

서론

목표 설정

본론

데이터 수집

데이터 정리

데이터 탐색

데이터 분석

결론

요약 정리

국민 청원 텍스트 분석

20141118019 송대훈

목표 설정

서론

목표 설정

본론

데이터 수집

데이터 정리

데이터 탐색

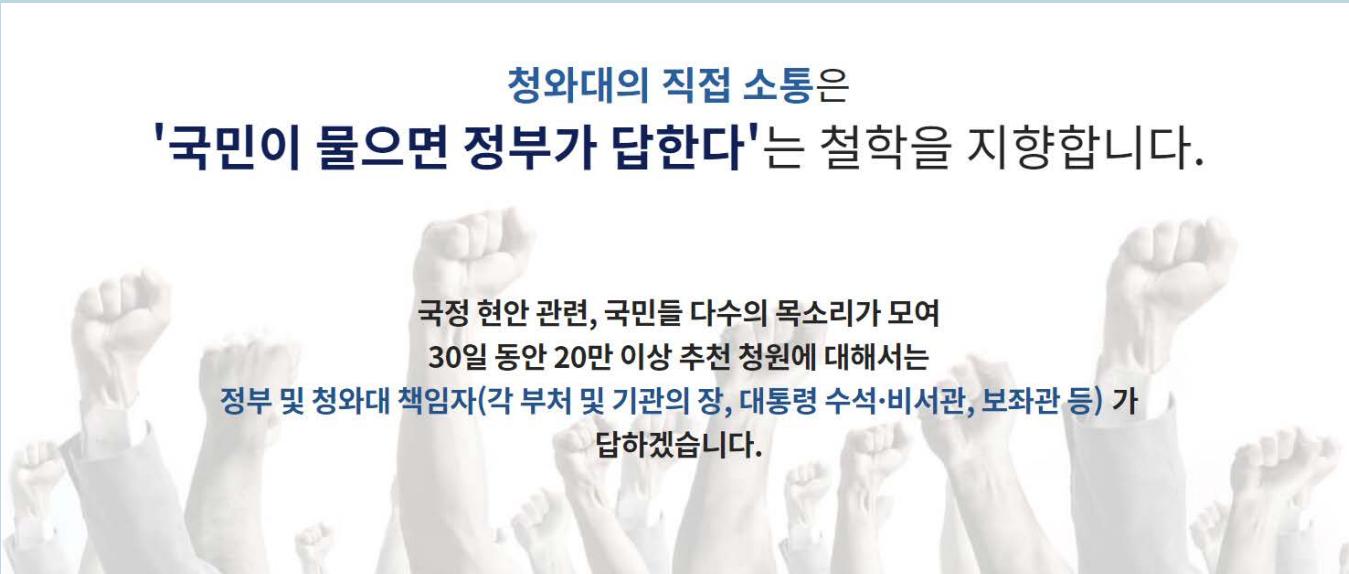
데이터 분석

결론

요약 정리

국민 청원

- 정부와 국민의 직접 소통 창구로 기능(<https://www1.president.go.kr/petitions>)



목표 설정

서론

목표 설정

본론

데이터 수집

데이터 정리

데이터 탐색

데이터 분석

결론

요약 정리

국민 청원

- 다양한 분야에 걸쳐서 사회 문제를 표출 가능

· 청원 분야별 보기

전체	정치개혁	외교/통일/국방	일자리	미래
성장동력	농산어촌	보건복지	육아/교육	안전/환경
저출산/고령화대책	행정	반려동물	교통/건축/국토	경제민주화
인권/성평등	문화/예술/체육/언론	기타		

목표 설정

서론

목표 설정

연구 동향

본론

데이터 수집

데이터 정리

데이터 탐색

데이터 분석

결론

요약 정리

- 국민 청원에 대한 이론적, 정보 기술적 접근
 - 김애니, 정소희, 최현빈, 김현희. (2018). 회귀분석과 텍스트마이닝을 활용한 미세먼지 비상저감조치의 실효성 및 국민청원 분석. *한국정보처리학회*, 7(11), 427-434.
 - 김주환, 허예림. (2018). 청와대 국민청원 게시판 분석을 통한 사회 이슈 개진 경향에 관한 연구. *한국PR 학회 학술대회*, 2018(11), 67.
 - 김찬우. (2019). 국민 청원 데이터를 통해 본 주요 개혁 이슈. *예술인문사회융합멀티미디어논문지*, 9(2), 823-832.
 - 김태은, 모은정. (2018). 청와대 국민청원 사이트에 참여하는 이용자의 심리적 요인에 관한 연구: 어떤 심리적 속성이 참여 의도를 높이는가. *한국정책학회 춘계학술발표논문집*, 2018, 1-20.
 - 조애리, 김희진, 윤재형. (2018). 효율적인 정치참여를 위한 국민청원 데이터 시각화 서비스 제안. *한국디자인학회 학술발표대회 논문집*, 2018(11), 292-293.

목표 설정

서론

목표 설정

본론

데이터 수집

데이터 정리

데이터 탐색

데이터 분석

결론

요약 정리

분석의 목표

- 국민 청원은 다른 SNS에 비해서 사회 문제에 밀접한 연관성을 가지는 데이터 보유
 - 국민 청원에 올리는 글은 상대적으로 noise가 적음
- 머신러닝을 통해 청원 내용에 맞는 분야를 자동 분류
 - 데이터와 라벨을 모두 가지고 있어 학습하기에 용이
- 특정 분야에 속한 청원 내용 중 현재 관심을 가지는 주제 파악
 - 토픽모델링을 활용해 청원 내용의 세부 주제 구성 파악

데이터 수집

서론

목표 설정

본론

데이터 수집

데이터 정리

데이터 탐색

데이터 분석

결론

요약 정리

대상 데이터 파싱

- 개별 청원 웹 페이지 URL은 "<https://www1.president.go.kr/petitions/고유번호>" 구성
 - 최신의 청원일수록 고유번호는 큰 숫자를 가짐
- HTML 파싱을 통해 청원상태, 제목, 참여인원, 카테고리, 청원시작/종료일, 청원내용 수집
 - 청원인은 *로 개인정보를 식별할 수 없으므로 수집 대상에서 제외
 - 청원동의 댓글은 주로 '동의합니다'이외의 정보가 없으므로 수집 대상에서 제외

데이터 수집

서론

목표 설정

대상 데이터 파싱

본론

데이터 수집
데이터 정리
데이터 탐색
데이터 분석

결론

요약 정리

The screenshot shows a petition page from the President's Office website. The main title is "1500만 등산인을 위한 등산앱 루가를 살려주세요" (Save the Lukka app for 15 million hikers). It displays 140 participants and includes sections for petition details, content, and signatures.

Petition Details:

- Category: 운동/여행/레저/환경
- Petition Date: 2019-05-31
- Petition End Date: 2019-06-30
- Petitioner: naver-***

Petition Content:

1500만 등산인을 위한 등산앱 루가를 살려주세요

2015년 산림청장을 수상하고
2016년 국무총리상을 수상하였으나
김성현은 수상에 눈물을 뜨지 않고
오늘 등산인들을 위해 일해온 루카앱이
수익성이 없다는 것만으로 1500만 등산인을 등지고 사장되는 위기에 처했습니다

5월1일자로 운영이 중단되자
등산인들의 어려움이 어떤까만이 아닙니다
인구의 1/4이 넘는 등산장을 위한 루카앱을 살려주세요

국민의 세금은 이번에 투자되어야 합니다

Signatures: 청원등록 140명

Side Panel:

- Top 5 Petitions:
 - 김무성 전 의원을 대간위로 다스려주십시오.
 - 축구클럽에 축구한다고 차운에 대여 보낸 아이가…
 - *** 불법 영웅 소비, 범죄 가짜 MVP 고객수사 확장…
 - 4.7월 10일 외국인 전기차와 단연 유통 절화 천원
 - 이전에서 빙어진 동물수간사건에 대한 강력한 처…
- Recent Petitions:
 - 대통령께서 <세월호참사 특별수사단>설치와…
 - 11.15조정기자진 피해배상 및 저작재산 학법제정…
 - 김여의 살인기자재개 신언론으로 및 관련자 일정수…
 - 정부아이돌봄서비스 아동돌보미 양육아동별 감액…
 - 소방공무원을 국가적으로 전환해주세오
 - 여러나라 살해한 총수운전사에게 송방이 차량이…
 - 비행수사 충장으로 블구라와 교육당사 명예까…
 - 안전하세요. 종인 윤자연입니다.
 - 경시***, 경상*** 외 ***에서 뇌출혈는지 호시부탁…
 - 중증연한 여성 대상 약을 범위 차별과 ***을 비롯한…
- Today's Trend:

건강하게 모래사는 소망

세계 2번째 EU 환경미래스토 등재
정부 R&D 2025년까지 연구 개발 지원
신약개발 지원 확대 및 혁신 기관 구축
창업 지원 강화 및 혁신 기관 협력 확장

데이터 수집

서론

목표 설정

본론

데이터 수집

데이터 정리

데이터 탐색

데이터 분석

결론

요약 정리

[코드] A_Crawling.py

- 12,582건의 청원 데이터 수집
 - 20,000건을 시작했지만, BeautifulSoup을 통한 파싱 과정에서 오류 발생한 데이터 예외 처리
 - 청원번호 579799인 4월 29일 청원 ~ 청원번호 559801인 3월 13일 청원 수집
- 수집된 데이터는 CSV 형식으로 petitions.csv에 저장
 - NO : 청원번호
 - TITLE : 청원제목
 - COUNT : 청원 참여 인원
 - STATE : 청원 상태
 - CATEGORY : 청원 분야 / DATE_START / DATE_END / CONTENT
 - DATE_START/DATE_END : 청원 시작일/종료일
 - CONTENT : 청원 내용

데이터 정리

서론

목표 설정

본론

데이터 수집

데이터 정리

데이터 탐색

데이터 분석

결론

요약 정리

한글 텍스트 정리

- 한글 텍스트를 정리하기 위해서 KoNLPy의 Komoran 형태소 분석기 사용
 - Komoran은 한국어 위키피디아를 통해 학습되며 HMM 기반의 품사 태깅 지원
- 데이터 탐색 및 분석 과정에서 한글 명사만을 자질(feature)로 선정
 - 딕셔너리와 불용어를 사전에 정의해서 한글 명사 처리의 정확성 높임

데이터 정리

서론

목표 설정

딕셔너리 정보 추가

단어	품사	단어	품사
가부장	NNG(일반명사)	사형제도	NNG(일반명사)
가부장제	NNG(일반명사)	소상공인	NNG(일반명사)
국민청원	NNG(일반명사)	심신미약	NNG(일반명사)
기간제	NNG(일반명사)	자한당	NNG(일반명사)
기초수급	NNG(일반명사)	적폐청산	NNG(일반명사)
대체복무	NNG(일반명사)	정규직	NNG(일반명사)
문재인	NNP(고유명사)	종북	NNG(일반명사)
발암물질	NNG(일반명사)	출퇴근	NNG(일반명사)
보이스피싱	NNG(일반명사)	포스코	NNP(고유명사)
불우이웃	NNG(일반명사)	훈밥	NNG(일반명사)
사법농단	NNG(일반명사)		

데이터 정리

서론

목표 설정

본론

데이터 수집

데이터 정리

데이터 탐색

데이터 분석

결론

요약 정리

불용어 제거

불용어	불용어	불용어	불용어
!!	때문	실상	저도
가나	라고	아냐	절대
가부	만원	요새	제가
경우	바로	요즘	조가
관련	보이	우리들의	지금
국민	부여	이번	천원
기간	부탁	인물	포스
니다	사람	입장	한곳
당장	사실	정도	해먹
등등	생각	자신	해주시
각종 한 글자 단어			

데이터 탐색

서론

목표 설정

본론

데이터 수집

데이터 정리

데이터 탐색

데이터 분석

결론

요약 정리

단어 네트워크 분석

- 공기 분석을 통해 단어와 단어의 연관성을 네트워크로 시각화
 - 시각화 도구인 gephi 활용
- 공기 분석을 통해 연관된 단어를 파악하고, 특정 분야에 대한 세부 주제 파악 가능
 - 가장 높은 빈도의 분야인 '정치개혁'에 대해서 네트워크 분석

데이터 탐색

서론

목표 설정

본론

데이터 수집

데이터 정리

데이터 탐색

데이터 분석

결론

요약 정리

[코드] B_CoWordFrequencyAnalysis.py

- '정치개혁' 분야의 1,500건의 청원 데이터를 분석
 - 명사 단어의 빈도 분석과 명사 단어에 대한 공기 분석 진행
- 명사에 대한 빈도 분석 결과를 petition_frequency.csv에 저장
- 공기 분석의 결과를 petition_coword.graphml에 저장
 - 6,200개의 Node와 478,108개의 Edge로 구성

단어	품사
대통령	490
나라	465
정부	334
국회의원	321
대한민국	261
세금	259
사건	257
부동산	253
장관	244
문재인	229

데이터 탐색

서론

목표 설정

[gephi] politic.gephi

본론

데이터 수집

데이터 정리

데이터 탐색

데이터 분석

결론

요약 정리

- petition_coword.graphml 파일을 바탕으로 네트워크 그래프 시각화
 - 단어와 단어의 공기 빈도에 대한 네트워크 분석
- Density, Degree, Modularity 정보 분석
 - Density는 그래프의 밀도로 0.025의 값을 지님
 - Degree는 Node가 연결된 정도로 평균 154.228의 값을 지님 (긴꼬리를 가진 그래프)
 - Modularity는 데이터를 몇 개의 Module로 나눠주며 총 7개의 Module 생성
- Edge Weigh가 10.0이상이고, Degree가 100이상인 데이터로 필터링
 - Node : 2392/6200 (38.58%)
 - Edge : 23,630/478,108 (4.94%)

데이터 탐색

서론

목표 설정

[gephi] politic.gephi

본론

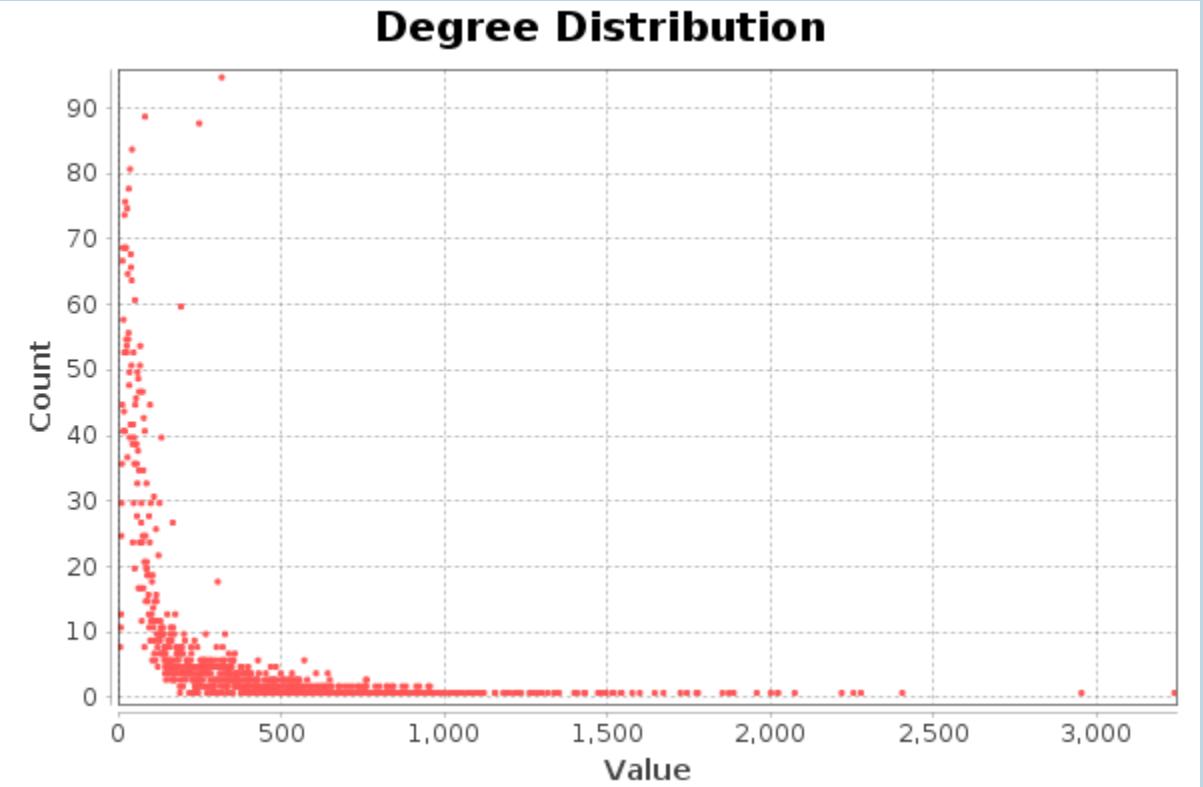
데이터 수집

데이터 정리

데이터 탐색

데이터 분석

- Degree 정보



결론

요약 정리

데이터 탐색

서론

목표 설정

[gephi] politic.gephi

본론

데이터 수집

데이터 정리

데이터 탐색

데이터 분석

- Modularity 분포 정보



데이터 탐색

서론

목표 설정

본론

데이터 수집

데이터 정리

데이터 탐색

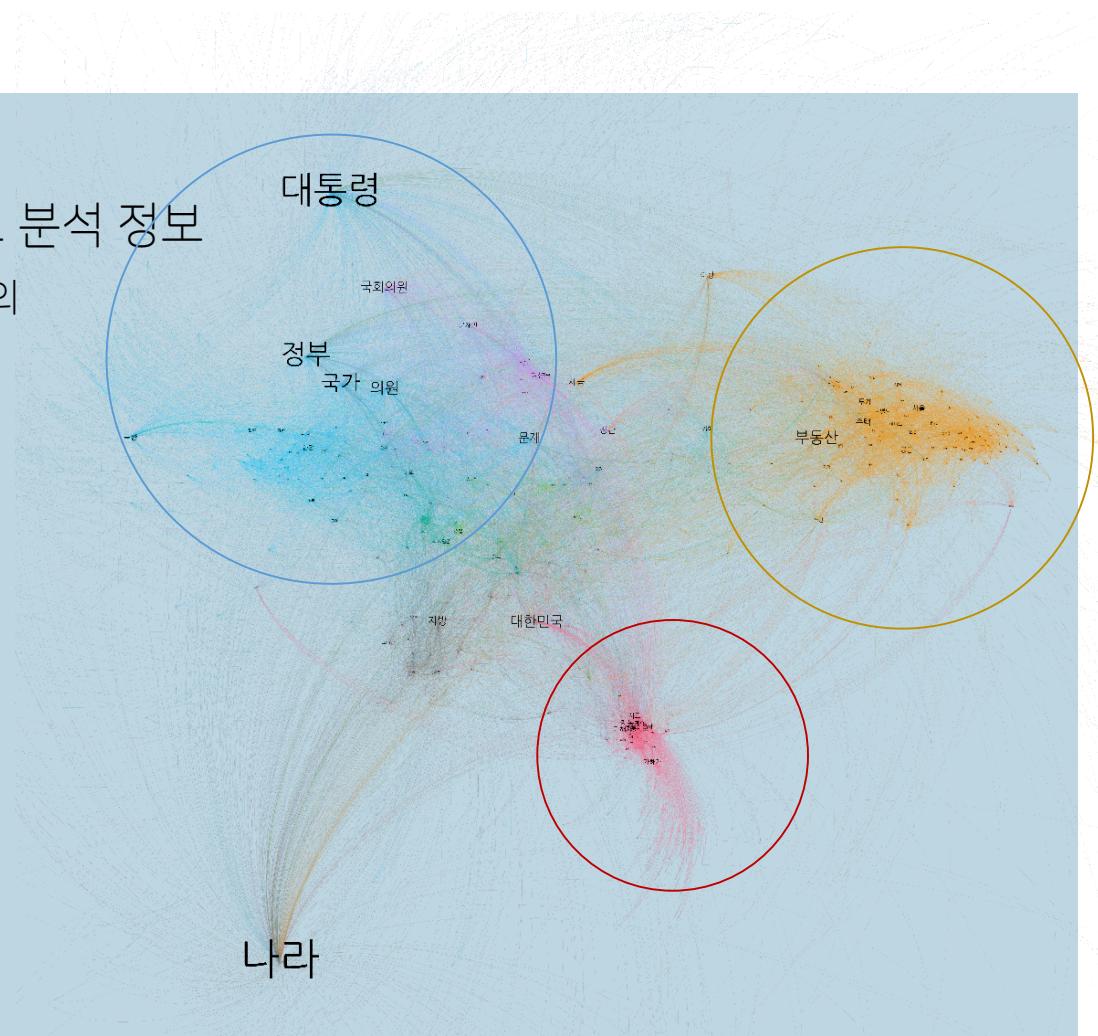
데이터 분석

결론

요약 정리

[gephi] politic.gephi

- Modularity를 바탕으로 한 네트워크 분석 정보
 - Filtering을 거쳤을 때 크게 3 덩어리의 Community 주목
 - 의회 민주주의 Community > 부동산 Community > 범죄와 관련된 Community
 - 실제 Topic Modeling 결과와 비교



문헌 자동 분류

서론

목표 설정

본론

데이터 수집

데이터 정리

데이터 탐색

데이터 분석

결론

요약 정리

기계학습을 통한 문헌 자동 분류

- 지도학습을 통한 분류기 생성
 - 청원 내용을 입력으로하고, 청원의 분야를 라벨로 사용
- 청원 분야에서 주요한 키워드 추출
 - Chi-Square 검정을 통해 분야에 맞는 핵심 키워드 추출
- 분야별 최대 청원수를 정해서 특정 분야가 독점적으로 학습에 사용되는 것 방지
- 10-fold Validation을 거쳐서 얻은 분야별 Precision을 사용해 분류기의 성능 평가

문헌 자동 분류

서론

목표 설정

[코드] C_ClassificationNB.py

본론

데이터 수집

데이터 정리

데이터 탐색

데이터 분석

- Naïve Bayes를 통해 생성한 분류기
 - 청원 내용을 입력으로하고, 청원의 분야를 라벨로 사용
- 분야별 최대 청원수를 조정함에 따라서 결과가 달라짐 (300 | 500 | 1,000)

결론

요약 정리

문헌 자동 분류

서론

목표 설정

본론

데이터 수집

데이터 정리

데이터 탐색

데이터 분석

결론

요약 정리

[코드] C_ClassificationNB.py

- 최대 청원수에 따른 분야별 청원 빈도 (300)

분야	청원 빈도	분야	청원 빈도
경제민주화	188	안전/환경	300
교통/건축/국토	300	외교/통일/국방	300
기타	300	육아/교육	300
농산어촌	40	인권/성평등	300
문화/예술/체육/언론	300	일자리	294
미래	300	저출산/고령화대책	52
반려동물	61	정치개혁	300
보건복지	300	행정	300
성장동력	79	합계	4,014

문헌 자동 분류

서론

목표 설정

본론

데이터 수집

데이터 정리

데이터 탐색

데이터 분석

결론

요약 정리

[코드] C_ClassificationNB.py

- 최대 청원수에 따른 분야별 핵심 단어 (300)

분야	핵심 단어(Unigram)	핵심 단어(Bigram)
경제민주화	횡령 배임 주주 소액 공매도	거래 정지 정지 감사 배임 거래 횡령 배임 소액 주주
교통/건축/국토	교통 기관사 부동산 투기 집값	청량리 서교 교통 불편 조정 지역 집값 안정 부동산 투기
기타	개인정보 교회 소방관 댓글 수괴	독립 기구 가격 가격 주사파 정권 댓글 조작 조작 대통령
농산어촌	농민 농지 농촌 농어촌 농산물	돈이 없어 관리 지역 인권 보장 행정 처리 서울 원전
문화/예술/체육/언론	유튜브 차태현 연예인 방송 선수	연예인 연예인 연예인 방송 사회 물의 차태현 김준호 방송 하차
미래	어의 구더기 구역질 순경 미래	자살 사건 문재인 탄핵 이희진 부모 대통령 응원 서울 원전

문헌 자동 분류

서론

목표 설정

본론

데이터 수집

데이터 정리

데이터 탐색

데이터 분석

결론

요약 정리

[코드] C_ClassificationNB.py

- 최대 청원수에 따른 분야별 핵심 단어 (300)

분야	핵심 단어(Unigram)	핵심 단어(Bigram)
반려동물	식용 고양이 반려 강아지 동물	가짜 유공자 국회 의원 동물 생명 동물 학대 반려 동물
보건복지	가입자 치료 국민연금 병원 환자	생리대 가격 공무원 연금 일시 반환 연금 수급 지역 가입자
성장동력	단일 내수 바이오 저장 특허	후보자 내정 신용 불량자 세금 서민 주도 성장 세금 안내면
안전/환경	관리자 공기 해상 플라스틱 미세 먼지	국토부 직원 노동부 국토부 해상 원전 관리자 작업 안전 관리자
외교/통일/국방	북미 군대 군인 김정은 북한	북미 회담 북한 주민 김정은 북한 수석 대변인 서해 수호
육아/교육	학교 수업 어린이집 학생 교육	어린이집 학대 우리나라 교육 수업 시간 보육 교사 아동 학대

문헌 자동 분류

서론

목표 설정

본론

데이터 수집

데이터 정리

데이터 탐색

데이터 분석

결론

요약 정리

[코드] C_ClassificationNB.py

- 최대 청원수에 따른 분야별 핵심 단어 (300)

분야	핵심 단어(Unigram)	핵심 단어(Bigram)
인권/성평등	여자 여성 남성 남자 인권	남성 가족 인권 유린 학생 화장 영장 기각 수사 청원
일자리	취업 채용 정규 외국인 일자리	청년 일자리 불법 외국인 채용 공고 불법체류 추방 정규 전환
저출산/고령화대책	시험관 출산 신혼부부 결혼 저출산	아이 아이 아들 아들 저출산 문제 출산 장려 결혼 결혼
정치개혁	정당 박영선 잔당 장관 해산	자유 한국당 장관 후보자 문제 정권 이명박 박근혜 박근혜 잔당
행정	통화 판사 유흥업소 공무원 경찰	공무원 공무원 불가 통화 개인 회생 구속 수사 통화 불가

문헌 자동 분류

서론

목표 설정

본론

데이터 수집

데이터 정리

데이터 탐색

결론

요약 정리

[코드] C_ClassificationNB.py

- 최대 청원수에 따른 분야별 Precision (300) [alpha = 3.0]

문헌 자동 분류

서론

목표 설정

본론

데이터 수집

데이터 정리

데이터 탐색

데이터 분석

결론

요약 정리

[코드] C_ClassificationNB.py

- 최대 청원수에 따른 분야별 Precision (300) [alpha = 3.0]

	1	2	3	4	5	6	7	8	9	10
안전/환경	0.3000	0.5000	0.5000	0.6333	0.5667	0.4667	0.4333	0.0667	0.6000	0.4667
외교/통일/국방	0.6667	0.4667	0.6000	0.5333	0.7000	0.7667	0.6000	0.7333	0.6667	0.7667
육아/교육	0.8333	0.7667	0.8333	0.9333	0.7667	0.7333	0.7667	0.9000	0.9667	0.8333
인권/성평등	0.4333	0.3667	0.2667	0.3000	0.4667	0.6333	0.4667	0.4333	0.5000	0.5333
일자리	0.6333	0.7667	0.7333	0.6333	0.5172	0.7586	0.7241	0.7241	0.5862	0.6897
저출산/고령화 대책	-	-	-	-	-	-	-	-	-	-
정치개혁	0.5333	0.4000	0.5667	0.4667	0.5333	0.6333	0.5667	0.4667	0.5667	0.4333
행정	0.1667	0.1667	0.1667	0.1667	0.4000	0.2000	0.3000	0.2667	0.1667	0.2333

문헌 자동 분류

서론

목표 설정

[코드] C_ClassificationNB.py

본론

데이터 수집

데이터 정리

데이터 탐색

데이터 분석

결론

요약 정리

- 최대 청원수에 따른 분야별 청원 빈도 (500)

분야	청원 빈도	분야	청원 빈도
경제민주화	188	안전/환경	453
교통/건축/국토	373	외교/통일/국방	372
기타	500	육아/교육	304
농산어촌	40	인권/성평등	500
문화/예술/체육/언론	356	일자리	294
미래	433	저출산/고령화대책	52
반려동물	61	정치개혁	500
보건복지	336	행정	303
성장동력	79	합계	5,144

문헌 자동 분류

서론

목표 설정

본론

데이터 수집

데이터 정리

데이터 탐색

데이터 분석

결론

요약 정리

[코드] C_ClassificationNB.py

- 최대 청원수에 따른 분야별 핵심 단어 (500)

분야	핵심 단어(Unigram)	핵심 단어(Bigram)
경제민주화	횡령 배임 소액 주주 공매도	거래 정지 정지 감사 배임 거래 횡령 배임 소액 주주
교통/건축/국토	기관사 교통 부동산 투기 집값	집값 안정 조정 지역 무단 횡단 공시 지가 현실화 부동산 투기
기타	셨다운 조작 문가 수괴 댓글	주사파 정권 문재인 수괴 사기 문재인 댓글 조작 조작 대통령
농산어촌	농민 농지 농촌 농어촌 농산물	토지 매수 인권 보장 행정 처리 쓰레기 소각 서울 원전
문화/예술/체육/언론	수신료 유튜브 방송 연예인 선수	문제 연예인 방송 하차 연예인 연예인 해외 사이트 사회 물의
미래	군부 비상 기운 미래 부역자	문재인 탄핵 청원 청원 북한 신경 대통령 응원 서울 원전

문헌 자동 분류

서론

목표 설정

본론

데이터 수집

데이터 정리

데이터 탐색

데이터 분석

결론

요약 정리

[코드] C_ClassificationNB.py

- 최대 청원수에 따른 분야별 핵심 단어 (500)

분야	핵심 단어(Unigram)	핵심 단어(Bigram)
반려동물	식용 고양이 반려 강아지 동물	보호 강화 눈물 호소 동물 생명 동물 학대 반려 동물
보건복지	가입자 치료 국민연금 병원 환자	공무원 연금 생리대 가격 연금 수급 지역 가입자 일시 반환
성장동력	매출 바이오 저장 우주 특허	세금 서민 후보 단일 후보자 내정 주 도 성장 세금 안내면
안전/환경	플라스틱 조두순 경유 공기 미세 먼지	국토부 직원 안전 관리 노동부 국토부 관리자 작업 안전 관리자
외교/통일/국방	북미 군대 김정은 비핵화 북한	북미 회담 주한 미군 서해 수호 통일 부 장관 수석 대변인
육아/교육	교사 학생 수업 교육 어린이집	어린이집 학대 수업 시간 우리나라 교육 보육 교사 아동 학대

문헌 자동 분류

서론

목표 설정

[코드] C_ClassificationNB.py

본론

데이터 수집

데이터 정리

데이터 탐색

데이터 분석

결론

요약 정리

- 최대 청원수에 따른 분야별 핵심 단어 (500)

분야	핵심 단어(Unigram)	핵심 단어(Bigram)
인권/성평등	사건 인권 수사 여성 장자연	사건 수사 여성 가족 장자연 사건 장자연 수사 수사 청원
일자리	채용 취업 외국인 정규 일자리	청년 일자리 불법 외국인 채용 공고 불법체류 추방 정규 전환
저출산/고령화대책	시험관 출산 신혼부부 결혼 저출산	아들 아들 아이 아이 저출산 문제 출산 장려 결혼 결혼
정치개혁	후보자 박영선 청문회 야당 장관	치인 노조 서민 정치인 프랑스 서민 장관 후보자 박근혜 잔당
행정	통화 행정 포항시 신종 공무원	개인 회생 단속 경찰 불가 통화 구속 수사 통화 불가

문헌 자동 분류

서론

목표 설정

본론

데이터 수집

데이터 정리

데이터 탐색

결론

요약 정리

[코드] C_ClassificationNB.py

- 최대 청원수에 따른 분야별 Precision (500) [alpha = 3.0]

문헌 자동 분류

서론

목표 설정

본론

데이터 수집

데이터 정리

데이터 탐색

데이터 분석

결론

요약 정리

[코드] C_ClassificationNB.py

- 최대 청원수에 따른 분야별 Precision (500) [alpha = 3.0]

	1	2	3	4	5	6	7	8	9	10
안전/환경	0.5217	0.5435	0.6522	0.5778	0.6444	0.7111	0.6000	0.7111	0.5556	0.4667
외교/통일/국방	0.3947	0.2632	0.3784	0.5405	0.3784	0.4054	0.3784	0.3784	0.5676	0.4595
육아/교육	0.6129	0.4839	0.4194	0.5806	0.4000	0.6000	0.5000	0.6000	0.5333	0.5667
인권/성평등	0.7000	0.4800	0.7400	0.7600	0.7800	0.7200	0.8600	0.7400	0.6800	0.9000
일자리	0.2667	0.3000	0.3000	0.2333	0.2069	0.3448	0.3793	0.3103	0.2069	0.2759
저출산/고령화 대책	-	-	-	-	-	-	-	-	-	-
정치개혁	0.6888	0.5600	0.6600	0.7000	0.7400	0.6800	0.7200	0.7000	0.6200	0.6400
행정	0.1667	0.1667	-	0.0333	0.0333	0.0333	0.0333	-	-	-

문헌 자동 분류

서론

목표 설정

[코드] C_ClassificationNB.py

본론

데이터 수집

데이터 정리

데이터 탐색

데이터 분석

결론

요약 정리

- 최대 청원수에 따른 분야별 청원 빈도 (1,000)

분야	청원 빈도	분야	청원 빈도
경제민주화	188	안전/환경	453
교통/건축/국토	373	외교/통일/국방	372
기타	1,000	육아/교육	304
농산어촌	40	인권/성평등	628
문화/예술/체육/언론	356	일자리	294
미래	433	저출산/고령화대책	52
반려동물	61	정치개혁	1,000
보건복지	336	행정	303
성장동력	79	합계	6,272

문헌 자동 분류

서론

목표 설정

본론

데이터 수집

데이터 정리

데이터 탐색

데이터 분석

결론

요약 정리

[코드] C_ClassificationNB.py

- 최대 청원수에 따른 분야별 핵심 단어 (1,000)

분야	핵심 단어(Unigram)	핵심 단어(Bigram)
경제민주화	횡령 배임 소액 주주 공매도	거래 정지 정지 감사 횡령 배임 배임 거래 소액 주주
교통/건축/국토	기관사 교통 부동산 투기 집값	집값 안정 무단 횡단 조정 지역 공시 지가 현실화 부동산 투기
기타	댓글 구글 페이스북 차단 유튜브	유튜브 유튜브 구글 차단 유튜브 페이스북 댓글 조작 조작 대통령
농산어촌	농민 농지 농산물 농촌 농어촌	토지 매수 인권 보장 행정 처리 쓰레기 소각 서울 원전
문화/예술/체육/언론	출연료 수신료 연예인 방송 선수	물의 연예인 연예인 방송 방송 하차 연예인 연예인 사회 물의
미래	재갈 비상 기운 미래 부역자	연금 개혁 탄핵 탄핵 북한 신경 대통령 응원 서울 원전

문헌 자동 분류

서론

목표 설정

[코드] C_ClassificationNB.py

- 최대 청원수에 따른 분야별 핵심 단어 (1,000)

분야	핵심 단어(Unigram)	핵심 단어(Bigram)
반려동물	안락사 고양이 반려 강아지 동물	보호 강화 눈물 호소 동물 생명 동물 학대 반려 동물
보건복지	가입자 치료 병원 국민연금 환자	자격 심사 공무원 연금 연금 수급 지역 가입자 일시 반환
성장동력	저장 독재자 난방 우주 특허	후보 단일 대표 대통령 주도 성장 운영 자금 세금 안내면
안전/환경	경유 안전 플라스틱 공기 미세먼지	국토부 직원 안전 관리 노동부 국토부 관리자 작업 안전 관리자
외교/통일/국방	군인 군대 김정은 비핵화 북한	나라 청춘 통일부 장관 북미 회담 주한 미군 수석 대변인
육아/교육	과목 수업 교사 교육 어린이집	교사 어린이집 어린이집 학대 우리나라 교육 보육 교사 아동 학대

문헌 자동 분류

서론

목표 설정

본론

데이터 수집

데이터 정리

데이터 탐색

데이터 분석

결론

요약 정리

[코드] C_ClassificationNB.py

- 최대 청원수에 따른 분야별 핵심 단어 (1,000)

분야	핵심 단어(Unigram)	핵심 단어(Bigram)
인권/성평등	수사 인권 사건 여성 장자연	여성 가족 수사 연장 수사 청원 장자연 수사 장자연 사건
일자리	채용 취업 정규 외국인 일자리	청년 일자리 불법 외국인 채용 공고 불법체류 추방 정규 전환
저출산/고령화대책	시험관 출산 신혼부부 결혼 저출산	아들 아들 아이 아이 저출산 문제 출산 장려 결혼 결혼
정치개혁	야당 잔당 정당 장관 국회의원	프랑스 서민 정치인 노조 서민 정치인 이명박 박근혜 박근혜 잔당
행정	책임자 통화 행정 포항시 공무원	공무원 공무원 유착 의혹 구속 수사 불가 통화 통화 불가

문헌 자동 분류

서론

목표 설정

본론

데이터 수집

데이터 정리

데이터 탐색

결론

요약 정리

[코드] C_ClassificationNB.py

- 최대 청원수에 따른 분야별 Precision (1,000) [alpha = 3.0]

문헌 자동 분류

서론

목표 설정

본론

데이터 수집

데이터 정리

데이터 탐색

데이터 분석

결론

요약 정리

[코드] C_ClassificationNB.py

- 최대 청원수에 따른 분야별 Precision (1,000) [alpha = 3.0]

문헌 자동 분류

서론

목표 설정

본론

데이터 수집

데이터 정리

데이터 탐색

데이터 분석

결론

요약 정리

[코드] C_ClassificationNB.py

- 주제 조정에 따른 Precision 결과 (500)

분야	청원 빈도	분야	청원 빈도
정치개혁	500	인권/성평등	500
안전/환경	453	합계	1,453

- 최대 청원수에 따른 분야별 핵심 단어 (1,000)

분야	핵심 단어(Unigram)	핵심 단어(Bigram)
정치개혁	대변인 후보자 정권 부동산 장관	장관 후보 장관 임명 박근혜 잔당 부동산 투기 장관 후보자
안전/환경	원전 중국 공기 안전 미세먼지	중국 미세먼지 조두순 출소 건설 현장 지진 피해 안전 관리자
인권/성평등	인권 남자 미세먼지 여성 장자연	연장 수사 수사 연장 장자연 사건 수사 청원 장자연 수사

문헌 자동 분류

서론

목표 설정

본론

데이터 수집

데이터 정리

데이터 탐색

데이터 분석

결론

요약 정리

[코드] C_ClassificationNB.py && D_ClassificationSVM.py

- 주제 조정에 따른 Precision 결과 (500) [alpha = 30] by Naïve Bayes

	1	2	3	4	5	6	7	8	9	10
정치개혁	0.7800	0.6600	0.7400	0.8000	0.8400	0.8000	0.8200	0.8200	0.8000	0.8200
안전/환경	0.6087	0.6304	0.7826	0.6000	0.7778	0.7556	0.6667	0.7333	0.6444	0.4889
인권/성평등	0.8600	0.6800	0.8200	0.8800	0.8400	0.8400	0.9000	0.8200	0.8800	0.9400

- 주제 조정에 따른 Precision 결과 (500) by Support Vector Machine

	1	2	3	4	5	6	7	8	9	10
정치개혁	0.8800	0.6600	0.7400	0.8600	0.8400	0.8000	0.8000	0.8400	0.8400	0.8600
안전/환경	0.5870	0.6957	0.8478	0.6444	0.7778	0.7000	0.8600	0.7778	0.6889	0.5778
인권/성평등	0.8200	0.5400	0.7600	0.8400	0.6600	0.8400	0.6400	0.8200	0.7600	0.9200

토픽 모델링

서론

목표 설정

LDA를 통한 토픽 추출

본론

데이터 수집

데이터 정리

데이터 탐색

데이터 분석

- 청원 수가 많은 분야에 대해서 세부적인 주제 파악
 - 정치개혁 분야와 기타 분야를 대상으로 분석
- 토픽의 수를 조정하면서 데이터 분석(3, 5, 7)
 - 가장 세부 분야를 잘 설명해주는 토픽의 수로 조정

결론

요약 정리

토픽 모델링

서론

목표 설정

[코드] E_TopicModeling.py

- 기타 분야에 대한 1,152개의 텍스트를 대상으로 LDA 적용

Topic(3)	Labeling	Frequency
사건 경찰 수사 청원 처벌 조사 나라 대한민국	경/검찰 수사 관련 청원	465
대통령 나라 문재인 정부 대출 전화 문제 개인	문재인 정부 관련 청원	487
유튜브 차단 사이트 구글 페이스북 연금 비정규직 해외	노동 및 인터넷 관련 청원	200

본론

데이터 수집

데이터 정리

데이터 탐색

데이터 분석

결론

요약 정리

토픽 모델링

서론

목표 설정

본론

데이터 수집

데이터 정리

데이터 탐색

데이터 분석

결론

요약 정리

[코드] E_TopicModeling.py

- 기타 분야에 대한 1,152개의 텍스트를 대상으로 LDA 적용

Topic(5)	Labeling	Frequency
사건 수사 경찰 대통령 청원 대한민국 처벌 나라	경/검찰 수사 관련 청원	393
대출 게임 문재인 사기 대통령 정부 전화 서민	금융 사기 관련 청원	286
유튜브 차단 사이트 구글 페이스북 연금 비정규직 실행	인터넷 관련 청원	183
나라 병원 경찰 정권 대통령 보수 정부 수술	의료 관련 청원	141
회사 시간 신고 문제 조사 기업 계약 상황	기업 관련 청원	149

토픽 모델링

서론

목표 설정

[코드] E_TopicModeling.py

본론

데이터 수집

데이터 정리

데이터 탐색

데이터 분석

결론

요약 정리

- 기타 분야에 대한 1,152개의 텍스트를 대상으로 LDA 적용

Topic(7)	Labeling	Frequency
학생 학교 시간 청원 가해자 조두순 안녕하세요 피해자	미성년자 범죄 관련 청원	158
대출 전화 서민 정부 사기 개인 은행 제도	금융 사기 관련 청원	190
유튜브 차단 사이트 구글 페이스북 연금 비정규직 실행	인터넷 관련 청원	170
병원 정권 수술 중국 기록 보수 경찰 제대혈	의료 관련 청원	63
회사 계약 시간 개신교 신고 전화 국내 공감	기업 관련 청원	99
대통령 문재인 나라 대한민국 국가 문제 게임 탄핵	대통령 관련 청원	251
사건 경찰 수사 조사 처벌 대한민국 장자연 승리	성매매 범죄 관련 청원	221

토픽 모델링

서론

목표 설정

[코드] E_TopicModeling.py

본론

데이터 수집

데이터 정리

데이터 탐색

데이터 분석

- 정치개혁 분야에 대한 1,745개의 텍스트를 대상으로 LDA 적용

Topic(3)	Labeling	Frequency
국회의원 대통령 나라 의원 국가 정부 자유 대표	정당정치 관련 청원	719
부동산 세금 투기 주택 서울 공급 서민 상승	부동산 관련 청원	195
대통령 사건 나라 대한민국 정권 장관 수사 문재인	대통령 관련 청원	831

결론

요약 정리

토픽 모델링

서론

목표 설정

본론

데이터 수집

데이터 정리

데이터 탐색

데이터 분석

결론

요약 정리

[코드] E_TopicModeling.py

- 정치개혁 분야에 대한 1,745개의 텍스트를 대상으로 LDA 적용

Topic(5)	Labeling	Frequency
자유 국가 의원 대통령 대한민국 정부 나라 한국당	정당정치 관련 청원	387
부동산 주택 투기 서울 공급 상승 서민 그린벨트	부동산 관련 청원	151
대통령 나라 사건 대한민국 문재인 정권 정부 수사	대통령 관련 청원	626
경찰 검찰 대통령 국회의원 자한당 수사 공수 사건	검/경찰 관련 청원	239
국회의원 세금 장관 대표 나라 여성 가족 후보자	장관 관련 청원	342

토픽 모델링

서론

목표 설정

본론

데이터 수집

데이터 정리

데이터 탐색

데이터 분석

결론

요약 정리

[코드] E_TopicModeling.py

- 정치개혁 분야에 대한 1,745개의 텍스트를 대상으로 LDA 적용

Topic(7)	Labeling	Frequency
자유 한국당 청원 구글 의원 정부 민주당 일본	정당정치 관련 청원	274
부동산 주택 서울 공급 투기 상승 그린벨트 집값	부동산 관련 청원	131
대통령 나라 사건 문재인 정권 대한민국 수사 정부	대통령 관련 청원	564
경찰 검찰 대통령 자한당 수사 정부 국내 사건	검/경찰 관련 청원	194
장관 후보자 나라 정부 청문회 투기 부동산 인사	장관 청문회 관련 청원	212
국회의원 세금 여성 대표 가족 의원 국회 비례	국회의원 관련 청원	194
대한민국 공수 피해자 정치인 친일파 쓰레기 의원 공개	국회의원 관련 청원	176

정리

서론

목표 설정

본론

데이터 수집

데이터 정리

데이터 탐색

데이터 분석

결론

요약 정리

데이터 분석 평가

- 국민청원 데이터를 자동 분류 하기 위해서는 모든 분야를 걸쳐서 고르게 많은 데이터 필요
 - 특정 분야에 집중될 경우 편향되어서 분류를 하는 경우가 상당히 많음
 - 분류기 자체 보다는 양질의 데이터를 사용하는 것이 분류 성능을 높이는 데 기여
- 청원의 내용을 세부적인 주제로 파악 가능
 - 뉴스 데이터와 연계해서 적용할 경우 세부적인 주제 파악 가능할 것이라 예상