국민대학교 텍스트데이터분석

국민청원 동의인원에 영향을 미치는 단어 찾기(자동청원생성기)

20162540 조혁준

목차

01 도입 – 요약

- 국민청원의 참여인원의 많고 적음을 결정하는 요소가 주제인지, 어휘 인지를 알아본다. 예를 들어 '사랑', '부탁'과 같은 일상 단어가 참여인원이 많은 청원에 자주 등장하는지에 대해서 검증한다.
- 감성분석을 통해서 일정 수의 이상의 인원이 동의한 청원과 그렇지 않은 청원의 어휘 상 차이점을 알아본다.
- 검증 결과 청원 참여인원 수에 영향을 미치는 것은 단어보다 주제임을 알게 되었다.
- 청원내용의 어휘는 참여인원에 영향을 미치지 않는다고 판단하고, 청원 내용 자동 생성기를 통해 청원내용을 자동으로 생성했다.

01 도입 – 서론

목적

- 청와대 국민청원에 참여인원이 많은 청원들은 어휘 상 공통적인 특징이 있는지 알아본다. (예를 들어, '고마운','슬픈'과 같은 청원주제와 관련 없는 일반적 어휘)
- 공통된 특징이 없다면, 사람들이 청원에 동의하게 되는 동기는 글의 주제나, 제목에서 발현하는 것으로 보고, 청원의 내용 작성에 공을 들일 필요가 없다고 판단한다.
- 따라서 청원의 내용을 자동으로 작성해주는 (GPT-2를 이용한) 청원내용 자동 생성기를 통해 청원내용을 자동화하여 작성하는 것을 목표로 한다.

가설

- 국민청원에는 주제를 초월하여 참여인원에 영향을 끼치는 어휘가 있다.

가설 검증 방법

- 단어 별 가중치를 통해 청원 참여인원이 많은 청원은 어휘에서 특별한 차별성이 있는지 1차적으로 검증 한다.
- 학습이 완료된 모델로 학습데이터와 분포가 비슷한 testdata를 평가 후 정확도가 낮으면 어휘 상 공통점이 없는 것으로 판단하고 가설을 기각한다.

02 본론 – 데이터 수집

- 데이터는 청와대 국민청원에서 크롤링.
- 약 3만개의 글을 수집.
- Lxml과 selenium을 활용.

443893	안전/환경	텔레그램 n번방 용의자 신상공개 및 포토라인 세워주세요	2020-04-17	2,715,626명
443892	안전/환경	텔레그램 n번방 가입자 전원의 신상공개를 원합니다	2020-04-19	2,026,252명
443891	정치개혁	자유 한국당 정당해산 청원	2019-05-22	1,831,900명
443890	기타	문재인 대통령님을 응원 합니다!	2020-03-27	1,504,597명
443889	정치개혁	문재인 대통령 탄핵을 촉구합니다.	2020-03-05	1,469,023명
443888	인권/성평등	신천지 예수교 증거장막성전(이하, 신천지)의 강제 해체(해산)을 청원합니다.	2020-03-23	1,449,521명

02 본론 – 데이터 요약

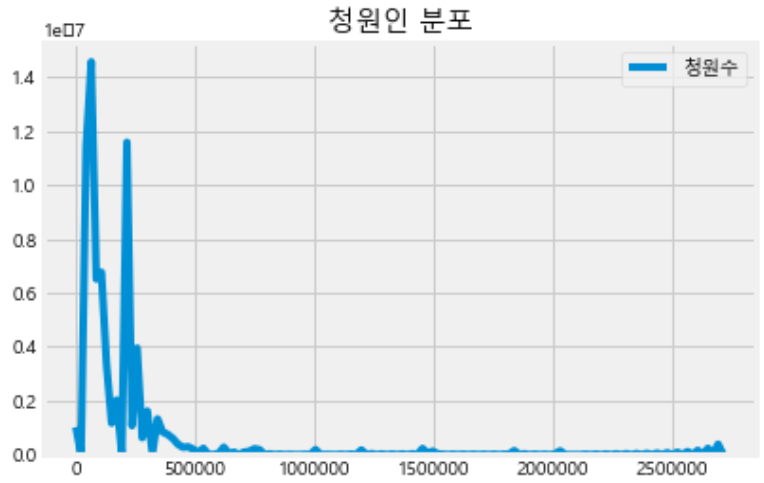
	제목	내용	청원수	카테고리	청원시작일	청원마감일	청원인
0	텔레그램 n번방 용의자 신상공개 및 포토라인 세워주세요	오늘 검거되었다고 합니다 타인의 수치심과 어린 학생들을 지옥으로 몰아넣은 가해자를...	2715626	안전/환경	청원시작 2020-03-18	청원마감 2020-04-17	청원인 naver - ***
1	텔레그램 n번방 가입자 전원의 신상공개를 원합니다	안녕하세요, 텔레그램 n번방에 대한 수사가 진행되고 일부의 용의자가 검거되어 다행...	2026252	안전/환경	청원시작 2020-03-20	청원마감 2020-04-19	청원인 naver - ***
2	자유 한국당 정당해산 청원	민주당과 정부에 간곡히 청원합니다 자유한국당은 국민의 막대한 세비를 받는 국회의원...	1831900	정치개혁	청원시작 2019-04-22	청원마감 2019-05-22	청원인 kakao - ***
3	문재인 대통령님을 응원 합니다!	코로나 바이러스19로 인해 대한민국 모든 국민이 힘든 시기에 있습니다. 하지만 국...	1504597	기타	청원시작 2020-02-26	청원마감 2020-03-27	청원인 naver - ***
4	문재인 대통령 탄핵을 촉구합니다.	국민의 한 사람으로서 문재인 대통령의 탄핵을 촉구합니다. 이번 우한 폐렴(신종 코...	1469023	정치개혁	청원시작 2020-02-04	청원마감 2020-03-05	청원인 naver - ***

크롤링한 32427개의 데이터프레임

제목/내용/청원 수/카테고리/청원 시작, 마감일/청원인으로 이루어져있음.

02 본론 – 데이터 정보

정치개혁	3993
인권/성평등	3672
기타	3494
보건복지	3238
육아/교육	3143
안전/환경	2948
교통/건축/국토	2357
행정	1767
문화/예술/체육/언론	1761
외교/통일/국방	1593
경제민주화	1257
일자리	1178
반려동물	696
미래	633
저출산/고령화대책	265
성장동력	247
농산어촌	185

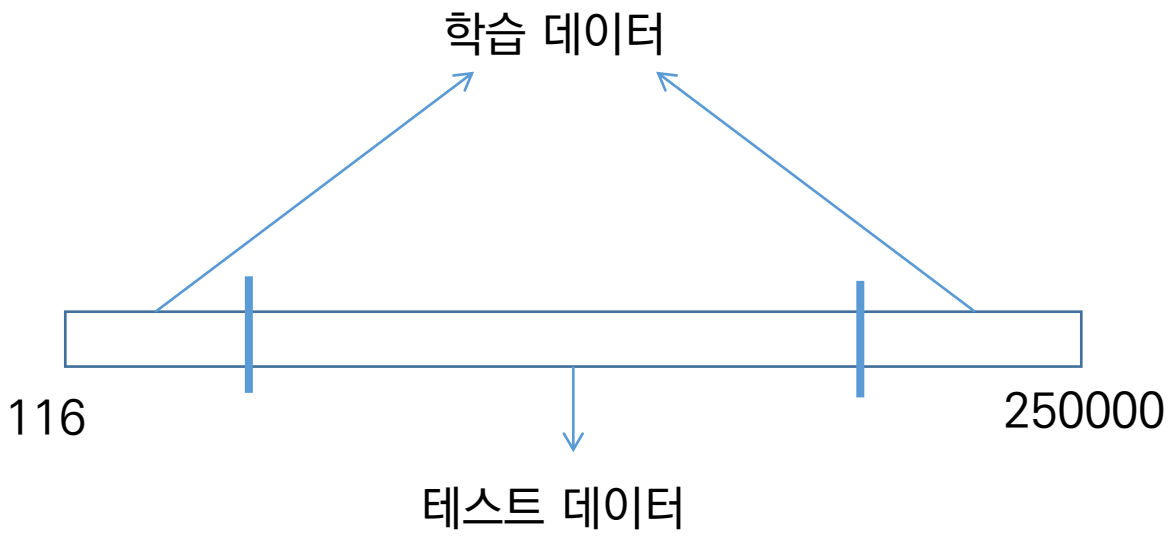


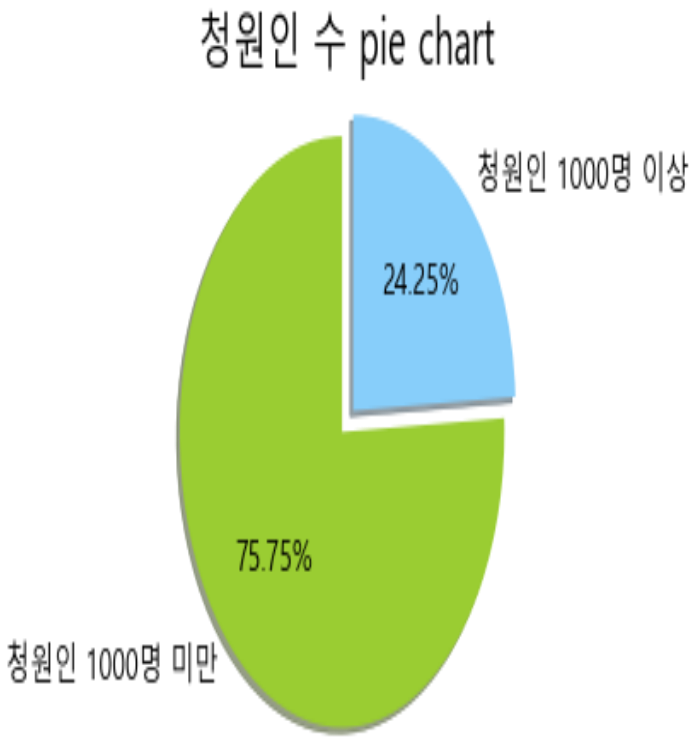
카테고리 별 청원글은 정치개혁이 3993개로 가장 많다.

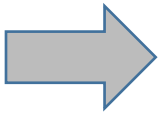
위의 표를 통해 국민들이 대체로 어느 부분에서의 개혁을 원하는지 파악할 수 있다.

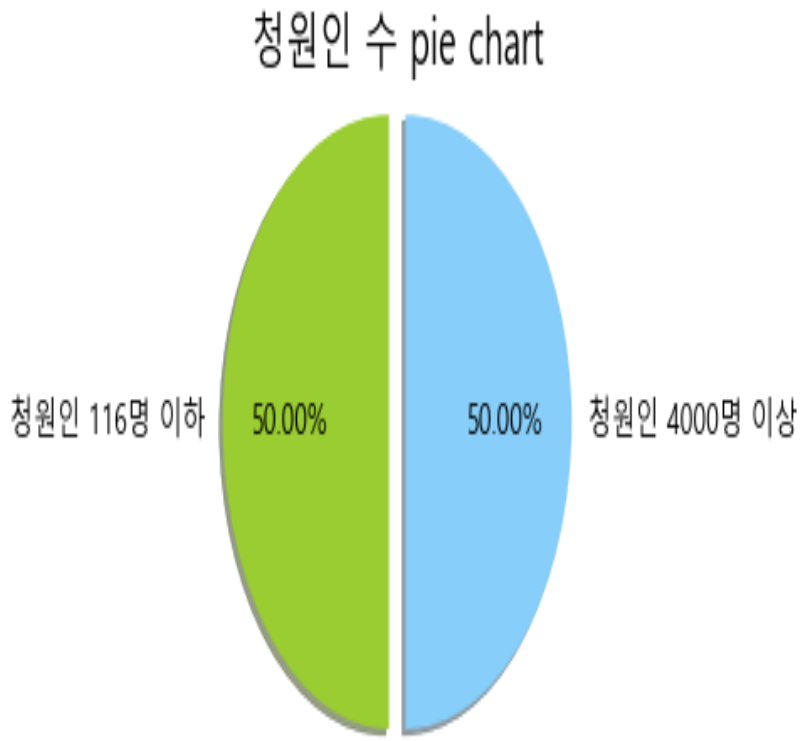
청원인 분포: 20만명을 넘는 청원은 극히 소수이다.

- 1천명 이상이 동의한 청원을 1, 그렇지 않은 청원을 0으로 분류하여 이진라벨링을 했다.
- 두 집단간 차이적 특징을 더 두드러지게 학습하기 위해서 양극단 값을 사용했다.









학습데이터의 라벨의 분포가 불균형해서 분포를 일정하게 정제하고,
Baseline을 50%로 설정했다.

데이터 분석을 진행하기에 앞서서, 신경망언어모형으로 base model을 만든 결과, 청원내용이 너무 길어서 학습이 제대로 되지 않는 문제 발생!

↓

Gensim의 summarize 모듈 사용을 통해 내용요약!

' 코로나 바이러스19로 인해 대한민국 모든 국민이 힘든 시기에 있습니다. 하지만 국민건강을 위해 대통령님을 비롯한 대한민국 정부 각 부처의 모든 분들이 밤낮 없이 바이러스 퇴치에 온갖 힘을 쏟고 계십니다. 하지만 신천치 라는 생각치도 못한 사이비 종교의 무분별한 바이러스 확산으로 인해,코로나 19청정지역이었던 대한민국인 단 일주일 사이 급속도록 확진자들이 불어 나고 있으며,국민들 모두 힘들어 하는 상황 까지 오게 되었습니다. 정부의 협조 요청에도 묵묵부담으로 일삼고 있는 사이비 종교 신천지. 이러한 악 조건 속에서도 대통령님은 밤낮없이 오직 국민들의 안전을 위해 노력 하고 계시며, 신천지 바이러스의 근원지가 되어 버린 대구&경북 지역을 위해 무척이나 애쓰시고 계십니다. 수많은 가짜 뉴스가 대통령님 및 질병관리본부 그리고 대한민국 각 부처을 힘들게 하고 있지만 수많은 대한민국 국민들은 “문재인 대통령”님을 믿고 응원하고 있습니다. 이 어려운 시기는 대통령님과 함께 반드시 이겨낼 것이며, 대한민국 국민 대다수는 정부에 대한 신뢰로 함께 극복해나갈 거라 믿어 의심치 않습니다. 문재인 대통령님 언제나 응원 합니다!! 문재인 대통령님이 있는 대한민국은 반드시 이 어려운 상황을 극복 해나갈 것입니다.!!
'

청원의 기본적 내용을 유지하면서, 글자 수를 줄일 수 있었다.

'코로나 바이러스19로 인해 대한민국 모든 국민이 힘든 시기에 있습니다. 하지만 국민건강을 위해 대통령님을 비롯한 대한민국 정부 각 부처의 모든 분들이 밤낮 없이 바이러스 퇴치에 온갖 힘을 쏟고 계십니다. 하지만 신천치 라는 생각치도 못한 사이비 종교의 무분별한 바이러스 확산으로 인해,코로나 19청정지역이었던 대한민국인 단 일주일 사이 급속도록 확진자들이 불어 나고 있으며,국민들 모두 힘들어 하는 상황 까지 오게 되었습니다. 이 어려운 시기는 대통령님과 함께 반드시 이겨낼 것이며, 대한민국 국민 대다수는 정부에 대한 신뢰로 함께 극복해나갈 거라 믿어 의심치 않습니다. 문재인 대통령님 언제나 응원 합니다!! 문재인 대통령님이 있는 대한민국은 반드시 이 어려운 상황을 극복 해나갈 것입니다.!!'

참여인원이 많은 청원의 단어구름

참여인원이 적은 청원의 단어구름

- 분석 전 간단하게 살펴보는 단어구름
- 하지만, 외관상 참여인원이 많은 청원과 적은 청원 사이에 어휘상 차이가 없어 보인다.

	토큰	가중치
298	마스크	0.924823
883	코	0.894464
79	개학	0.716676
885	코로나19	0.694563
457	상황	0.626867
985	확진자	0.610153
535	신천지	0.598306
817	진행	0.526529
689	일본	0.523091
346	발생	0.499583
994	후	0.492674
727	전국	0.476649
790	중	0.459652
250	대구	0.458888
376	병원	0.437495
877	치료	0.433751
482	성범죄	0.433529
668	이유	0.428735
43	n번방	0.428339
56	가해자	0.426128

단어별 가중치

	토큰	가중치
210	난민	-0.882103
207	나라	-0.666766
266	대한항공	-0.587472
167	국회의원	-0.561044
162	국민연금	-0.539404
273	돈	-0.433577
161	국민들	-0.398840
725	적폐	-0.345571
912	폐	-0.336180
765	조사	-0.327005
126	공매	-0.325823
772	조폭	-0.318107
211	난민들	-0.307934
325	뭐	-0.301055
441	사퇴	-0.299183
764	조두순	-0.291401
415	비리	-0.286870
168	국회의원들	-0.286120
153	구속	-0.282162
701	자기들	-0.280584

– 참여인원이 많은 청원(1)과 그렇지 않은 청원(0)에서 어휘 상 차이가 발견된다.
– 하지만 검증하려고 하는 글에서 일반적으로 자주 사용되는 어휘인 ‘정말’,’너무합니다’ 등이 아닌, 글의 주제와 관련된 ‘코로나’, ‘난민’ 등의 어휘라서 더 검증이 필요하다.

Fasttext, LSTM, 전이학습 모델, Conv1d 모델을 모두 구성하고, 가장 성능이 좋은 모델을 튜닝하여 감성분석을 진행.

fasttext

```
Model: "sequential_7"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense_29 (Dense)             (None, 1024)              103424
_________________________________________________________________
dropout_10 (Dropout)         (None, 1024)              0
_________________________________________________________________
dense_30 (Dense)             (None, 512)               524800
_________________________________________________________________
dropout_11 (Dropout)         (None, 512)               0
_________________________________________________________________
dense_31 (Dense)             (None, 256)               131328
_________________________________________________________________
batch_normalization_3 (Batch (None, 256)               1024
_________________________________________________________________
dense_32 (Dense)             (None, 128)               32896
_________________________________________________________________
dense_33 (Dense)             (None, 1)                 129
=================================================================
Total params: 793,601
Trainable params: 793,089
Non-trainable params: 512
_________________________________________________________________
```

ACC : 0.677

Dropout과 Batchnorm 층을 추가

LSTM

```
Model: "sequential_6"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding_3 (Embedding)      (None, None, 8)           16008
_________________________________________________________________
lstm_1 (LSTM)                (None, 8)                 544
_________________________________________________________________
dense_14 (Dense)             (None, 1)                 9
=================================================================
Total params: 16,561
Trainable params: 16,561
Non-trainable params: 0
_________________________________________________________________
```

ACC : 0.711
(시간이 매우 오래걸림)

신경망 언어모형의 가중치를 이용한 전이학습 모델

```
Model: "sequential_4"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
sequential_3 (Sequential)    (None, None, 8)           16008
_________________________________________________________________
dense_15 (Dense)             (None, None, 512)         4608
_________________________________________________________________
global_average_pooling1d_1 ( (None, 512)               0
_________________________________________________________________
dense_16 (Dense)             (None, 256)               131328
_________________________________________________________________
dropout_5 (Dropout)          (None, 256)               0
_________________________________________________________________
dense_17 (Dense)             (None, 128)               32896
_________________________________________________________________
dense_18 (Dense)             (None, 1)                 129
=================================================================
Total params: 184,969
Trainable params: 184,969
Non-trainable params: 0
_________________________________________________________________
```

ACC : 0.6725

Conv1d

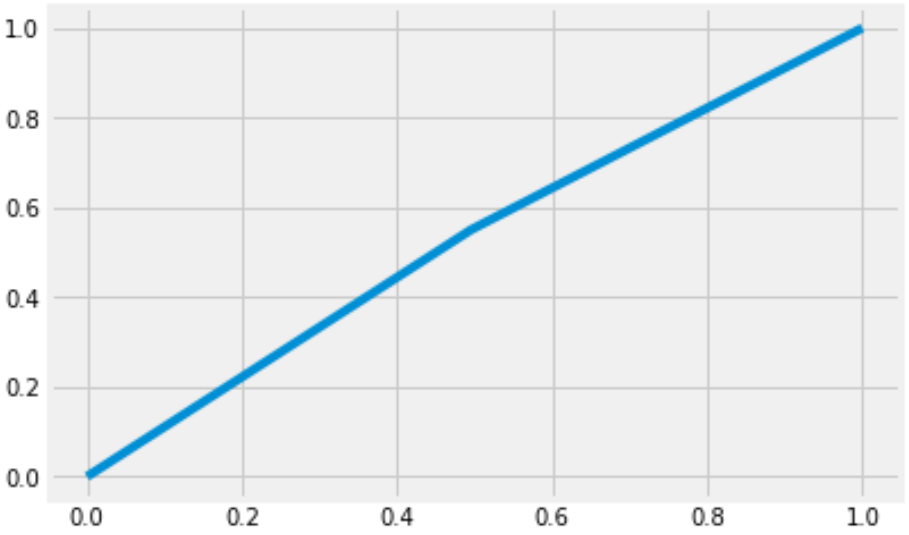
```
Model: "sequential_9"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding_5 (Embedding)      (None, 3718, 8)           16008
_________________________________________________________________
conv1d_2 (Conv1D)            (None, 3716, 3)           75
_________________________________________________________________
max_pooling1d_2 (MaxPooling1 (None, 743, 3)            0
_________________________________________________________________
conv1d_3 (Conv1D)            (None, 741, 3)            30
_________________________________________________________________
max_pooling1d_3 (MaxPooling1 (None, 148, 3)            0
_________________________________________________________________
flatten_1 (Flatten)          (None, 444)               0
_________________________________________________________________
dense_17 (Dense)             (None, 1)                 445
=================================================================
Total params: 16,558
Trainable params: 16,558
Non-trainable params: 0
_________________________________________________________________
```

ACC : 0.6805

02 결론 – 감성분석 결과 해석

학습시간이 짧고, 성능이 준수한 Conv1d 모델을 hyper opt 패키지를 통해 튜닝하고, test data로 평가한 결과 해석

ACC : 0.5221 (baseline: 0.5)



Roc-auc curve를 통해서도 성능이 굉장히 안 좋은 것을 확인할 수 있다.

Test data로 성능을 평가했지만, 성능이 좋지 못하다.

만약 참여인원이 많은 청원에 공통적인 어휘상 특징이 있다면, test data로 평가를 했을 때, 일정 수준 이상(학습데이터의 validation accuracy인 0.68 이상)의 성능을 도출해야 하지만, 모델 성능을 통해 그렇지 않다는 것을 알 수 있다.

따라서 이번 프로젝트를 통해 생성된 모델 성능은 청원 참여인원에 영향을 주는 어휘는 없는 것을 시사한다.

그러므로 “국민청원에는 주제를 초월하여 참여인원에 영향을 끼치는 어휘가 있다.” 는 가설을 기각한다.

이를 통해 국민청원의 이용자들은 청원의 내용보다는 제목이나, 주제와 같은 글의 전체적인 성격을 보고 청원에 동의한다는 것을 유추해 볼 수 있다.

02 결론 – 청원내용 자동 생성기

앞선 결과 해석에서 청원내용을 구성하고 있는 단어와 참여인원은 관련이 없다는 것을 알았다.
그렇다면 청원 내용을 작성하는데 시간과 정성을 쏟기 보다는 청원의 주제나 제목 작성에 노력을 기울이는 것이 더 나은 선택일 것이다.
이를 보조할 수 있는 GPT-2를 사용한 청원내용 자동 생성기를 만든다.

I petition for the abolition of juvenile law. Teenagers are committing countless crimes by exploiting juvenile law.

입력

청원내용 자동 생성기

I petition for the abolition of juvenile law. Teenagers are committing countless crimes by exploiting juvenile law. They commit many criminal crimes, including robbery of property and aggravated assaults. Juvenil is not a juvenile's only responsibility; it also is a child-child responsibility of the juvenile courts, who must make sure that all juveniles are properly treated as children in their circumstances and that their rights, as adults in their situation were respected, respected, protected. This bill would make juvenile court proceedings a mandatory and fair procedure for the adjudicate, the decision and decision making of all parties involved with a case. This is the right and duty that should be given every citizen, whether they live in this state of Texas and have legal guardians, have their children and their parents and all the children who are not adults in the situation that the law allows them. I urge Senator Denny to sign the petition. Please contact your Senator or the Senator's Committee on Criminal and Criminal Procedure, and you could also ask your Representative to consider this issue, if not for your representative in your State of residence and your Senator's office. I hope this petition helps to bring the law and its provisions in order to ensure a better future to all citizens of the State who want to live free of the juvenile justice and adult criminal justice laws. I would be grateful and proud

출력

02 결론 – 청원내용 자동 생성기

I petition for the abolition of juvenile law. Teenagers are committing countless crimes by exploiting juvenile law. They commit many criminal crimes, including robbery of property and aggravated assaults. Juvenil is not a juvenile's only responsibility; it also is a child-child responsibility of the juvenile courts, who must make sure that all juveniles are properly treated as children in their circumstances and that their rights, as adults in their situation were respected, respected, protected. This bill would make juvenile court proceedings a mandatory and fair procedure for the adjudicate, the decision and decision making of all parties involved with a case. This is the right and duty that should be given every citizen, whether they live in this state of Texas and have legal guardians, have their children and their parents and all the children who are not adults in the situation that the law allows them.\n: I urge Senator Denny to sign the petition. Please contact your Senator or the Senator's Committee on Criminal and Criminal Procedure, and you could also ask your Representative to consider this issue, if not for your representative in your State of residence and your Senator's office. I hope this petition helps to bring the law and its provisions in order to ensure a better future to all citizens of the State who want to live free of the juvenile justice and adult criminal justice laws. I would be grateful and proud

나는 이 청원이 청소년 사법 및 성인 형사 사법법으로부터 벗어나 살고자 하는 모든 국가의 시민들에게 더 나은 미래를 보장하기 위해 법과 규정을 가져오는 것에 도움이 되기를 바란다.

‘소년법 폐지를 청원한다. 십대들은 소년법을 악용하여 무수한 범죄를 저지르고 있다.’는 구절을 청원내용 자동 생성기에 입력해서, 위와 같은 준수한 청원내용을 자동으로 생성할 수 있었다.

한계점

- GPT-2에서 한글을 지원하지 않아서 한글 생성기를 만들지 못했다. SK에서 만든 koGPT가 있었으나, 환경 및 실력 부족으로 만들지 못한 것이 아쉽다.

느낀점

- 텍스트데이터분석을 하면서, 텍스트 안에 숨어있는 특징들을 찾는 과정이 재미있다고 느꼈다. 흥미를 느끼고 앞으로 스스로 더욱 공부할 동기가 생겼다.

감사합니다!