

# TF-IDF 기반 핵심 키워드의 추출 시스템 설계 및 구현

국민청원 데이터를 중심으로

서해준

Jan, 29. 2020

## 요 약

2017년 8월 서비스된 청와대 국민청원제도는 기존의 국민청원제도의 한계를 타파했다는 점, 직접민주주의의 실현에 한 발짝 더 나아갔다는 계기가 되었다는 점에서 주목할 만한 서비스라고 할 수 있다.

본 연구는 이러한 청와대 국민청원제도에 대해 핵심 키워드를 추출하는 시스템을 제안한다. 목적은 다음과 같다. 첫째, 청원과 관련된 내용을 빠르게 파악하여 청원의 요지가 무엇인지 파악하는 것. 둘째, 특정 이익집단의 문제해결도구로 사용됨을 방지하는 것. 셋째, 청원의 답변을 받지 못하더라도 정보를 얻을 수 있게끔 하는 것이 그 목적이다.

이를 위해 청원 데이터를 수집하여 형태소 분석을 통해 명사를 추출해내고, TF-IDF 통계 모델을 이용해 가중치가 높은 키워드를 추출해낸다. 추가적으로, 핵심 키워드를 통해 핵심 문장 및 관련 청원을 제공하는 시스템을 제안한다.

## 제 1 장

# 서론

### 1.1 연구배경

“국민이 물으면 정부가 답한다” 라는 국정철학 아래 만들어진 청와대 국민청원 제도는 크게 두 가지 문제점이 있다. 첫째, 높은 청원 수와 방문자 수에 비해 매우 낮은 답변률을 갖는다는 것. 둘째, 특정 이익집단의 문제해결 요소로 이용될 수 있다는 것이 그 문제점이다.

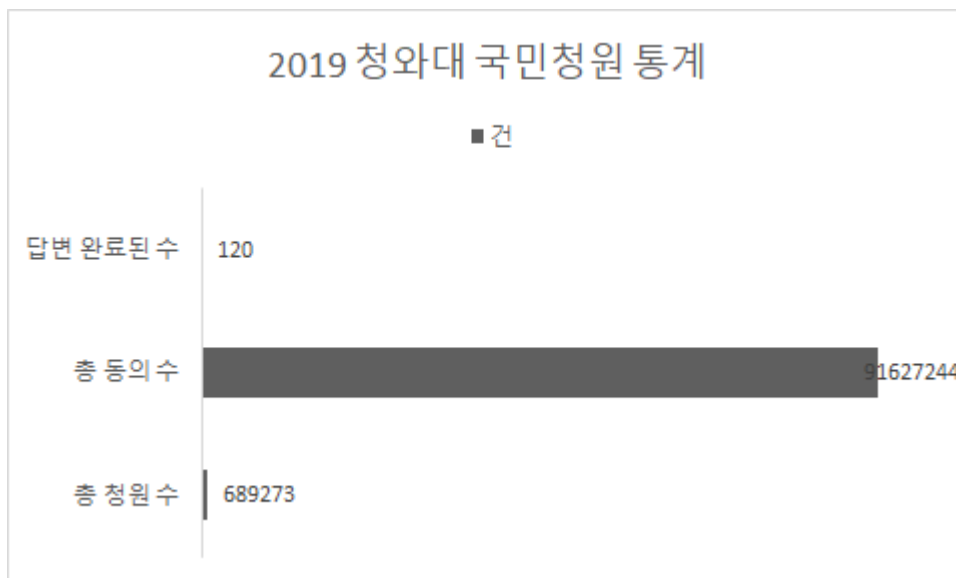


그림 1.1: 2019 청와대 국민청원 통계

이에 청원 데이터의 핵심 키워드 분석을 통해 청원의 주제 및 관련 청원을 제공하고자 하였으며, 이를 통해 청원에 대한 쉽고 빠른 이해와 국민 각자가 숙고한 판단을 하여 특정 이익집단에게 휘둘리지 않는 숙고한 판단을 도울 수 있을 것이다.

## 1.2 연구방법

연구에 사용된 데이터는 2017년 8월 19일부터 2019년 2월 4일까지의 만료된 국민청원 데이터셋이다. 총 395,547 개이며, 각 청원을 하나의 문서(Document)로 정의하였다.

각 문서별로 명사를 추출한 뒤, TF-IDF를 계산하는 방식으로 핵심 키워드를 추출하되, 데이터 정규 및 불용어 처리를 위해 기존의 TF-IDF 모델을 변형한 새로운 통계 모델을 제시한다. 이를 통해 불용어가 제거된 핵심 키워드만을 추출할 수 있게 되고, 추가적으로 추출된 키워드 기반 핵심 문장 및 관련 청원을 제공하는 방법을 소개하도록 하겠다.

데이터 수집 및 서버 구성을 위한 언어는 EcmaScript <sup>1</sup> 를 이용하였으며, 수집한 데이터는 서버에 CSV 파일로 저장하여 이용하였다.

## 1.3 관련 연구

### 1.3.1 TF-IDF

TF-IDF는 단일 문서 내 키워드의 중요도를 전체 문서군에 대해 상대적으로 평가하기 위한 통계 수치 모델이다. 이는 단일 문서에서의 키워드 출현 빈도에 전체 문서에서의 키워드 출현 빈도의 역수를 서로 곱한 값으로, 꽤나 전통적인 통계 모델이다. 그러나 인터넷 뉴스 기사에서 키워드를 추출하거나 [8], 최근의 이직 추천 시스템 설계 [7] 와 같은 경우에도 종종 사용되는 모델이다.

TF-IDF 통계 모델은 현재 통계의 주를 이루고 있는 기계학습과는 달리 학습이 필요하지 않다는 점, 전처리 과정이 복잡하지 않다는 점에 있어 범용적으로 적용할 수 있다는 장점이 있다 [8] [4].

---

<sup>1</sup>ECMA-262 표준 JavaScript

### 1.3.2 국민청원제도

새로운 정부가 출범하며, 2017년 8월부터 인터넷을 통해 서비스되고 있는 국민청원제도는 기존 청원제도의 한계를 타파했다는 점에서 매우 주목할만한 가치가 있다. 이에 [6] 과 같은 청원 데이터 기반 이슈 분석 연구나, [9] 과 같은 국민청원 제도에 대한 새로운 방향으로의 정비 방안. 그리고, [5] 와 같은 국민청원제도 게시판이 실제 시민들의 정치효능감에 어떠한 영향을 미치는지에 대한 연구 등 후속연구가 뒤따르고 있다.

본 논문에서는 국민청원 데이터에서 핵심 키워드를 추출하는 방법을 TF-IDF 통계 모델의 변형을 통해 제안한다. 기존의 TF-IDF 통계 모델은 방대한 데이터에서는 적절한 수치를 제공하지 못한다는 단점이 있다 [4]. 40만 개에 가까운 데이터를 이용하는 본 연구에서는 이를 위해 TF-IDF 모델에 방대한 데이터셋에서도 적절한 값을 유추해낼 수 있는 정규화를 적용한 TF-IDF 모델을 제시하며, 이를 통해 국민청원을 분석하고 핵심 키워드를 추출하는 방법을 소개하고자 한다.

## 제 2 장

# 핵심 키워드 추출 시스템 설계

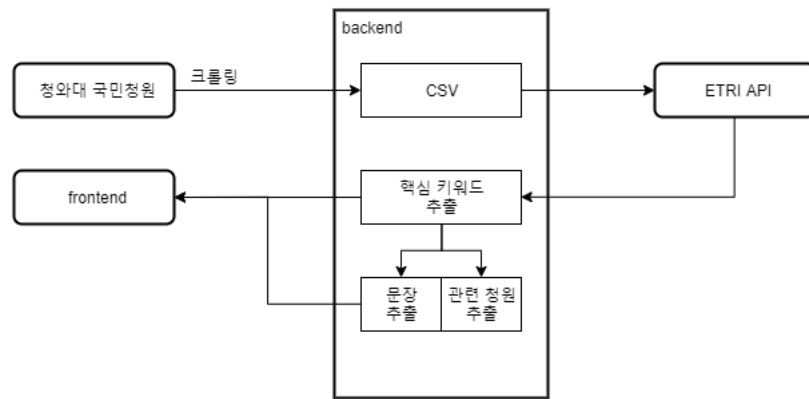


그림 2.1: 시스템 개략도

효과적으로 청원과 관련된 데이터를 제공하였는지 여부는 무엇에 의해 좌우되는가? 이는 핵심 키워드를 어떻게 추출하였는지에 따라 달라질 것이다. 이를 위해 핵심 키워드 추출 시스템을 설계하였고, 본 논문에서 이를 소개하도록 하겠다.

먼저 청원 글 수집을 위해 청와대 홈페이지의 국민청원 게시판에서 웹 크롤링을 진행한다. 크롤링되는 청원 글의 대상은 정확도를 위하여 완료된 청원 글만을 대상으로 진행하며, 각 청원 글은 향후 가중치 부여를 위해 제목과 내용을 분리하여 서버에 CSV 파일로 저장하게 된다.

그 다음으로, TF-IDF를 통해 청원 글에서 핵심 키워드를 추출하기 이전에

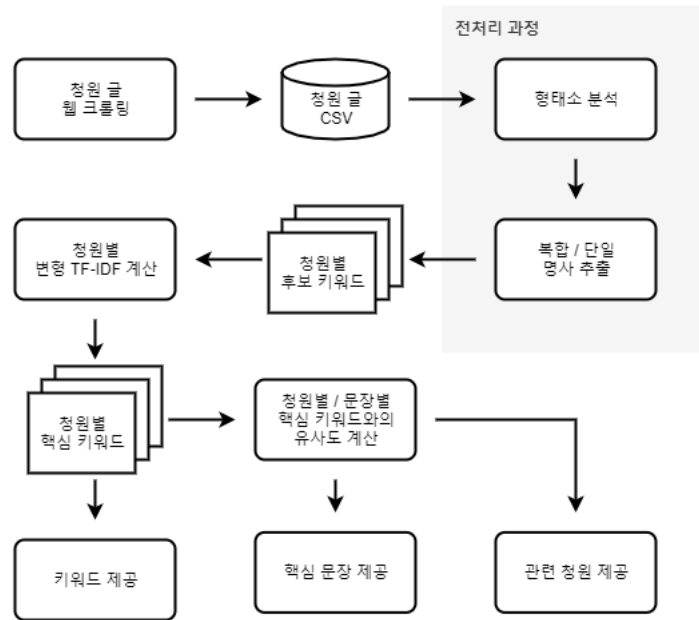


그림 2.2: 핵심 키워드 추출 시스템

전처리 과정을 진행하게 된다. 먼저 ETRI<sup>1</sup>의 형태소 분석 API [2]를 이용해 청원 글을 형태소 단위로 분리한다. 이 결과로 하나의 청원 글이 여러개의 형태소로 분리되게 되며, 이를 기반으로 단일 및 복합 명사를 추출해 “후보 키워드”를 추출하는 작업을 진행한다. 이후 TF-IDF 계산 및 가중치 계산을 진행하게 된다.

계산된 TF-IDF 가중치를 내림차순으로 정렬하여 가장 높은 가중치를 갖는 5개의 후보 키워드를 “핵심 키워드”라고 하며, 이는 향후 핵심 문장 및 관련 청원 글 파악에도 이용된다.

## 2.1 청원 데이터 수집 및 전처리 과정

HTTP Request를 통해 가져와지는 HTML 태그를 바탕으로 필요한 부분만을 추출해내는 웹 크롤링 방식을 통해 데이터 수집을 진행하였다. 단, 현재 청와대 청원 게시판은 크롤링을 막아놓은 상태이며, 이를 위해 가상으로 브라우저를 생성해내는 라이브러리인 Puppeteer [3]를 사용하였으며, 크롤링된 HTML 태그를

<sup>1</sup>한국전자통신연구원

파싱하기 위해 Cheerio [1] 라이브러리를 이용해 크롤링 프로그램을 개발하였다.

2019년 2월 4일까지의 청원이 종료된 모든 청원 글을 대상으로 진행하였으며, 각 청원 글은 제목과 내용으로 분리하여 표 2.1과 같은 형태로 CSV 파일에 저장된다.

— 청원종료 —

## 기간제교사의 정규직화를 반대합니다.

참여인원 : [ 1,933명 ]



### 청원내용

대통령님, 안녕하세요. 저는 임용을 준비하고 수험생입니다. 처음 기간제 정규직화 된다고 했을 때, 공채 채용과정이 있는데, 설마 그러겠어 라고 생각하고 있었습니다. 그리고 기간제는 그 채용과정이 공개채용에 비해 불투명하고, 아이들을 가르칠 수 있는 전문성이 떨어지는데, 어떻게 된다는 거지? 그런 생각이라 사실 그냥 뜬 소문이라고 생각했습니다.

수년간 임용을 준비하는 수험생으로서, 여러 기회비용을 포기하고, 오로지 교단에서 서서 아이들에게 내가 알고 있는 지식을 제대로 가르 치겠다는 신념으로 공부해왔습니다. 그런데 오늘 보니 기간제가 정규직화되는 것이 거의 확정적으로 되는 것같아, 사실 너무 상대적 허탈 감을 느낍니다. 기간제 선생님들은 임용고시를 일단 보류해두고 현장에 나아간 것이고, 임용 준비생은 일단 임용고시 합격을 목표로 그 한 해 기회비용까지 써가면서 공부하는 것이니까요. 그래서 전문적인 지식을 더 갖추고 있는 임용수험생들이 현장에 일하지 않았다는 이유로 기간제 교사에게 기회까지 박탈당하면서 자리를 내주어야 하나 의구심이 듭니다. 존경하는 대통령님 기간제 정규직화 다시 한번 생각해주시기 바랍니다. 양질의 아이들에게 교육을 제공하기 위해서라도, 기간제 정규직화는 안됩니다. 똑같은 기회의 평등을 제공해주십시오. 전문적인 지식은 갖춰그 교사들이 경험하고있는 다양한 청탁을 경청하기 바랍니다. 신임임용은 투명하고 공정한 임용 절차가 필요합니다. 미래에

그림 2.3: 원본 청원 글

article_id	start	end	answered	votes	category
46	2017-08-19	2017-09-18	0	1933	육아/교육
title			contents		
기간제교사의 정규직화를 반대합니다.			대통령님 안녕하세요. 저는 임용을 ...		

표 2.1: 청원 글 데이터 모델

구현에 사용된 청원 글 수는 대략 40만 개 정도이며, 수집한 청원 글을 사용

하기 위한 전처리 과정은 두 가지가 있다. 첫 번째로, 청원 글에서 명사를 추출해내기 위해 형태로 분석을 진행한다. 이는 언급한 ETRI 형태소 분석 API [2]를 사용하며, “JSON” 형태로 반환된다. 이를 도식화하면 그림 2.4와 같은 형태소 분석 결과가 나타나게 된다.

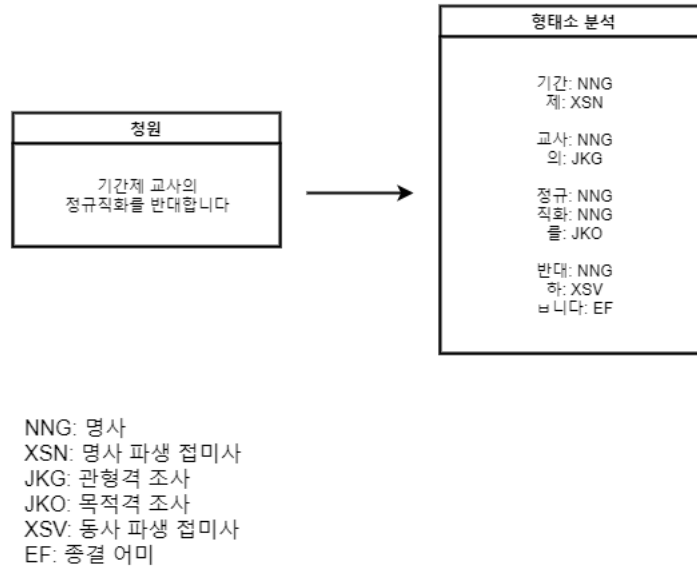


그림 2.4: 형태소 분석 결과

이를 기반으로 단일명사 및 복합명사를 추출한다. 단일 명사만을 추출하였을 경우 “기간제”, “교사”와 같은 의미를 표현하기 어렵기에, “기간제 교사”와 같은 복합명사를 추가적으로 추출하는 것이며, 다음과 같은 오토마타를 통해 단일 및 복합명사를 추출하게 된다.



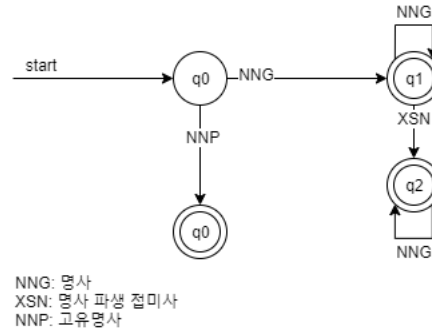


그림 2.5: 명사 추출 오토마타

추출 명사	기간제 교사	정규직화	반대	교사	대통령님 안녕	임용	준비
-------	--------	------	----	----	---------	----	----

표 2.2: 추출 명사 결과

## 2.2 TF-IDF 계산을 통한 핵심 키워드 도출

그러나 이렇게 추출된 “후보 키워드”는 각 청원 글에 대한 키워드 집합일 뿐이며, 아직 어떠한 키워드가 청원 글을 적절히 나타내는지에 대한 가중치 값은 계산되지 않았다. 가중치 값을 계산해 “핵심 키워드”를 추출함으로써 시스템 이용자는 청원 글에 대한 요지를 파악할 수 있기에, 이는 반드시 필수적인 절차라고 할 수 있다.

### 2.2.1 TF-IDF

핵심 키워드는 TF-IDF 통계 모델을 이용해 계산되나, 기존의 TF-IDF 모델의 경우 방대한 데이터셋에 대해서는 적절한 가중치를 계산하지 못하기에 [4], 본 연구에서는 이를 변형한 TF-IDF 모델을 이용하도록 하겠다.

먼저 기존의 TF-IDF 모델의 경우 (2.1)과 같은 수식을 갖는다.

$$\begin{aligned}
 TF(i, d) &= \frac{n(i, d)}{\sum_j n(j, d)} \\
 IDF(i, D) &= \log \frac{|D|}{|\{i \in d : d \in D\}|} \\
 TFIDF(i, d, D) &= TF(i, d) \times IDF(i, D)
 \end{aligned} \tag{2.1}$$

$n(i, d)$ : 키워드  $i$ 가 문서  $d$ 에서 나타나는 빈도 수

$\sum_j n(j, d)$ : 문서  $d$ 에서 나타나는 모든 키워드의 빈도 수

$D$ : 문서 집합

$|D|$ : 문서 집합의 길이

$|\{i \in d : d \in D\}|$ : 키워드  $i$ 가 나타나는 문서의 개수

TF(Term-Frequency)는 키워드  $i$ 에 대해 문서  $d$ 에서 나타나는 모든 키워드가 나타나는 빈도 수로 나누어 정규화하는 함수이다. 그러나 TF 값만을 가지고 문서를 표현하기에는 불용어 수준의 단어가 나타날 가능성이 있기 때문에 [8],  $IDF$  함수를 도입한다.

IDF(Inverse Document-Frequency)는 키워드  $i$ 가 포함된 문서 수  $|\{i \in d : d \in D\}|$ 를 전체 문서 집합의 수  $|D|$ 로 나눈 것에 역수를 취한 것이다. 이 둘을 서로 곱함으로써 전체 문서에서는 적게 언급되면서도 특정 문서에서는 언급 빈도가 높은 키워드만을 추출해낼 수 있다.

### 2.2.2 NTF-IDF

기존의 TF-IDF 통계 모델 중,  $TF$  값은 그 값이 너무 작아지거나 적절하게 나오지 않는 문제가 있기에 [8] (식 2.2)와 같은  $TF$ 의 변형을 사용하곤 한다.

$$\begin{aligned}
 BTF(k, D) &= \sum_{j=1}^{|D|} n(k, j) \\
 NTF(i, T, D) &= \frac{BTF(i, D)}{\max_{k=1}^{|T|} BTF(k, D)} \\
 NTFIDF(i, T, D) &= NTF(i, T, D) \times IDF(i, D)
 \end{aligned} \tag{2.2}$$

$T$ : 카테고리 내에서 나타나는 모든 키워드의 집합

$|T|$ : 카테고리 내에서 나타나는 모든 키워드 집합의 길이

$\max_k BTF(k, D)$ : 가장 큰  $BTF$  값

BTF(Basic Term-Frequency)는 전체 문서에서 키워드  $i$ 의 빈도 수의 합을 의미하며, NTF(Normalized Term-Frequency)는 키워드  $i$ 에 대해 가장 큰 빈도수를 갖는 카테고리의 BTF 값으로 나누어 정규화를 진행한 값을 의미한다. 이 과정을 통해 선행연구 [8] 에서 불용어가 제거됨을 볼 수 있다.

### 2.2.3 TF-IDF의 변형

단, 이는 단일 문서가 아닌 특정 카테고리의 비중을 고려하여  $TF$ 를 변형한 알고리즘이며, 본 논문에서는 이를 단일 문서에 대해 알고리즘을 적용할 수 있도록 변형하여 진행하였다.

$$NTF(i, K, d) = \frac{n(i, d)}{\max_{k=1}^{|K|} n(k, d)} \quad (2.3)$$

$$NTFIDF(i, d, K, D) = (NTF(i, K, d) + 1) \times IDF(i, D)$$

$n(i, d)$ : 키워드  $i$ 가 문서  $d$ 에서 나타나는 빈도 수

$K$ : 문서에서 나타나는 모든 키워드의 집합

$|K|$ : 문서에서 나타나는 모든 키워드 집합의 길이

$\max_k n(k, d)$ : 키워드  $i$ 에 대해 문서  $d$ 에서 나타나는 가장 큰 빈도 수

NTF(Normalized Term-Frequency)는  $n(i, d)$ 가 1을 넘지 않도록  $TF$ 를 정규화 한 값이며,  $NTFIDF$ 는  $NTF$ 와  $IDF$ 의 곱으로, 특정 청원 글에 집중적으로 언급되는 키워드의 식별을 가능하게 한다. 즉, 해당 키워드가 높은 값을 가질수록 해당 청원 글을 잘 나타낸다고 할 수 있는 것이다.

이는 기존의 TF-IDF보다 불용어 처리에 있어 더 향상된 성능을 보이며, 이는 아래 “3.3 기존 TF-IDF와 변형 TF-IDF 간의 비교”에서 그 차이를 보이도록 하겠다.

## 2.3 핵심 문장 및 관련 청원 제공

도출된 핵심 키워드를 기반으로 어렵지 않게 핵심 문장 및 관련 청원의 제공이 가능하다. 이는 키워드를 더욱 확장된 용도로써의 이용이 가능함으로 의미함과 동시에, 향후 청원 분석뿐만이 아닌 타 분석 연구에도 도입할 수 있다는 것을 의미한다.

### 2.3.1 핵심 문장 제공 방법

청원 내용을 문장 단위로 나누는 전처리 작업이 선행되어야 한다. ETRI API [2]에서 제공하는 기능 중 하나인 형태소 분석을 마찬가지로 이용하여 문장을 나누게 되며, ETRI API 자체에서 제공해주는 기능이다. 이 결과 표 2.3과 같은 형태로 문장이 나누어지게 된다.

도출된 문장들 중, 핵심 키워드가 많이 출현하는 문장. 즉, 핵심 키워드의 TF 값이 높은 문장을 선택함으로써 핵심 문장을 도출해낼 수 있게 된다.

문장
기간제교사의 정규직화를 반대합니다.
대통령님 안녕하세요.
저는 임용을 준비하고 수험생입니다.
처음 기간제 정규직화 된다고 했을 때, 공채 채용과정이 있는데, ...
그리고 기간제는 그 채용과정이 공개채용에 비해 불투명하고, ...

표 2.3: 추출 문장

### 2.3.2 관련 청원 제공 방법

마찬가지로 도출된 핵심 키워드를 이용한다. 모든 청원 문서에 대해 핵심 키워드가 해당 문서를 잘 나타내는지 여부를 결정하는 NTF-IDF 값을 계산한 뒤, 이를 높은 값을 갖는 순서대로 도출하여 관련 청원을 도출해낼 수 있게 된다.

이와 같이 어렵지 않은 방법으로 도출되는 핵심 문장과 관련 청원이지만, “3.2 변형 TF-IDF를 통한 핵심 키워드 추출에 관한 실험”에서 보여지듯이, 상당히 의미있는 결과를 가져다 줄 수 있다.

## 제 3 장

# 실행 및 결과

### 3.1 실행 데이터

2017년 8월 19일부터 2019년 2월 4일까지, 총 395,547 개의 만료된 청와대 국민 청원 데이터를 수집하였다. 수집된 데이터는 CSV 파일로 저장하여 사용하였으며, 그 구조는 아래와 같다.

article_id	start	end	answered	votes	category
46	2017-08-19	2017-09-18	0	1933	육아/교육
title			contents		
기간제교사의 정규직화를 반대합니다.			대통령님 안녕하세요. 저는 임용을 ...		

표 3.1: 청원 글 데이터 모델

일반적으로 ‘title’에서 청원과 관련된 핵심 내용이 언급되기에, ‘title’과 ‘contents’를 서로 나누어 저장하였으며, 이를 이용해 ‘title’과 관련된 키워드에는 추가적인 가중치를 부여할 수 있게 된다.

### 3.2 변형 TF-IDF를 통한 핵심 키워드 추출에 관한 실행

언급했듯이, TF-IDF는 해당 문서를 가장 잘 나타내는 키워드를 도출해내기 위해 사용하는 통계 모델이다. 아래의 표 3.2는 46번 청원 글 “기간제교사의 정규직화를 반대합니다”에 대한 변형 TF-IDF 가중치 모델을 적용해 나온 키워드와 TF-IDF 값을 나타낸 표이며, 이를 통해 시스템 사용자는 46번 청원 글은 기간제 교사의 정규직화를 반대한다는 것을 짐작할 수 있게 된다.

키워드	정규직화	기간제	정규직	기간제교사	반대
TF-IDF	16.9	16.5	13.6	12.5	8.0

표 3.2: 46번 청원 글의 키워드 및 TF-IDF

표 3.3 및 표 3.4는 추출된 핵심 문장과 관련 청원이다. 핵심 문장은 TF 값이 높은 다섯 개의 문장을 추출해내며, 이를 문장 순서대로 재구성하여 제공한다. 관련 청원의 경우 추출된 핵심 키워드 기반 NTF-IDF 값을 계산하여 이를 기준으로 제공하게 된다.

문장	TF	Index
기간제교사의 정규직화를 반대합니다.	5	0
처음 기간제 정규직화 된다고 했을 때, 설마 그러겠어라고 생각하고 있었습니다.	3	3
그런데 오늘 보니 기간제가 정규직화되는 것이 거의 확정적으로 되는 것같아, 사실 너무 상대적 허탈감을 느낍니다.	3	7
존경하는 대통령님 기간제 정규직화 다시 한번 생각해주시기 바랍니다.	3	10
양질의 아이들에게 교육을 제공하기 위해서라도, 기간제 정규직화는 안됩니다.	3	11

표 3.3: 46번 청원 글의 핵심 문장

청원 번호	청원 제목	NTF-IDF
43	기간제 교사의 정규직화를 반대합니다.	524.5
42	기간제 교사의 정규직화를 반대합니다.	513.9
653	기간제교사 정규직화 전환을 반대합니다. 상대적 박탈감이 너무 크네요ㅠㅠ	469.0
394	차라리 기간제 교사 제도를 없애주세요.	451.8
683	(기간제교사의 정규직화 반대) 노력으로 빛을 볼수 있는 사회를 만들어주세요.	375.8

표 3.4: 46번 청원 글과 관련된 타 청원

### 3.3 기존 TF-IDF와 변형 TF-IDF 간의 비교

기존의 TF-IDF 통계 모델은 문서의 수가 많아질수록 정규화가 적절하게 진행되지 않아 적절한 수치를 제공할 수 없다 [4]. 따라서 기존의 모델을 그대로 사용할

경우 불용어를 포함하는 경우가 잦았으며, 이에 정규화 수식을 포함하는 새로운 TF-IDF 모델을 설계하게 되었다.

표 3.5에서 보여지듯이 기존 TF-IDF 통계 모델의 경우 상위 다섯 개의 키워드를 추려보면, 주제 “기간제 교사의 정규직화 반대”와는 상관이 없는 불용어가 포함되어 있음을 볼 수 있다. 이는 많은 데이터셋에서 주로 나타나는 TF-IDF 통계 모델의 문제이며, 표 3.6의 변형 TF-IDF 통계 모델의 경우, 주제와 상관없는 불용어는 제거되었음을 볼 수 있다.

키워드	기간제	정규직화	정규직	대통령	채용과정
-----	-----	------	-----	-----	------

표 3.5: 기존 TF-IDF 통계 모델을 사용한 키워드 추출 결과

키워드	정규직화	기간제	정규직	기간제교사	반대
-----	------	-----	-----	-------	----

표 3.6: 변형 TF-IDF 통계 모델을 사용한 키워드 추출 결과

이 결과로 변형 TF-IDF 모델을 사용하는 것이 더 적절한 결과를 가져다줄 수 있게 된다.



## 제 4 장

# 결론

본 연구를 통해 TF-IDF 기반 청와대 국민청원제도에 대한 핵심 키워드를 추출해 보았고, 이를 이용해 핵심 문장 및 관련 청원 글을 추출해 보았다. 그러나 청원은 각 개인의 생각이 온전히 담겨져 있는 주장인 동시에, 한 인간의 삶이 담겨있는 글이라고도 할 수 있다. 따라서, 어느 한 요소를 배제함으로써 청원의 요지를 모두 파악할 수 있다는 주장에는 현실적인 한계가 있다.

이러한 한계에 대응하기 위해 최대한 많은 데이터셋에 대해 TF-IDF 계산을 진행했으나, 추출된 키워드 및 문장이 청원의 주장을 온전히 나타냄을 보장하지는 않을 것이다. 이는 향후 추가적인 연구를 통해 키워드 추출 시스템의 성능과 추가적인 가중치 제공 방안 및 검증 단계를 마련해야만 할 것이다.

통계 모델인 TF-IDF 뿐만 아니라, 기계 학습을 통해 데이터셋을 분석함으로써 더욱 정교한 처리를 할 수도 있을 것으로 기대한다.

## 참고문헌

- [1] *Cheerio*. URL: <https://github.com/cheeriojs/cheerio> (visited on 01/16/2020).
- [2] *ETRI Open API*. URL: [http://aiopen.etri.re.kr/guide\\_wiseNLU.php](http://aiopen.etri.re.kr/guide_wiseNLU.php) (visited on 01/16/2020).
- [3] *Puppeteer*. URL: <https://github.com/puppeteer/puppeteer> (visited on 01/16/2020).
- [4] Juan Ramos et al. “Using tf-idf to determine word relevance in document queries”. In: *Proceedings of the first instructional conference on machine learning*. Vol. 242. Piscataway, NJ. 2003, pp. 133–142.
- [5] 김선희. “국민청원게시판의 특성이 시민들의 정치효능감에 미치는 효과”. In: (2019).
- [6] 소프트웨어정책연구소. “국민청원 데이터에 기반한 SW분야 이슈 분석”. In: (2019).
- [7] 신동현. “TF-IDF 기반 이직 대상 기업 추천 시스템 설계 및 구현”. In: (2018).
- [8] 이성직 and 김한준. “TF-IDF의 변형을 이용한 전자뉴스에서의 키워드 추출 기법”. In: *한국전자거래학회지* 14.4 (2009), pp. 59–73.
- [9] 한국행정연구원. “국민참여와 소통 확대를 위한 행정절차법제 정비 방안”. In: (2018).