



# 국민청원 데이터 분석

3조(상부상조) 김진욱 | 김정민 | 박주현



# INDEX

- |                      |                        |                            |
|----------------------|------------------------|----------------------------|
| 1. 주제 선정 배경          | 4. 문제 해결 및 중간 계획 수행 현황 | 8. Document Classification |
| 2. 국민청원 데이터의 특성 및 수집 | 5. Text Rank 타이틀 분석    | 9. Topic Modeling          |
| 3. 선행 연구 분석          | 6. Sentiment Analysis  | 10. 요약 및 결론                |
|                      | 7. Document Clustering | 11. 코드 수정사항                |

# 1. 주제 선정 배경

# 국민이 물으면 정부가 답한다

국민청원이 새롭게 찾아왔습니다

“

100명 이상 사전동의를 받은 청원만 청원 게시판에 공개됩니다.  
내가 올린 청원을 ‘내 청원 보기’에서 확인하실 수 있습니다.  
답변이 어려운 청원, 숨김처리 되는 청원 등 FAQ를 통해 궁금증을 풀어드립니다.

국민과 함께 만드는 국민 청원 더 많은 참여와 관심 부탁드립니다.

청 와 대

<https://www1.president.go.kr/articles/5872>

공군 이어 육군도… “여단장이 폭언했다” 국민청원 YTN PICK | 8시간 전 네이버뉴스 |

지난 11일 공군에서 발생한 황제 군 복무‘ 의혹이 국민청원 게시판을 통해 공론화된 데 이어 또 군 관련 청원 글이 게재됐다. 지난 16일 육군 지상작전사령부 1군단 사령부 1공병여단 소속 A 일병은 청와대 국민청원...

- ↳ **青 국민청원에 또 군 소원수리… “여단장이 폭언했다”** 서울신문 PICK | 11시간 전 네이버뉴스
- ↳ **국민청원에 또 군 소원수리… “여단장이 폭언했다”** 연합뉴스 PICK | 11시간 전 네이버뉴스
- ↳ **공군 ‘황제 병사’ 이어… 육군 여단장 ‘폭언’** 매일신문 PICK | 7시간 전 네이버뉴스
- ↳ **국민청원 또… “공군 황제 병사” 이어 육군 여단장 폭언** 머니투데이 PICK | 8시간 전 네이버뉴스

관련뉴스 13건 전체보기 >

“칼치기 차량에 버스 쿵… 19살 동생 전신마비” 언니의 청원

중앙일보 PICK | 10시간 전 네이버뉴스 |

진주에서 시내버스에 타고 있던 여고생이 교통사고로 전신마비가 된 사고와 관련해 피해 가족이 청와대 국민청원을 올렸다. 사과와 함께 차별을 받을 수 있도록 교통사고 처벌법을...

- ↳ “칼치기 차량에 동생 전신마비”… 진주 여고생 부산일보 PICK | 9시간 전 네이버뉴스
- ↳ “여고생 교통사고로 전신마비”… 국민청원 올라 KBS | 12시간 전 네이버뉴스
- ↳ “동생은 전신마비인데 가해자는 사과 없었…” 서울경제 PICK | 8시간 전 네이버뉴스
- ↳ “칼치기 차량에 동생은 사지마비”… 진주 여고생 동아일보 PICK | 8시간 전 네이버뉴스

관련뉴스 9건 전체보기 >

‘등록금 반환’ 국민청원 빛발… 나들리라 교육부 비난도

연합뉴스 PICK | 1일 전 네이버뉴스 |

당정이 신종 코로나바이러스 감염증(코로나19) 사태에 따른 대학생들의 등록금 반환 요구와 관련해 예산 반영 등 대책을 마련하기로 한 가운데 청와대 국민청원에도 등록금 반환을 요구하는 목소리가 크다. 17일 청와대...

- ↳ “우리 등록금 반환 정말 어렵나” 青 국민청원 아주경제 | 1일 전

전신마비 여학생 가족 청와대 국민 청원

경남일보 | 30분 전 네이버뉴스 |

청와대 국민청원 홈페이지에 가해자 엄벌과 함께 교통사고 처벌법 개정을 촉구하는 청원을 올렸다. A씨는 “교통사고로 사지마비가 된 제 동생의 억울함을 알리고, 사고 후 6개월이 되도록 단 한 번도 진심 어린 사과를 하지...

- ↳ 권익위 “황제복무‘ 青 청원인, 공의신고자 인정 어렵다” 문화일보 PICK | 1일 전 네이버뉴스 |
- ↳ 이른바 ‘황제 군 복무’ 의혹을 청와대 국민청원 게시판에 폭로한 청원인이 공의신고자로 인정받기 어렵다는 유권해석이 나왔다. 17일 미래통합당 지성호 의원실에 따르면 국민권익위원회는 ‘황제 군 복무’ 국민...
- ↳ [레이더P] “공군 황제복무 青 게시판에 올라…” 매일경제 PICK | 1일 전 네이버뉴스
- ↳ 권익위 “황제복무‘ 青 청원인, 공의신고자… 연합뉴스 PICK | 1일 전 네이버뉴스

“청주 투기조정대상지역 재고해달라” 청와대 국민청원 올라

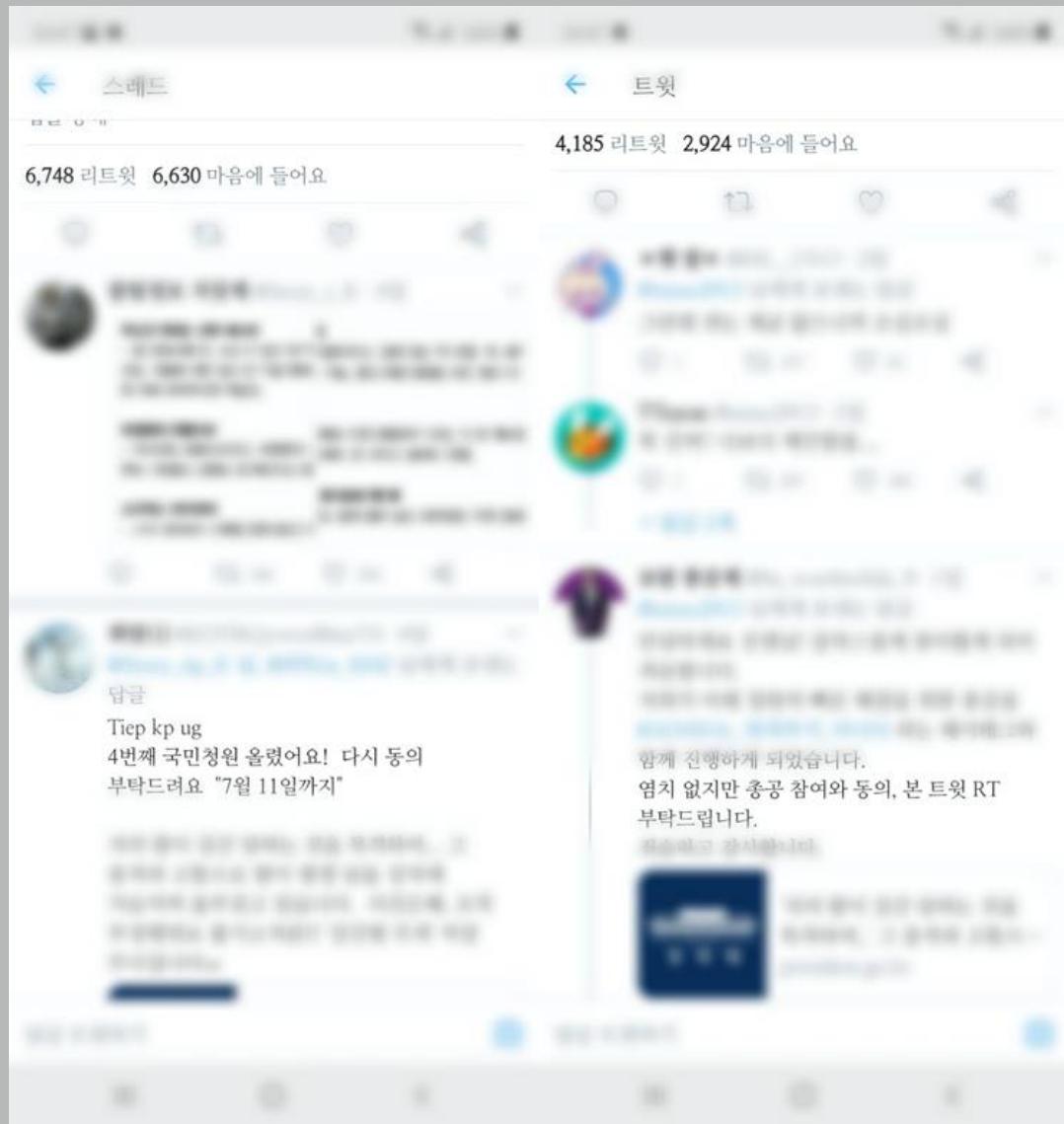
뉴스1 | 3시간 전 네이버뉴스 |

정부가 발표한 ‘6.17 부동산 대책’에 청주시가 조정대상지역에 포함되자 “청주 조정지역을 재고해 달라”는 내용의 국민 청원이 청와대 국민청원에 올랐다. 18일 청와대 홈페이지에 올라온 해당 청원은 이날 오후 7시 기준...

- ↳ 청와대 국민청원 “청주 투기조정대상지역… 시사뉴스 | 3시간 전

<국민청원 관련 기사들>

청와대 국민 청원은 2017년 신설된 이후 사회 이슈의 척도로 활용될 만큼 그 중요도와 영향력이 커졌다. 사회적 이슈가 발생할 경우 그에 대한 해결을 촉구하는 청원을 그 즉시 찾아볼 수 있으며, 반대로 청원의 내용이 사회적 이슈가 되기도 한다.



<트위터 청원 동참 요청 예시>

트위터 등의 sns에서도 그 영향력을 확인할 수 있는데, 사람들 간에 공유가 많이 된 게시물의 경우 어김없이 청원에 동참해 줄 것을 호소하는 댓글이 달린다. 인스타그램 등에서도 사람들이 주변인에게 청원 동참을 부탁하는 글을 어렵지 않게 찾아볼 수 있다. 이는 사람들이 정부에 자신의 뜻을 전달할 가장 확실한 방법 중 하나로 국민청원을 인식하고 있음을 보여준다.

이에 따라 국민의 의견 및 사회 이슈를 확인할 수 있는 데이터로서 국민청원 데이터가 매우 중요하며, 분석에 의미가 있을 것으로 판단했다. 또한, 막연하게 먼 내용이 아닌 사람들에게 가깝고 주변에서 쉽게 접할 수 있는 데이터라는 점에서 더욱 분석하는 의미가 있는 데이터라고 생각한다.

데이터의 정리가 비교적 잘 되어 있다는 점 역시 주제 선정의 이유가 되었다. Sns, 블로그 등의 정제되지 않은 데이터에 비해 국민청원 데이터는 주제별로 분류가 이루어져 있으며 제목, 청원 수 등의 비교적 정리된 데이터가 존재하기 때문에 분석에 용이할 것으로 생각되었다.

## 2. 국민청원 데이터의 특징 및 수집

## · 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

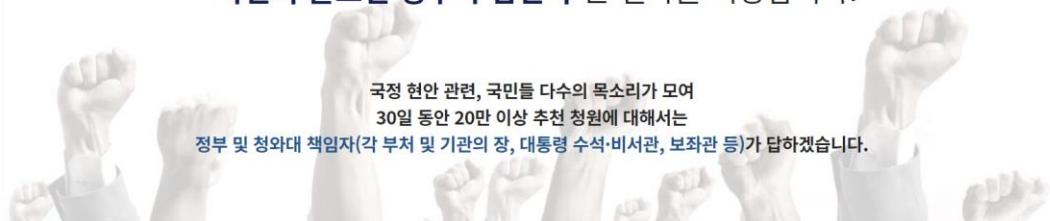
인권/성평등

문화/예술/체육/언론

기타

### 청와대의 직접 소통은

'국민이 물으면 정부가 답한다'는 철학을 지향합니다.



국정 현안 관련, 국민들 다수의 목소리가 모여  
30일 동안 20만 이상 추천 청원에 대해서는  
정부 및 청와대 책임자(각 부처 및 기관의 장, 대통령 수석·비서관, 보좌관 등)가 답하겠습니다.

지금 청원하기

내 청원 보기

공개된 청원 답변은 청와대 홈페이지 > 국민청원 > 답변된 청원 또는 청와대 트위터, 페이스북, 유튜브를 통해 확인하실 수 있습니다.  
청원 관련 문의 : webmaster@president.go.kr

#### ① 이런 청원은 삭제·숨김 처리될 수 있습니다

동일한 내용으로 중복 게시된 청원은 최초 1개 청원만 남기고 '숨김' 처리 또는 삭제될 수 있습니다.

욕설 및 비속어를 사용한 청원은 관리자에 의해 삭제 또는 일부 내용 '숨김' 처리될 수 있습니다.

폭력적, 선정적, 또는 특정 집단에 대한 혐오 표현 등 청소년에게 유해한 내용을 담은 청원은 관리자에 의해 삭제될 수 있습니다.

개인정보, 허위사실, 타인의 명예를 해손하는 내용이 포함된 청원은 관리자에 의해 삭제 또는 일부 내용 '숨김' 처리될 수 있습니다.

#### ② 이런 청원은 답변이 어려울 수 있습니다

재판이 진행 중이거나, 입법부·사법부의 고유 권한과 관련한 내용으로 삼권분립의 정신을 해손할 소지가 있는 청원에는 답변이 어려울 수 있습니다.

지방자치단체 고유 업무에 해당하는 내용 등 중앙 정부의 역할과 책임 범위를 벗어난 경우 답변이 어려울 수 있습니다.

청원 주요 내용이 허위사실로 밝혀진 경우 답변이 어려울 수 있습니다.

인종, 국적, 종교, 나이, 지역, 장애, 성별 등 특성과 관련 있는 개인, 집단에 대한 차별 및 비하 등 위험적 요소가 포함된 청원에는 답변이 어려울 수 있습니다.

청와대 및 정부에 대한 민원·제안 및 공익신고·고발 등을 국민권익위원회의 국민신문고([www.epeople.go.kr](http://www.epeople.go.kr))를 이용해 주시길 부탁드립니다.

국민청원 FAQ 전체보기

국민청원분야는 전체 뿐 아니라 17개의 카테고리별로 나누어져 있어 주제별 데이터의 분석도 가능

## 답변 대기 중인 청원

분류	제목	청원 만료일	참여인원
육아/교육	저의 아들이 6년간다닌어린이집에서 남자원장으로부터 3년간 성폭행을 당했습니다	20.03.08	271,123명
육아/교육	저희 25개월딸이 초등학생 5학년에게 성폭행을 당했습니다	20.04.19	533,883명
인권/성평등	초등학생을 성폭행하고 찍은 불법촬영물로 협박해 금품을 갈취하여 인생을 망가뜨린 고등학생의 엄중한 처벌을 요구합니다	20.04.19	354,260명
기타	수출용 코로나19 진단키트 이름을 독도로 해주세요	20.04.24	385,617명
인권/성평등	<텔레그램 n번방 사건 특별조사팀을 서지현 검사를 필두로 한 80%이상 여성 조사팀으로 만들어 주십시오>	20.04.23	286,101명
인권/성평등	N번방 담당판사 오덕식을 판사자리에 반대,자격박탈을 청원합니다.	20.04.26	466,900명
교통/건축/국토	민식이 법 개정을 청원합니다.	20.04.22	354,857명
안전/환경	박사방 회원 중 여아살해모의한 공익근무요원 신상공개를 원합니다.	20.04.28	519,948명
인권/성평등	"오늘 너 킀(KILL)한다"라며 술을 먹이고 제 말을 할풍 강간한 미성년자들을 고발합니다.	20.04.28	400,474명
기타	렌트카 훔쳐 사망사고를 낸 10대 엄중 처벌해주세요	20.05.02	1,007,040명

진행 중 청원		만료된 청원	
전체 추천순 TOP 5			
번호	분류	제목	청원 만료일
1. <a href="#">안전/환경</a>	텔레그램 n번방 용의자 신상공개 및 포토라인 세워주세요	2020-04-17 2,715,626명	
2. <a href="#">안전/환경</a>	텔레그램 가입자 친원의 신상공개를 원합니다	2020-04-19 2,026,252명	
3. <a href="#">정치개혁</a>	자유 한국당 정당해산 청원	2019-05-22 1,831,900명	
4. <a href="#">기타</a>	문재인 대통령님을 응원 합니다!	2020-03-27 1,504,597명	
5. <a href="#">정치개혁</a>	문재인 대통령 탄핵을 촉구합니다.	2020-03-05 1,469,023명	

전체 목록		지금 청원하기	최신순 보기 ▾
번호	분류	제목	청원 만료일
831. <a href="#">인권/성평등</a>	저희 아파트 경비아저씨의 억울함을 풀어주세요...	2020-06-10 382,174명	
830. <a href="#">육아/교육</a>	등교 개학 시기를 미루어주시기 바랍니다.	2020-05-24 214,804명	
829. <a href="#">육아/교육</a>	울산 초등학교 1학년 아이에게 펜티벌기 속제내고 학생 사진에 '섹시팬티', '공주님 수줍게 클리어', '매력적이고 섹시한 00'이라고 성희롱한 남교사를 파면해 주세요.	2020-05-28 208,376명	
828. <a href="#">교통/건축/국토</a>	서울 강남구 재건축 지역에 탈북자 세터인 아파트 의무비율로 법제화 시켜주세요.	2020-05-16 146,708명	
827. <a href="#">성장동력</a>	다목적 방사광 가속기'는 꼭 호남권에 유치되어야 합니다.	2020-05-27 107,253명	
826. <a href="#">보건복지</a>	정부, 美日한국전 참전국에 마스크지원시 일본 지원 반대합니다	2020-05-20 81,193명	
825. <a href="#">문화/예술/체육/언론</a>	SBS 거짓보도에 공식사과 요청 및 이** 기자 처벌요청	2020-06-03 76,623명	

진행 중 청원		만료된 청원	
전체 추천순 TOP 5			
번호	분류	제목	청원 만료일
1. <a href="#">인권/성평등</a>	저희 아파트 경비아저씨의 억울함을 풀어주세요...	2020-06-10 382,174명	
2. <a href="#">육아/교육</a>	등교 개학 시기를 미루어주시기 바랍니다.	2020-05-24 214,804명	
3. <a href="#">육아/교육</a>	울산 초등학교 1학년 아이에게 펜티벌기 속제내고 학생 사진에 '섹시팬티', '공주님 수줍게 클리어', '매력적이고 섹시한 00'이라고 성희롱한 남교사를 파면해 주세요.	2020-05-28 208,376명	
4. <a href="#">교통/건축/국토</a>	서울 강남구 재건축 지역에 탈북자 세터인 아파트 의무비율로 법제화 시켜주세요.	2020-05-16 146,708명	
5. <a href="#">성장동력</a>	다목적 방사광 가속기'는 꼭 호남권에 유치되어야 합니다.	2020-05-27 107,253명	

전체 목록		지금 청원하기	최신순 보기 ▾
번호	분류	제목	청원 만료일
831. <a href="#">인권/성평등</a>	저희 아파트 경비아저씨의 억울함을 풀어주세요...	2020-06-10 382,174명	
830. <a href="#">육아/교육</a>	등교 개학 시기를 미루어주시기 바랍니다.	2020-05-24 214,804명	
829. <a href="#">육아/교육</a>	울산 초등학교 1학년 아이에게 펜티벌기 속제내고 학생 사진에 '섹시팬티', '공주님 수줍게 클리어', '매력적이고 섹시한 00'이라고 성희롱한 남교사를 파면해 주세요.	2020-05-28 208,376명	
828. <a href="#">교통/건축/국토</a>	서울 강남구 재건축 지역에 탈북자 세터인 아파트 의무비율로 법제화 시켜주세요.	2020-05-16 146,708명	
827. <a href="#">성장동력</a>	다목적 방사광 가속기'는 꼭 호남권에 유치되어야 합니다.	2020-05-27 107,253명	
826. <a href="#">보건복지</a>	정부, 美日한국전 참전국에 마스크지원시 일본 지원 반대합니다	2020-05-20 81,193명	
825. <a href="#">문화/예술/체육/언론</a>	SBS 거짓보도에 공식사과 요청 및 이** 기자 처벌요청	2020-06-03 76,623명	

국민청원은 또 현재 상태에 따라 답변 대기중, 진행중, 만료된 청원으로 나뉜다. 답변대기중인 청원은 청원의 동의 수가 20만명이 넘었지만 아직 정부의 답변을 듣지 못한 청원을 의미하고, 청원의 기간이 한달이기 때문에 한달이 지난 청원은 만료된 청원으로 분류한다.

전체 목록

지금 청원하기

최신순 보기 ▾

번호

분류

제목

청원 만료일

참여인원

검색

1 2 3 4 5 6 7 8 9 10 Next

### 청원 FAQ

---

**1 국민청원 어떻게 참여하나요?**

- 누구라도 트위터, 페이스북, 네이버, 카카오 중 보유 계정으로 소셜로그인을 통해 직접 청원글을 작성하거나, 진행 중인 청원에 동의하실 수 있습니다. 한 번 작성하거나 동의한 청원에 대해서는 삭제나 철회하실 수 없습니다.
- 홈페이지에 공개된 청원은 누구나 자유롭게 열람할 수 있습니다.

**2 국민청원 게시글 어떻게 작성하나요?**

- 새 청원글 작성 전, 진행 중인 청원을 먼저 검색하면서 유사한 청원이 게시되었는지 확인하실 수 있습니다. 새로운 청원글을 작성하는 대신 기존유사 청원에 동참하시면 국민의 뜻을 한 곳으로 모으는데 도움이 됩니다.
- 새 청원글 작성 시, 내용을 대표하는 제목 및 관련 분야를 선택하신 후 내용을 입력해주시면 됩니다.
- 청원 내용과 관련 있는 웹사이트 또는 영상 URL이 있다면 첨부해주시기 바랍니다.
- 청원 내용의 주요 키워드를 태그관에 입력해주시면 다른 참여자들이 여러분들의 청원을 쉽게 찾을 때 도움이 됩니다.
- 마지막으로, 입력하신 모든 내용을 신중하게 검토한 뒤 청원을 등록해주시기 바랍니다. 한번 작성된 청원은 수정 및 삭제가 불가능합니다. 최초 청원 취지와 다른 내용으로 변경되는 것을 방지하기 위한 것이니 신중한 작성 부탁드립니다.

**청원 작성 요청 자세히 보기**

1) 검색  
새 청원글 작성 전 유사한 내용의 청원이 진행 중인지 먼저 확인해주세요. 새 청원글을 작성하는 것보다 유사한 기존 청원에 동의해주시면 국민의 힘을 모으는데 도움이 됩니다.

2) 제목 입력  
청원 내용을 대표하는 짧은 제목(100자 이하)을 입력해주세요. 주요 단어를 제목에 포함시켜 주시면 검색을 통한 노출이 잘 이뤄질 수 있습니다.

3) 카테고리 선택  
청원 내용과 관련된 분야를 선택해주세요. 참여자들이 '분야별 청원' 메뉴를 통해 접근할 수 있습니다.

4) 내용 입력  
사람들이 쉽게 읽고 이해할 수 있는 내용으로 작성해주시면 됩니다. 글자 수 제한은 없습니다.

5) 링크 첨부  
청원 내용과 관련된 링크 주소를 입력해주세요. 다만 청원 내용과 관련이 없거나 부적절한 링크는 관리자에 의해 승강처리 될 수 있습니다.

6) 태그  
3개 이하의 태그를 입력해주세요. 참여자들이 손쉽게 청원을 검색하고, 내용을 이해하는 데 도움을 줄 수 있습니다.

7) 검토 및 게시  
한번 작성된 청원은 수정 및 삭제가 불가능합니다. 최초 청원 취지와 다른 내용으로 변경되는 것을 방지하여 청원작성자의 의견을 보호하기 위한 조치이나 신중하게 게시해주시기 바랍니다.

**3 국민청원 게시판에 공개되기 위해서는 '100명의 사전 동의'가 필요합니다.**

- 청원글 작성 원료 시, 여러분에게 '사전동의 링크(URL)'가 부여됩니다.
- 30일 이내에 여러분의 청원을 지지하는 100명의 사전 동의를 받으시아 청원게시판에 청원 내용이 공개되어 더 많은 국민께서 청원에 동참하실 수 있게 됩니다.
- '사전동의 링크(URL)'를 SNS 등에 공유해 사전 동의를 받아주세요. 사전 동의는 제공된 '사전동의 링크(URL)'를 통해서만 가능합니다. 100명의 사전 동의를 받은 청원은 관리자의 검토를 거쳐 청원게시판에 공개됩니다. 단, 국민청원 요건에 맞지 않는 청원은 100명의 동의를 받더라도 게시판에 공개되지 않거나 관리자에 의해 일부 내용이 '숨김' 처리될 수 있습니다.
- '사전동의 링크(URL)'를 잊으셨거나 등록한 청원 등의 수가 급급하시다면 청와대 홈페이지 > 국민청원 > 내 청원 보기 를 통해 확인하실 수 있습니다.

**청원 요건 자세히 보기**

동일한 내용으로 중복 게시된 청원은 최초 1개 청원만 남기고 '숨김' 처리 또는 삭제될 수 있습니다.

욕설 및 비속어를 사용한 청원은 관리자에 의해 삭제 또는 일부 내용 '숨김' 처리될 수 있습니다.

폭력적, 성장적, 또는 특정 집단에 대한 혐오 표현 등 청소년에게 유해한 내용을 담은 청원은 관리자에 의해 삭제될 수 있습니다.

개인정보, 허위사실, 타인의 명예를 해손하는 내용이 포함된 청원은 관리자에 의해 삭제 또는 일부 내용 '숨김' 처리될 수 있습니다.

**4 국민청원, 이렇게 답변드립니다.**

- 청와대 홈페이지에 공개된 시점으로부터 30일 이내 20만 명 이상의 국민이 동의한 청원에 정부 및 청와대 관계자(각 부처 장관, 대통령 수석 비서관, 특별보좌관 등)가 답변하겠습니다.
- 다만, 청원 요건에 맞지 않는 경우 답변이 어려울 수 있습니다.
- 공개된 청원 답변은 청와대 홈페이지 > 국민청원 > 답변된 청원, 청원, 청와대 트위터, 페이스북, 유튜브를 통해 확인하실 수 있습니다.

**청원 요건 자세히 보기**

재판이 진행 중이거나, 일법부·사법부의 고유 권한과 관련한 내용으로 상권분권의 정신을 해손할 소지가 있는 청원에 대해서는 답변이 어려울 수 있습니다.

지방자치단체 고유 업무에 해당하는 내용 등 중앙 정부의 역할과 책임 범위를 벗어난 경우 답변이 어려울 수 있습니다.

청원 주요 내용이 허위사실로 밝혀진 경우 답변이 어려울 수 있습니다.

인종·국적·종교·나이·지역·장애·성별 등 특성과 관련 있는 개인이나 집단에 대한 차별 및 비하 내용 등 위헌적 요소가 포함된 청원에 대해서는 답변이 어려울 수 있습니다.

청와대 및 정부에 대한 민원·제안 및 공익신고·고발 등을 국민권익위원회의 국민신문고([www.epeople.go.kr](http://www.epeople.go.kr))를 이용해 주시길 부탁드립니다.

검색을 통해 이미 존재하는 유사한 청원은 올리지 않도록 하고 있으며 제목 입력을 요구한다.  
 카테고리는 이용자가 설정하도록 되어있다.  
 재판이 진행중이거나, 차별의 소지가 있는 청원 등은 동의 수가 20만이 넘더라도 답변이 어려울 수 있는 점을 안내한다.

soynlp의 의미는 콩nlp가 아닙니다. 오래전 스페인어를 공부하였을 때, gmail 계정을 soy.lovit으로 만들었습니다. ‘lovit’은 자주 쓰는 저의 필명입니다. ‘soy lovit’은 스페인어이며, 영어로 번역하면 ‘I am lovit’입니다. 이후 프로젝트 이름 앞에 soy를 붙였습니다. 굳이 번역하자면 “저는 NLP입니다”입니다.



Hyunjoong Kim  
lovit

Data scientist / Natural Language Processing / Machine Learning // soy.lovit@gmail.com

<https://lovit.github.io>



lovit / petitions\_scrapers

Code Issues 0 Pull requests 0 Actions Projects 0 Security 0 Insights

Join GitHub today

GitHub is home to over 50 million developers working together to host and review code, manage projects, and build software together.

Sign up

Dismiss

petitions scrapper

83 commits 1 branch 0 packages 0 releases 1 contributor

Branch: master New pull request

lovit Fix minor Latest commit 810f15f on 6 Oct 2019

data Commit scraped data with content 2 years ago

output\_sample Renaming folder 2 years ago

petitions\_scrapers Handling closed petitions 7 months ago

Find file Clone or download

데이터수집기는 GitHub를 통해 Lovit이란 사람의 petition\_scrapers라는 수집기를 이용

# 청와대 국민청원 수집기

청와대 국민청원 (<https://www1.president.go.kr/petitions>) 홈페이지에 올라온 청원의 내용 / 공감 개수 / 댓글을 수집하는, 파이썬 (Python)으로 구현된 청와대 국민청원 크롤러입니다. 2019년 3월 이후 청원 게시판이 "청원중"과 "청원완료"로 나뉘어짐에 따라 "청원완료"된 글들만 수집하도록 `scraping_petitions.py` 스크립트를 수정하였습니다. 단 `parse_page` 함수는 현재 청원중인 글에 대해서도 적용이 가능합니다.

## Usage

### 하나의 청원 페이지에서 정보 가져오기

`parse_page` 함수에 청원 페이지의 url 을 입력하면 아래의 정보들을 수집할 수 있습니다.

항목	설명
begin	청원 시작일
category	청원 카테고리 (외교, 국방, 경제 등)
content	청원 내용
crawled_at	수집 시각
end	청원 종료일
num_agree	수집 시각의 청원 동의 수
replies	청원 댓글
status	현재 청원 진행 상황 (청원시작, 청원진행중, 청원종료, 브리핑)

## 청원 카테고리 살펴보기

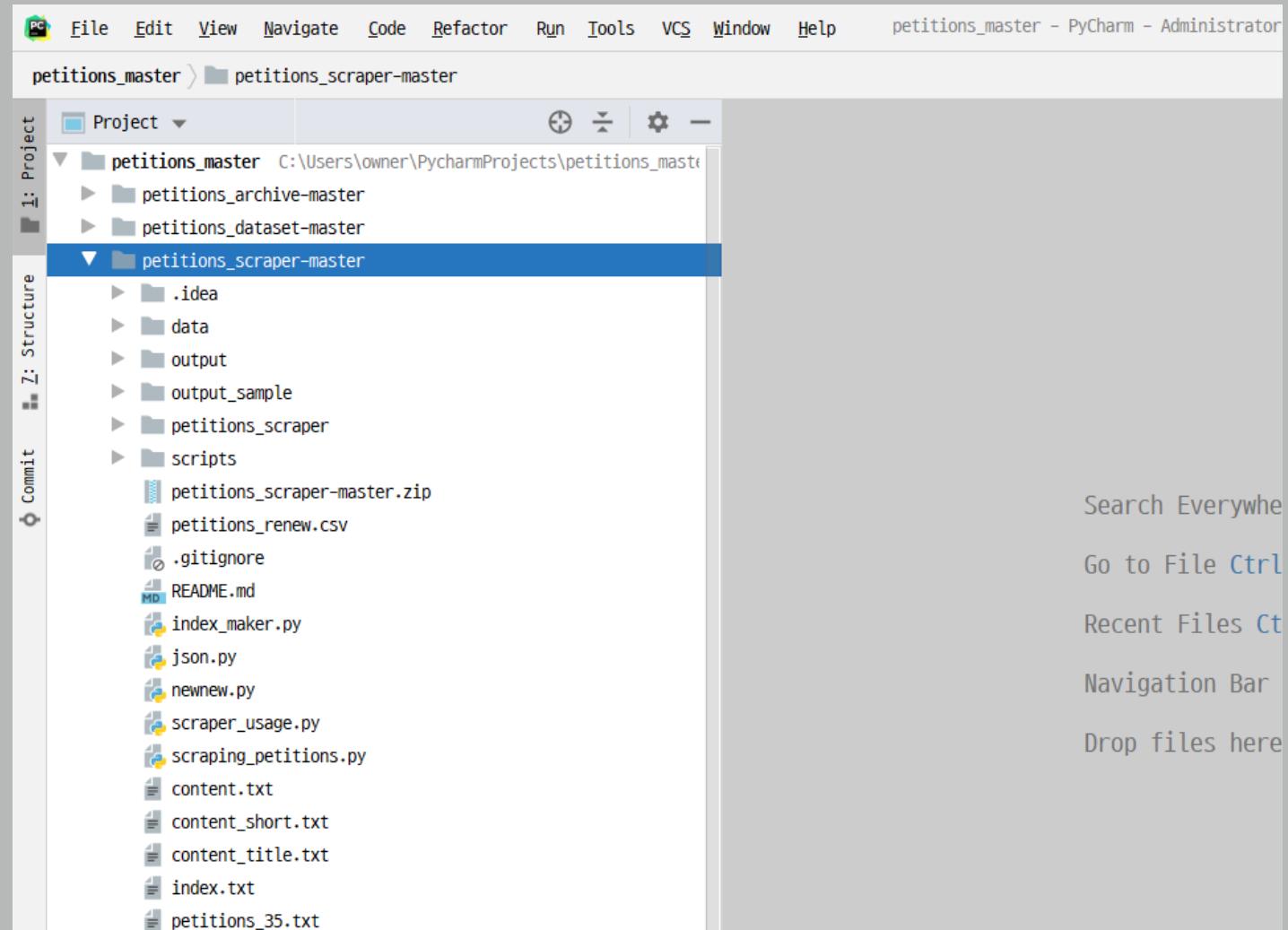
현재 청원 게시판에 등록되어 있는 카테고리를 살펴볼 수 있습니다.

```
from petitions_scrapper import show_categories  
  
show_categories()
```

```
idx = 0 , name = 전체  
idx = 35 , name = 정치개혁  
idx = 36 , name = 외교/통일/국방  
idx = 37 , name = 일자리  
idx = 38 , name = 미래  
idx = 39 , name = 성장동력  
idx = 40 , name = 농산어촌  
idx = 41 , name = 보건복지  
idx = 42 , name = 육아/교육  
idx = 43 , name = 안전/환경  
idx = 44 , name = 저출산/고령화대책  
idx = 45 , name = 행정  
idx = 46 , name = 반려동물  
idx = 47 , name = 교통/건축/국토  
idx = 48 , name = 경제민주화  
idx = 49 , name = 인권/성평등  
idx = 50 , name = 문화/예술/체육/언론  
idx = 51 , name = 기타
```

이 수집기를 통해 청원시작일 종료일, 카테고리 등으로 구분해서 정보를 가져 올 수 있었다.

대부분 '청원합니다'나 '동의합니다' 밖에 없었기 때문에 댓글은 제외했다.



scraping\_petition 파일을 실행하면 index.txt파일을 기반으로 크롤링이 진행  
수집한 결과 8000개가 넘은 json파일이 만들어졌다.

The screenshot shows a PyCharm interface with several windows open:

- scraping\_petitions.py**: A Python script with code for handling petitions. It includes functions for main(), show\_last\_index\_func(directory), update\_target(first\_index, last\_index), load\_index(path), and save\_index(path, index). A conditional block checks if \_\_name\_\_ == '\_\_main\_\_' and calls main().
- index\_maker.py**: A Python script that prints indices from 579682 to 587827 to a file named index.txt.
- index.txt**: A text file containing the output of index\_maker.py, listing indices and their status.
- File Explorer**: A window showing a list of JSON files (579682 to 579694) with their details: date (2020-04-25), time (e.g., 08:20:00), type (JSON), and size (e.g., 1KB, 4KB, 4KB, 5KB, 6KB, 2KB, 1KB, 6KB, 17KB, 31KB, 2KB, 13KB, 2KB).
- 579684... 속성**: Properties dialog for the file 579684.json. It shows the file is a JSON file (type: JSON 원본 파일 형식의 모든 파일), located at C:\Users\owner\PycharmProjects\petitions\_main (위치), and has a size of 26.5MB (27,852,685 바이트) and a disk usage of 42.6MB (44,756,992 바이트). It also lists the file's modification date and time.
- 579682.json**: Content of the JSON file 579682.json. It contains a single object with fields: crawled\_at (2020-04-25 08:20:00), category (정치개혁), begin (2019-04-22), end (2019-05-22), content (empty string), num\_agree (1831900), petition\_idx (579682), status (브리핑), and title (자유 한국당 정당해산 청원).
- 579682.json**: Content of the JSON file 579682.json. It contains a single object with fields: crawled\_at (2020-04-25 08:20:00), category (정치개혁), begin (2019-04-22), end (2019-05-22), content (empty string), num\_agree (1831900), petition\_idx (579682), status (브리핑), and title (자유 한국당 정당해산 청원).
- 585017.json**: Content of the JSON file 585017.json. It contains a single object with fields: crawled\_at (2020-04-25 10:57:58), category (안전/환경), begin (2020-02-07), end (2020-03-08), content (우한 폐렴 신종 코로나 1급), num\_agree (1088), petition\_idx (585017), status (청원종료), and title (우한 폐렴 신종 코로나 1급).
- 587827.json**: Content of the JSON file 587827.json. It contains a single object with fields: crawled\_at (2020-04-25 10:56:54), category (문화/예술/체육/언론), begin (2020-04-07), end (2020-05-07), content (최근 MBC의 잇따른 보도에), num\_agree (259809), petition\_idx (587827), status (청원진행중), and title (방송통신위원회는 방송의 공적).

json파일의 'crawled at' 부분은  
청원의 상태가 시간에 따라 변할 수  
있으므로 언제 크롤링했는지를  
확인하게 해주는 기능이다.

json.py

```

1 # -*- encoding: utf-8 -*-
2 import json
3 import os
4
5 id_file = "petitions_end.txt"
6 path = []
7 ls = []
8
9 ...
10
11 idx = 'br_ing.txt'
12
13 with open(id_file, "r", encoding="utf-8") as f:
14     for root, dirs, files in os.walk(
15         'C:/Users/owner/PycharmProjects/petitions_master/petitions_scraper-master/output/'):
16         for fname in files:
17             path.append(os.path.join(root, fname))
18
19 for i in range(len(path)):
20     with open(path[i], "r", encoding='utf-8') as json_file:
21         json_data = json.load(json_file)
22         if json_data.get('num_agree') >= 20000 and json_data.get('status') == "청원진행중":
23             ls.append(json_data.get('category'))
24             ls.append(json_data.get('petition_idx'))
25             ls.append(json_data.get('begin'))
26             ls.append(json_data.get('end'))
27             ls.append(json_data.get('status'))
28             ls.append(str(json_data.get('num_agree')))
29             ls.append(json_data.get('title'))
30             ls.append(json_data.get('content'))
31
32             with open(idx, 'a', encoding='utf-8') as ee:
33                 ee.write('\t'.join(ls))
34                 ee.write('\n')
35             ls = []
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83

```

A - H (16) — I - P (22) —

- br\_35 pet\_all
- br\_36 petition\_title
- br\_37 petitions\_35
- br\_41 petitions\_36
- br\_42 petitions\_37
- br\_43 petitions\_38
- br\_45 petitions\_39
- br\_46 petitions\_40
- br\_47 petitions\_41
- br\_49 petitions\_42
- br\_50 petitions\_43
- br\_51 petitions\_44
- br\_all petitions\_45
- br\_br petitions\_46
- br\_end petitions\_47
- br\_ing petitions\_48
- petitions\_49 petitions\_49
- petitions\_50 petitions\_50
- petitions\_51 petitions\_51
- petitions\_brief petitions\_brief
- petitions\_end petitions\_end
- petitions\_ing petitions\_ing

크롤링한 json파일을 텍스트 파일로 바꿔주는 부분이다. Br은 동의 수가 20만명을 충족한 청원들이고, petitions는 전체 청원을 모두 포함했다. 35~51은 카테고리 별 청원을 의미하고 그 외에도 답변진행중이거나 만료된 청원 등을 추가적으로 분류했다.

The screenshot displays three windows side-by-side:

- DevTools - www1.president.go.kr/petitions/585290**: Shows the HTML structure of a petition page. A sidebar on the right shows CSS styles for a class named `:hover`. The main pane shows the full HTML code.
- parser.py**: An IDE window titled "View\_write" containing Python code. The code defines several functions: `parse_page`, `_as_json`, `is_closed_petition`, `parse_meta`, `parse_content`, `parse_number_of_agree`, and `parse_status`. The `parse_content` function is highlighted with a yellow background. It uses BeautifulSoup to select elements with classes `petitions_view_write` and `view_write`.
- petitions\_brief.txt**: A text file containing a list of 12 petitions, each with a number, title, date, status, and a brief description.

	제목	번호	제출일	마감일	작성자	내용
1	정치개혁	579682	2019-04-22	2019-05-22	브리핑	183190 자유 한국당 정당해산 청원
2	정치개혁	579708	2019-04-24	2019-05-24	브리핑	210344 국회의원도 국민이 직접 소환할 수 있어야 합니다.
3	정치개혁	579779	2019-04-29	2019-05-29	브리핑	337964 더불어 민주당 정당해산청구!!
4	정치개혁	579840	2019-04-30	2019-05-30	브리핑	250219 문재인 대통령의 탄핵을 청원합니다.
5	외교/통일/국방	579927	2019-05-03	2019-06-02	브리핑	224852 김무성 전 의원을 내려놔 다스려주십시오.
6	반려동물	580325	2019-05-20	2019-06-19	브리핑	217483 이전에서 벌어진 동물수간사건에 대한 강력한 처벌과 동물학대에
7	유아/교육	580402	2019-05-24	2019-06-23	브리핑	213025 축구클럽에 축구한다고 차량에 태워 보낸 아이가 돌아오지 않았습
8	인권/성평등	580452	2019-05-27	2019-06-26	브리핑	216862 2019제*****호 가해용의자의 상강 취소 및 강학금 환수와 구속
9	인권/성평등	580639	2019-06-04	2019-07-04	브리핑	348417 우리딸을 성폭행한 후 잔인하게 목졸라 죽인 극악무도한 살인마를
10	안전/환경	580707	2019-06-07	2019-07-07	브리핑	223006 불쌍한 우리 형님을 찾아주시고, 살인범 ***의 사형을 청원합니다
11	육아/교육	580884	2019-06-14	2019-07-14	브리핑	240298 아동 성폭행범을 감형한 *** 판사 파면하라
12	인권/성평등	581300	2019-07-08	2019-08-07	브리핑	263792 리얼돌 수입 및 판매를 금지해주세요

코드 실행 결과, 카테고리, 분류기호, 청원날짜 등으로 나누어 정렬되는 모습을 확인할 수 있다.

노란색으로 표시한 `view_write`부분에 유의가 필요하다. `Parse_content`를 해도 내용이 나오지 않는 청원들이 있었는데, 이 부분은 확인 결과, 청원에 답변이 되어 해당 부분에 내용이 아닌 동영상이 들어있었다.

crawled\_at 2020-04-25

category	N	Sum	청원종료	청원진행중	>=200000 브리핑	아직	Avr_Agr
정치개혁	35	688	644	31	15	13	2 18,065
외교/통일/국방	36	396	358	37	1	1	- 3,995
일자리	37	288	267	21	1	-	1 3,596
미래	38	88	82	6	-	-	2,296
성장동력	39	68	60	8	-	-	2,020
농산어촌	40	84	83	1	-	-	887
보건복지	41	1,337	1,233	101	3	3	- 3,865
육아/교육	42	765	668	94	5	3	2 6,269
안전/환경	43	781	728	46	8	7	1 12,633
저출산/고령화대책	44	42	41	1	-	-	823
행정	45	528	484	43	2	1	1 4,417
반려동물	46	124	119	2	3	3	- 13,741
교통/건축/국토	47	659	640	19	1	-	1 2,040
경제민주화	48	247	227	20	-	-	- 1,637
인권/성평등	49	643	577	54	18	12	6 16,212
문화/예술/체육/언론	50	384	363	15	7	6	1 7,541
기타	51	953	876	72	9	5	4 8,377
Sum		8,075	7,450	571	73	54	19 7,729

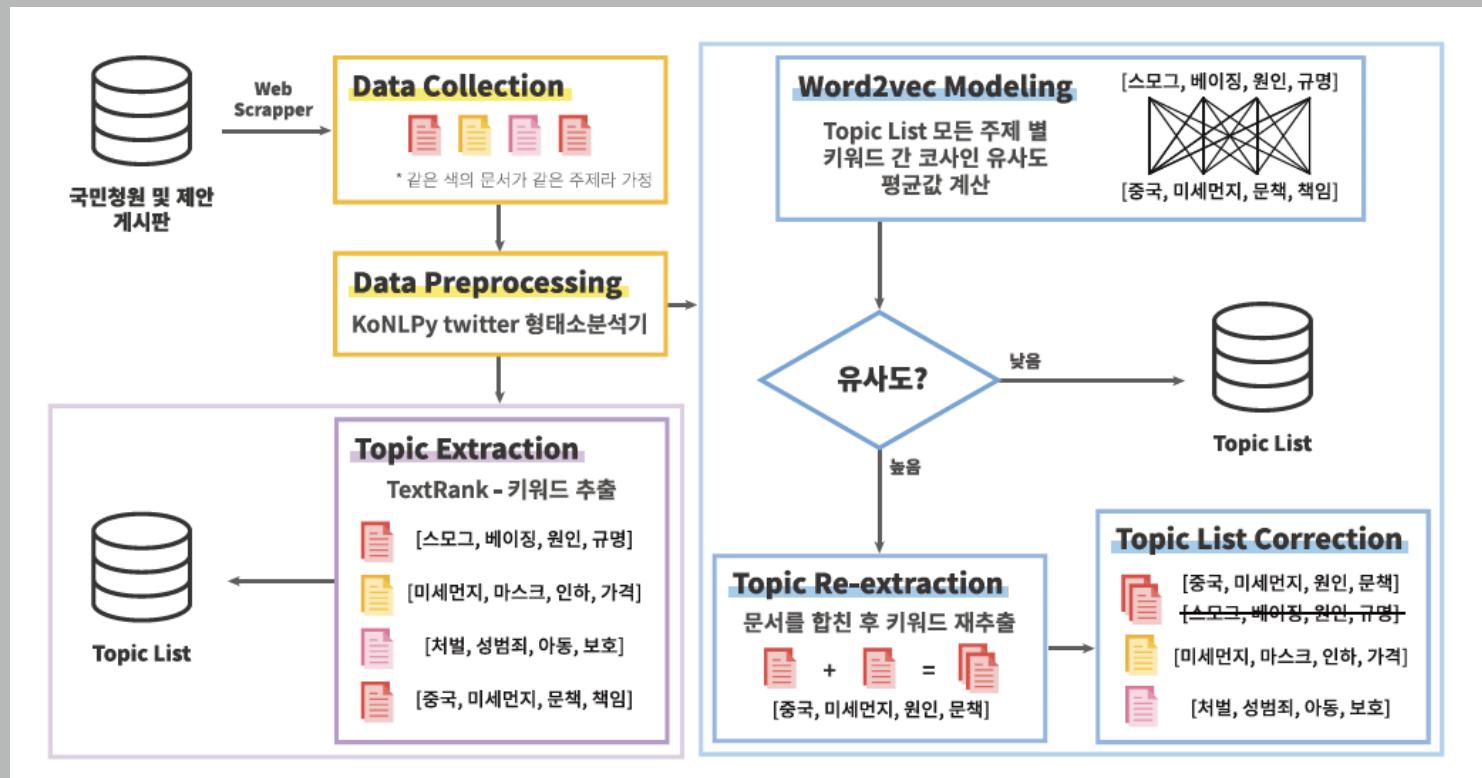
<카테고리별 청원 빈도수>

년	월	Sum	청원종료	청원진행중	>=200000 브리핑	아직	Avr_Agr
	2019 4	161	157	-	4	4	- 20,772
	2019 5	714	710	-	4	4	- 4,883
	2019 6	572	569	-	3	3	- 5,156
	2019 7	610	604	-	6	6	- 5,422
	2019 8	638	631	-	7	7	- 7,430
	2019 9	501	499	-	2	2	- 5,075
	2019 10	460	452	-	8	8	- 7,489
	2019 11	436	433	-	3	3	- 5,944
	2019 12	448	446	-	2	2	- 4,770
	2020 1	402	399	-	3	3	- 5,811
	2020 2	1,216	1,209	-	10	7	3 8,712
	2020 3	1,635	1,341	289	19	5	14 11,325
	2020 4	282	-	282	2	-	2 8,585
	Sum	8,075	7,450	571	73	54	19 7,729

<날짜별 청원 빈도수>

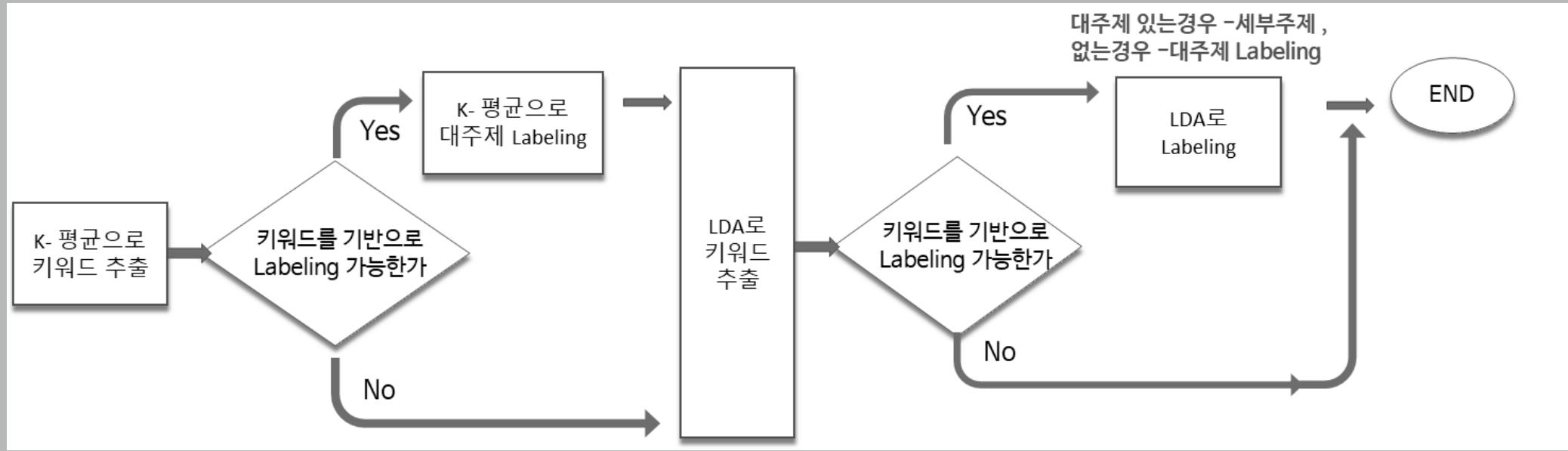
### 3. 선행 연구 분석

# 1. Textrank와 word2vec을 활용한 국민청원 게시판 중복주제 탐지. 대한산업공학회 춘계공동학술대회 논문집



중복 청원 글을 탐지하기 위하여 textrank를 이용하여 청원 글 별 키워드 집합 추출, word2vec 모형을 이용하여 키워드 간 평균 코사인 유사도 연산을 통해 중복 주제 여부를 판단

## 2. 국민청원 주제 분석 및 딥러닝 기반 답변 가능 청원 예측



국민 청원의 주제를 분석하고 딥러닝을 활용하여 답변 가능한 청원을 예측하는 모델을 제안

추천순으로 1,500개의 청원글을 수집하고 K-means 클러스터링(10)을 적용, 청원글을 군집하여 대주제 정의

구체적인 세부 주제를 정의하기 위하여 토픽 모델링을 실시

LSTM을 활용하여 답변 가능한 청원 예측 모델을 생성

20만의 청원동의를 얻는 청원을 예측하기 위한 모델을 개발

### 3. 대 국민청원에서는 무엇이 일어나는가? : 자연어 처리를 활용한 청와대 국민청원 분석

| 표 1 | 청와대 국민청원 게시판에서 추출된 토픽 목록

연번	토픽	최다빈출단어	독점성 단어	비중	빈도
1	범죄	처벌, 피해자, 범죄, 사건, 가해자, 피해, 폭행	가해자, 무고, 조두순, 물카, 폭행, 무고죄, 미투운동	0,095	6148
2	부동산	주택, 부동산, 집값, 아파트, 서민, 정책, 투기	집값, 부동산, 폭등, 무주택, 보유세, 주택, 임대	0,064	5625
3	조세/준조세	국민, 세금, 공무원, 연금, 국민연금, 폐지, 서민	연금, 국민연금, 누진세, 에어컨, 누진, 기세, 고갈	0,053	5283
4	국회비판	국회의원, 국민, 국회, 의원, 선거, 정치, 자유한국	국회의원, 국회, 선거, 세비, 특활비, 해산, 자유한국	0,052	5237
5	대통령	대통령, 국민, 정부, 문재인, 정권, 청와대, 당신	문재인, 대통령, 촛불, 박근혜, 당신, 정권, 지지율	0,047	5022
6	난민	난민, 외국인, 한국, 국민, 자국민, 우리나라, 이슬람	난민, 자국민, 이슬람, 제주도, 예멘, 무슬림, 체류	0,045	4806
7	검찰수사	사건, 조사, 수사, 검찰, 경찰, 비리, 검사	이재명, 특검, 드루킹, 진실, 조폭, 검찰, 김경수	0,045	4177
8	외교/안보	북한, 우리, 일본, 평화, 역사, 미국, 전쟁	독도, 남한, 비핵화, 북한, 천안함, 김정철, 폭침	0,044	3958
9	인터넷/언론	청원, 기사, 언론, 내용, 방송, 뉴스, 사이트	알베, 네이버, 사이트, 게시, 링크, 방송, 언론사	0,044	3784
10	노동	근무, 최저임금, 회사, 근로자, 임금, 직원, 시간	근로자, 비정규직, 임금, 근로시간, 최저임금, 최저시급, 연차	0,044	3703
11	스포츠	선수, 올림픽, 감독, 경기, 축구, 스포츠, 국가대표	선수, 축구, 국가대표, 빙상연맹, 노선영, 축구협회, 월드컵	0,035	3630
12	금융/가상화폐	공매도, 주식, 투자, 기업, 거래, 기상화폐, 시장	공매도, 주식, 가상화폐, 증권, 거래소, 코인, 블록체인	0,034	3256
13	교육/입시	학생, 교육, 학교, 시험, 교사, 대학, 공부	영어, 수능, 수시, 사교육, 과목, 교육부, 학생	0,033	2932
14	성별갈등	여성, 남성, 남자, 여자, 군대, 사회, 평등	여성기족부, 페미니즘, 페미니스트, 여성, 국방의무, 성차별, 남성	0,03	2931
15	보육	아이, 교사, 어린이집, 부모, 학교, 시간, 선생님	어린이집, 보육교사, 보육, 교복, 유치원, 아동학대, 유아	0,029	2845
16	시법불신	판결, 판사, 재판, 현법, 법률, 사법부, 법원	판사, 사법부, 판결, 대법원, 정형식, 재판, 사법	0,027	2777

연번	토픽	최다빈출단어	독점성 단어	비중	빈도
17	갑질	사용, 판매, 불법, 대한항공, 나이, 청소년, 게임	대한항공, 생리대, 고객, 항공사, 한진, 조현민, 신분증	0,027	2657
18	음주운전	사고, 차량, 경찰, 안전, 소방관, 발생, 음주운전	운전자, 음주운전, 차량, 주차, 운전, 차주, 소방	0,026	2576
19	이웃분쟁/동물	동물, 흡연, 담배, 사람, 강아지, 반려동물, 학대	동물, 강아지, 반려견, 식용, 동물학대, 인락사, 충간소음	0,026	2485
20	종교/이념	대한민국, 나라, 국민, 국가, 우리, 사회, 우리나라	나라, 대한민국, 자유, 다운, 교회, 종교, 동성애	0,025	2406
21	환경/에너지	미세먼지, 중국, 원전, 사용, 환경, 정부, 문제	미세먼지, 냉사, 발전소, 발암물질, 원전, 태양광, 낚시	0,025	2261
22	의료	병원, 치료, 환자, 의료, 간호사, 수술	병원, 환자, 간호사, 의료사, 치과, 수술, 대학병원	0,024	2172
23	저출산	지원, 소득, 가정, 혜택, 결혼, 자녀, 아이	부부, 양육비, 난임, 한부모, 자녀, 이혼, 건강보험료	0,022	2077
24	아파트	사업, 권리, 계약, 공사, 아파트, 진행, 업체	조합, 택배, 계약, 조합원공사, 건축, 입찰	0,022	1724
25	생활민원	지역, 주민, 서울, 시민, 지방, 계획, 개발	개인회생, 변제, 지역, 미을, 주민, 노선, 회생	0,019	1695
26	정책의견	정책, 문제, 제도, 정부, 국가, 현재, 경제	도입, 감소, 분야, 변화, 증가, 병식, 병향	0,019	1530
27	장애인복지	활동, 장애인, 단체, 사회, 시설, 센터, 복지	장애인, 특수, 활동, 집회, 애인, 복지사, 장애	0,018	516
28	더미	사람, 생각, 보고, 마음, 자기, 정도, 하나	얘기, 가요, 소리, 사람, 생각, 그때, 화가	0,012	0

감정이 정치적 참여와 판단에 미치는 영향을 경험적으로 분석  
자연어 처리(토픽모델링(28)과 word2vec)를 통해 국민청원 페이지에 나타난 의제와 감정들(분노, 슬픔 중심)을 파악한 후, 음이향 회귀분석을 통해 글의 구성요소와 동의수의 관계를 파악  
슬픔보다는 분노가 동의에 큰 효과

#### 4. 국민청원 데이터 시각화 분석 <http://voiceofpeople.kr/about>



무분별한 청원의 대량생산, 같은 주제에 대한 청원과 동의의 분산, 기간별 청원 흐름 파악의 어려움 등을 문제점으로 꼽음  
시계열 분석, 동의수를 기준으로 다르게 표현했다는 점에서 활용가능성 시각화 면에서는 정리가 잘 되어 있으나 분석에 있어 단순 키워드만이 기준이 된다는 한계

## 5. 국민청원 데이터에 기반한 SW분야 이슈분석

〈표 4〉 SW분야 국민청원 클러스터링 결과 및 주제별 핵심 키워드

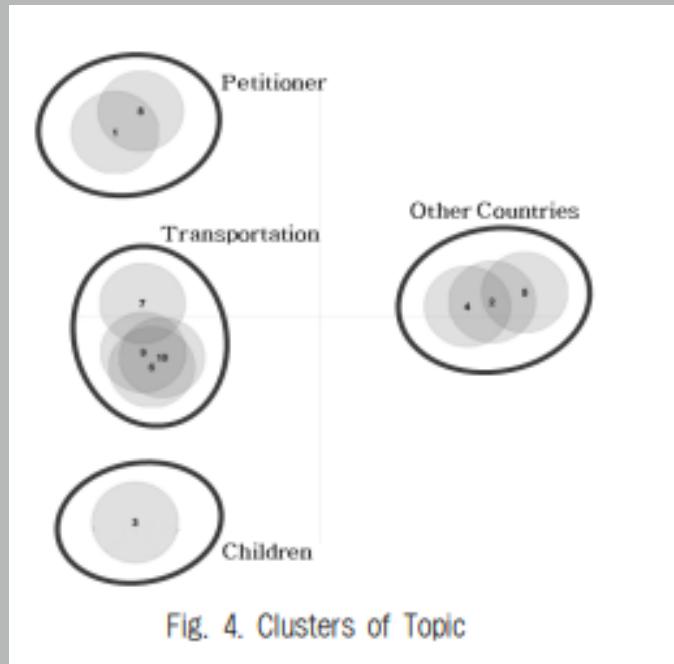
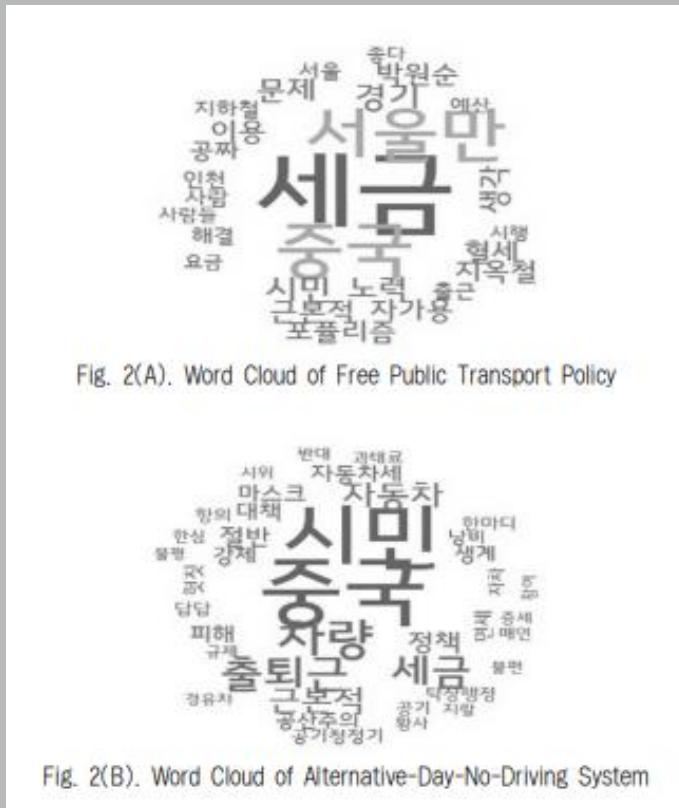
가상화폐		보안		게임산업		교육		기업지원		생태계	
키워드	인접 중앙성	키워드	인접 중앙성	키워드	인접 중앙성	키워드	인접 중앙성	키워드	인접 중앙성	키워드	인접 중앙성
블록체인	84.9%	개인정보	85.9%	게임	91.7%	교육	87.2%	기업	80.5%	독점	92.9%
규제	80.2%	해킹	83.8%	아이템	73.3%	학교	77.3%	지원	69.6%	자본	92.9%
비트코인	78.6%	보안	81.7%	청소년	68.8%	소프트웨어	75.6%	데이터	69.1%	플랫폼	92.9%
투자	77.5%	프로그램	79.8%	모바일게임	67.7%	코딩	73.9%	경쟁력	66.5%	네트워크	89.7%
미래	77.5%	인증	76.1%	사행성	67.7%	프로그래밍	69.4%	일자리	64.4%	생태계	89.7%

9) 클러스터내 주요한 키워드를 의미하는 지표 중 하나로, 연결 그래프에서 정점과 다른 모든 정점 사이의 최단경로길이의 역수로 계산

소프트웨어라는 분야를 6가지 대주제로 분류하고 유사도 측정 기법을 활용해 각 주제에 잘 부합하는 청원을 찾는 방식  
TF-IDF기법을 활용한 전처리나 gephi 활용등에서 수업과 연관.  
각 주제별 핵심단어 중심성 분석

## 6. 회귀분석과 텍스트 마이닝을 활용한 미세먼지 비상저감조치의 실효성과 국민청원 분석

<https://www.kci.go.kr/kciportal/ci/sereArticleSearch/ciSereArtiView.kci?sereArticleSearchBean.artild=ART002409905>



미세먼지 관련 시민들의 생각을 알기 위한 워드클라우드 작성  
미세먼지를 검색어로 LDA기반의 토픽모델링을 적용해 10개의 토픽과 토픽 별 30개의 단어 추출.  
각 토픽의 주요단어의 유사성을 파악하여 겹치는 토픽끼리 묶어 크게 네 가지의 주제로 다시 분류하는 방식

## 4. 문제 해결 및 중간 계획 수행 현황

## 브리핑(답변된 청원) content 수집 불가 문제: 해결

채널A·TV조선 재승인 취소 청원

국민청원

채널A·TV 재승인 취소

한상혁 방송통신위원회

청원내용

최근 MBC의 잇따른 보도에 따르면 종편 채널A 기자가 현직 검사장과의 친분을 내세우며 여권 인사의 비위를 털어놓으라고 취재원을 압박

div.View\_write 830 x 508

DevTools - www1.president.go.kr/petitions/587827

Elements Console Sources Network Performance Memory

```
<div class="petitionsView_left">
  <div class="petitionsView_left_pg">
    <!-- 청원 원쪽 -->
    <div class="petitionsView_progress">...</div>
    <h3 class="petitionsView_title">방송통신위원회는 방송의 공고 언론이기를 포기한 채널A와 TV조선의 재승인을 취소하라</h3>
    <h2 class="petitionsView_count">...</h2>
    <div class="petitionsView_info">...</div>
    <div class="petitionsView_grapay">...</div>
    <!-- petitionsView_grapay end -->
    <!-- 청원 본문 -->
    <div class="petitionsView_write">
      <div class="petitionsView_writeHead">...</div>
      <div class="View_write">...</div> == $0
      <div class="petitionsView_writeHead">...</div>
      <div class="View_write" style="word-break:break-all">...
        <ul class="View_write_link">...</ul>
      </div>
      <!-- 청원 본문 -->
      <div class="pr_tk25" style="text-align:left">...</div>
      <!-- 청원 동의 -->
      <div id="petitionsReply_area" class="petitionsReply_area">
        <!-- petitionsReply_area end -->
        <!-- //청원 동의 -->
        <!-- //청원 원쪽 -->
      </div>
      <!-- petitionsView_left_pg end -->
    </div>
    <!-- // petitionsView_left -->
  </div>
```

html body #wrap #contents #cont\_view div div div div div div div.View\_write

Styles Event Listeners DOM Breakpoints Properties Accessibility

Filter :hov .cls + ▲

```
petitions_scraper-master > petitions_scraper > parser.py
```

```
parser.py
```

```
81
82     def parse_content(soup):
83         content = soup.select(
84             'div[class=petitionsView_write] div[class=View_write]')
85
86         if not content:
87             return ''
88
89         # return normalize_text(content[1].text) 브리핑만
90         return normalize_text(content[0].text)
```

브리핑(답변된 청원)은 영상이 첨부되면서, div class="View\_write"를 추가했기 때문에 발생한 문제였다. 따라서 petitions\_scraper를 쓰면, 영상이 들어있기 때문에 텍스트가 수집되지 않았던 것이었다.

브리핑만 수집할 때에는 parser.py 파일에서 리스트 soup.select의 인덱스를 바꾸었다.

## 인코딩 문제: 해결

kkma

```
result = pipeline.processCorpus(corpus)
File "C:\Users\owner\PycharmProjects\pyTextMiner\pyTextMiner\__init__.py", line 65, in processCorpus
    inst = apply(p, c, inst)
File "C:\Users\owner\PycharmProjects\pyTextMiner\pyTextMiner\__init__.py", line 57, in apply
    return [apply(p, a[1:], i) for i in inst]
File "C:\Users\owner\PycharmProjects\pyTextMiner\pyTextMiner\__init__.py", line 57, in <listcomp>
    return [apply(p, a[1:], i) for i in inst]
File "C:\Users\owner\PycharmProjects\pyTextMiner\pyTextMiner\__init__.py", line 55, in apply
    return p(inst)
File "C:\Users\owner\PycharmProjects\pyTextMiner\pyTextMiner\tokenizer\__init__.py", line 86, in __call__
    return self.inst.pos(args[0])
File "C:\Program Files\Python37\lib\site-packages\konlpy\tag\kkma.py", line 70, in pos
    morphemes.append((morpheme.getString(), morpheme.getTag()))
UnicodeDecodeError: 'utf-8' codec can't decode byte 0xed in position 0: invalid continuation byte

Process finished with exit code 1
```

komoran

```
result = pipeline.processCorpus(corpus)
File "C:\Users\owner\PycharmProjects\pyTextMiner\pyTextMiner\__init__.py", line 65, in processCorpus
    inst = apply(p, c, inst)
File "C:\Users\owner\PycharmProjects\pyTextMiner\pyTextMiner\__init__.py", line 57, in apply
    return [apply(p, a[1:], i) for i in inst]
File "C:\Users\owner\PycharmProjects\pyTextMiner\pyTextMiner\__init__.py", line 57, in <listcomp>
    return [apply(p, a[1:], i) for i in inst]
File "C:\Users\owner\PycharmProjects\pyTextMiner\pyTextMiner\__init__.py", line 55, in apply
    return p(inst)
File "C:\Users\owner\PycharmProjects\pyTextMiner\pyTextMiner\tokenizer\__init__.py", line 64, in __call__
    return self.inst.pos(args[0])
File "C:\Program Files\Python37\lib\site-packages\konlpy\tag\komoran.py", line 69, in pos
    result = [(token.getMorph(), token.getPos()) for token in result]
File "C:\Program Files\Python37\lib\site-packages\konlpy\tag\komoran.py", line 69, in <listcomp>
    result = [(token.getMorph(), token.getPos()) for token in result]
UnicodeDecodeError: 'utf-8' codec can't decode byte 0xed in position 0: invalid continuation byte

Process finished with exit code 1
```

Komoran과 Kokoma는 UTF-8을  
벗어나는 emoji 등의 특수한 텍스트를  
처리하지 못한다. 수많은 시행착오 끝에  
이모지가 있는 텍스트를 처리 못함을  
발견했다.

## 인코딩 문제: 해결

kkma

```
result = pipeline.processCorpus(corporus)
File "C:\Users\owner\PycharmProjects\pyTextMiner\pyTextMiner\__init__.py", line 65, in processCorpus
    inst = apply(p, c, inst)
File "C:\Users\owner\PycharmProjects\pyTextMiner\pyTextMiner\__init__.py", line 57, in apply
    return [apply(p, a[1:], i) for i in inst]
File "C:\Users\owner\PycharmProjects\pyTextMiner\pyTextMiner\__init__.py", line 57, in <listcomp>
    return [apply(p, a[1:], i) for i in inst]
File "C:\Users\owner\PycharmProjects\pyTextMiner\pyTextMiner\__init__.py", line 55, in apply
    return p(inst)
File "C:\Users\owner\PycharmProjects\pyTextMiner\pyTextMiner\tokenizer\__init__.py", line 86, in __call__
    return self.inst.pos(args[0])
File "C:\Program Files\Python37\lib\site-packages\konlpy\tag_kkma.py", line 70, in pos
    morphemes.append((morpheme.getString(), morpheme.getTag()))
UnicodeDecodeError: 'utf-8' codec can't decode byte 0xed in position 0: invalid continuation byte

Process finished with exit code 1
```

komoran

```
result = pipeline.processCorpus(corporus)
File "C:\Users\owner\PycharmProjects\pyTextMiner\pyTextMiner\__init__.py", line 65, in processCorpus
    inst = apply(p, c, inst)
File "C:\Users\owner\PycharmProjects\pyTextMiner\pyTextMiner\__init__.py", line 57, in apply
    return [apply(p, a[1:], i) for i in inst]
File "C:\Users\owner\PycharmProjects\pyTextMiner\pyTextMiner\__init__.py", line 57, in <listcomp>
    return [apply(p, a[1:], i) for i in inst]
File "C:\Users\owner\PycharmProjects\pyTextMiner\pyTextMiner\__init__.py", line 55, in apply
    return p(inst)
File "C:\Users\owner\PycharmProjects\pyTextMiner\pyTextMiner\tokenizer\__init__.py", line 64, in __call__
    return self.inst.pos(args[0])
File "C:\Program Files\Python37\lib\site-packages\konlpy\tag_komoran.py", line 69, in pos
    result = [(token.getMorph(), token.getPos()) for token in result]
File "C:\Program Files\Python37\lib\site-packages\konlpy\tag_komoran.py", line 69, in <listcomp>
    result = [(token.getMorph(), token.getPos()) for token in result]
UnicodeDecodeError: 'utf-8' codec can't decode byte 0xed in position 0: invalid continuation byte

Process finished with exit code 1
```

## 대한애국당 천막들 군인과 경찰들 동원해서 강제로 철거 실시하라! #跪

대한애국당 자진 철거를 거부해서 지금까지 불법적으로 천막 버티고 있습니다.

여러분 광화문광장에 대한애국당 천막이 불법적으로 설치된 모습이 매우 불쾌하고, 광장에 통행이 불쾌할거 뻔합니다

그러므로 문재인 대통령 및 박원순 시장님께 경고하는 의미로 군인과 경찰들 동원해서 당장 대한애국당 천막들 철거 실시하라!

Komoran과 Kokoma는 UTF-8을  
벗어나는 emoji 등의 특수한 텍스트를  
처리하지 못한다. 수많은 시행착오 끝에  
이모지가 있는 텍스트를 처리 못함을  
발견했다.

## 인코딩 문제: 해결

```
7     def remove_emoji(text):
8         only_BMP_pattern = re.compile("["
9                         u"\u00010000-\u0010FFFF"
10                        "]+", flags=re.UNICODE)
11         return only_BMP_pattern.sub(r'', text)
12
13     ta = json_data.get('title')
14     ta_after = remove_emoji(ta)
15     ls.append(ta_after)
16
17     ca = json_data.get('content')
18     ca_after = remove_emoji(ca)
19     ls.append(ca_after)
```

Title과 Content에만 이모지가 있기 때문에, 해당 부분을 정규표현식을 통해 없앴다.

## 데이터파일의 크기: 해결

메모리 용량의 문제를 해결하기 위해 2가지 방법을 사용했다.

1. 전체 데이터 외에도, 처리할 txt 데이터를 분류별로 나눠서 만들었다. 이를 통해 메모리가 부족한 기기에서도 분석을 할 수 있었다.
2. (김진욱) 메모리를 교체했다. 32GB 램 2개로 교체했으며, 이를 통해 메모리 관련 오류는 완전히 없어졌다.

## 특정 주제에 관한 유사한 청원에 동의한 사람 수 활용: 일부

본래 특정 주제에 관한 유사한 청원에 동의한 사람 수의 활용 방안을 찾아보고자 했으나, 시간은 유한했고, 선택과 집중을 하기 위해 텍스트마이닝 자체에 집중하기로 했다.

프로젝트에서 주된 내용이 되기보다는 필요한 곳에 쓰기로 했다.

## 감정분석, 토픽모델링을 통한 심층분석

수행했으며, 자세히 소개할 예정

## 5. Text Rank 태이틀 분석

Keyword가 Title에 포함되어 있는 것만 걸러내어, 각 Category 별로 모았다. 키워드의 배열 순서는 문헌 별 키워드의 TextRank 수치 합계가 기준이며, 배경색이 있는 것은 그 카테고리에서만 나타난 단어이다.

정치 개혁	외교 통일 국방	일자리	미래	성장 동력	농산 어촌	보건 복지	육아 교육	안전 환경	저출산 고령화 대책	행정	반려 동물	교통 건축 국토	경제 민주화	인권 성평등	문화 예술 체육 언론	기타
국민	국민	세요	처벌	해외 촉구	세요	세요	세요	예정	합니다	세요	세요	국민	합니다	세요	세요	
합니다	장애	국민	기준	정부	제한	코로나	학교	합니다	가정	세요	국민	합니다	합니다	국민	청원	국민
청원	합니다	코로나	한국	한국	세금	국민	학원	안전	위해	코로나	정부	국민	대통령	청원	입니다	합니다
반대	금지	청원	세요	세요	정부	입니다	청원	입니다	청원	입니다	처벌	청원	세요	피해	합니다	처벌
국회	청원	입니다	경찰	경제	국민	청원	교사	국민	아닌	국민	사건	입니다	공사	한국	코로나	지원
세요	나라	직원	청원	나라		합니다	교육	청원	국민	기업	대통령	구청	국가	범죄	조사	조사
처벌	한국	병원	입니다	원전		보험	즉각	운행	라고	대통령	합니다	조사	요청	사건	금지	가족
시장	세요	사태	서울	국민		지원	국민	제한		지원		지역	부정	입니다	촉구	나라
대통령	폐지	업체	국민			활동	인하	등급		조사		처벌	따른	병원	운영	연기

(전체 이미지 생략)

공통적으로 나타나는 키워드가 많기에, 해당 카테고리를 잘 나타낼 수 있게 하기 위해 의미를 가진 단어만을 남겼다. 그 카테고리에만 있는 단어가 없는 ‘반려동물’ 같은 카테고리나, 남긴 단어만으로는 알 수 없는 ‘미래’, ‘농산어촌’, ‘저출산고령화대책’ 카테고리는 불가피하게 제거한 단어를 포함했다. counter.WordCounter()를 통해 본 상위빈도 단어들에 비해, 애매한 결과가 나왔다. 4개 카테고리에 문제가 있어 이를 활용하기에는 어려웠기 때문에, 분석 데이터에 대한 이해를 통해, 차후 분석을 위한 발판으로 삼았다.

정치개혁	국회 채널 국회의원 대표 문재인 대통령 민주당 소년법 선거
외교통일국방	폐지 자유 참가 일본
일자리	직원 상장
미래	처벌 기준 한국 경찰 청원 서울 국민 수사
성장동력	해외 원전
농산어촌	축구 제한 세금 정부 국민
보건복지	활동 건강 기관 환자 바이러스 생계 재난 실시 부모 방지 책임 진행 치료 장애인 시행
육아교육	학원 즉각 인하 강의 생활 보육 어린이집 대학생 학년 개학 근무 위탁 중고 교육부 수시 미성년자 아동 유치원 학대
안전환경	운행 등급 주민 청소년 대구 행사 법안 길거리 사기 복무 환경
저출산고령화대책	예정 가정 아닌 국민
행정	판매 사용 격리 일반 피해자 종교 공무
반려동물	국민 정부 처벌 사건 대통령
교통건축국토	구청 아파트 조정 인사 잘못 행위 주택 가격 역사 버스 도로 환수 선고 불법 마음
경제민주화	공사 부정 사업
인권성평등	개인 이름 자유한국당 벌금 권력 초등 제기 회장
문화예술체육언론	축구 방송 절차 시민 관리
기타	연기 표준 배달 전화 탄핵 인천 영업

# 6. Sentiment Analysis

국민청원은 문재인정부가 국민과의 소통을 위해 내세운 정책으로 유사정책들과 비교해볼 때 접근성과 신속성이 높다는 것이 장점으로 꼽힌다. 이를 통해 국민들은 해결하지 못한 억울한 일이나 사회적 논의가 필요한 주제를 공론화할 수 있다.

국민청원 데이터를 감성분석하기 위해 먼저 국민청원 데이터에 대해 알 필요가 있다고 생각했다. 청원의 사전적 의미는 일이 이루어지도록 원하고 청한다는 뜻으로, 국민들은 국민청원이라는 웹사이트를 통해 개인적으로 해결하지 못한 억울한 일이나 사회적 논의가 필요한 주제를 공론화할 수 있다. 또한 사람들은 즐거운 일보다 억울한 일에 더 많은 반응을 보이는 경향이 있기 때문에 이러한 청원 데이터의 경우 부정적인 감정이 높을 수밖에 없다고 생각했다. 따라서 기본적으로 청원데이터가 부정적이라고 가정을 하고 실제로 얼마나 부정적인지 그 정도를 파악하고자 했다.

두번째로 특정 주제의 경우 사람들의 반응이 더 격해질 수 있다고 생각했다. 예를 들어 같은 범죄라도 그 대상이 성인이 아닌 어린아이일 경우 국민들은 더 많은 분노를 표출할 가능성이 있고 특정 나라와의 관계가 좋지 못할 경우 사소한 분쟁에도 많은 청원이 올라올 수 있다. 따라서 카테고리별로 감성점수의 절대값이 큰 문장들을 확인해 보았다.

감정분석을 위해 감성 사전과 점수처리방식에도 고려할 부분이 있다고 생각했다.

예를 들어 'ㅠㅠ'라는 표현은 현재 'Sentiword\_dict'에서 다른 유사한 표현들과 함께 -1의 감성점수를 가지고 있다. 그러나 리뷰같은 데이터의 경우 '너무 재밌어요 ㅠㅠ' 와 같은 긍정적인 의미로 해석이 가능하다. 즉, 부정적인 의미라기 보다는 '너무' 나 '완전' 처럼 앞 뒤 맥락에 따라 달라지는 경우로 해석하는 것이 맞다고 생각했다. '너무'나 '완전'이 0의 중립점수를 가진 것을 고려할 때 기본적인 'ㅠㅠ'의 점수도 0이 적절하다고 여겨진다. 다만 청원 데이터의 경우 텍스트가 부정적일 것이라고 가정했기 때문에 'ㅠㅠ'에 부정적인 점수를 그대로 유지했다. 그 외에도 '억울'이라는 단어가 감정사전에 긍정적인 단어로 정의되어 있거나, '도와주'라는 단어가 일반적으로 도와주다 라는 긍정적인 의미로 많이 쓰이지만 청원 데이터에서는 도와주세요 라는 부정적인 어조로 많이 쓰이기 때문에 부정적인 점수가 필요하다고 생각하여 다음과 같이 사전의 점수처리방식을 조정하였다.

도와주/VV;시/EP,1,0,1,0,0,0,NEG,1

억장/NNG;이/JKS;무너/VV,1,0,1,0,0,0,NEG,1

청원/NNG;하/XSA,1,0,1,0,0,0,NEG,1처벌/NNG;하/XSA,1,0,1,0,0,0,NEG,1

에서/JKB;억울/XR,1,0,1,0,0,0,NEG,1

에서/JKB;억울/XR;하/XSA,1,0,1,0,0,1,NEG,1이역만리/NNG;에서/JKB;억울/XR,1,0,1,0,0,0,NEG,1억울/XR,1,0,1,0,0,0,NEG,1

억울/XR;하/XSA,1,0,1,0,0,0,NEG,1

억울/XR;하/XSA;게/EC,1,0,1,0,0,0,NEG,1

1	category	no	date	end	status	agree	title	content
2	정치개혁	583525	19.11.11	19.12.11	청원종료	1952	민노총 폐: [단독] 어른 -	0.6363636363636360
3	인권/성평:	584524	20.01.20	20.02.19	청원종료	564	미성년 성·미성년 성·-	0.4545454545454540
5	정치개혁	585185	20.02.17	20.03.18	청원종료	346	중앙선거권 한국당 "선 -"	0.4444444444444440
7	기타	582768	19.09.20	19.10.20	청원종료	998	화성 연쇄·화성 연쇄 -	0.4444444444444440
8	안전/환경	584722	20.02.03	20.03.04	청원종료	163	코로나 병·요즘에 코 -	0.4400000000000000
9	기타	580501	19.05.29	19.06.28	청원종료	32081	민주노총·연이은 경 -	0.4166666666666660
10	인권/성평:	580670	19.06.05	19.07.05	청원종료	554	물카 범죄·물카 범죄 -	0.4133333333333330
13	정치개혁	580600	19.06.03	19.07.03	청원종료	1761	518가짜유 518가짜유 -	0.4000000000000000
14	정치개혁	580689	19.06.07	19.07.07	청원종료	1318	파스트 트·국민 다수 -	0.4000000000000000
16	기타	581708	19.07.29	19.08.28	청원종료	315	***에셋 치   <a href="https://twi">https://twi</a> -	0.3636363636363630
17	정치개혁	587474	20.03.31	20.04.30	청원진행중	225	마스크 배·활교안·ழ -	0.3636363636363630
18	인권/성평:	587197	20.03.26	20.04.25	청원종료	6427	텔레그램   [국민청원] -	0.3636363636363630
19	기타	584768	20.02.03	20.03.04	청원종료	224	중국 우한 현재 중국 -	0.3529411764705880
21	정치개혁	582222	19.08.22	19.09.21	청원종료	14073	문재인대통 김문수·자 -	0.3478260869565210
23	기타	582330	19.08.28	19.09.27	청원종료	972	강원 초등·또 다시 10 -	0.3442622950819670
27	기타	585940	20.03.02	20.04.01	청원종료	595	코로나19·신천지로 -	0.3414634146341460
28	인권/성평:	580125	19.05.13	19.06.12	청원종료	322	'장애인 출 저는 자체 -	0.3369565217391300
34	정치개혁	579910	19.05.02	19.06.01	청원종료	7571	정의당 해산 갑질정당 -	0.3333333333333330
35	안전/환경	584756	20.02.03	20.03.04	청원종료	2243	중국 입국 우리 나라 -	0.3260869565217390
38	안전/환경	584741	20.02.03	20.03.04	청원종료	484	우한폐렴·우한폐렴· -	0.3250000000000000
42	인권/성평:	583632	19.11.18	19.12.18	청원종료	518	남자 성기·여성 자위 -	0.3170731707317070
43	인권/성평:	584582	20.01.22	20.02.21	청원종료	280	정치인의·정치인의 -	0.3136094674556210
45	정치개혁	583696	19.11.22	19.12.22	청원종료	9347	문재인대통 헌기총·전 -	0.3103448275862060
47	인권/성평:	586294	20.03.05	20.04.04	청원종료	574	무서운 임·힘든 상인 -	0.3033707865168530
48	인권/성평:	587189	20.03.25	20.04.24	청원종료	17588	N번방 운동***과 26만 -	0.2987012987012980
49	인권/성평:	586316	20.03.05	20.04.04	청원종료	1185	이** 교주(사이비교주) -	0.2978723404255310
50	정치개혁	580494	19.05.29	19.06.28	청원종료	1953	대한애국동 대한애국동 -	0.2962962962962960
53	기타	583113	19.10.15	19.11.14	청원종료	7004	연예인(fx)·연예인(fx) -	0.2941176470588230
56	기타	583114	19.10.15	19.11.14	청원종료	3475	가수설리·가수 설리 -	0.2905982905982900
60	안전/환경	584532	20.01.20	20.02.19	청원종료	309	성폭행 살·안녕하세요 -	0.2886597938144320
62	기타	584452	20.01.14	20.02.13	청원종료	504	불법 성매·고양시 덕· -	0.2876712328767120
66	인권/성평:	587012	20.03.23	20.04.22	청원종료	1314	n번방 이동n번방 이동 -	0.2857142857142850
68	인권/성평:	583137	19.10.16	19.11.15	청원종료	2002	설리의 사·14일에 사 -	0.2857142857142850
69	기타	583711	19.11.24	19.12.24	청원종료	29425	사이버 범·누군가를 -	0.2822580645161290
70	인권/성평:	584480	20.01.16	20.02.15	청원종료	214	경찰과 짜·경찰과 짜 -	0.2754491017964070
74	안전/환경	584061	19.12.16	20.01.15	청원종료	2051	불법 체류·우리나라 -	0.2746113989637300
75	인권/성평:	580725	19.06.10	19.07.10	청원종료	30426	충북 **대·2019년 6월 -	0.2745098039215680
77	정치개혁	580148	19.05.13	19.06.12	청원종료	1103	가짜뉴스·저는 대한 -	0.2727272727272720
78	기타	585282	20.02.21	20.03.22	청원종료	3176	윤석열 검·뉴스타파0 -	0.2692307692307690
80	기타	586093	20.03.02	20.04.01	청원종료	854	가짜청원처 청와대 청 -	0.2692307692307690
83	정치개혁	585693	20.02.26	20.03.27	청원종료	3112	소상공인을 제발 살려 -	0.2678571428571420
84	인권/성평:	580526	19.05.30	19.06.29	청원종료	301	공항 직원·얼마전 공 -	0.2635658914728680
85	안전/환경	580655	19.06.05	19.07.05	청원종료	152	재난배상·재난안전 -	0.2632978723404250
90	인권/성평:	587036	20.03.23	20.04.22	청원종료	3726	텔레그램·이번 텔레 -	0.2624113475177300
91	아저/하녀	584907	20.02.04	20.03.05	청원종료	207	으하페 려·으하페 려 -	0.2619047619047610

(표의 일부)

먼저 청원데이터의 텍스트들이 어느정도 부정적으로 나타나는지를

확인해보고자 카테고리 구분을 하지 않고 전체 청원의 감성점수를 분석했다.

결과는 6106개의 청원이 부정점수를 가지고 있었고, 342개의 청원이 중립,

1627개의 청원이 긍정점수를 가지고 있어 생각보다는 긍정적인 점수를

가진 청원이 많았다. 전체데이터의 감성점수를 확인한 결과 약 0.6에서 -0.6

사이의 감성점수를 가지고 있음을 확인할 수 있었다.

## 감성점수가 부정적인 상위 20개의 청원

정치개혁	583525	19-11-11	19-12-11	청원종료	1952 민노총 폐지 동의하시길 부탁드립니다	-	0.6363636363636360
보건복지	581179	19-07-02	19-08-01	청원종료	3282 오늘부터 제 아내는 '장애의 정도가 심한 장애인' 입니다.	-	0.6048387096774190
육아/교육	581345	19-07-11	19-08-10	청원종료	132 등교시간	-	0.5000000000000000
인권/성평·	584524	20-01-20	20-02-19	청원종료	564 미성년 성폭행, 살인 형량을 늘려주세요.	-	0.4545454545454540
기타	582768	19-09-20	19-10-20	청원종료	998 화성 연쇄살인사건 범인을 처벌해주세요!!	-	0.4444444444444440
정치개혁	585185	20-02-17	20-03-18	청원종료	346 중앙선거관리위원회 조해주 구속 동의하시길 부탁드립니다	-	0.4444444444444440
안전/환경	584722	20-02-03	20-03-04	청원종료	163 코로나 병이 도는데 학교를 가야하나요?	-	0.4400000000000000
성장동력	581584	19-07-22	19-08-21	청원종료	648 [규정신설] 가짜 관내출장 달고 세금 뽑아 먹은 공무원, 50배 환수 규정 만들어주세요	-	0.4285714285714280
보건복지	584709	20-02-03	20-03-04	청원종료	117 신종 코로나바이러스 백신	-	0.4285714285714280
기타	580501	19-05-29	19-06-28	청원종료	32081 민주노총 해산 청원	-	0.4166666666666660
인권/성평·	580670	19-06-05	19-07-05	청원종료	554 몰카 범죄자 강력처벌	-	0.4133333333333330
보건복지	584840	20-02-04	20-03-05	청원종료	2862 우한교민 전안에서 격리 반대한다고 진천으로 옮기는 것을 반대 합니다	-	0.4090909090909090
보건복지	584595	20-01-23	20-02-22	청원종료	21242 우한폐렴 중국인관광객 막아주세요	-	0.4000000000000000
정치개혁	580600	19-06-03	19-07-03	청원종료	1761 518가짜유공자 정치인 퇴출시킵시다	-	0.4000000000000000
보건복지	582889	19-09-30	19-10-30	청원종료	1729 구충제 성분, 펜벤다졸의 암 치료에 대한 임상실험 시작해 주세요.	-	0.4000000000000000
정치개혁	580689	19-06-07	19-07-07	청원종료	1318 패스트 트랙을 취소 해주세요	-	0.4000000000000000
보건복지	584476	20-01-16	20-02-15	청원종료	3160 아주대 의료원장의 이국종 교수님에대한 육설과 갑질을 수사해 주십시오	-	0.3846153846153840
보건복지	586633	20-03-13	20-04-12	청원종료	523 대구·경북 코로나19 특별재난지역 선포를 촉구합니다.	-	0.3666666666666660
인권/성평·	587197	20-03-26	20-04-25	청원종료	6427 릴레그램 N번방. 미성년자를 성폭행한 범인의 신상 공개를 요청합니다.	-	0.3636363636363630
기타	581708	19-07-29	19-08-28	청원종료	315 **에셋 처벌해주시길 바랍니다.	-	0.3636363636363630

코로나와 N번방 사건 관련 청원들의 감성 점수가 매우 높은 것을 확인할 수 있다. N번방 사건은 앞에서 언급했듯이 미성년이 피해자였던 사건이기 때문에 사회적으로 더 많은 분노가 표출된 것 같다. 코로나는 중국에 대한 여론을 악화시켰고 우한과 중국에 대한 부정적인 청원들이 높은 감성점수를 가진 것을 확인할 수 있었다.

## 감성점수가 긍정적인 상위 20개의 청원

안전/환경	582082	19-08-16	청원종료	241	포항시 검은수돗물	0.6666666666666666
일자리	580439	19-05-27	청원종료	3772	엘리베이터 유령 노동자를 살려주세요.	0.5000000000000000
안전/환경	583410	19-11-04	청원종료	324	5등급 경유차 4개월 운행중단에 따른 환경개선부담금, 보유세, 보험금 반환해주세요	0.3529411764705880
안전/환경	586636	20-03-13	청원종료	1031	콜센터 당분간 휴업했으면 좋겠어요	0.2857142857142850
보건복지	585669	20-02-26	청원종료	319	대구.경북에 헬기 방역을	0.2857142857142850
문화/예술	581379	19-07-11	청원종료	285	평양 대동강 맥주를 즐기게 해 주세요!	0.2777777777777770
문화/예술,	580642	19-06-04	청원종료	15778	학교체육을 말살하는 스포츠혁신위원회의 2차권고안 철회를 청원합니다.	0.2500000000000000
안전/환경	580809	19-06-12	청원종료	558	길거리 흡연에 관한 법을 강화해주세요	0.2500000000000000
정치개혁	583369	19-10-31	청원종료	812	국회의원 정수 50% 감축, 2번까지 만 할수 있다	0.2380952380952380
교통/건축,	581191	19-07-03	청원종료	201	개선해주세요	0.2352941176470580
반려동물	580686	19-06-07	청원종료	744	유기된 동물들의 따뜻한 보금자리가 필요합니다!!	0.2278481012658220
교통/건축,	582964	19-10-04	청원종료	193	강남이 좋으냐 묻지 마시고, 강북 인프라 투자 강화 요청	0.2250000000000000
안전/환경	585679	20-02-26	청원종료	2585	신천지 교인 명단 공개 및 신천지 비밀 포교 장소 공개요청	0.2142857142857140
기타	586889	20-03-20	청원종료	2038	의정부지청 감사하라	0.2142857142857140
보건복지	585509	20-02-24	청원종료	1408	코로나19 확산방지위해 재택근무 가능한 직장인은 재택근무 하도록 명하라!	0.2121212121212120
교통/건축,	583781	19-11-28	청원종료	1348	국가에 강제수용되는 토지에 관한 양도세 100% 감면	0.2121212121212120
반려동물	580451	19-05-27	청원종료	1400	200여마리 유기견을 무분별하게 안락사시킨 ***에게 엄중처벌 바랍니다	0.2105263157894730
보건복지	586133	20-03-02	청원종료	8747	마스크국민께공평신속배급을 동사무소에서세대별로공급하기	0.2083333333333330
교통/건축,	580549	19-05-30	청원종료	234	자동차 레몬법의 강제적, 포괄적 시행을 청원합니다.	0.2058823529411760
정치개혁	582350	19-08-28	청원종료	13349	조국 법무부 장관 후보와 가족들에 무차별적으로 음혜와 허위 사실을 만들어 퍼트리	0.2040816326530610

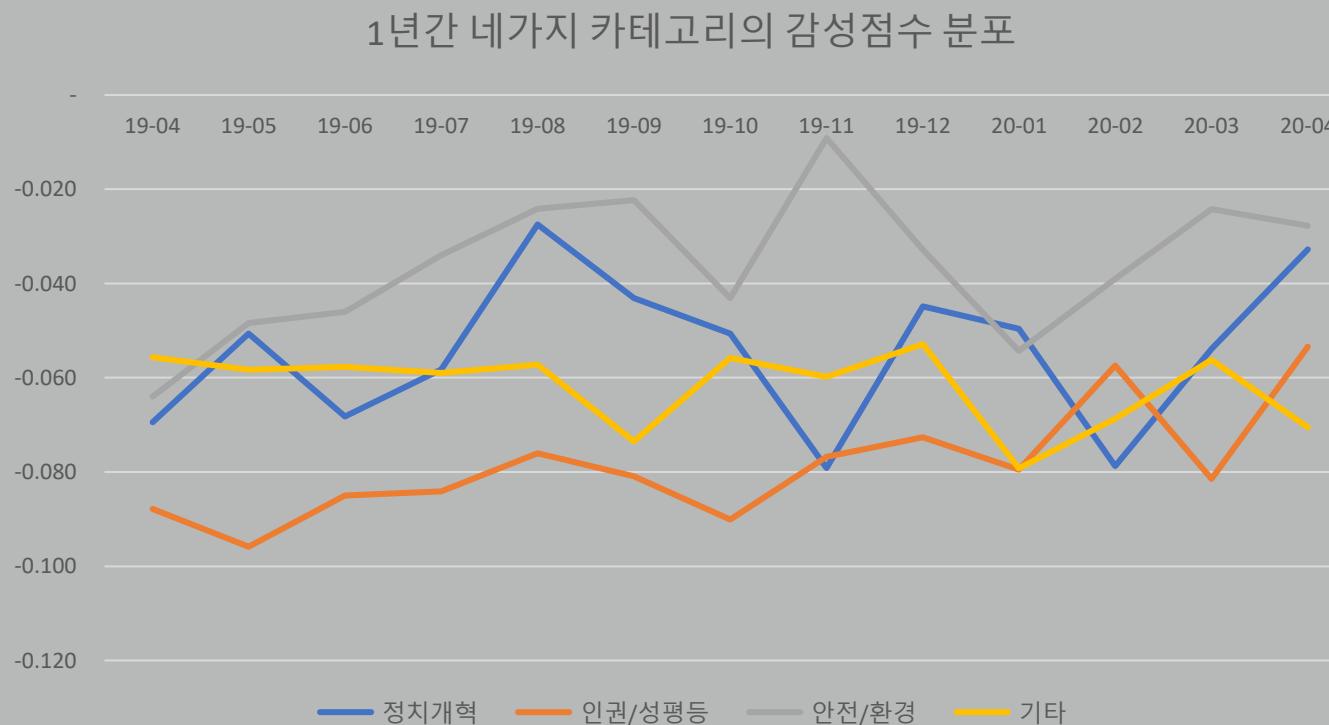
감성점수가 높은 청원 들에서는 부정적인 청원들과 달리 일부 주제에 특정되는 경향이 약한 것을 확인할 수 있었다. 대구, 경북에 대한 방역요청이 신천지 교인 공개요청 등 코로나 단어가 직접 포함되지는 않아도 코로나와 관련된 청원들이 눈에 띄었다. 긍정적인 청원에서는 '감사.' '개선' 등 비교적 완곡하거나 긍정적인 표현들이 많이 보였다.

## 동의 수가 많은 상위 20개의 청원

안전/환경	586819	20-03-18	2715626 텔레그램 n번방 응의자 신상공개 및 포토라인 세워주세요	- 0.18181818181810
안전/환경	586880	20-03-20	2026252 텔레그램 n번방 가입자 전원의 신상공개를 원합니다	- 0.1813953488372090
정치개혁	579682	19-04-22	1831900 자유 한국당 정당해산 청원	0.0156250000000000
기타	585683	20-02-26	1504597 문재인 대통령님을 응원 합니다!	- 0.0683760683760683
정치개혁	584936	20-02-04	1469023 문재인 대통령 탄핵을 촉구합니다.	- 0.0573248407643312
인권/성평	585290	20-02-22	1449521 신천지 예수교 증거장막성전(이하, 신천지)의 강제 해체(해산)을 청원합니다.	- 0.0122950819672131
기타	587624	20-04-02	981638 렌트카 훔쳐 사망사고를 낸 10대 엄중 처벌해주세요	- 0.1219512195121950
보건복지	584593	20-01-23	761833 중국인 입국 금지 요청	- 0.1363636363636360
정치개혁	582190	19-08-21	757730 청와대는 조국 법무부장관 후보자의 임명을 반드시 해주십시오!!	0.0533333333333333
안전/환경	586885	20-03-20	650579 가해자 n번방박사,n번방회원 모두 처벌해주세요	- 0.2238805970149250
육아/교육	586860	20-03-20	533883 저희 25개월딸이 초등학생 5학년에게 성폭행을 당했습니다	- 0.1745120551090700
안전/환경	587335	20-03-29	516729 박사방 회원 중 여아살해모의한 공익근무요원 신상공개를 원합니다.	- 0.0937500000000000
보건복지	585682	20-02-26	491136 코로나19의 확산방지를 위해 애써주시는 문재인대통령님과 질병관리본부 및 정부부처 관계지	-
정치개혁	582351	19-08-28	481076 기밀누설죄를 범한 윤석열 총장을 처벌해 주십시오.	-
인권/성평	587323	20-03-27	466265 N번방 담당판사 오덕식을 판사자리에 반대, 자격박탈을 청원합니다.	- 0.1509433962264150
인권/성평	586879	20-03-20	457487 N번방 대회 참여자들도 명단을 공개하고 처벌해주세요	- 0.1529411764705880
기타	586609	20-03-12	439648 코로나19로 위기에 처한 국민들을 위해 국회의원들의 웰빙반납 또는 삭감을 건의합니다	- 0.0448979591836734
안전/환경	583543	19-11-11	415691 어린이들의 생명안전법안 통과를 촉구해주시길 간곡히 부탁드립니다	- 0.0350404312668463
인권/성평	587352	20-03-29	393153 “오늘 너 킬(KILL)하다”라며 술을 먹이고 제 딸을 학동 강간한 미성년자들을 고발합니다.	- 0.0528634361233480
기타	587169	20-03-25	385617 수출용 코로나19 진단키트 이름을 독도로 해주세요	- 0.0153846153846153

동의수가 많은 청원은 그만큼 많은 수의 사람들이 청원에 공감하고 관련된 주제에 분노를 느꼈다는 것으로 해석할 수 있다. 동의 수를 기준으로 보았을 때 특정 주제에 대한 집중도가 더 강해졌음을 알 수 있었다. 20개의 청원에서 미성년자와 관련된 청원은 9개나 차지하고 있고, 코로나, N번방 관련 청원들도 매우 많은 비율을 차지하고 있었다. 동의수가 높은 청원들의 감성점수를 살펴보면 거의 -0.2~0.2사이에 모두 포함되는 것을 알 수 있으며 중립인 청원들도 볼 수 있다. 분석 전에는 감성점수가 큰 청원들이 주제에 대한 국민들의 강한 반응을 통해 많은 동의 수를 얻을 수 있을 것이라고 생각했지만 오히려 그 반대로 객관적이고 차분하게 상황을 설명하고 논리적으로 전달하는게 많은 동의 수를 얻을 수 있었던 게 아닌가 생각된다.

감성점수와 동의수 별로 청원을 살펴보면서 눈에 띄었던 점은 같은 주제에 속하는 청원들일지라도 여러 카테고리로 나누어 졌다는 것이다. N번방 관련청원은 안전/환경, 인권/성평등, 기타 등으로 나누어 졌고 코로나와 관련된 청원들도 보건/복지, 인권/성평등, 기타 등 다양한 카테고리로 묶여 있음을 알 수 있었다. 이러한 점에서 국민청원의 카테고리에 대해 개선할 필요성을 느끼게 되었고 상위20개의 청원에서 특히 많이 출현한 인권/성평등, 정치/개혁, 안전/환경, 기타 네가지의 카테고리를 가지고 시간의 흐름에 따른 감정점수 분포를 나타내 보았다.



분포 결과 안전/환경 카테고리는 오히려 예상보다 중립에 가까운 분포를 많이 보였다. 가장 낮은 부정적인 점수 분포를 보인 주제는 인권/성평등이다.

category	no	date	agree	title	score
인권/성평·	587197	20-03-26	6427	텔레그램 N번방. 미성년자를 성폭행한 범인의 신상 공개를 요청합니다.	- 0.3636363636363630
인권/성평·	586294	20-03-05	574	무서운 임대인<**시설관리공단>	- 0.3033707865168530
인권/성평·	587189	20-03-25	17588	N번방 운영자 *** 사형해주세요	- 0.2987012987012980
인권/성평·	586316	20-03-05	1185	이** 교주에게 국민 피해보상 청구합시다	- 0.2978723404255310
인권/성평·	587012	20-03-23	1314	n번방 이용자 신상 모조리 밝혀주세요.	- 0.2857142857142850
인권/성평·	587036	20-03-23	3726	텔레그램 n번방 사건으로 피해입은 74명의 성폭력 피해자 분들과 추가될 수 있는 피해자분들	- 0.2624113475177300
인권/성평·	586961	20-03-23	1231	여성으로써 살아가기 무섭습니다. 부디 저의 호소를 들어주세요.	- 0.2516778523489930
인권/성평·	586209	20-03-03	281	포털사이트 모든 기사에 대한 댓글을 막아주세요.	- 0.2475247524752470
인권/성평·	587205	20-03-26	3416	n번방 전 운영자 와치맨의 형량을 검토해주세요.	- 0.1813725490196070
인권/성평·	587456	20-03-30	2661	범죄피해여성의 성착취 사진.동영상촬영 유포.특수강간및 강간치상.장애인간음.성매매강요	- 0.1800535475234270
인권/성평·	587328	20-03-27	56592	N번방 담당 오덕식 판사의 권한 자격 박탈을 요청합니다	- 0.1690140845070420
인권/성평·	587478	20-03-31	4964	05년생 집단 폭행 사건	- 0.1643835616438350
인권/성평·	586225	20-03-03	479	계약직 우체국택배 위탁원입니다. 부정부패한 관리자의 보복으로 재계약 못할 위기에 놓여	- 0.1634615384615380
인권/성평·	586470	20-03-09	971	억울합니다	- 0.1598440545808960
인권/성평·	586879	20-03-20	457487	N번방 대화 참여자들도 명단을 공개하고 처벌해주시오	- 0.1529411764705880
인권/성평·	587180	20-03-25	1783	청각장애학생의 온라인 강의 수강에 따른 학습권을 보장해주세요	- 0.1527093596059110
인권/성평·	587323	20-03-27	466265	N번방 담당판사 오덕식을 판사자리에 반대,자격박탈을 청원합니다.	- 0.1509433962264150
인권/성평·	586605	20-03-12	1217	질병치료목적—직장휴가불인정	- 0.1445783132530120
인권/성평·	587117	20-03-24	2806	온라인 성매매 특별법 만들어 주세요!	- 0.1443298969072160

비교적 최근에 낮은 부정점수를 보인 20년 3월의 데이터이다.

감성점수가 낮은 청원들을 살펴보니 대부분 N번방과 관련된 주제임을 알 수 있었다. 미성년자가 피해자였던 사건이고 사회적으로 많은 분노를 일으켰기 때문에 상당히 부정적인 감정이 점수에 표현된 것 같다.

category	no	date	agree	title	score
정치개혁	585185	20-02-17	346	중앙선거관리위원회 조해주 구속 동의하시길 부탁드립니다	- 0.4444444444444440
정치개혁	585693	20-02-26	3112	소상공인을 살려주세요!! 하루하루가 지옥입니다	- 0.2678571428571420
정치개혁	585812	20-02-27	10261	문재인 대통령 탄핵을 반대하는 국민청원을 반대합니다	- 0.2558139534883720
정치개혁	585612	20-02-25	7879	중국폐렴 중국말만듣는 한국대통령탄핵 찬성합니다	- 0.2205882352941170
정치개혁	584919	20-02-04	4829	중국인에게 우한폐렴 치료비 무상지원 안됩니다.	- 0.2076923076923070
정치개혁	585607	20-02-25	2145	신천지 생활비 지원금지 청원	- 0.2000000000000000
정치개혁	584851	20-02-04	240	중국우한폐렴 진천 인재개발원 격리수용 반대합니다.	- 0.1791044776119400
정치개혁	585617	20-02-25	23419	길게 말 안합니다. 문재인 코로나폐렴 사태 책임 지고 하야 하십시오	- 0.1650485436893200
정치개혁	585675	20-02-26	8511	문재인 대통령을 탄핵청원하는 이들을 우리는 탄핵한다!	- 0.1578947368421050
정치개혁	585813	20-02-27	5442	박능후 보건복지부 장관, 즉각 '사퇴할 것을 촉구'합니다!	- 0.1470588235294110
정치개혁	585687	20-02-26	123848	문재인대통령의 탄핵을 반대합니다!!~	- 0.1428571428571420
정치개혁	585138	20-02-14	6396	이인영 구속 동의하시길 부탁드립니다	- 0.1428571428571420
정치개혁	584918	20-02-04	3670	문재인 대통령은 국민앞에 사죄하고 하야하십시오	- 0.1428571428571420
정치개혁	585722	20-02-27	8154	문재인 대통령을 탄핵해주세요. 탄핵을 못하겠으면 중국인 입국금지라도 해주세요.	- 0.1363636363636360
정치개혁	585852	20-02-28	12145	문재인 대통령의 하야를 촉구합니다	- 0.1304347826086950
정치개혁	585734	20-02-27	1017	대구 봉쇄 조치 반대	- 0.1250000000000000
정치개혁	584909	20-02-04	1462	어르신들의 주말 서울 도심 집회 불허 청원	- 0.1176470588235290
정치개혁	585808	20-02-27	2933	박능후 보건복지부 장관의 사퇴를 요구합니다.	- 0.1139896373056990
정치개혁	585083	20-02-11	210801	전자개표기 폐지 동의하시길 부탁드립니다	- 0.1111111111111110

정치/개혁 주제는 월별로 점수의 분포 폭이 매우 컸다. 특정 정치 이슈에 따라 사람들의 반응이 바뀔 수 있다는 점이 영향을 끼친 게 아닐까 추측해보고 있다. 위는 가장 최근에 부정적인 점수를 받은 2020년 2월 데이터이다.

2020년 2월은 코로나가 본격적으로 확산되던 시기로 코로나와 관련된 주제들이 대부분을 차지하고 있었다. 앞의 인권/성평등과 마찬가지로 사람들이 분노할 만한 주제가 크게 이슈가 될 경우 많은 사람들이 부정적인 감정을 드러내는 청원을 많이 쓴 것을 알 수 있다. 여기에서는 중국, 신천지 등에 대한 부정적인 견해나 대통령과 보건부 장관이 사퇴를 요구하는 등 사람들이 코로나와 관련해서 많은 부정적인 감정을 가지고 있었다는 것을 확인 할 수 있었다.

# 7. Document Clustering

# 국민청원 주제 분석 및 딥러닝 기반 답변 가능 청원 예측

우 윤 희<sup>†</sup> · 김 현 희<sup>††</sup>

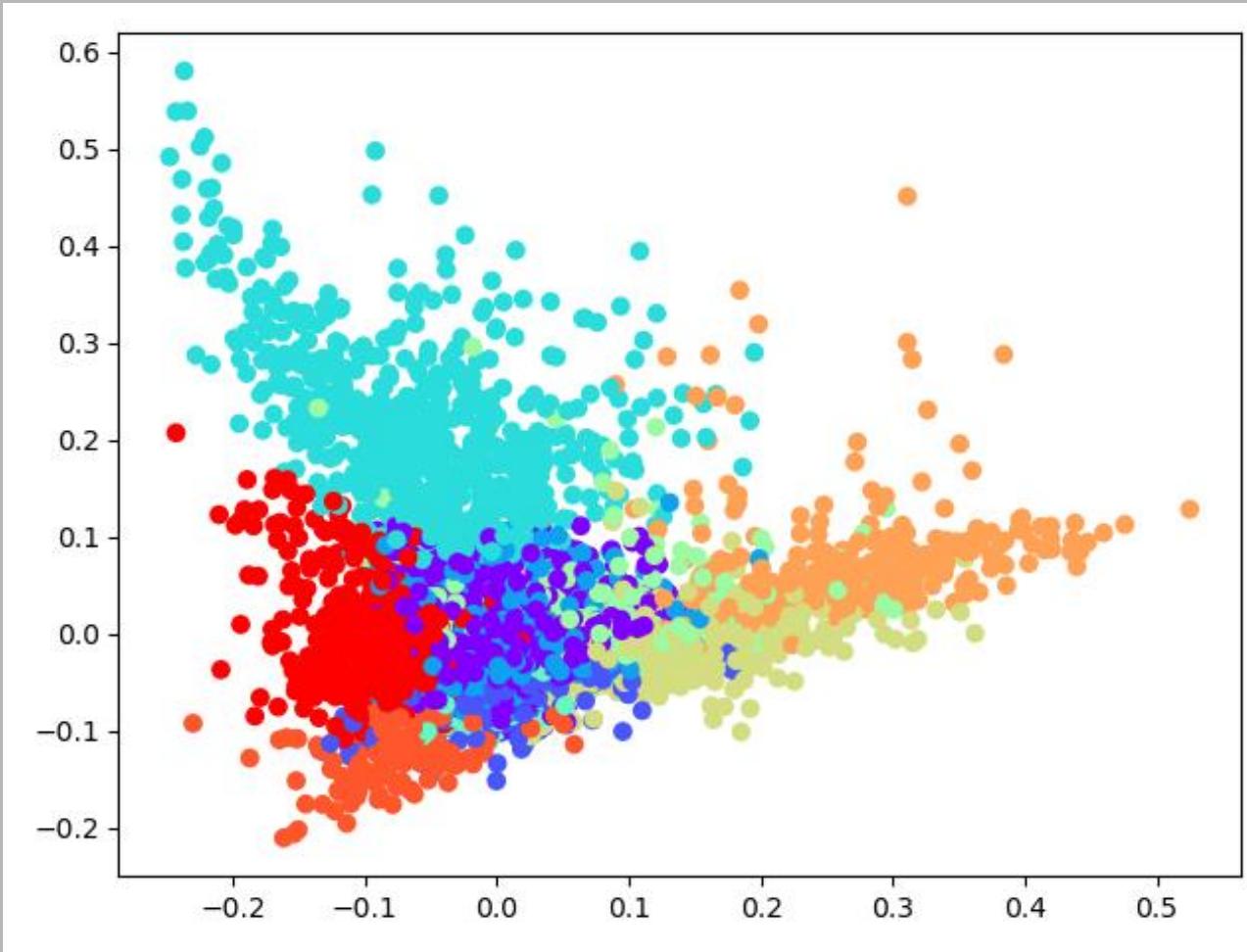
논문에 따르면, 군집의 개수가 10일 때 청원글이 가장 효과적으로 분류되었다고 한다.

상대적으로 K-Means는 시간이 오래 걸리지 않기 때문에, max\_iter를 1만으로 정했다.

```
== ==
cluster_namek-means:
cluster_name k-means
Clustering sparse data with KMeans(max_iter=10000, n_clusters=10, n_init=1)
done in 2.276s

k-means
Cluster 0: 회사 직원 보험 업체 근무 계약 업무 임금 기업 지급
Cluster 1: 아이 어린이집 교사 보육 선생 부모 아동 유치원 엄마 원장
Cluster 2: 코로나 병원 신천지 확진 환자 중국 입국 국민 치료 격리
Cluster 3: 피해자 사건 가해자 처벌 경찰 범죄 수사 여성 폭행 성범죄
Cluster 4: 국민 일본 정부 분양 도시 주민 아파트 지역 나라 국가
Cluster 5: 마스크 구매 판매 가격 약국 코로나 공급 국민 구입 정부
Cluster 6: 검찰 국민 수사 국회의원 대통령 의원 국회 장관 총장 개혁
Cluster 7: 지원 동물 대출 코로나 자영업 학원 소득 국민 재난 정부
Cluster 8: 학교 학생 개학 수업 교육 교사 아이 연기 온라인 코로나
Cluster 9: 시험 응시 수험 연기 필기 자격증 자격 실기 출제 국가
```

ScatterPlot으로 근접한 것끼리 색이 비슷한 것을 보아, 클러스터링이 잘 되었다는 것을 확인



```

124      # agglo를 그대로 가져온것
125      for document_id, cluster_label in enumerate(cluster_labels):
126          if cluster_label not in clusters:
127              clusters[cluster_label] = []
128              clusters[cluster_label].append(document_id)
129              print(str(cluster_label) + " -- " + str(document_id))
130      # agglo를 그대로 가져온것
DocumentClustering > print_results() > if self.name == 'k-means'

```

시각화만으로는 전체적인 것 만을 볼 수 있고, 문헌 각각이 어떻게 분류되는지는 알 수 없었다. 그것이 가능한 agglo에 있는 코드를 그대로 따와서 적용했다.

petitionKmean.txt			
[6 4 4 ... 2 3 4]	1	6	579682
6 -- 0	2	4	579683
4 -- 1	3	4	579684
4 -- 2	4	4	579685
4 -- 3	5	3	579686
3 -- 4	6	7	579687
7 -- 5	7	4	579688
4 -- 6	8	4	579689
4 -- 7	9	6	579690
6 -- 8	10	1	579691
1 -- 9			
1 -- 10			
0 -- 11			

차후 클러스터링 결과를 활용하기 위해, 각 청원의 Category를 Cluster로 대체한 파일을 만들었다.

# 8. Document Classification

프로젝트에서 쓴 모든 Classification은

LinearSVC가 가장 성능이 좋았기 때문에, Classifier Model은 전부 LinearSVC(max\_iter=1000)를 사용했다. 시간이 오래 걸리므로, max\_iteration을 최소한인 1000으로 하였다.

## LinearSVC

차원 축소

특징 선택(Feature Selection):

가장 좋은 특징을 선택하고 나머지 제거

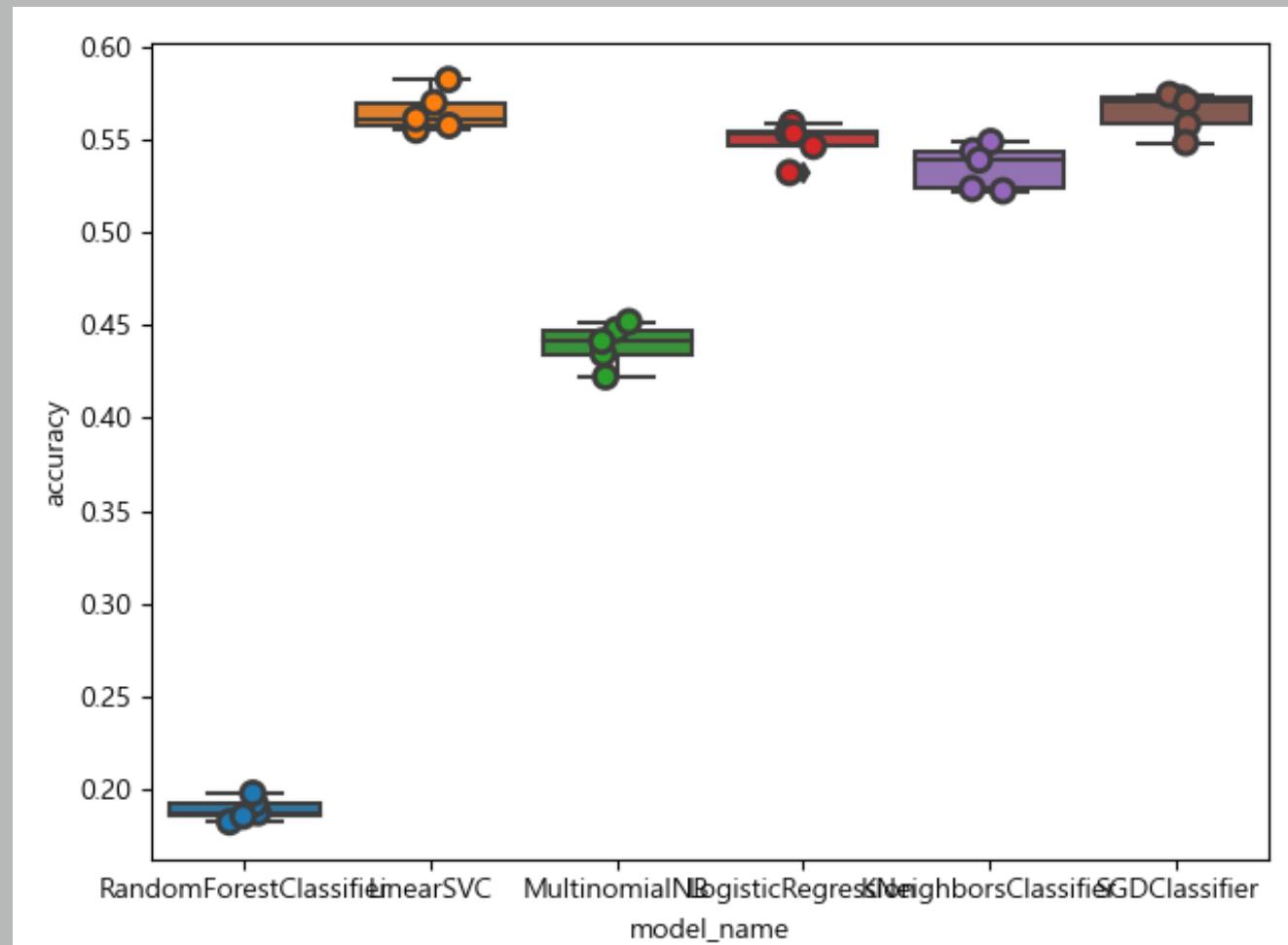
래퍼(Wrapper) 기법: 순차적 특징 선택

- 가장 좋은 결과를 선택하기 전에 다양한 조합을

## 샘플링

- 특징의 부분집합으로 여러 개의 미니 모델을

생성하고 가장 좋은 결과를 점수화



· 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		

C:\Program Files\Python37\lib\site-packages\

sklearn\metrics\\_classification.py:1221:

UndefinedMetricWarning:

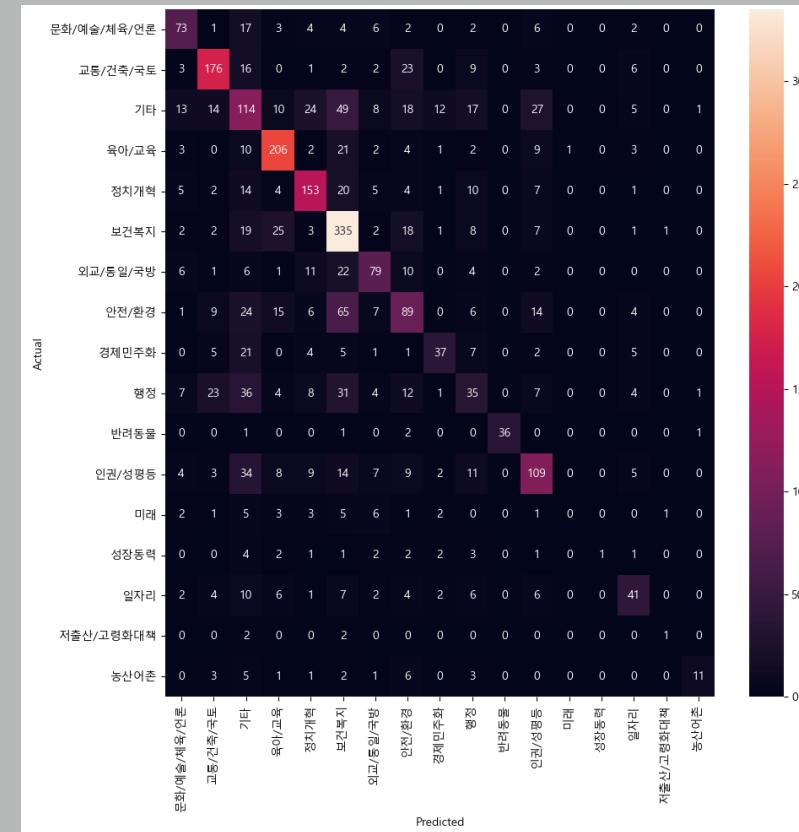
Precision and F-score are ill-defined and being set to 0.0 in labels with no predicted samples.

Use `zero\_division` parameter to control this behavior.

\_warn\_prf(average, modifier, msg\_start, len(result))

	precision	recall	f1-score	support
문화/예술/체육/언론	0.62	0.50	0.55	120
교통/건축/국토	0.73	0.73	0.73	241
기타	0.29	0.42	0.34	312
육아/교육	0.72	0.74	0.73	264
정치개혁	0.65	0.66	0.66	226
보건복지	0.55	0.83	0.66	424
외교/통일/국방	0.61	0.50	0.55	142
안전/환경	0.43	0.38	0.41	240
경제민주화	0.70	0.34	0.46	88
행정	0.36	0.16	0.22	173
반려동물	1.00	0.71	0.83	41
인권/성평등	0.56	0.49	0.52	215
미래	0.00	0.00	0.00	30
성장동력	0.00	0.00	0.00	20
일자리	0.63	0.43	0.51	91
저출산/고령화대책	0.00	0.00	0.00	5
농산어촌	0.00	0.00	0.00	33
accuracy			0.55	2665
macro avg	0.46	0.41	0.42	2665
weighted avg	0.54	0.55	0.53	2665

## 대한민국 청와대 국민청원(청와대에서 정한) 청원 분야 17개를 기준으로 Train을 진행



Warning 메시지를 보면, 현 분류 방식이 문제가 많다는 것을 알 수 있다. HeatMap에서 대각선만을 보면 알 수 있듯이, '미래', '성장동력', '저출산/고령화대책', '농산어촌'을 제대로 분류하지 못한다. F-Measure를 살펴보면 평균적으로 50%만큼 맞게 분류를 했다는 것을 알 수 있다. 카테고리가 17개나 됨에도 50%를 맞게 분류한 것은 자동으로 분류하는 것을 감안하면 나쁘지 않다. 그러나 실제로 괜찮게 분류하기 위해서는 80% 이상은 제대로 분류를 해야 한다.

자동분류를 하기 위한 방안으로, 데이터의 범위 및 양을 바꾸거나, 카테고리를 새로 만드는 것 중에서 카테고리를 바꾸는 것으로 정했다.

1		효율적인 정치참여를 위한 국민청원 데이터 시각화 서비스 조애리(AERE CHO); 김희진(HEEJIN KIM)
		한국HCI학회 학술대회, 2019, Volume 2019, Issue 2 학술지논문 Full Text Online
		미리보기 ▾  관련 추천논문 ▶

2		효율적인 정치참여를 위한 국민청원 데이터 시각화 서비스 제안 조애리(Aere Cho); 김희진(Heejin Kim); 유재영(Jae Young Yun)
		한국디자인학회 학술발표대회 논문집, 2018, Volume 2018, Issue 11 학술지논문 Full Text Online
		미리보기 ▾

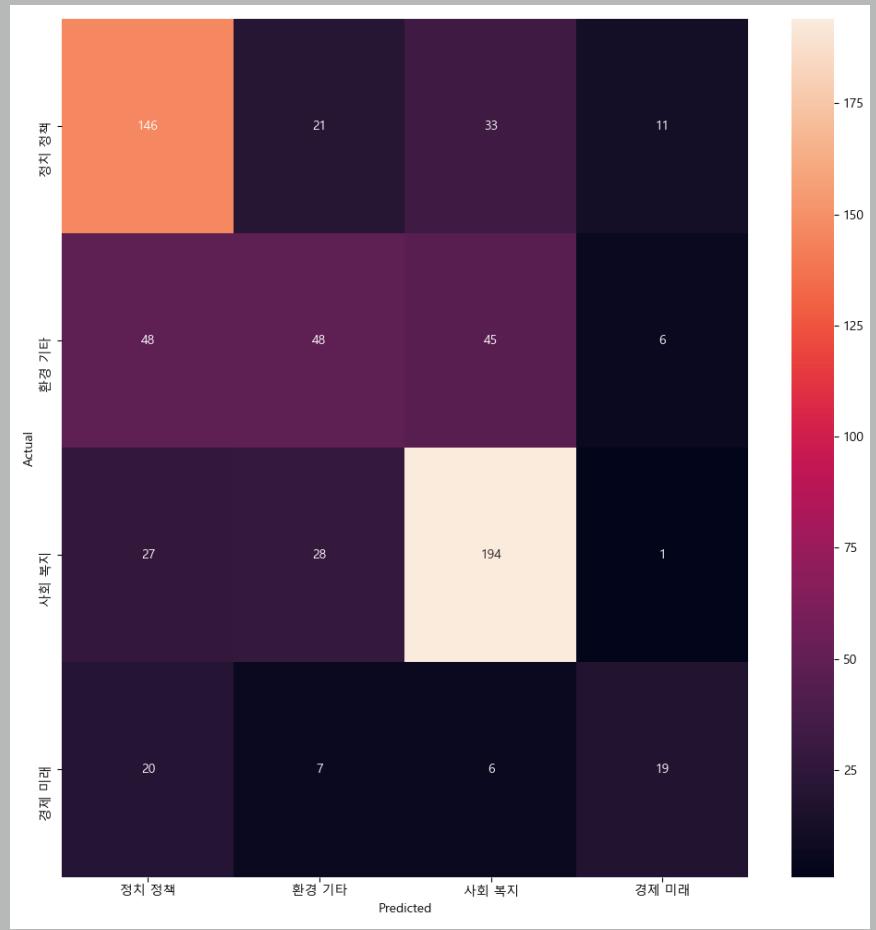


<http://voiceofpeople.kr/guide>에서 새로 분류한 기준에 따라, 각 카테고리를 4개로 대체하여 Classification을 진행했다.

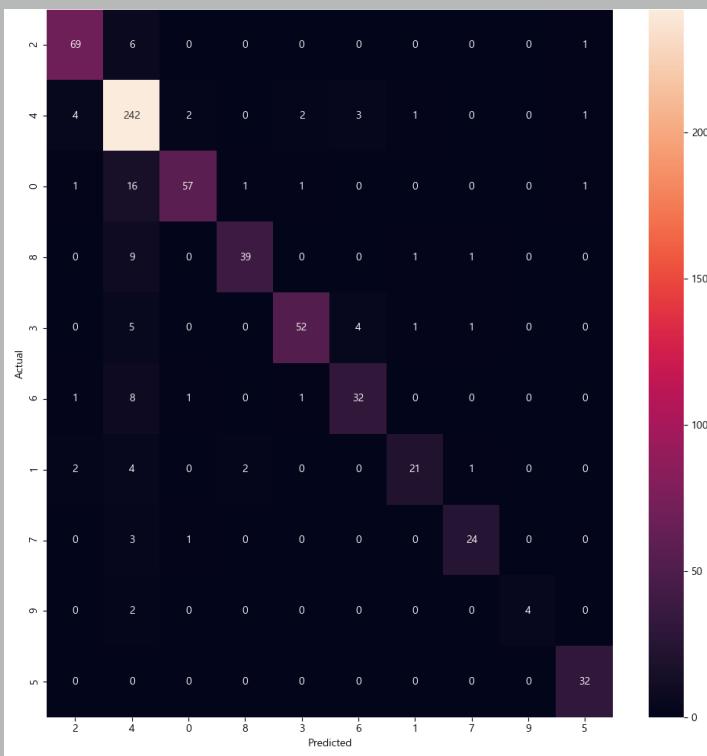
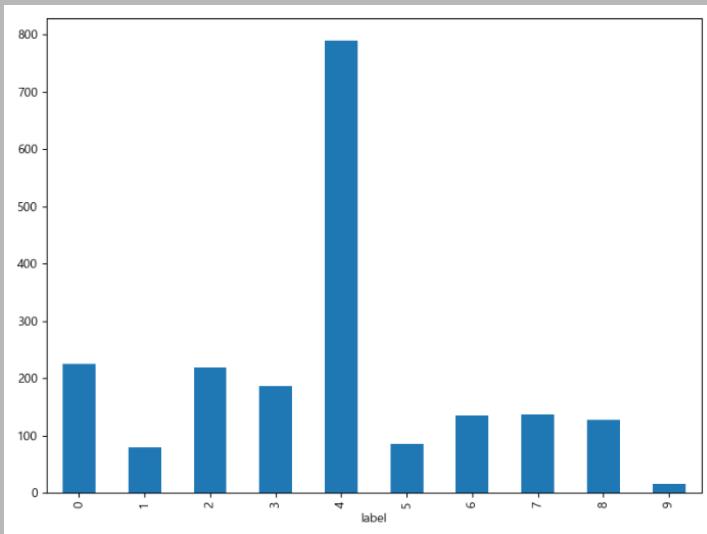
petition_final_voiceofpeople.txt			
1	정치 정책	579682	20
2	사회 복지	579683	20
3	정치 정책	579684	20
4	환경 기타	579685	20
5	정치 정책	579686	20
6	정치 정책	579687	20
7	정치 정책	579688	20
8	환경 기타	579689	20
9	정치 정책	579690	20
10	사회 복지	579691	20
11	사회 복지	579692	20
12	사회 복지	579693	20
13	정치 정책	579694	20
14	사회 복지	579695	20
15	환경 기타	579696	20
16	환경 기타	579697	20
17	환경 기타	579698	20
18	환경 기타	579699	20
19	경제 미래	579700	20

		precision	recall	f1-score	support
정치 정책	0.61	0.69	0.65	211	
환경 기타	0.46	0.33	0.38	147	
사회 복지	0.70	0.78	0.73	250	
경제 미래	0.51	0.37	0.43	52	
		accuracy		0.62	660
		macro avg	0.57	0.54	660
		weighted avg	0.60	0.62	660

원래 17개 카테고리와 달리 올바른 분류가 0인 분류는 없어졌다. 이는 단지 카테고리 개수가 줄어들어서 성능이 올라간 것뿐이므로, 만족스럽지 못하다. 명확히 대분류한 기준이 없어서인지, 목표인 80%에 미치지 못했다.



	precision	recall	f1-score	support
2	0.90	0.91	0.90	76
4	0.82	0.95	0.88	255
0	0.93	0.74	0.83	77
8	0.93	0.78	0.85	50
3	0.93	0.83	0.87	63
6	0.82	0.74	0.78	43
1	0.88	0.70	0.78	30
7	0.89	0.86	0.87	28
9	1.00	0.67	0.80	6
5	0.91	1.00	0.96	32
accuracy			0.87	660
macro avg	0.90	0.82	0.85	660
weighted avg	0.87	0.87	0.86	660



(k=10) K-Means Cluster에 따라 Classification 해보면, 매우 분류가 잘 되었음을 알 수 있다. 1번과 6번을 제외하면 80%를 넘었으며, 평균적으로도 목표치인 80%를 넘었다.

Train에서 나아가 Predict 하기에 문제가 없어 보였기 때문에, 새로운 데이터를 수집했다

— 청원종료 —

# 저희 아파트 경비아저씨의 억울함을 풀어주세요...

참여인원 : [ 446,434명 ]



파일 920, 폴더 0

종류:

## JSON 파일 형식의 모든 파

위치

C:\Users\kimsixsue\PyCh

39

3.08MB (3,234,377 바이트)

디스

4.92MB (5,169,152 바이트)

petition\_idx 587828 ~ 588752 총 920개(기간은 1달)를 수집했다.

원래 Dataset 8000여 개는 4월 25일에 수집했고, 30일 이상 후에 새 데이터를 수집했기 때문에, 데이터 선정에 문제가 없다.

```

cluster_name k-means
Clustering sparse data with KMeans(max_iter=10000, n_clusters=10, n_init=1)
done in 2.276s

```

```

k-means
Cluster 0: 회사 직원 보험 업체 근무 계약 업무 임금 기업 지급
Cluster 1: 아이 어린이집 교사 보육 선생 부모 아동 유치원 엄마 원장
Cluster 2: 코로나 병원 신천지 확진 환자 중국 입국 국민 치료 격리
Cluster 3: 피해자 사건 가해자 처벌 경찰 범죄 수사 여성 폭행 성범죄
Cluster 4: 국민 일본 정부 분양 도시 주민 아파트 지역 나라 국가
Cluster 5: 마스크 구매 판매 가격 약국 코로나 공급 국민 구입 정부
Cluster 6: 검찰 국민 수사 국회의원 대통령 의원 국회 장관 총장 개혁
Cluster 7: 지원 동물 대출 코로나 자영업 학원 소득 국민 재난 정부
Cluster 8: 학교 학생 개학 수업 교육 교사 아이 연기 온라인 코로나
Cluster 9: 시험 응시 수험 연기 필기 자격증 자격 실기 출제 국가

```

새로 수집한 청원 920개의 카테고리와 분류된 Cluster를 비교해봤다.

오른쪽 표는 카테고리와 간의 관계를 나타낸 분류표이다. 파란색 배경은 카테고리와 클러스터 두가지 측면에서 동시에 높은 비율을 차지하고 있고, 노란색 배경은 한쪽 측면만 높은 비율을 차지하고 있는 것을 나타낸 것이다. 반려동물과 클러스터0을 예시로 들면, 클러스터 0 중에 반려동물 카테고리에 속하는 토픽이 가장 많으며, 반려동물 카테고리 중에서도 클러스터0으로 분류된 토픽이 가장 많다는 뜻이다. 해당 1달동안 올라온 청원은 5가지 클러스터로만 분류가 되었다. 또한 4가지 카테고리가 특정 클러스터에 분류되는 경향이 큰 것을 확인할 수 있었다. (0-반려동물, 2-보건복지, 4-안전/환경+육아/교육)

	0	2	3	4	8
경제민주화		6		28	
교통/건축/국토	1	33		10	
기타	4	23		82	1
농산어촌		4			
문화/예술/체육/언론		9		17	
미래		2		8	
반려동물	10	3		8	
보건복지	4	69	2	65	
성장동력		10			2
안전/환경	6	25	1	28	
외교/통일/국방	1	16	1	33	
육아/교육	3	50	2	87	
인권/성평등	4	8		74	
일자리	2	18		24	
저출산/고령화대책					1
정치개혁	3	10		43	
행정	1	44		34	

## 청와대 국민청원에서는 무엇이 일어나는가? :

자연어 처리를 활용한 청와대 국민청원 분석\*

송준모\*\* · 박영득\*\*\*

### 논문 요약

청와대 국민청원은 단순한 분노의 발산 통로인가, 직접 민주주의의 장인가. 본 연구는 청와대 국민청원에 대한 상반된 가치판단을 넘어서, 감정이 정치적 참여와 판단에 미치는 영향을 경험적으로 분석하고자 한다. 이를 위해 청와대 국민청원 페이지 개설 시점에서부터 연구 시작 시점까지 1년 이상 축적된 게시물을 전수 수집하였다(증복 제기 후 총 88,113 건). 그리고 자연어 처리(구조적

다음은 predict에서 사용한 920개의 청원 중에서 가장 추천수가 높은 48개를 추출해

제대로 분류가 되는지를 살펴보았다. 키워드 10개만을 기준으로 본 결과, 48개 중

19개의 청원만이 제대로 분류되었다. Train단계에서 F1-Score가 높아

predict단계에서도 좋은 성능을 보일 것으로 예상했지만, Predict단계에서는 제대로 분류되지 않은 것을 알 수 있었다.

predict단계에서 아쉬운 성능을 보인 것은 학습에 쓴 데이터가 8천건 밖에 안 되었던 점이 영향을 미쳤다고 생각했다. 또한 코로나라는 특수한 상황이 수집 청원 기간에서

많은 비율을 차지해 과적합(overfit) 문제가 생긴 것도 하나의 원인으로 보았다. 선행

연구에서는 청원을 88,000건 이상 수집했던 것을 보면 더 많은 데이터를 분석할 수

있었다면 조금 더 나은 결과를 가져올 수 있었을 것으로 생각된다. 다음에 다시 분석할

기회가 생긴다면 이번에는 국민청원 2.0이후가 아닌 전체 국민청원 데이터를 대상으로 분석해보고자 했다.

분류	Cluster	판단 이유	title	content
안전/환경	3	O: 범죄	밝은미래의 청소년들을 문제라고 생각합니다	
반려동물	2	O: 병원	동물병원은 반려인을 선에 역행하는 법 개정이	
인권/성평등	4	O: 아파트	저희 아파트 경비아저씨 같은데요, 연예계 종사	
기타	8	O: 온라인	유튜브 크리에이터 스톤이들을 처벌해주세요	
보건복지	4	O: 일본	정부, 美日한국전 참전·합니니다 절대 반대합니다	
외교/통일/국방	4	O: 일본	일본에게 마스크를 보내한 결정 부탁드립니다	
일자리	0	O: 일자리	코로나19 경제위기속 청년입니다 감사합니다	
행정	0	O: 임금	대구 봉사 의료진 체불임금과 연대 호소합니다	
인권/성평등	2	O: 치료	화마속 10명구한 불법 필요하다고 생각합니다.	
농산어촌	2	O: 코로나	쌀 수출 제한을 해주세요! 미리 대비 해야합니다	
문화/예술/체육/언론	2	O: 코로나	피트니스, 요가, 필라테대출/지원 업무 이관 및	
미래	2	O: 코로나	준비된 전문가 정은경(부 질병예방센터 센터장	
보건복지	2	O: 코로나	생활지원사들의 위치추적. 저희들은 업무 특성상	
외교/통일/국방	2	O: 코로나	군인의 연가보상비 삭감처럼 다니고 싶습니다	
육아/교육	2	O: 코로나	등교 개학 시기를 미루기로 미루어주시기를 청	
일자리	2	O: 코로나	**전자*** 기사들 사람용합니다. 한마디로 출근	
정치개혁	2	O: 코로나	전국민 긴급재난지원금에서 먼저 기부해주시면	
행정	2	O: 코로나	재난 지원금 소득하위 기에 이 청원을 올립니다	
외교/통일/국방	3	O: 피해자	*** 항공사로 부터 대행에 의해 수정되었습니다	
경제민주화	4	X	**자산운용 ***** 원유/갈동안 5일간 진행)의 으	
교통/건축/국토	4	X	악용될 소지가 다분한 폐지되기를 청원합니다.	
기타	4	X	간통죄 처벌 할수있게 적으로 해왔습니다. 그로	
문화/예술/체육/언론	4	X	SBS 거짓보도에 공식시고 총작 직인파일도 그를	
미래	4	X	증권시장을 교란하는 제도 마련 부탁드립니다	
반려동물	4	X	반려동물 억울한 죽음 의사 분께 전화 발신하이	
성장동력	4	X	자영업자 지원 은행이 살펴주심을 기원합니다	
안전/환경	4	X	경남 진주 일가족 살해? 이 사건은 직계비속0	
육아/교육	4	X	울산 초등학교 1학년 이 민감해야 하며, 성인지	
일자리	4	X	설계사 시험을 볼수있가해 주시면 좋겠습니다.	
저출산/고령화대책	4	X	출산을 감소 해결 및 교통시에 해결할 수 있다.	
정치개혁	4	X	정**교수님 보석허가 히 의해 수정되었습니다]	
행정	4	X	화재로부터 10여명의 죽 것이라고 생각합니다	
보건복지	3	X	긴급돌봄 촉소, 신청 기 생각해주시길 바랍니다	
육아/교육	3	X	보육교사 근무환경 개선지도록 조치해주세요	
경제민주화	2	X	수익형호텔 바로 잡아 점도 한 끝하고 있다. 정	
교통/건축/국토	2	X	서울 강남구 재건축 지! 표합니다. 감사합니다	
기타	2	X	중앙선거관리위원회는 평원합니다. 감사합니다	
성장동력	2	X	다목적 방사광 가속기'가 공현한 것도 방사광 기	
안전/환경	2	X	평택시 도일동 소각장 라고 하고 있다고 하면서	
교통/건축/국토	0	X	**과적단속반의 부당함! 행 하자 후 바로 신호등	
기타	0	X	평택시 **번째 코로나현 현재 대구에서 피땀흘	
반려동물	0	X	사기, 동물학대를 일삼! 의해 수정되었습니다	
보건복지	0	X	덴탈마스크 공적판매로 정스러울 것 같습니다	
안전/환경	0	X	대학가들이 장기적인 텔 바랍니다. 감사합니다	
외교/통일/국방	0	X	2021 도쿄올림픽에서 하면 옥일기가 계속하여	
육아/교육	0	X	개학연기 !태 전파가 될것입니다	
인권/성평등	0	X	세월호 희생자들을 모욕 중1의 글이었습니다)	
정치개혁	0	X	대구시장 탁핵을 요청하고 대구를 제일 아끼	

# 9. Topic Modeling

나의 큰 O는 log x야

## [토픽 모델링] DMR의 하이퍼 파라미터 추정

그냥 공부  
by 적분  $\int 2tdt=t^2+c$  · 2017. 10. 3. 13:29

1 2

앞서 여러 글에서 밝혔듯이 LDA(Latent Dirichlet Allocation, 잠재 디리클레 할당)은 베이즈 추론을 아주아주 잘 확장해서 특정한 단어들이 한 문헌에 등장할때 문헌 집합 내의 각 단어들이 어떤 주제에 속했는지를 계산하는 생성 모형을 제공해줍니다. (이에 대한 자세한 설명은 [\[잠재 디리클레 할당 파헤치기\] 2. 디리클레 분포와 LDA](#) 게시물을 참조해주세요!) LDA에서는 모든 문헌을 동등하게 취급했는데요, 각각 문헌의 특성을 반영하지 못한다는 한계를 극복하기 위해 2012년 D Mimno는 DMR(Dirichlet Multinomial Regression) 토픽 모델링을 제안했습니다. DMR 토픽 모델링에 대한 개략적인 설명은 [\[토픽 모델링\] 확장된 LDA 기법들 - ATM, DMR](#) 게시물에 잘 나와있지만, 여기에서는 실제로 DMR의 추론 과정을 이해하지 못해서 다 날려먹고 대충 대충 적었는데요, 드디어 조금 이해를 할 수 있게된 고로 글을 새로 파서 정리해보았습니다.

## 다항 토픽 모델링

### • 다항 토픽 모델

- 문헌과 주제분포를 기반으로 저자, 발행처, 참고문헌, 날짜 정보 등의 문헌의 메타데이터 특성(feature)을 제3의 파라미터로 설정하여 토픽 결과를 도출하는 LDA 기반 토픽 모델링 기법

방법을 정하기 위해 수업 PPT와 bab2min 블로그를 참고했다. 시계열 분석을 위해, 각 문헌의 특성을 반영하는 Dirichlet-Multinomial Regression 토픽 모델링을 주로 이용했으며, 시계열과 상관 없는 곳에는 LDA를 이용했다.

The screenshot shows the PyCharm IDE interface with the following details:

- File Bar:** File, Edit, View, Navigate, Code, Refactor, Run, Tools, VCS, Window, Help.
- Project Bar:** pyTextMiner > pyTextMiner > \_\_init\_\_.py
- Code Editor:** The \_\_init\_\_.py file is open, showing Python code for reading a field-delimited file with a year column. The code defines a class `CorpusFromFieldDelimitedFileWithYear` and its constructor `\_\_init\_\_`. It uses a try-except block to handle index errors and prints out-of-index messages. The code also initializes `self.docs` and `self.pair\_map`.
- Sidebar:** Project (1: Project), Structure (I: Structure), Favorites (2: Favorites).

```
102     class CorpusFromFieldDelimitedFileWithYear(Corpus):
103         def __init__(self, file, doc_index=1, year_index=0):
104             array = []
105             id = 0
106             pair_map = {}
107             with open(file, encoding='utf-8') as ins:
108                 for line in ins:
109                     fields = line.split('\t')
110                     try:
111                         array.append(fields[doc_index])
112                         # pair_map[id] = fields[year_index]
113                         pair_map[id] = fields[year_index][0:7] # 프로젝트 용
114
115                         id += 1
116                     except IndexError:
117                         print("out of index " + str(id))
118
119             self.docs = array
120             self.pair_map = pair_map
```

시계열 분석을 위해 기준을 세웠다.  
수집한 청원의 청원시작일은 2019년 4월 22일부터  
2020년 4월 7일까지이다. 약 1년에 가까운 기간이기  
때문에, 기준을 한 달씩으로 잡았다. 이를 통해 시작하는  
달은 2019년 4월, 끝나는 달은 2020년 4월이 되었다.

topic	perplexity
17	3,641.5
34	3,223.2
51	3,086.4
68	2,961.3
85	2,877.9
102	2,858.0
119	2,831.9
136	2,801.8
153	2,800.4
170	2,783.2
187	2,768.0
200	2,743.6
204	2,769.4
221	2,716.0
238	2,698.1
239	2,741.2
240	2,717.6

241	2,755.9
242	2,793.8
243	2,787.4
247	2,729.5
248	2,806.9
249	2,776.5
250	2,710.9
251	2,786.1
252	2,761.2
253	2,720.6
254	2,706.5
<b>255</b>	<b>2,685.1</b>
256	2,761.9
260	2,805.3
272	2,718.0
300	2,752.3
400	2,839.3
500	2,922.7
1,000	4,259.5

이진탐색 알고리즘을 이용해 topic number를 수정했고, Dirichlet-Multinomial Regression을 기준으로 토픽 개수는 255개가 최적임을 알아냈다.  
 선행연구들처럼 분류를 한 후에 각 분류마다 토픽모델링을 한 것이 아니기 때문에, 이처럼 전체에서 토픽개수가 많이 나왔다.  
 이는 카테고리를 따지지 않고, 국민의 목소리가 다른 주제가 매우 폭넓다는 것을 나타낸다.

The screenshot shows the PyCharm IDE interface. The top menu bar includes File, Edit, View, Navigate, Code, Refactor, Run, Tools, VCS, Window, and Help. The title bar says "pyTextMiner - pyTextMinerTopicModel.py". The left sidebar has sections for Project, Structure, and I: Structure. The main code editor window displays the following Python code:

```
623     # ranking the candidates of labels for a specific topic
624     labeler = tp.label.FoRelevance(mdl, cands, min_df=5, smoothing=1e-2, mu=0.25)
625     for k in range(mdl.k):
626         print("== Topic #{0} ==".format(k))
627         print("Labels:", ', '.join(label for label, score in labeler.get_topic_labels(k, top_n=5)))
628         for word, prob in mdl.get_topic_words(k, top_n=10):
629             print(word, prob, sep='\t')
630         print()
631
pyTextMinerTopicModel > dmr_model()
```

The screenshot shows a browser window displaying the tomotopy.label API documentation. The URL is bab2min.github.io/tomotopy/v0.8.1/kr/label.html. The page content includes:

```
def get_topic_labels(self, k, top_n=10)
```

토픽 `k`에 해당하는 레이블 후보 상위 `n`개를 반환합니다.

**Parameter**

`k` : int 토픽을 지정하는 정수 `top_n` : int 토픽 레이블의 개수

Get\_topic\_words는 해당 토픽의 키워드임을  
직관적으로 알 수 있었으나,  
get\_topic\_labels가 무엇인지는 직관적으로  
알 수 없었다.  
다행히 get\_topic\_labels가 무엇인지 설명이  
있었기에, 토픽의 라벨을 정할 수 있었다.

Topic	LABEL 선택	LABEL ALL											
133	회사 대표	회사 회사, 회사 운영, 경영진, 회사 대표, 회사 정상	회사	직원	대표	이사	기업	개인	퇴사	본사	본인	진행	
41	활성 정책	기대 효과, 효과 기대, 적극 활용, 활성 정책, 구조 개선	필요	가능	정책	개선	효과	해결	경제	국가	현재	제안	

토픽 각각의 후보 라벨 중에 라벨을 선택했다. 엉뚱하게 뭉쳐진 토픽 일부를 제외하고, 대부분 키워드를 통해 라벨을 하나로 정할 수 있었다.

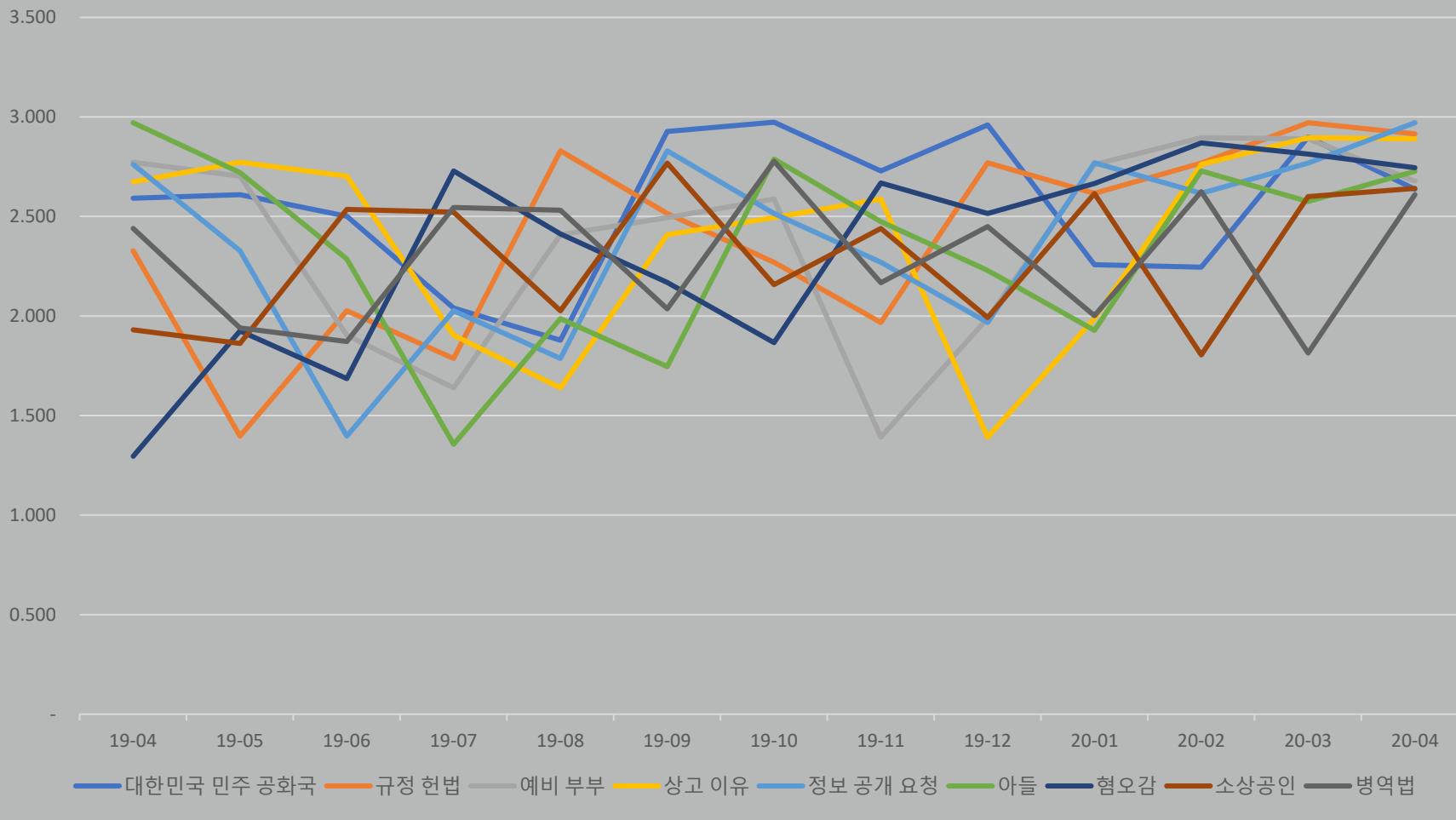
강화도, 증구 청, 평화 소녀, 본토, 소요 비용	공항	인천	관광객	청년	면세점	증구	인력거	방문	설치	평화
-----------------------------	----	----	-----	----	-----	----	-----	----	----	----

토픽 255개 중, 216개의 라벨을 하나로 정할 수 있었다.

LABEL 선택											
대한민국 민주 공화국	국가	대한민국	경제	분노	분열	정치	행위	세력	권력	반	
규정 헌법	헌법	자유	대한민국	민주	기본	보장	재판소	행위	침해	평등	
예비 부부	결혼	결혼식	선원	선박	부부	위약금	무궁화	예비	국제	신혼	
상고 이유	법원	판결	소송	변호사	재판	판사	대법원	청구	제기	결정	
정보 공개 요청	정보	공개	개인	제공	해당	확인	내용	보호	정확	자료	
아들	아들	부모	엄마	언니	가족	자식	모습	아빠	아버지	서울	
혐오감	광고	악플	인터넷	광고물	댓글	노출	선정	자살	연예인		
소상공인	공인	자영업	소상	매출	운영	상공	영업	현실	임대료	직원	
병역법	복무	요원	의무	병역	국방	강제	국가	기간	현역	지급	

1년 전체, 수치의 합계가 가장 큰 토픽 9개이다. 특정 기간에 갑자기 불타오르거나 식지 않고, 국민의 관심이 가장 많은 토픽이다.

## 지속적으로 많이 나오는 토픽

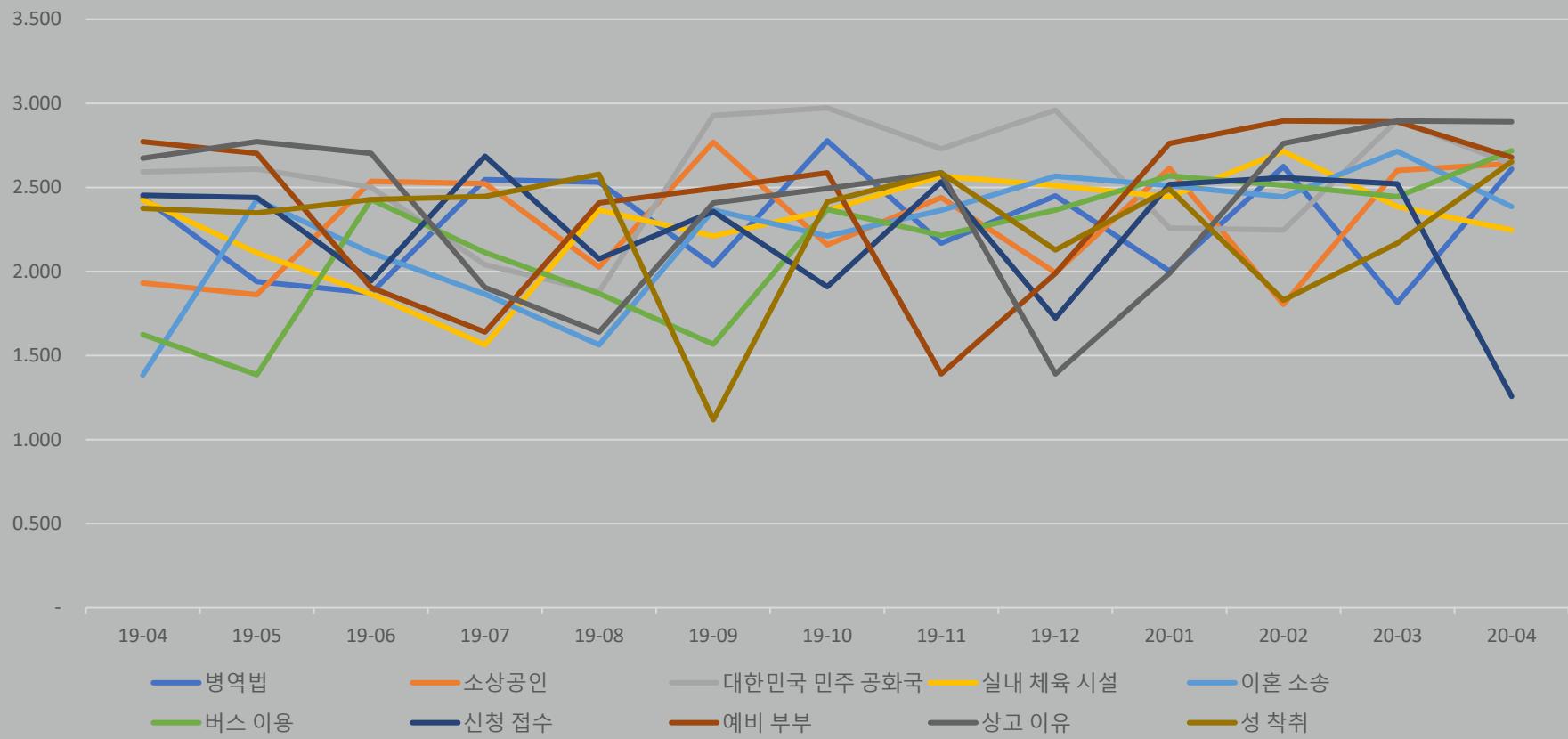


라벨을 기준으로 보면, 어느  
청원에나 들어갈만한  
'아들'이라는 가족 관련 단어  
토픽이 있다. 또한 꾸준히  
문제제기는 되지만, 해결하기  
어려운 토픽들이 대부분을  
차지하고 있다.

LABEL 선택										
<b>병역법</b>	복무	요원	의무	병역	국방	강제	국가	기간	현역	지급
<b>소상공인</b>	공인	자영업	소상	매출	운영	상공	영업	현실	임대료	직원
<b>대한민국 민주 공화국</b>	국가	대한민국	경제	분노	분열	정치	행위	세력	권력	
<b>실내 체육 시설</b>	시설	체육	운영	운동	권고	체육관	태권	휴업	업종	도장
<b>이혼 소송</b>	남편	아내	이혼	결혼	가정	가족	딸	아빠	소송	부부
<b>버스 이용</b>	버스	이용	시	지하철	노선	교통	철도	구간	운행	출퇴근
<b>신청 접수</b>	전화	신청	담당자	통화	연락	피해	접수	답변	해결	
<b>예비 부부</b>	결혼	결혼식	선원	선박	부부	위약금	무궁화	예비	국제	신혼
<b>상고 이유</b>	법원	판결	소송	변호사	재판	판사	대법원	청구	제기	결정
<b>성 착취</b>	처벌	텔레그램	성범죄	사건	N번방	피해자	착취	범죄		

지속적으로 많이 나오는 토픽과 비슷하지만 약간 다른 측면의 토픽들이다. 이번에는 국민들이 갖는 관심의 정도가 언제든 비슷한 토픽들 10개를 모아봤다.

## 갑자기 변하지 않는 일관적인 토픽



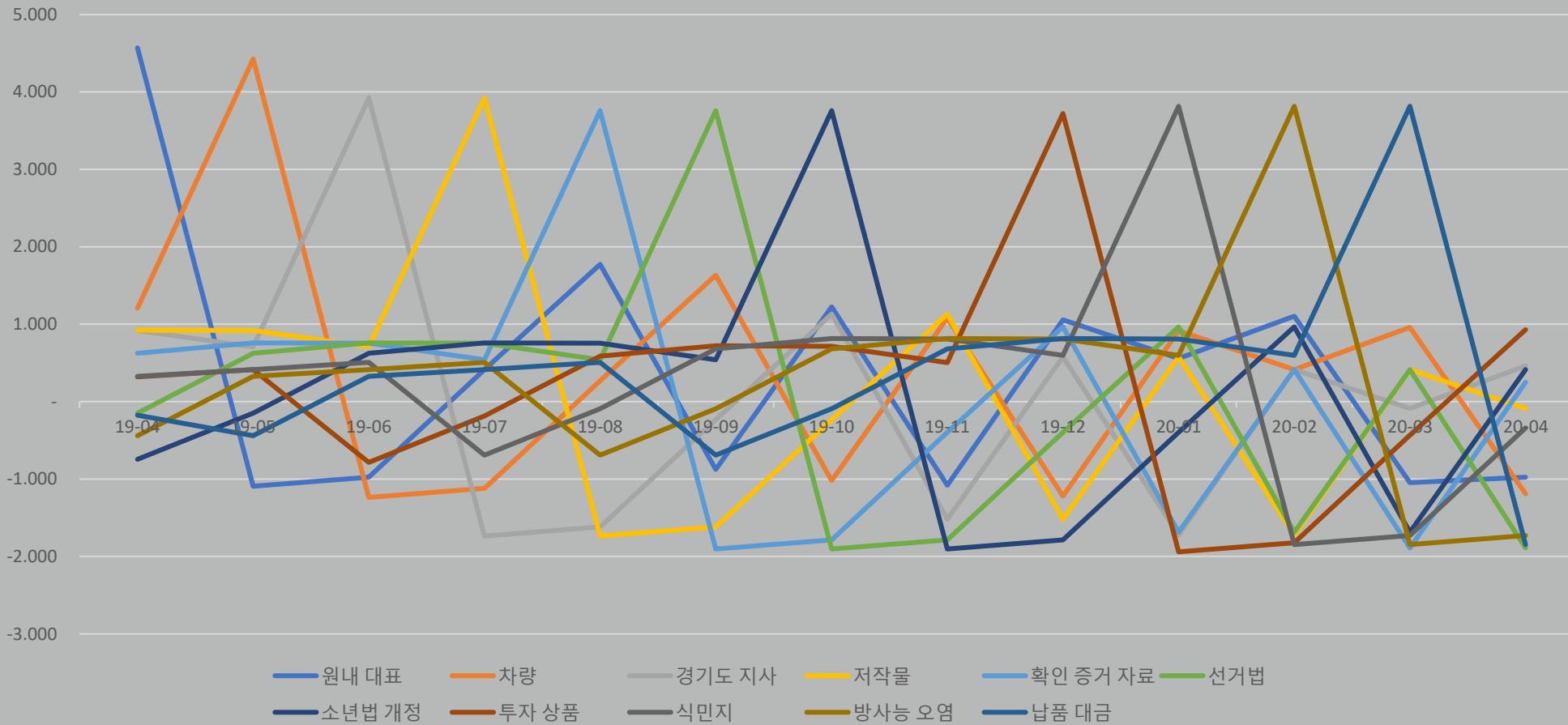
앞에서 본 5개를 제외하고,  
실내 체육 시설, 이혼 소송,  
버스 이용, 신청 접수, 성 착취,  
이렇게 5개의 주제가  
나타났다. 이혼 소송 및 신청  
접수, 버스 이용은 언제나  
제기되는 문제이다.  
이외 주제는 코로나와 관련해  
'실내 체육 시설', 범죄 처벌  
관련해 '성 착취' 이 두가지가  
나타났다. 청원에 참여한 추천  
수를 반영하지 않았기 때문에,  
편차가 적은 토픽에 들어갔다.

LABEL 선택										
원내 대표	의원	당	자유한국당	대표	민주당	정당	해산	국회의원	정치	
차량	차량	자동차	운행	수리	발생	비용	구입	경유	중고차	엔진
경기도 지사	경기도	이재명	지사	체납	공무원	관리	도민	기사	경기	헬기
저작물	작품	예술	작가	영화	아이디어	문화	만화	소재	주인공	드라마
확인 증거 자료	내용	확인	자료	제출	해당	작성	요청	서류	진행	요구
선거법	선거	투표	후보	선거법	총선	결정	공직	운동	실시	당선
소년법 개정	청소년	성인	미성년자	처벌	부모	소년법	소년	나이	연령	
투자 상품	금융	투자	거래	상품	도박	자본	범죄	설립	화폐	회원
식민지	일본	한국	역사	독도	위안부	운동	경제	전쟁	영토	불매
방사능 오염	일본	방사능	올림픽	원전	후쿠시마	안전	도쿄	오염	수입	위험
납품 대금	업체	납품	거래	유통	물건	업자	협력	제품	사업자	하도급

LABEL 선택	19-04	19-05	19-06	19-07	19-08	19-09	19-10	19-11	19-12	20-01	20-02	20-03	20-04
원내 대표	4.568	- 1.093	- 0.976	0.413	1.774	- 0.875	1.223	- 1.079	1.057	0.556	1.103	- 1.046	- 0.977
차량	1.208	4.425	- 1.236	- 1.119	0.270	1.631	- 1.018	1.079	- 1.222	0.914	0.413	0.960	- 1.189
경기도 지사	0.919	0.706	3.924	- 1.738	- 1.620	- 0.231	1.130	- 1.520	0.578	- 1.723	0.413	- 0.088	0.459
저작물	0.923	0.919	0.706	3.924	- 1.738	- 1.620	- 0.231	1.130	- 1.520	0.578	- 1.723	0.413	- 0.088
확인 증거 자료	0.624	0.758	0.753	0.541	3.758	- 1.903	- 1.785	- 0.397	0.965	- 1.685	0.413	- 1.889	0.248
선거법	- 0.149	0.624	0.758	0.753	0.541	3.758	- 1.903	- 1.785	- 0.397	0.965	- 1.685	0.413	- 1.889
소년법 개정	- 0.747	- 0.149	0.624	0.758	0.753	0.541	3.758	- 1.903	- 1.785	- 0.397	0.965	- 1.685	0.413
투자 상품	0.319	0.413	- 0.784	- 0.187	0.586	0.721	0.716	0.504	3.721	- 1.940	- 1.823	- 0.434	0.927
식민지	0.327	0.413	0.507	- 0.690	- 0.093	0.680	0.814	0.809	0.597	3.815	- 1.847	- 1.729	- 0.340
방사능 오염	- 0.442	0.327	0.413	0.507	- 0.690	- 0.093	0.680	0.814	0.809	0.597	3.815	- 1.847	- 1.729
납품 대금	- 0.178	- 0.442	0.327	0.413	0.507	- 0.690	- 0.093	0.680	0.814	0.809	0.597	3.815	- 1.847

파란배경은 해당 토픽이 가장  
핫했을 때를 뜻하고, 주황배경은  
가장 관심이 적었을 때를 뜻한다.

## 갑자기 이슈화된 토픽



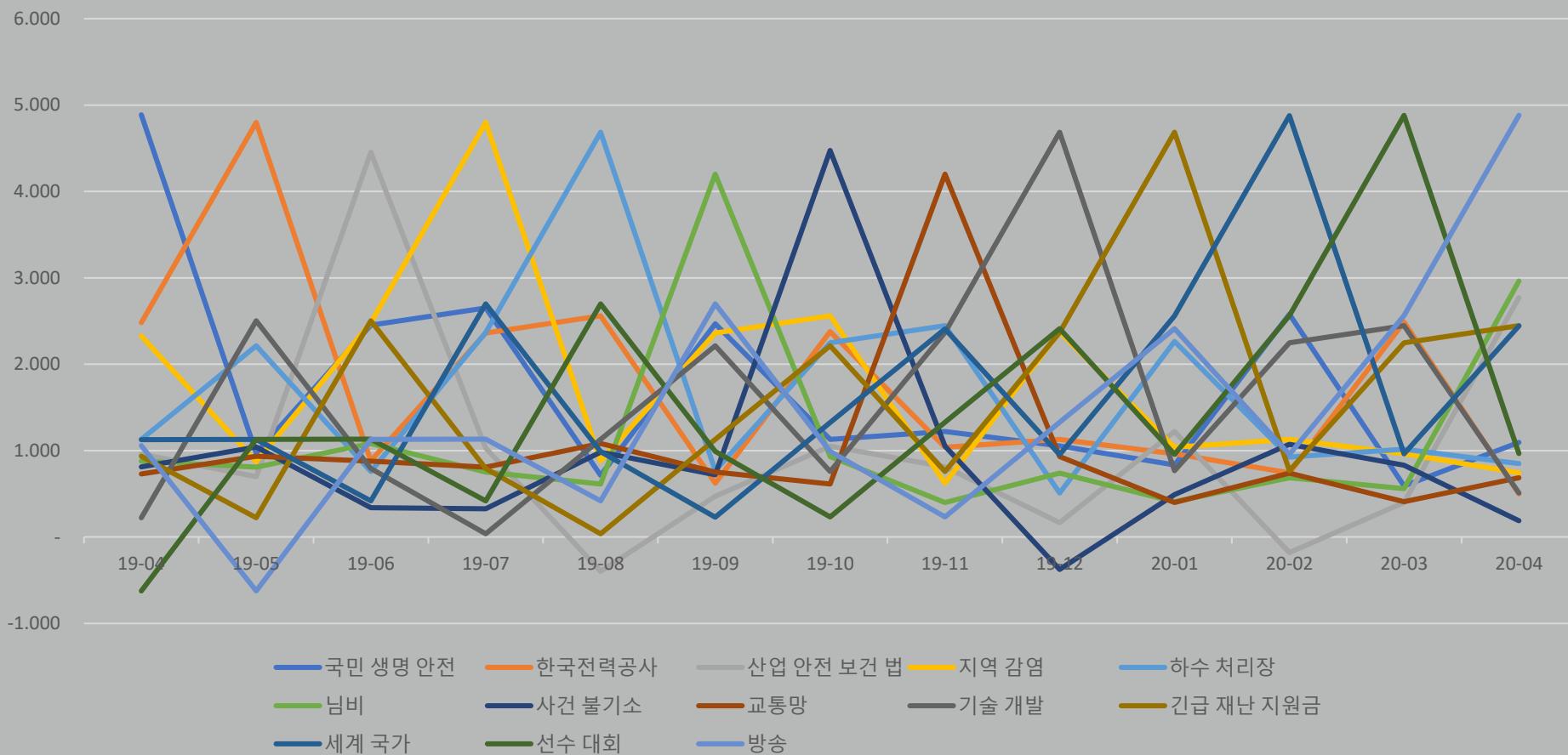
편차가 큰, 갑자기 이슈화가 되었지만, 바로 사그라든 토픽들 11개이다.

LABEL 선택											
국민 생명 안전	국가	안전	위험	생명	사태	가능	보호	조치	발생	필요	
한국전력공사	한전	전기	전력	요금	제이비	발전소	섬	심야	사용	공급	
산업 안전 보건 법	안전	사고	위험	발생	현장	작업	산업	관리	전문	예방	
지역 감염	코로나	감염	확진	확산	지역	바이러스	발생	조치	감염병	전국	
하수 처리장	물	오염	환경	수질	처리장	수돗물	발생	검출	오존	환경부	
님비	연금	소각장	차량기지	퇴직	은평구	신정	쓰레기	공단	은평	양천구	
사건 불기소	수사	사건	검찰	조사	검사	증거	처분	고소	검찰청	제출	
교통망	도시	서울	지역	교통	개발	지구	발표	수도	정신	노선	
기술 개발	기술	시스템	개발	과학	가능	혁신	산업	연구	수소	기관	
긴급 재난 지원금	지원	지급	소득	재난	지원금	혜택	기준	복지	긴급		
세계 국가	한국	세계	미국	해외	국내	중국	일본	국가	나라	대한민국	
선수 대회	협회	선수	대회	스포츠	경기	팀	축구	운동	감독	대표	
방송	방송	채널	유튜브	방송국	공영	유튜버	수신료	시청자	운영	유튜브	

LABEL 선택	19-04	19-05	19-06	19-07	19-08	19-09	19-10	19-11	19-12	20-01	20-02	20-03	20-04
국민 생명 안전	4.888	0.971	2.451	2.652	0.714	2.468	1.130	1.220	1.055	0.833	2.578	0.590	1.096
한국전력공사	2.480	4.798	0.881	2.361	2.561	0.623	2.377	1.039	1.130	0.964	0.742	2.487	0.500
산업 안전 보건 법	0.959	0.698	4.453	1.030	- 0.398	0.470	1.052	0.810	0.166	1.222	- 0.180	0.400	2.770
지역 감염	2.327	0.871	2.480	4.798	0.881	2.361	2.561	0.623	2.377	1.039	1.130	0.964	0.742
하수 처리장	1.130	2.213	0.758	2.366	4.684	0.767	2.248	2.448	0.510	2.264	0.926	1.016	0.851
님비	0.878	0.810	1.082	0.752	0.613	4.199	0.930	0.398	0.738	0.409	0.686	0.560	2.962
사건 불기소	0.810	1.041	0.339	0.327	0.980	0.719	4.473	1.051	- 0.377	0.491	1.072	0.831	0.187
교통망	0.730	0.933	0.878	0.810	1.082	0.752	0.613	4.199	0.930	0.398	0.738	0.409	0.686
기술 개발	0.222	2.503	0.794	0.036	1.130	2.213	0.758	2.366	4.684	0.767	2.248	2.448	0.510
긴급 재난 지원금	0.935	0.222	2.503	0.794	0.036	1.130	2.213	0.758	2.366	4.684	0.767	2.248	2.448
세계 국가	1.127	1.130	0.416	2.697	0.988	0.230	1.324	2.408	0.952	2.561	4.878	0.962	2.442
선수 대회	- 0.624	1.130	1.133	0.419	2.700	0.991	0.233	1.327	2.410	0.955	2.563	4.881	0.964
방송	1.056	- 0.624	1.130	1.133	0.419	2.700	0.991	0.233	1.327	2.410	0.955	2.563	4.881

총 13달 중, 가장 핫한 토픽들이다.  
 이전 토픽들과 비슷하게 해당 기간에  
 어떤 이슈가 있었는지 알 수 있다.

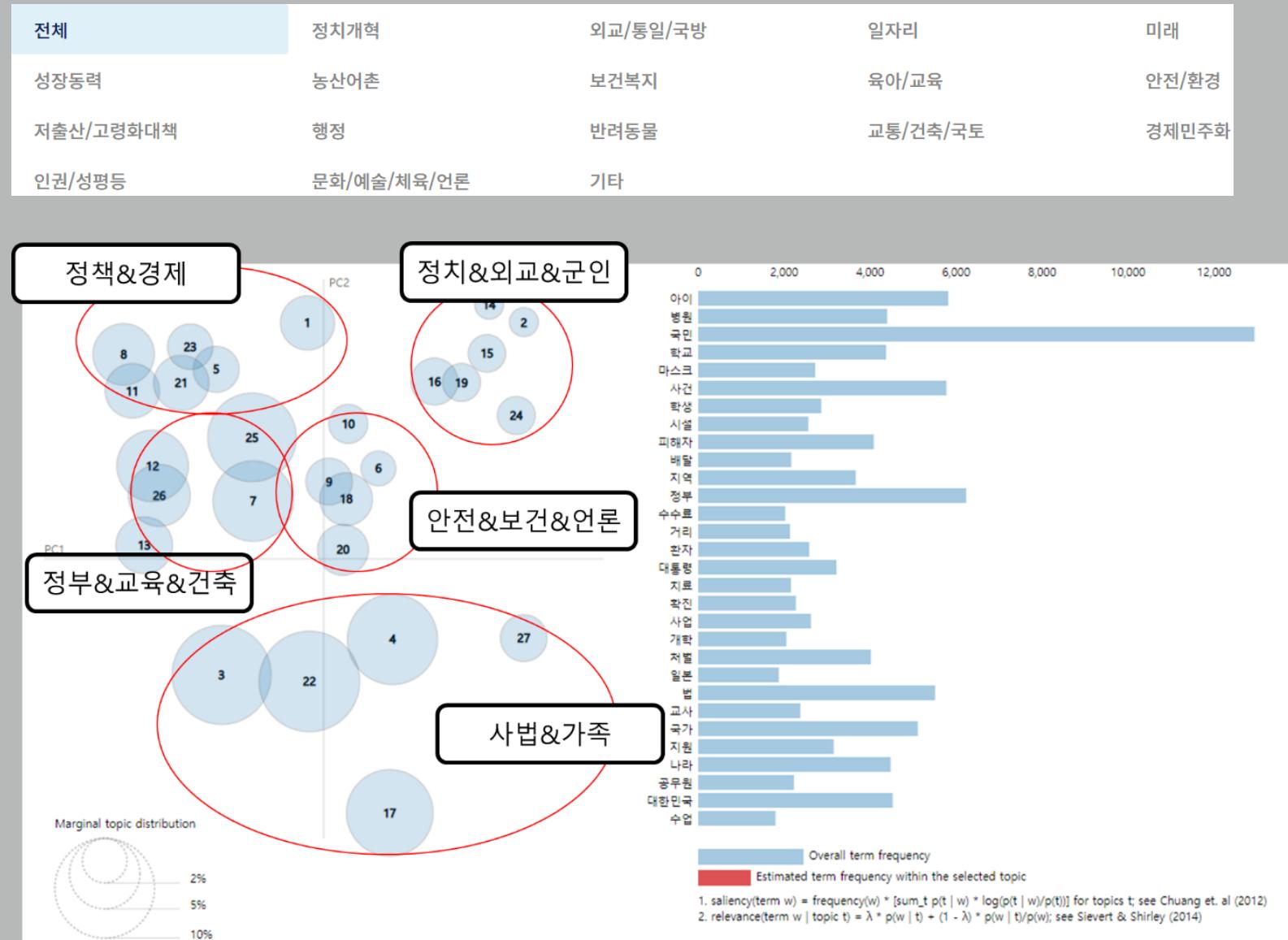
## 월별 가장 핫한 토픽



이전 토픽들과는 다르게, 기존에도 관심은 있었지만 갑자기 관심이 집중된 것을 알 수 있다.

매우 많은 토픽은 시계열 분석에는 적합했다. 그러나 토픽들을 묶어보기에는 어려움이 있었다. 따라서 국민청원 분류 카테고리에 들어있는 단어 개수만큼 임의로 Topic Number를 27개로 정했다.

pyLDAVis를 통해 시각화했다. 시각화 결과를 통해 각 토픽들의 키워드들을 중심으로, 크게 다섯가지 주제로 묶을 수 있었다. 각 토픽의 라벨은 묶인 주제들의 키워드들을 나타낼 수 있는 주제로 정했다.



**정치&외교&군인** - 외국인, 국회의원, 군인, 대통령, 일본, 나라, 복무, 사망, 북한

정치, 외교와 국방이 조금 동떨어져 있다고 생각할 수도 있지만 국방관련 청원에 자주 등장하는 주변국들이나 군인에 대한 주제들이 높은 연관성을 보인 것 같다.

**사법&가족** - 검찰, 사건, 처벌, 피해자, 조사, 판결, 남편, 아이, 부모, 폭행, 아버지, 가해자

가족과 관련된 청원에서 사법, 즉 가정내 폭력과 불화가 많은 부분을 차지하고 있었다.

**정부&교육&건축** - 교사, 부모, 어린이집, 선생, 분양, 공사, 아파트, 정부, 대한민국, 국가

교육과 건축관련 내용들이 많은 부분을 차지했고 이에 관해 정부에 건의하는 내용들이 주를 이루었다.

**정책&경제** - 수수료, 매출, 세금, 대출, 공무원, 시, 주민, 보험

경제관련 단어들이 많이 출현했고 이에 관한 정책들도 많이 보였다.

**안전&보건&언론** - 확진, 병원, 바이러스, 사고, 상황, 환자, 성, 건강, 마스크, 거리두기, 방송, 채널

최근에 코로나와 관련된 청원들이 다수 등장했는데 이 부분이 병원 등 보건주제와 같이 출현하는 경우가 많았다.

또한 언론에서 이 부분을 많이 다루었기 때문에 같이 묶을 수 있었다.

## 10. 요약 및 결론

## 요약 (1/4)

앞서 언급했듯 document classification에서는 predict 단계에서 아쉬운 성능을 보였다. 학습에 쓴 데이터가 8천건 밖에 안 되었던 점이 영향을 미친 것으로 보이며, 또한 코로나라는 특수한 상황이 수집 청원 기간에서 많은 비율을 차지해 과적합(overfit) 문제가 생긴 것도 하나의 원인으로 보았다. 선행 연구에서는 청원을 88,000건 이상 수집했던 것을 보면 더 많은 데이터를 분석할 수 있었다면 조금 더 나은 결과를 가져올 수 있었을 것으로 생각되므로 다음에 다시 분석할 기회가 생긴다면 이번에는 국민청원 2.0이후가 아닌 전체 국민청원 데이터를 대상으로 분석해보고자 했다.

또한 현재 국민청원 데이터의 카테고리 구분에 문제점을 인식하고 새로운 분류기준에 맞추어 document classification을 진행하였다. 이번에는 문헌이 아닌 문헌 안의 단어들을 기준으로 분류를 해봄으로써 키워드들의 분포와 흐름을 알아보고자 했다. 분류방법은 여러가지가 있었지만 시계열 분석을 위해, 각 문헌의 날짜 별 특성을 반영할 수 있는 Dirichlet-Multinomial Regression 토픽 모델링을 이용했다.

## 요약 (2/4)

청원데이터에서의 최선의 토픽 수를 찾기 위해서 perplexity를 바꾸어 가면서 토픽모델링을 진행해 보았다. Perplexity는 그 값이 작을수록 실제 문헌 결과를 잘 반영하므로 학습이 잘 되었다고 볼 수 있다. 반복 횟수가 증가할수록 같은 점점 감소하다가 일정 시점 이후에는 변화가 거의 없는 모습을 띠게 된다. 이때의 perplexity를 청원데이터의 최종 perplexity라고 생각했다.

청원데이터의 최종 perplexity값은 255였다. 이처럼 토픽 개수가 매우 많이 나온 것은 카테고리를 따지지 않고, 국민의 목소리가 다루는 주제가 매우 폭넓다는 것을 나타낸다. 기존의 카테고리가 17개였던 것을 생각해보면 청원의 다양한 주제와 키워드들을 담기에는 부족했을 수도 있을 것이다. 이 255개의 토픽을 기준으로 시계열 분석을 진행해보았다.

지속적으로 출현했던 키워드들은 나라와 사법 가족에 관한 내용들이 많았다. 갑자기 이슈화된 키워드들은 서로 그 관계성을 찾아보기 힘들 정도로 다양한 분야에서 출현했다. 이는 그 당시에 사회적으로 이슈가 되었던 특정 사건들에 영향을 받았기 때문이라고 생각된다.

## 요약 (3/4)

월별로 가장 핫 했던 키워드들은 앞에서 다룬 사회적 이슈가 된 키워드처럼 논의가 활발하게 되었던 주제를 알 수 있었다. 갑자기 이슈화된 키워드들과의 차이점은, 이슈화된 키워드들은 이전보다 훨씬 높은 비율로 상승했기 때문에 그전에 사회적 논의가 적었던 주제들이라면 가장 핫 했던 키워드들은 그 이전의 출현 빈도는 상관없다. 따라서 평소가 논의가 활발했던 주제라면 특정 월에 높은 빈도를 보일 경우 핫 했던 키워드에는 포함될 수 있지만 갑자기 이슈화된 키워드가 되기는 어려울 것으로 생각된다.

이렇게 perplexity값이 가장 낮은 255개의 토픽을 기준으로 시계열 분석을 해보았다. 그러나 perplexity값이 낮다는 것은 학습이 잘되었다는 뜻이지 그 결과의 해석이 편하다는 뜻은 아니다. 오히려 이처럼 너무 많은 토픽 수를 보일 경우에는 전체적으로 토픽 수를 분류하기에 어렵다고 느꼈다. 따라서 각 카테고리에 있는 주제들의 총 수인 27개를 토픽 수로 잡고 토픽모델링을 다시 진행해보았다.

## 요약 (4/4)

시각화 결과 토픽들의 분포를 볼 수 있었고 각 토픽별로 핵심적인 키워드들을 중심으로 크게 다섯 가지의 주제로 묶을 수 있었다. 토픽모델링에서의 토픽은 사실상 핵심적인 키워드들의 집합이고 이를 통해 주제를 정하는 것은 해석의 영역이라고 생각하여 5가지로 묶은 주제들의 키워드들을 기준으로 주제를 생각해보았다. 그 결과 정치/외교/국방, 사법/가족, 정부/교육/건축, 정책/경제, 안전/보건/언론 크게 이렇게 다섯 가지로 구분할 수 있었고 각 주제에 포함된 키워드들을 통해 주제가 이렇게 묶이게 된 이유와 청원들이 어떤 내용을 포함하고 있는지를 살펴볼 수 있었다.

## 요약의 요약

lovit / petitions\_scraper <- 브리핑 content 수집 문제

O K(10)-Means Cluster

Komoran & Kokoma <- UnicodeDecodeError: invalid continuation byte Predict 실패 -> 과적합(OverFit)

-> 코로나19라는 특수한 상황이 포함된 적은 데이터셋

### 국민청원 데이터 특성

카테고리가 많으며, 각각 청원 개수가 많이 다름

토픽 모델링

차원 축소는 특징 선택(LinearSVC)가 용이

DMR 시계열 분석 토픽 255개

감정 분석 -> 현 카테고리 분류 한계

-> 4가지 분석 ( 늘 많은 / 편차 적은 / 급상승급하락 / 월별 Top1 )

### 분류

X voiceofpeople

LDA 토픽 27개 -> 5가지로 묶음

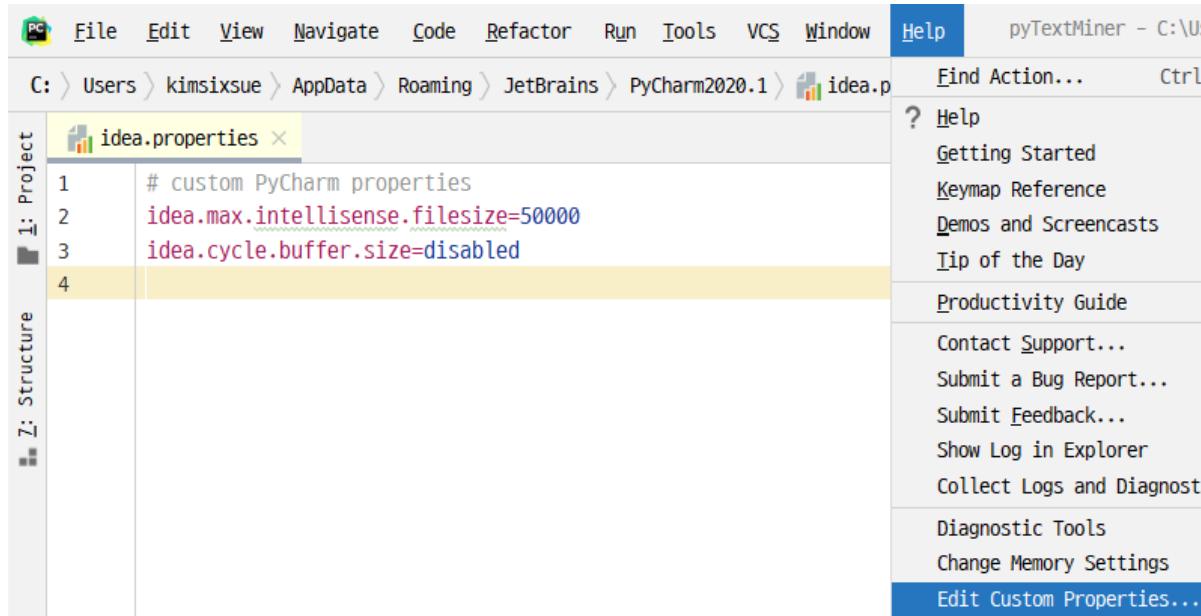
(하지만 토픽은 토픽일 뿐 카테고리가 될 수는 없음)

# 결론

- 데이터 선정에 문제가 있다.
- 현 카테고리가 적합하지 않다.
- 카테고리를 변경하면 좋을 것이다.
- 감정 분석과 토픽 모델링, 시계열 분석을 했다.
- 뚜렷한 연구 주제가 없었다. 하나로 묶을 수가 있어야 한다.
- 프로젝트를 기반으로 새로 전개해나갈 수 있다.
- 문헌 유사도 혹은 Jango를 활용할 수 있다.

# 11. 코드 수정사항

# 기타



PyCharm 내부에서  
50메가까지의 파일을  
편집하고, 출력 결과물이  
없어지지 않게 설정했다.

```
import winsound

# 도,레,미,파,솔,라,시 Hz
so1 = {'do': 261, 're': 293, 'mi': 329, 'pa': 349, 'sol': 391, 'ra': 440, 'si': 493}
mel2 = ['do', 'mi', 'mi', 'mi', 'sol', 'sol', 're', 'pa', 'pa', 'ra', 'si', 'si']
dur2 = [4, 4, 2, 4, 4, 2, 4, 4, 2, 4, 4, 2]
music2 = zip(mel2, dur2)
for melody, duration in music2:
    winsound.Beep(so1[melody], 1000 // duration)
```

대부분의 코드는 돌아가는데 오래 걸리기  
때문에, 언제 끝나는지를 알기 위해, 소리가  
나는 코드를 인터넷에서 구해왔다.

# Scraper json

[https://github.com/lovit/petitions\\_scraper](https://github.com/lovit/petitions_scraper)

이하 코드들은 Petitions\_scraper와 결합해 만든 것임을 알립니다.

## Index\_maker.py

The screenshot shows the PyCharm interface. On the left, there's a 'Run' tool window with a green play button icon and the text 'index\_maker'. The main area shows the command line output:

```
Run: index_maker
index status
587828 0
587829 0
587830 0
587831 0
587832 0
587833 0
587834 0
```

```
a = 587828
b = 588752
print('index
status')
for i in range(a,
b + 1):
    print(i, index status)
```

아주 단순한 코드로, 실행 출력 결과물을 복사해서, index.txt 파일에 그대로 붙여 넣으면 된다.

<https://www1.president.go.kr/petitions/587828> 부터

<https://www1.president.go.kr/petitions/588752>까지

두 청원을 포함, 그 사이에 있는 청원들까지 수집

원래 패키지에 있는 scraping\_petitions.py 를 실행하면 해당 청원들이 수집되어 Json 파일형태로 하나씩 Output 폴더 안에 생긴다.

## Emoji.py

```
import re
def remove_emoji(text):
    only_BMP_pattern = re.compile(r"["
        u"\U00010000-\U0010FFFF"
        "]+", flags=re.UNICODE)
    return only_BMP_pattern.sub(r'', text)
```

Input 값에 있는 EMOJI 등의 UTF-8을 벗어나는 등의 특수한 문자들을 없애준다.

정규표현식을 써본 적이 없었기 때문에, 검색을 상당히 많이 했었다.

Komoran과 kokoma에서 생기는 오류의 원인을 찾기 위해 상당한 노력을 들였다.

해당 코드를 다른 파이썬 코드 안에 함수로 집어넣어서 활용했다.

## Csc\_to\_dic.py

```
import csv

file = open('idxcsat.CSV', 'r')
csvfile = csv.reader(file)
lists = []
for item in csvfile:
    lists.append(item)
dicind = dict(lists)
```

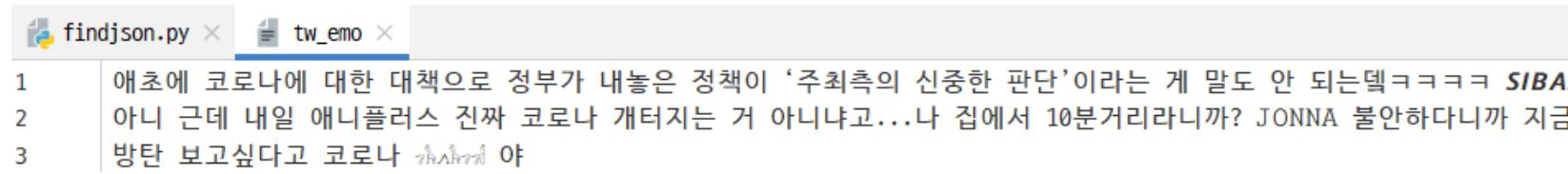
```
a = 579682
a = str(a)
```

```
if a in dicind:
    print(dicind[a])
```

579682	사법/가족
579683	정책/경제
579684	정책/경제
579685	정책/경제
579686	사법/가족
579687	사법/가족
579688	정치/외교/군인

idxcsat.CSV를 열어서, petition\_id에 맞게, 카테고리를 새로 부여한다.  
이후 나올 코드를 보면, 활용 방식을 볼 수 있다.

## Findjson.py



```
findjson.py x tw_emo x
1   애초에 코로나에 대한 대책으로 정부가 내놓은 정책이 '주최측의 신중한 판단'이라는 게 말도 안 되는뎁ㅋㅋㅋㅋ SIBAL
2   아니 근데 내일 애니플러스 진짜 코로나 개터지는 거 아니냐고...나 집에서 10분거리라니까? JONNA 불안하다니까 지금
3   방탄 보고싶다고 코로나 ㅋㅋㅋㅋ 야
```

```
import re

def remove_emoji(text):
    only_BMP_pattern = re.compile(r"[\U00010000-\U0010FFFF]+", flags=re.UNICODE)
    return only_BMP_pattern.sub(r'', text)

with open('tw0320.txt', 'r', encoding='utf-8') as txt:
    for i in txt:
        a = txt.readline().split('\t')
        # print(a[3])
        b = a[3]
        after = remove_emoji(b)

        with open('tw_emo', 'a', encoding='utf-8') as ee:
            if b != after:
                each = b + '\n'
                ee.write(each)
```

질문에 답변할 때, 활용한 코드이다.

해당 텍스트에 정규표현식에 해당하는 텍스트가  
있으면, 파일에 작성한다.

---

텍스트 문서 (35)

- content
- content\_short
- content\_title
- final\_20
- index
- petition\_final
- petition\_final\_voiceofpeople
- petition\_modif
- petition\_new\_920\_voiceofpeople
- petition\_removed\_emoji
- petitionKmean
- petitions\_35
- petitions\_36
- petitions\_37
- petitions\_38
- petitions\_39
- petitions\_40
- petitions\_41
- petitions\_42
- petitions\_43
- petitions\_44
- petitions\_45
- petitions\_46
- petitions\_47
- petitions\_48
- petitions\_49
- petitions\_50

---

텍스트 문서 (29)

- petitions\_51
- petitions\_brief
- petitions\_end
- petitions\_ing
- plz\_remove\_bom
- red\_blue\_20
- red\_blue\_br
- tw0320
- final\_20
- final\_35
- final\_36
- final\_37
- final\_38
- final\_39
- final\_40
- final\_41
- final\_42
- final\_43
- final\_44
- final\_45
- final\_46
- final\_47
- final\_48
- final\_49
- final\_50
- final\_51
- new\_920\_for\_predict
- petition\_final - 복사본
- petition\_final
- petition\_final\_8070
- petition\_final\_voiceofpeople
- petition\_new\_920\_voiceofpeople
- petition\_to\_27\_to\_5
- petition\_with\_emoji
- petitionKmean
- red\_blue\_20
- red\_blue\_br

다음과 같이 많은 양의 파일이 만들어졌다. 2개를 제외하고는 모두 다음 코드를 변형해서 만든 파일들이다. 시간이 많지 않았기 때문에, 최적의 코드로 만들지 못했다. 그때그때 조금씩 수정하면서 사용했다.

```
# -*- encoding: utf-8 -*- # 시행착오 중 하나입니다.
```

```
import json  
import os  
import re
```

## Json conve.py

```
def remove_emoji(text): # 제목과 내용에만 있는 것들을 Regex를 통해 제거합니다  
    only_BMP_pattern = re.compile("[  
        u"\U00010000-\U0010FFFF"  
    ]+", flags=re.UNICODE)  
    return only_BMP_pattern.sub(r'', text)  
  
id_file = 'petitions_end.txt' # 이게 왜 있는지 기억이 안 납니다.  
path = []  
with open(id_file, "r", encoding="utf-8") as f:  
    for root, dirs, files in os.walk( # 이 코드 또한 원래 짜본적이 없어서 고생한 기억이 납니다.  
        'C:/Users/kimsixsue/PycharmProjects/petitions_master/petitions_scraper-master/output'): # 이 코드 또한 원래 짜본적이 없어서 고생한 기억이 납니다.  
        for fname in files: # output 폴더에 있는 파일들을  
            path.append(os.path.join(root, fname)) # list에 추가합니다  
ls = []  
  
import csv  
  
file = open('idxcsat10.CSV', 'r') # 해당 idx에 맞게 부여한 새 카테고리가 들어있습니다.  
csvfile = csv.reader(file) # ex ) 579682,6  
lists = []  
for item in csvfile: # ex) 579683,4  
    lists.append(item)  
dicind = dict(lists)  
  
for i in range(len(path)): # json도 이번에 써본게 처음입니다. 어떻게 하는지 몰라서 주위에 많이 물어봤었습니다  
    with open(path[i], "r", encoding='utf-8') as json_file:  
        json_data = json.load(json_file)  
  
        tem = json_data.get('category') # json 파일에서 카테고리 값을 얻습니다.  
        ls.append(tem) # 카테고리를 바꾸기 용이하게 하기 위해, 일부러 2개로 나누었습니다
```

## Json conve.py

```
original_idx = str(json_data.get('petition_idx')) # 숫자가 할당이 안 되기 때문에 문자열로 바꿨습니다.
if original_idx in dicind:
    ls.append(original_idx)
ls[0] = (dicind[original_idx]) # 파일에 있는 idx에 해당하는 새 카테고리로 아까 할당한 카테고리를 대체합니다.

ls.append(json_data.get('begin'))
ls.append(json_data.get('end'))
ls.append(json_data.get('status'))
ls.append(str(json_data.get('num_agree'))) # 숫자가 할당이 안 되기 때문에 문자열로 바꿨습니다.

ta = json_data.get('title')
ta_after = remove_emoji(ta) # title에는 이모지가 있을 수 있습니다.
ls.append(ta_after)

ca = json_data.get('content')
ca_after = remove_emoji(ca) # content에도 이모지가 있을 수 있습니다.
ls.append(ca_after)

# for i in range(len(ls)): 이 부분은 오류의 원인을 몰라 넣은 코드입니다.
#     ls[i].encode('utf-8').decode('utf-8') 시행착오를 보여줍니다

idx = 'petitionKmean.txt'
with open(idx, 'a', encoding='utf-8') as ee: # 일부러 w 대신 a를 썼습니다. 각종 변경을 하기 용이합니다.
    # if int(ls[5]) >= 200000: 특정 조건을 넣어서, 해당하는 것만 집어넣게 합니다.
    each = '\t'.join(ls) + '\n' # tsv 형태로 만들어줍니다
    ee.write(each)
ls = [] # 초기화를 해줬습니다.
```

# pyTextMiner

## pyTextMiner\\_\_init\_\_.py

```
class Check_Bank(Corpus): # 토픽모델링에서 전처리를 하고 나니까
    def __init__(self, file, index): # 문헌 개수가 5개 줄어서 이유를 알기 위해 만든 코드입니다.
        array = []
        line_number = 0
        with open(file, encoding='utf-8') as ins:
            for line in ins.readlines():
                inside = line.split('₩t')
                line_number += 1
                print(len(inside[index])) # 길이가 가장 짧은 것을 보니까, 영어 혹은 하이퍼링크였습니다.
            try:
                in_in = inside[index]
            except IndexError:
                print(line_number, '라인 에러 확인, txt 파일 확인요망')
                array.append(in_in)
        self.docs = array
```

```

86     class CorpusFromFieldDelimitedFile(Corpus):
87         def __init__(self, file, index):
88             array = []
89             line_number = 0
90             with open(file, encoding='utf-8') as ins:
91                 for line in ins.readlines():
92                     inside = line.split('\t')
93                     line_number += 1 # 질문 답변에 답하기 위해 고생했던 기억이 납니다.
94                     try:
95                         in_in = inside[index]
96                     except IndexError:
97                         print(line_number, '라인 에러 확인, txt 파일 확인요망') # 데이터가 작으면, 이렇게 출력해도 됩니다
98                     array.append(in_in) # 물론 엄청 크면, 교수님처럼 개수만 세는게 낫습니다
99             self.docs = array # 막상 만들어보니 너무 간단했습니다.

CorpusFromFieldDelimitedFile > __init__()

102    class CorpusFromFieldDelimitedFileWithYear(Corpus):
103        def __init__(self, file, doc_index=1, year_index=0):
104            array = []
105            id = 0
106            pair_map = {}
107            with open(file, encoding='utf-8') as ins:
108                for line in ins:
109                    fields = line.split('\t')
110                    try:
111                        array.append(fields[doc_index])
112                        # pair_map[id] = fields[year_index]
113                        pair_map[id] = fields[year_index][0:7] # 프로젝트 용
114
115                        id += 1
116                    except IndexError:
117                        print("out of index " + str(id))
118
119            self.docs = array
120            self.pair_map = pair_map

CorpusFromFieldDelimitedFileWithYear > __init__() > with open(file, encoding='utf-8...

```

2020-02-02 같은 식으로 데이터가 들어있기 때문에 토픽모델링 시계열분석에서 좌측 7개만을 기준으로 하여, 연도와 월만을 이용했다.

## Id\_to\_category\_petition.json

```
{  
    "0": "정치개혁",  
    "1": "외교/통일/국방",  
    "2": "일자리",  
    "3": "미래",  
    "4": "성장동력",  
    "5": "농산어촌",  
    "6": "보건복지",  
    "7": "육아/교육",  
    "8": "안전/환경",  
    "9": "저출산/고령화대책",  
    "10": "행정",  
    "11": "반려동물",  
    "12": "교통/건축/국토",  
    "13": "경제민주화",  
    "14": "인권/성평등",  
    "15": "문화/예술/체육/언론",  
    "16": "기타"  
}
```

Lasso\_term\_extraction.py 를 이용하기 위해 만든 json 파일이다.  
그러나 프로젝트에 사용할 일은 없었다.

Komoran을 위해 만든 User\_dic.txt 이다. 그러나 주로 쓴 tokenizer는 Mecab이었다.

더불어민주당 NNP	쏘렌토 NNP	경찰서 NNG
엔터테인먼트 NNG	세월호 NNP	운정 NNP
자유한국당 NNP	박근혜 NNP	아이 NNG
대전국토청 NNP	민식이 NNP	휴원 NNG
국토교통부 NNP	나크리 NNP	펜션 NNG
텔레그램 NNP	공수처 NNP	팩트 NNG
지소미아 NNP	N번방 NNP	태움 NNG
화상학대 NNG	싱크홀 NNG	증시 NNG
초저출산 NNG	신도시 NNG	점용 NNG
소상공인 NNG	시복식 NNG	복판 NNG
메커니즘 NNG	선진화 NNG	방안 NNG
후베이 NNP	민주화 NNG	공적 NNG
자한당 NNP	대포차 NNG	사적 MM

## Documentclustering.py

```
110     self.X = self.X.toarray()
111     # self.X = self.X.toarray()
112     # 나는 numpy 1.18인데 문제가 있는데 이상하군. 아래와 같이 바꾸어서 임시적으로 문제를 해결할수 있을까 보자
113     # self.X = csr_matrix(self.X).todense()
114
115     pca_t = PCA().fit_transform(self.X)
116     print(self.clustering.labels_)
117     plt.scatter(pca_t[:, 0], pca_t[:, 1], c=self.clustering.labels_, cmap='rainbow')
118     plt.show()
119
120     cluster_labels = self.clustering.labels_
121     clustering_dict = self.clustering.__dict__
122     clusters = {}
123
124     # agglo를 그대로 가져온것
125     for document_id, cluster_label in enumerate(cluster_labels):
126         if cluster_label not in clusters:
127             clusters[cluster_label] = []
128             clusters[cluster_label].append(document_id)
129             print(str(cluster_label) + " -- " + str(document_id))
130     # agglo를 그대로 가져온것
131
DocumentClustering > print_results() > if self.name == 'k-means'
```

윗 부분의 문제는 컴퓨터를 완전 초기화를 하고 모든 것을 완벽하게 설치했음에도 해결할 수 없었다.

밑 부분은 각 문헌이 어디로 분류되었는지 보기위해

아래에 있는 agglo의 print result를 그대로 복사했다.

Bokeh plot에서 각각 라벨이 출력되도록 바꾸다가 그만두었다.

해당 코드에서는 전체 문헌이 아닌, 필터링된 천개 이하의 문헌만 남기기 때문이다.

## Lda.py

```
73     # Plot the Topic Clusters using Bokeh
74     # index_8070 = np.loadtxt('./8070.txt').reshape(8070, 1)
75     # tsne_lda = np.concatenate((tsne_lda, index_8070), axis=1)
76     # https://docs.bokeh.org/en/latest/docs/user\_guide/annotations.html
77     # source = ColumnDataSource(data=dict(x=tsne_lda[:, 0],
78     #                                     #                                     y=tsne_lda[:, 1],
79     #                                     #                                     names=tsne_lda[:, 2]))
80     n_topics = 4
81     mycolors = np.array([color for name, color in matplotlib.colors.cnames.items()])
82     plot = figure(title="t-SNE Clustering of {} LDA Topics".format(n_topics),
83                   plot_width=900, plot_height=700)
84     plot.scatter(x=tsne_lda[:, 0], y=tsne_lda[:, 1], color=mycolors[topic_num])
85
86     # labels = LabelSet(x='weight', y='height', text='names', level='glyph',
87     #                     #                     x_offset=5, y_offset=5, source=source, render_mode='canvas')
88     # plot.add_layout(labels)
89
90     show(plot)
```

LDAManager > run()