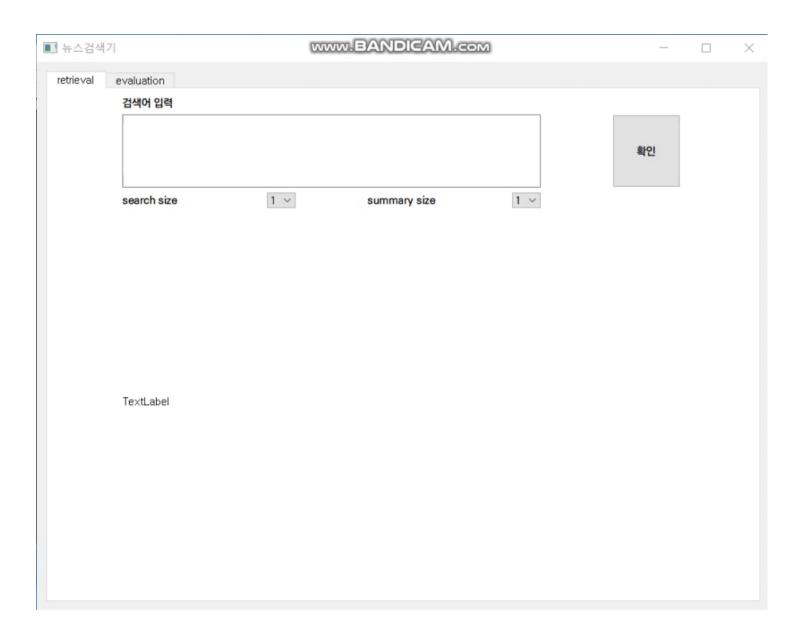
Crawling News Texts



Extracting Conversation Texts



Similar News Information, Similar News Summarization



Select some has categories that may lead to tabloid rumors

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



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



	제목	주소	본문	날짜/카테고리
0	훔친 차 몰다 사고 낸 고교생의식 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 달 아난 고교생 덜미	https://v.daum.net/v/20220501220916537	경찰 마크. 경향신문 자료사진 훔친 차량을 몰다 사고 를 낸 고교생이 의식을 잃은 동	20220501/society/affair
2	고교생, 훔친 차량 몰다 사고'의식불 명' 동승자 버리고 도주	https://v.daum.net/v/20220501220127479	훔친 차를 몰다 사고를 낸 고교생이 의식을 잃은 동승 자를 아무런 조처 없이 차량 밖	20220501/society/affair
3	고교생, 무면허로 훔친차 몰다 사고 후 도주동승자 '의식불명'	https://v.daum.net/v/20220501215259356	훔친 차를 몰다 사고를 낸 고교생이 의식을 잃은 동승 자를 아무런 조처 없이 차량 밖	20220501/society/affair
4	익산 태양광 발전시설 ESS서 화재2억 원 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 기획 전을 실시하고 관련 상품을 최	20220531/economic/consumer
69404	SSG닷컴, 6월 한 달간 '스마일클럽' 고 객 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 업사이클 링 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 방 송의 영향력이 감소하고, 주	20220531/economic/consumer
69407	횡성 국순당-칠성사이다 협업 '칠성막 사' 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

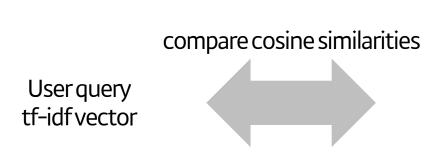
	제목	주소	본문	날짜/카테고리
0	훔친 차 몰다 사고 낸 고교생의식 잃은 동승자 버리고 도주	https://v.daum.net/v/20220501223102728	파주에서 무면허 상태로 훔친 차량을 몰다 사고를 낸 고교생이 의식 잃은 동송자를 버	20220501/society/affair
1	훔친차 몰다 사고낸 뒤 동승자 두고 달 아난 고교생 멀미	https://v.daum.net/v/20220501220916537	경찰 마크. 경향신문 자료사진 훔친 차량을 몰다 사고 를 낸 고교생이 의식을 잃은 동	20220501/society/affair
2	고교생, 훔친 차량 몰다 사고'의식불 명' 동승자 버리고 도주	https://v.daum.net/v/20220501220127479	훔친 차를 몰다 사고를 낸 고교생이 의식을 잃은 동승 자를 아무런 조처 없이 차량 밖	20220501/society/affair
3	고교생, 무면허로 훔친차 몰다 사고 후 도주동승자 '의식불명'	https://v.daum.net/v/20220501215259356	훔친 차를 몰다 사고를 낸 고교생이 의식을 잃은 동승 자를 아무런 조처 없이 차량 밖	20220501/society/affair
4	익산 태양광 발전시설 ESS서 화재2억 원 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, 리조트록·아웃도어룩 최대 83% 할인	https://v.daum.net/v/20220531060009417	W컨셉은 오는 6월 112일 여름 휴가를 콘셉트로 기획 전을 실시하고 관련 상품을 최	20220531/economic/consumer
69404	SSG닷컴, 6월 한 달간 '스마일클럽' 고 객 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 업사이클 링 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 방 송의 영향력이 감소하고, 주	20220531/economic/consumer
69407	횡성 국순당-칠성사이다 협업 '칠성막 사' 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	훔친 차 몰다 사고 낸 고교생의식 잃은 동 승자 버리고 도주	https://v.daum.net/v/20220501223102728	파주에서 무면허 상태로 훔친 차량을 몰다 사고를 낸 고교생이 의식 잃은 동 승자를 버	20220501/society/affair	['잃다', '도주', '몰다, 동승자', '내다', '훔치 다', '버리다',	['않다', '확인돼다', '위반', '주유소', '조 사', '고교생', '파주 시'
1	훔친차 몰다 사고낸 뒤 동승자 두고 달아 난 고교생 덜미	https://v.daum.net/v/20220501220916537	경찰 마크. 경향신문 자료 사진 훔친 차량을 몰다 사 고를 낸 고교생이 의식을 잃은 동	20220501/society/affair	['고교생', '덜미', '몰 다', '동승자', '사고내 다', '훔치다차', '달 아	['않다', '위반', '주유 소', '조사', '고교생', '파주시', '투숙객',
2	고교생, 홍친 차량 몰 다 사고…'의식불명' 동승자 버리고 도주	https://v.daum.net/v/20220501220127479	홍친 차를 몰다 사고를 낸 고교생이 의식을 잃은 동 승자를 아무런 조처 없이 차량 밖	20220501/society/affair	['고교생', "사고'의식 불명'", '도주', '몰다, '동승자', '차량', '	['않다', '위반', '주유 소', '조사', '고교생', '파주시', '투숙객',
3	고교생, 무면허로 훔 친차 올다 사고 후 도 주동승자 '의식불명'	https://v.daum.net/v/20220501215259356	홍친 차를 몰다 사고를 낸 고교생이 의식을 잃은 동 승자를 아무런 조저 없이 차량 밖	20220501/society/affair	['고교생', '무면혀', 몰다', '훔치다차', '도 주동승자', "의식불 명"	['타고', '않다', '미숙', '발견', '경기', '고양', '출구', '주차
4	익산 태양광 발전시 설 ESS서 화재2억 원 재산 피해	https://v.daum.net/v/20220501215023318	1일 오후 2시 50분께 전복 익산시 망성면 어량리의 한 태양광 발전설비 에너 지저장	20220501/society/affair	['피해', '재산', '발전 시설', '익산', '태양 광']	['오후', '현재', '망성 면', '불길', '경찰', '내 다', '꺼지다', '
69170	W컨셉, 리조트록·아 웃도어록 최대 83% 할인	https://v.daum.net/v/20220531060009417	W컨셉은 오는 6월 112일 여름 휴가를 콘셉트로 기 획전을 실시하고 관련 상 품을 최	20220531/economic/consumer	['할인', '최대']	['콘셉트', '브랜드', '휴양지의', '매주', '데 이즈데이즈', '헤어', '
69171	SSG닷컴, 6월 한 달 간 '스마일클럽' 고객 대상 프로모션 실시	https://v.daum.net/v/20220531080001392	뉴스1 SSG닷컴은 6월 한 달간 스마일클럽 가입 고 객 대상 특별 프로모션을 열	20220531/economic/consumer	["'스마일클럽'", '대 상', '하다', '실시', '프 로모션', '고객']	['프리미엄', '결제하 다', '등록되다', '연 다', '브랜드', '월요 일', '
69172	업사이클링 제품 300 여개 한눈에	https://v.daum.net/v/20220531030621223	30일 경기 성남시 현대백 화점 판교점에서 업사이를 링 제품을 한데 모은 하우 투 리그	20220531/economic/consumer	['업사이클링', '한눈'. '제품']	['행사', '경기', '현대 백화점', '한국환경산 업협회', '브랜드', '제 품'
69173	중년층 이탈에 TV 홈쇼핑 업체들 'TV' 탈 출 작전'	https://v.daum.net/v/20220531030218046	TV홈쇼핑 업체들의 탈 TV 가 가숙화하고 있다. TV 방 송의 영향력이 감소하고, 주	20220531/economic/consumer	['이탈', '탈출', '중년 충', "작전'"]	['설립하다', '않다', '진출하다', '중장년 충', '따르다', '브랜 드', '
69174	횡성 국순당-칠성사 이다 협업 '칠성막사' 내일 출시	https://v.daum.net/v/20220531001042395	70년 전통 막걸리와 사이 다가 만나면 어떤 맛을 낼 까. 횡성 향토기업 국순당 이 칠성	20220531/economic/consumer	["'칠성막사", '내일', 횡성', '협업', '국순당 칠성사', '출시']	['전통', '말다', '만나 다', '제품', '구현하 다', '느끼다', '막거 리'
9175 ı	rows × 6 columns	afterte	vt propro	ressing ar	nd tolon	nizatio

after text preprocessing and tokenization

Implement news retrieval system by learn tf-idf



1 ^{훔친치}	물다 사고 낸 고교생의식 잃은 동승자 버리고 도주 다 물다 사고낸 뒤 동승자 두고 달 아난 고교생 멀미	https://v.daum.net/v/20220501223102728 https://v.daum.net/v/20220501220916537	파주에서 무면허 상태로 훔친 차량을 몰다 사고를 낸 고교생이 의식 잃은 동승자를 버	20220501/society/affair
,		https://u.doum.pot/s/20220504220046527		
7.77		https://v.uaum.neuv/20220501220916557	경찰 마크. 경향신문 자료사진 훔친 차량을 몰다 사고 를 낸 고교생이 의식을 잃은 동	20220501/society/affair
2	생, 훔친 차량 몰다 사고'의식불 명' 동승자 버리고 도주	https://v.daum.net/v/20220501220127479	훔친 차를 몰다 사고를 낸 고교생이 의식을 잃은 등승 자를 아무런 조처 없이 차량 밖	20220501/society/affair
3 고교생	ti, 무면허로 훔친차 몰다 사고 후 도주동승자 '의식불명'	https://v.daum.net/v/20220501215259356	훔친 차를 몰다 사고를 낸 고교생이 의식을 잃은 동승 자를 아무런 조처 없이 차량 밖	20220501/society/affair
4 익산 태	l양광 발전시설 ESS서 화재2억 원 재산 피해	https://v.daum.net/v/20220501215023318	1일 오후 2시 50분께 전북 익산시 망성면 어량리의 한 태양광 발전설비 에너지저장	20220501/society/affair
₹				
69403 W	컨셉, 리조트룩·아웃도어룩 최대 83% 할인	https://v.daum.net/v/20220531060009417	W컨셉은 오는 6월 112일 여름 휴가를 콘셉트로 기획 전을 실시하고 관련 상품을 최	20220531/economic/consumer
69404 SSG딧	단컴, 6월 한 달간 '스마일클럽' 고 객 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 업사이클 링 제품을 한데 모은 하우 투 리그	20220531/economic/consumer
69406 중년충	충 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 '칠성막 사' 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

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	remov	ve one news of th	acpairs	날짜/카테고리	제목토큰	본문토큰
	훔친 차 몰다 사고 낸 고교생의식 잃은 동 승자 버리고 도주	https://v.daum.net/v/20220501223102728	파주에서 무면허 상태로 훔친 차량을 몰다 사고를 낸 고교생이 의식 잃은 동 승자를 버	202205	01/society/affair		['오전', '주차돼다', '잃다', '파주', '크게', '인근', '위반', '
1	훔친차 몰다 사고낸 뒤 동승자 두고 달아 난 고교생 덜미	https://v.daum.net/v/20220501220916537	경찰 마크. 경향신문 자료 사진 훔친 차량을 몰다 사 고를 낸 고교생이 의식을 잃은 동	202205	501/society/affair	['덜미', '사고내다', '고교생', '몰다', '두 다', '훔치다차', '동 승자	['오전', '고양지역', '주차돼다', '잃다', '크게', '인근', '알려 지다
2	고교생, 훔친 차량 몰 다 사고'의식불명' 동승자 버리고 도주	nttps://v.daum.net/v/20220501220127479	훔친 차를 몰다 사고를 낸 고교생이 의식을 잃은 동 승자를 아무런 조처 없이 차량 밖	202205	01/society/affair	['도주', "사고'의식 불명"', '고교생', '몰 다', '차량', '동승자', '	['오전', '주차돼다', '잃다', '크게', '인근', '말하다', '위반',
3	고교생, 무면허로 훔 친차 몰다 사고 후 도 주동승자 '의식불명'	nttps://v.daum.net/v/20220501215259356	훔친 차를 몰다 사고를 낸 고교생이 의식을 잃은 동 승자를 아무런 조처 없이 차량 밖	202205	01/society/affair	['무면허', '고교생', '몰다', '도주동승자', '사고', '훔치다차', ""	['택시', '오전', '없이', '화중로', '병원', '숨 다', '있다', '잃
4		nen near-similar news appo https://v.daum.net/v/20220501215023318	eared like this전복 의산시 망성면 어량리의 한 태양광 발전설비 에너 지저장	202205	01/society/affair	['익산', '피해', '발전 시설', '재산', '태양 광']	['잡다', '진압', '망성 면', '불길', '전기', '배 터리', '정확하다',

Retrieval Examples

input query: '검찰청장 수사'

```
default
                                링 크
1 https://v.daum.net/v/20220502161620767
                                                            오 전 서울
 https://v.daum.net/v/20220502104017060
 https://v.daum.net/v/20220502104021065
                                            울산경찰청장이 2일 오전 서울
  https://v.daum.net/v/20220520164645528
                                                       20일 오후 서울
  https://v.daum.net/v/20220520164617511 이정수 서울중앙지검장이 20일 오후 서울 서초구 서울중앙지검에서
when similar threshold = 0.9
                                 링 크
1 https://v.daum.net/v/20220502161620767 김지용 대검찰청
                                                   사 부 장 이 지 난 4월 29일 서울
2 https://v.daum.net/v/20220502104017060 황운하 전
                                              산 경 찰 청 장 이 2일 오 전 서울
                                                                    서 초 구
                                                       20일 오후 서
3 https://v.daum.net/v/20220520164645528 이정수 서
                                                        20일 오후 서울
  https://v.daum.net/v/20220520164657532 이정수
  https://v.daum.net/v/20220520164610505 이정수 서울중앙지검장이 20일 오후 서울 서초구 서울중앙지검에서
when similar threshold = 0.8
                                링 크
1 https://v.daum.net/v/20220502161620767 김지용
                                                                          대 검 찰 청
 https://v.daum.net/v/20220502104017060 황운하
                                                       20일 오후
                                                               서 울
 https://v.daum.net/v/20220520164645528
                                                                    서 초 구
  https://v.daum.net/v/20220502161720826
                                   경찰이 더불어민주당
                                                    대선 후보였던 이재명
  https://v.daum.net/v/20220513100554611 이성윤 서울고등검찰청
                                                      검사장이 13일 오전 서울 서초구 중앙지방법원에서
```

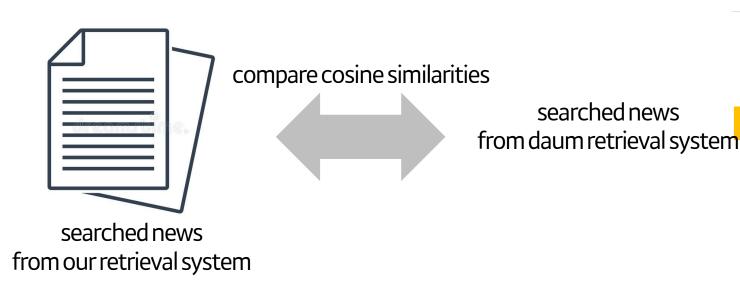
Evaluation for retrieved news

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

Evaluation for retrieved news

crawling searched daum news when given an input query





4. Process: Summary

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 B 양이 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. 춘천 지법 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 저질 렀다. 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 서울중앙 지법으로 들어가고 있다. 뉴시스 우리은행 회삿돈을 빼돌린 혐의를 받는 직원의 동생도 공범으로 함께 구속됐다. 서울중앙지법 허정인 판사 는 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

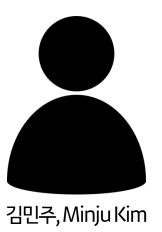


	검색어 입력	
	루나 CEO 장소 적힌 경찰 내부보고서 유출	확인
	search size 4 V summary size	3 ~
	페이스북 폭락 사태가 벌어진 한국산 코인 루나테라USD 발행업체 테 신변보호를 요청한 가문데 이들의 개인정보가 담긴 내부 보고서가 외 경찰에 따르면 서울경찰청은 권 대표 아내의 신변보호 내용이 담긴 비 진상조사에 착수했다. 권도형 테라폼랩스 대표 <u>링크</u>	부로 유출돼 경찰이 조사에 착수했다. 14일
	14일 경찰에 따르면 서울경찰청은 권도형 테라폼랩스 대표 배우자의 유출된 사건의 사실 관계를 파악 중이다. 기사내용 요약 경찰 구체적(발행사 대표의 가족이 경찰에 신변보호를 요첨한 가운데 이동의 개인 경찰이 조사에 착수했다. 전날 권 대표의 배우자는 한 남성 A씨가 집 경찰에 신고했다 링크	인 유출 배경 등 확인 중 암호화폐 루나테라 성보가 담긴 내부 보고서가 외부로 유출돼
	14일 경찰에 따르면 서울경찰청은 권 대표 가족의 개인정보가 담긴 L 사실관계를 확인 중이다. 권도형 테라폼랩스 대표 뉴스1 99 폭락한 일 대표 가족이 경찰의 신변 보호 대상자로 지정된 가운데, 해당 내용이 대표 배무자는 즉시 경찰에 신고했고 긴급 신변 보호를 요청해 신변 ! 링크	발호화폐 루나테라 발행사 테라폼랩스 권도형 담긴 경찰 내부 보고서가 유출됐다. 이에 권
	12일 서울 종로구 글로벌센터에서 열린 CPTPP회견에서 참석자들이 있다. 링크	CPTPP 가입 중단이 적힌 손팻말을 듣고

5. Team Roles

5. Team Roles







- Crawling for data gathering & evaluation
- -Text preprocessing and tokenization
- Remove similar news from searched results
- Retrieval news through tf-idf
- -Evaluation for retrieved news

- Summarization for retrieved news
- Service visualization through GUI

NLP& Information Retrieval Class

Team Project

SoftNLP 유승욱, 김민주, 정훈석

Thanks!