NLP

자연어처리 모델을 이용한 법조문 조회 프로그램 구현

3팀(삼삼오오) 팀장 김찬희, 박유정, 정새하, 정한슬

2022. 02.14 ~ 3.24

자연어처리 모델을 이용한 법조문 조회 프로그램 구현 (2022. 02.14 ~ 3.24)

자연어처리 모델을 이용해 감사 지적사항 및 관련근거 판별에 필요한 법을 제시

Work Team & Member

3팀(삼삼오오) 김찬희, 박유정, 정새하, 정한슬

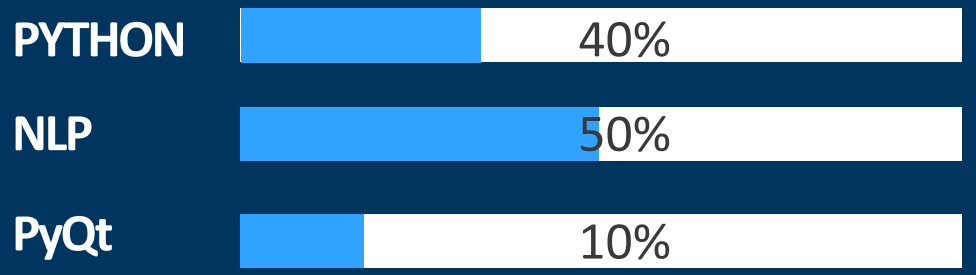
Work Schedule

2.14~2.19 기획
2.20~2.26 데이터 수집
2.20~3.12 모델, 형태소 분석 성능 테스트 및 선정
2.27~3.13 유사도 분석
3.06~3.25 시각화

Work Dataset

법조문 145,279개

Skills



목차 a table of contents

1 프로젝트 개요

2 데이터 수집 및 모델, 형태소 분석기선정

3 모델 구축 및 학습

4 시각화 및 결과

5 개선 사항 및 소감

6 참고자료

Part 1

프로젝트 개요

1.문제 정의 및 목적
2.사용 툴 및 기술
3.역할 분담
4.시스템 순서도
5.진행 과정 및 소요 기간
6.프로젝트 GitHub

1.1 프로젝트 정의 및 목적

문제 정의

ESG는 기업의 비재무적 요소인 '환경(Environment), 사회(Social), 지배구조(Governance)'의 약자

2025년부터 기업이 의무적으로 공시해야하는 평가 지표로 기업은 당장 ESG 점수를 높여야 한다는 과제를 떠안게 됐다

목적

자연어처리 모델을 이용해 감사 지적사항 및 관련근거 판별에 필요한 법을 제시해 감사인력 및 소요시간을 줄이도록 함

1.2 사용 툴 및 기술

IDE

jupyter

PC

Team Collaboration Tool

GitHub

slack

Library

BeautifulSoup

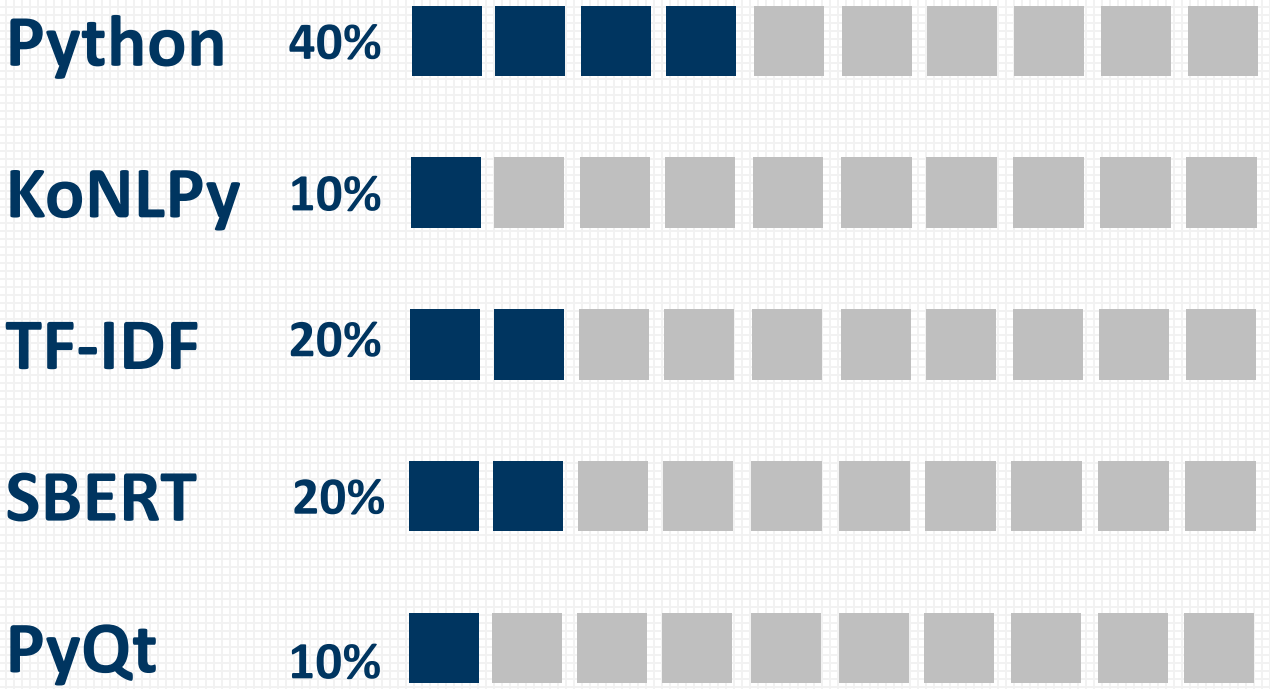
Pandas

NumPy

KoNLPy

Qt

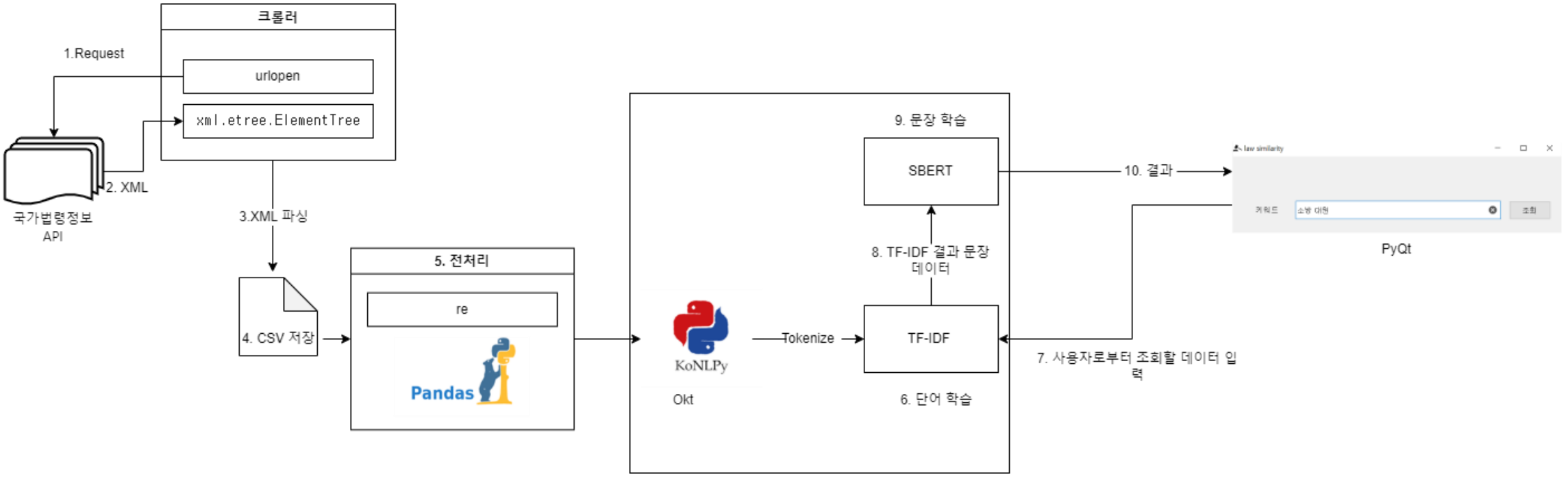
scikits learn



1.3 역할 분담

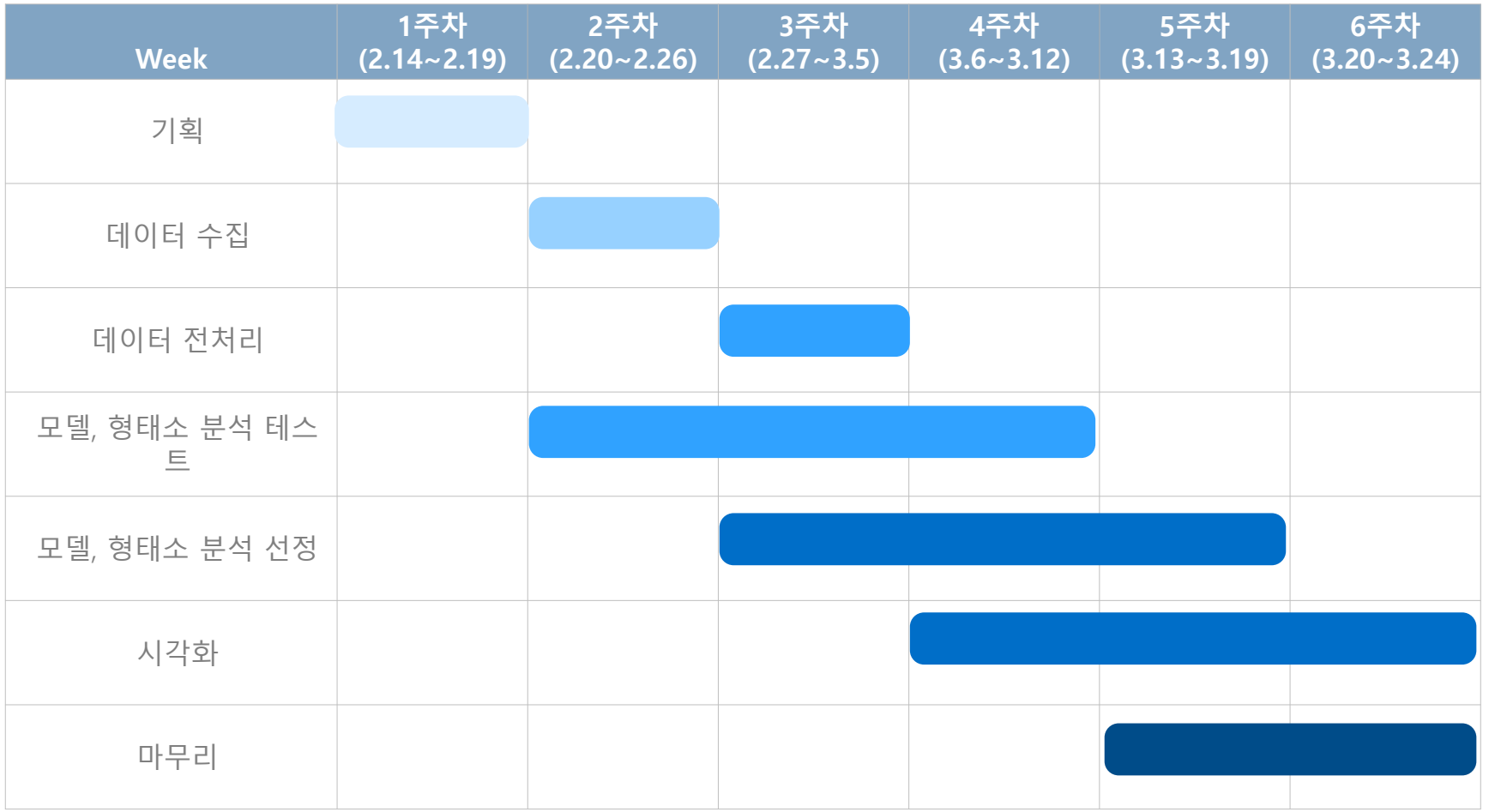
김찬희[PL/Developer]	박유정[DA/Developer]	정새하[AA/Developer]	정한슬[QA/Developer]
데이터 수집 데이터 전처리 SBERT 모델 설계 모델 적합성 판단 팀 일정 관리	데이터 수집 데이터 전처리 형태소 분석기 비교 Doc2Vec 모델 설계 TF-IDF 모델 설계	데이터 수집 데이터 전처리 NLP 기초 모델구축 모델 성능 테스트 SBERT 모델 성능 테스트 프로젝트 코드 통합 PPT 제작 및 발표	데이터 수집 데이터 전처리 PyQt 프로그램 구현 Mecab 테스트 팀 일정 관리 및 기록

1.4 시스템 순서도


크롤러
1.Request
urlopen
xml.etree.ElementTree
2. XML
국가법령정보
API
3.XML 파싱
4. CSV 저장
5. 전처리
re
Pandas
9. 문장 학습
SBERT
10. 결과
8. TF-IDF 결과 문장
데이터
KoNLPy
Okt
Tokenize
TF-IDF
6. 단어 학습
7. 사용자로부터 조회할 데이터 입
력
PyQt

1.5 진행 과정 및 소요 기간

PROJECT TIMELINE 2022년 2-3월



1.6 프로젝트 GitHub

GitHub – Model Code

https://github.com/Growing3Team/Law_NLP_Project

d2v.model	doc2vec 적용 코드
law_similarity_project.ipynb	final
tf-idf_law_similarity.ipynb	tf-idf_law_similarity

GitHub – 법령정보 API 크롤러 Code

https://github.com/Growing3Team/law_similarity

clean_laws	Add files via upload
README.md	Initial commit
law_data_chanhee.zip	Add files via upload
law_data_hanseul.zip	law data part of hanseul from crawler
law_data_saeha.zip	Add files via upload
law_data_yujeong.zip	law data part of yujeong
law_parsing.py	Add files via upload
lawdata_scraper_saeha.ipynb	법령데이터 크롤러 (데이터 확인 최종)

Part 2

데이터 수집 및 모델, 형태소 분석기 선정

1.데이터 수집 및 전처리
2.후보 모델
3.모델 성능 테스트
4.최종 모델 선정
5.SBERT 한국어 모델 성능 테스트
6.형태소 분석기 비교 및 선정

2.1 데이터 수집 및 전처리

국가법령정보 API를 이용한 데이터 수집

국가법령정보 공동활용
LAW OPEN DATA

판 연 ☐ **제1조(목적)** 이 법은 건강한 가정생활의 영위와 가족의 유지 및 발전을 위한 국민의 권리 · 의무와 국가 및 지방자치단체 등의 책임을 명백히 하고, 구현에 기여하는 것을 목적으로 한다.

판 ☐ **제2조(기본이념)** 가정은 개인의 기본적인 욕구를 충족시키고 사회통합을 위하여 기능할 수 있도록 유지 · 발전되어야 한다.

판 연 ☐ **제3조(정의)** 이 법에서 사용하는 용어의 정의는 다음과 같다.

1. "가족"이라 함은 혼인 · 혈연 · 입양으로 이루어진 사회의 기본단위를 말한다.
2. "가정"이라 함은 가족구성원이 생계 또는 주거를 함께 하는 생활공동체로서 구성원의 일상적인 부양 · 양육 · 보호 · 교육 등이 이루어지는 생활
3. "건강가정"이라 함은 가족구성원의 욕구가 충족되고 인간다운 삶이 보장되는 가정을 말한다.
4. "건강가정사업"이라 함은 건강가정을 저해하는 문제(이하 "가정문제"라 한다)의 발생을 예방하고 해결하기 위한 여러 가지 조치와 가족의 부양

판 연 ☐ **제4조(국민의 권리와 의무)** ①모든 국민은 가정의 구성원으로서 안정되고 인간다운 삶을 유지할 수 있는 가정생활을 영위할 권리를 가진다.
②모든 국민은 가정의 중요성을 인식하고 그 복지의 향상을 위하여 노력하여야 한다.

2.1 데이터 수집 및 전처리

XML 파싱으로 진행


```
▼<조문단위 조문키="0001001">
  <조문번호>1</조문번호>
  <조문여부>조문</조문여부>
 ▼<조문제목>
    <![CDATA[ 목적 ]]>
  </조문제목>
  <조문시행일자>20150101</조문시행일자>
  <조문이동이전>000000</조문이동이전>
  <조문이동이후>000000</조문이동이후>
  <조문변경여부>N</조문변경여부>
 ▼<조문내용>
    <![CDATA[ 제1조(목적) 이 법은 건강한 가정생활의 영위와 가족의 유지 및 발전을 위한 국민의 권리·의무와 국가 및 지방자치단체 등의 책임을 명백히 하고,
    구현에 기여하는 것을 목적으로 한다. ]]>
  </조문내용>
 </조문단위>
▼<조문단위 조문키="0002001">
  <조문번호>2</조문번호>
  <조문여부>조문</조문여부>
 ▼<조문제목>
    <![CDATA[ 기본이념 ]]>
  </조문제목>
  <조문시행일자>20150101</조문시행일자>
  <조문이동이전>000000</조문이동이전>
  <조문이동이후>000000</조문이동이후>
  <조문변경여부>N</조문변경여부>
 ▼<조문내용>
    <![CDATA[ 제2조(기본이념) 가정은 개인의 기본적인 욕구를 충족시키고 사회통합을 위하여 기능할 수 있도록 유지·발전되어야 한다. ]]>
  </조문내용>
 </조문단위>
▼<조문단위 조문키="0003001">
  <조문번호>3</조문번호>
  <조문여부>조문</조문여부>
 ▼<조문제목>
    <![CDATA[ 정의 ]]>
  </조문제목>
  <조문시행일자>20150101</조문시행일자>
  <조문이동이전>000000</조문이동이전>
  <조문이동이후>000000</조문이동이후>
  <조문변경여부>N</조문변경여부>
 ▼<조문내용>
    <![CDATA[ 제3조(정의) 이 법에서 사용하는 용어의 정의는 다음과 같다. ]]>
  </조문내용>
```

2.1 데이터 수집 및 전처리

XML 크롤링 및 파싱 코드 작성

```
link_front = "https://www.law.go.kr"

decrees_list = []
for index_num in range(0,len(urls_data['법령상세링크'])):
    url = link_front+urls_data['법령상세링크'][index_num]
    #get xml
    response = urlopen(url).read()
    xtree = ET.fromstring(response)
    decrees_count = len(xtree.find('조문'))
    #조문번호, 조문내용 가져오기
    for j in range(0, decrees_count):
        decrees_info_list =[]
        decress_tree =  xtree.find('조문')[j]
        decress_tree_str = ET.tostring(decress_tree, encoding='unicode')
        if "조(" not in decress_tree_str:
            continue
        else:
            decrees_num = xtree.find('조문')[j].find('조문번호').text
            decrees_contents= xtree.find('조문')[j].find('조문내용').text

            #항, 호 추출
            for object in decress_tree.iter("항"):
                for i in range(0, len(object)):
                    xml_str = ET.tostring(object[i], encoding='unicode')
                    if "<항내용>" in xml_str:
                        hang_content = object.find("항내용").text
                        decrees_contents = decrees_contents+ hang_content
                    elif "<호>" in xml_str:
                        #호를 인식해도 해당 내용을 못가져오는 경우가 있음..
                        # print(object.find("호").find("호내용").text)
                        xml_str_split = xml_str.split("<호내용>")
                        xml_str_replace1=xml_str_split[1].replace("</호>","").strip()
                        xml_str_replace2=xml_str_replace1.replace("</호내용>","")
                        if "<목>" in xml_str_replace2:
                            m_split = xml_str_replace2.split("<목>")
                            xml_str_replace2 = m_split[0].strip()
                            for k in range(0, len(m_split)):
                                if "<목내용>" in m_split[k]:
                                    mc_split = m_split[k].split("<목내용>")
                                    mc_str= mc_split[1].replace("</목내용>","").strip()
                                    mc_str = mc_str.replace("</목>","")
                                    #print(mc_str)
                                    xml_str_replace2 = xml_str_replace2+" "+ mc_str

                        if ("&lt;" in xml_str_replace2) or ("&gt;" in xml_str_replace2):
                            xml_str_replace2=xml_str_replace2.replace("&lt;","<")
                            xml_str_replace2=xml_str_replace2.replace("&gt;",">")
                        decrees_contents = decrees_contents+xml_str_replace2

            decrees_info_list.append(urls_data['법령명'][index_num])
            decrees_info_list.append(urls_data['법령MST'][index_num])
            decrees_info_list.append(urls_data['법령ID'][index_num])
            decrees_info_list.append(urls_data['시행일자'][index_num])
            decrees_info_list.append(urls_data['공포번호'][index_num])
            decrees_info_list.append(urls_data['법령구분명'][index_num])
            decrees_info_list.append(decrees_num)
            decrees_info_list.append(decrees_contents)
            decrees_list.append(decrees_info_list)
```

조문 내용 중 항, 호, 목 이라는 목록을 나타내는 데이터가 따로 있어서 분리하는 코드

```
#조문번호, 조문내용 가져오기
for j in range(0, decrees_count):
    decrees_info_list =[]
    decress_tree =  xtree.find('조문')[j]
    decress_tree_str = ET.tostring(decress_tree, encoding='unicode')
    if "조(" not in decress_tree_str:
        continue
    else:
        decrees_num = xtree.find('조문')[j].find('조문번호').text
        decrees_contents= xtree.find('조문')[j].find('조문내용').text

        #항, 호 추출
        for object in decress_tree.iter("항"):
            for i in range(0, len(object)):
                xml_str = ET.tostring(object[i], encoding='unicode')
                if "<항내용>" in xml_str:
                    hang_content = object.find("항내용").text
                    decrees_contents = decrees_contents+ hang_content
                elif "<호>" in xml_str:
                    #호를 인식해도 해당 내용을 못가져오는 경우가 있음..
                    # print(object.find("호").find("호내용").text)
                    xml_str_split = xml_str.split("<호내용>")
                    xml_str_replace1=xml_str_split[1].replace("</호>","").strip()
                    xml_str_replace2=xml_str_replace1.replace("</호내용>","")
                    if "<목>" in xml_str_replace2:
                        m_split = xml_str_replace2.split("<목>")
                        xml_str_replace2 = m_split[0].strip()
                        for k in range(0, len(m_split)):
                            if "<목내용>" in m_split[k]:
                                mc_split = m_split[k].split("<목내용>")
                                mc_str= mc_split[1].replace("</목내용>","").strip()
                                mc_str = mc_str.replace("</목>","")
                                #print(mc_str)
                                xml_str_replace2 = xml_str_replace2+" "+ mc_str
```

2.1 데이터 수집 및 전처리

XML 파싱 결과 CSV로 저장

총 145,279개 조문 데이터

```
columns = ['법령명', '법령MST', '법령ID', '시행일자', '공포번호', '법령구분명', '조문번호', '조문내용']
law_df = pd.DataFrame(decrees_list, columns = columns)
law_df.head()
```

	법령명	법령 MST	법령 ID	시행일자	공포번호	법령구분명	조문번호	조문내용
0	서해 5도 지원 특별법 시행령	205328	11337	2018.12.13	제29323호	대통령령	1	제1조(목적) 이 영은 「서해 5도 지원 특별법」에서 위임된 사항과 그 시행에 필요...
1	서해 5도 지원 특별법 시행령	205328	11337	2018.12.13	제29323호	대통령령	2	제2조(종합발전계획의 중요한 사항의 변경) 「서해 5도 지원 특별법」(이하 "법"이...
2	석면안전관리법	211537	11384	2020.5.27	제16606호	법률	3	제3조(서해 5도 지원위원회의 구성 및 운영)\n ① 법 제7조제1항에 따른 서해 ...
3	석면안전관리법	211537	11384	2020.5.27	제16606호	법률	4	제4조(국고보조율) 법 제8조제2항에 따라 지방자치단체가 종합발전계획과 연도별 시행...
4	석면안전관리법	211537	11384	2020.5.27	제16606호	법률	5	제5조(지방교부세 특별지원) 행정안전부장관은 「지방교부세법」 제6조 · 제9조 · 제9조의...

법령명,법령MST,법령ID,시행일자,공포번호,법령구분명,조문번호,조문내용
서해 5도 지원 특별법 시행령,205328,11337,2018.12.13,제29323호,대통령령,1,"제1조(목적) 이 영은 「서해 5도 지원 특별법」에서 위임된 사항과 그 시행에 필요한 사항을 규정함을
"
서해 5도 지원 특별법 시행령,205328,11337,2018.12.13,제29323호,대통령령,2,"제2조(종합발전계획의 중요한 사항의 변경) 「서해 5도 지원 특별법」(이하 ""법""이라 한다) 제5조제1
"
서해 5도 지원 특별법 시행령,205328,11337,2018.12.13,제29323호,대통령령,3,"제3조(서해 5도 지원위원회의 구성 및 운영)
① 법 제7조제1항에 따른 서해 5도 지원위원회(이하 ""위원회""라 한다)의 위원은 다음 각 호의 사람이 된다. <개정 2013.3.23, 2014.11.19, 2015.7.24, 2017.7.26>
1. 기획재정부장관 · 교육부장관 · 통일부장관 · 국방부장관 · 행정안전부장관 · 문화체육관광부장관 · 농림축산식품부장관 · 산업통상자
2. 삭제 <2015.7.24>
② 삭제 <2015.7.24>
③ 위원장은 위원회의 업무를 총괄하고, 위원회를 대표한다.
④ 위원회의 회의에 부칠 사항은 회의 개최 10일 전까지 위원에게 알려야 한다. 다만, 긴급한 경우에는 그러하지 아니하다.
⑤ 위원장이 부득이한 사유로 직무를 수행할 수 없을 때에는 위원장이 미리 지명한 위원이 그 직무를 대행한다.
⑥ 위원회의 회의는 재적위원 과반수의 출석으로 개의(開議)하고, 출석위원 과반수의 찬성으로 의결한다.
⑦ 위원회의 사무를 처리하기 위하여 간사 1명을 두며, 간사는 행정안전부차관으로 한다. <개정 2013.3.23, 2014.11.19, 2017.7.26>
⑧ 삭제 <2015.7.24>
⑨ 제1항 및 제3항부터 제7항까지에서 규정한 사항 외에 위원회의 운영에 필요한 사항은 위원회의 심의를 거쳐 위원장이 정한다. <개

2.1 데이터 수집 및 전처리

정규 표현식을 이용한 기호 제거

조문 내용에 영향이 덜 가도록 하되 텍스트만 남길 수 있도록 진행

```
import re

clean_laws_list = []
for sentence in jomuns:

    sentence = re.sub('[①②③④⑤⑥⑦⑧⑨⑩⑪⑫⑬⑭⑮⑯⑰⑱⑲⑳]','',sentence)
    sentence = re.sub('[\t\n]','',sentence)
    sentence = re.sub('[-=+,#/\?:^$@*\"※~&%·!』「」\\‘|\(\)\[\]`\'…》]','',sentence)
    sentence = re.sub('[/一-龥/]', '', sentence)
    sentence = re.sub('<.*?>','', sentence) #<개정 2022.4.23>
    sentence = re.sub('[0-9][.]','', sentence) #숫자.
    sentence = re.sub('[.]','',sentence)
    clean_laws_list.append(sentence)
```

2.1 데이터 수집 및 전처리

기호 제거 전

법령명,법령MST,법령ID,시행일자,공포번호,법령구분명,조문번호,조문내용
서해 5도 지원 특별법 시행령,205328,11337,2018.12.13,제29323호,대통령령,1,"제1조(목적) 이 영은 「서해 5도 지원 특별법」에서 위임된 사항과 그 시행에 필요한 사항을 규정함을
"
서해 5도 지원 특별법 시행령,205328,11337,2018.12.13,제29323호,대통령령,2,"제2조(종합발전계획의 중요한 사항의 변경) 「서해 5도 지원 특별법」(이하 ""법""이라 한다) 제5조제1
"
서해 5도 지원 특별법 시행령,205328,11337,2018.12.13,제29323호,대통령령,3,"제3조(서해 5도 지원위원회의 구성 및 운영)
① 법 제7조제1항에 따른 서해 5도 지원위원회(이하 ""위원회""라 한다)의 위원은 다음 각 호의 사람이 된다. <개정 2013.3.23, 2014.11.19, 2015.7.24, 2017.7.26>
1. 기획재정부장관・교육부장관・통일부장관・국방부장관・행정안전부장관・문화체육관광부장관・농림축산식품부장관・산업통상자
2. 삭제 <2015.7.24>
② 삭제 <2015.7.24>
③ 위원장은 위원회의 업무를 총괄하고, 위원회를 대표한다.
④ 위원회의 회의에 부칠 사항은 회의 개최 10일 전까지 위원에게 알려야 한다. 다만, 긴급한 경우에는 그러하지 아니하다.
⑤ 위원장이 부득이한 사유로 직무를 수행할 수 없을 때에는 위원장이 미리 지명한 위원이 그 직무를 대행한다.
⑥ 위원회의 회의는 재적위원 과반수의 출석으로 개의(開議)하고, 출석위원 과반수의 찬성으로 의결한다.
⑦ 위원회의 사무를 처리하기 위하여 간사 1명을 두며, 간사는 행정안전부차관으로 한다. <개정 2013.3.23, 2014.11.19, 2017.7.26>
⑧ 삭제 <2015.7.24>
⑨ 제1항 및 제3항부터 제7항까지에서 규정한 사항 외에 위원회의 운영에 필요한 사항은 위원회의 심의를 거쳐 위원장이 정한다. <개

기호 제거 후

{"laws": ["제1조목적 이 영은 서해 5도 지원 특별법에서 위임된 사항과 그 시행에 필요한 사항을 규정함을 목적으로 한다", "제2조종합발전계획의 중요한 사항의 변경 사
을 지원보조율로 한다 다만 기준보조율 보조금 관리에 관한 법률 제10조에 따른 차등보조율 또는 다른 법률에 따른 보조율이 100분의 80 이상인 경우에는 그 보조율을
각 호의 구분에 따른 사람에게 정주생활지원금을 지급하여 줄 것을 신청하는 경우에는 그 사람에게 지급할 수 있다 제1항제1호 또는 제2항의 지급대상자인 경우 주민
한 지역현안 수요의 발생으로 특별한 재정수요가 있다고 요청하는 경우에는 법 제15조제3항에 따라 지방교육재정교부금 중 특별교부금을 그 운영재원의 범위에서 심사
제18조에 따라 영농영어를 위하여 대출받은 정책자금에 대하여 2년의 범위에서 농림축산식품부장관 및 해양수산부장관이 정하여 고시하는 바에 따라 대출상환 유예 및
면으로 인한 환경과 국민건강의 피해를 예방하기 위하여 석면의 안전관리에 필요한 시책을 수립하고 시행하여야 한다 사업자는 사업 활동에 따라 발생할 수 있는 석면
등 필요한 사항은 대통령령으로 정한다", "제7조실태조사 환경부장관 관계 중앙행정기관의 장 및 시·도지사는 기본계획과 시행계획을 효율적으로 수립·추진하기 위하여
석면등의 석면 함유 여부를 스스로 확인·조사하거나 산업안전보건법 제119조제2항 각 호 외의 부분 본문에 따른 석면조사기관이하 석면조사기관이라 한다으로 하여금
신고를 한 자가 제4항에 따른 사업장 주변의 석면배출허용기준을 지키지 아니하였을 때에는 대통령령으로 정하는 바에 따라 작업중지를 명할 수 있다 제3항에 따라

2.2 후보 모델 선정

단어 인식

Word2Vec

단어 간 유사성을 고려하기 위해 단어의 의미를 벡터화 시켜주는데, 이러한 방법을 워드투벡터라고 한다. Word2vec은 추론 기반 기법으로, 데이터의 일부를 사용하여 순차적으로 학습하는 미니배치 학습을 바탕으로 한다.

TF-IDF

TF-IDF(Term Frequency-Inverse Document Frequency)는 단어의 빈도와 역 문서 빈도(문서의 빈도에 특정 식을 취함)를 사용하여 DTM 내의 각 단어들마다 중요한 정도를 가중치로 주는 방법

문장 인식

Doc2Vec

Word2Vec 에 이어 2014년 구글 연구팀이 발표한 **문서 임베딩 모델**
타겟 단어와 이전 단어 개가 주어졌을 때, 이전 단어들 + 해당 문서의 아이디로 타겟 단어를 예측하는 과정에서 문맥이 비슷한 문서 벡터와 단어 벡터가 유사하게(코사인 유사도) 임베딩 된다.

SBERT

SBERT는 기본적으로 BERT의 문장 임베딩의 성능을 우수하게 개선시킨 모델

2.3 모델 성능 테스트

Word2Vec VS TF-IDF

Word2Vec를 이용해 조문 데이터 1만개 학습 후 input 값에 해당하는 조문이 어떻게 나오는지 테스트

```
laws_data_list_cut = laws_data_list[0:10000]
len(laws_data_list_cut)
```

```
10000
```

```
laws_word_list=[]
for i in laws_data_list_cut:
    words = okt.morphs(i)
    laws_word_list.append(words)
```

```
model = Word2Vec(sentences=laws_word_list, vector_size=40, window=5, min_count=5, workers=4, sg=0)
```

```
model_name = 'w2v.model'
model.save(model_name)
```

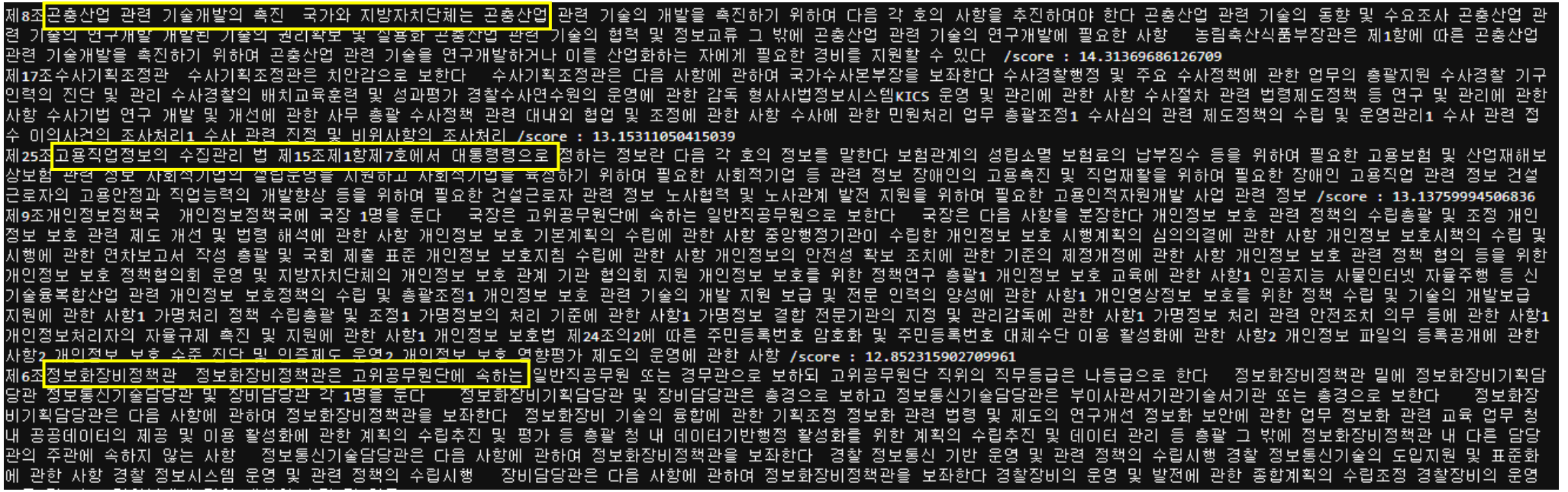
```
load_model= Word2Vec.load("w2v.model")
test_string = "소방 대원 관련"
tokened_test_string = okt.nouns(test_string) # input을 형태소 분석을 실시
tokened_test_string
```

```
['소방', '대원', '관련']
```

2.3 모델 성능 테스트

Word2Vec VS TF-IDF

소방 대원 관련 이라는 input 값을 넣었지만 곤충 산업 관련 기술 개발 등 input 값과 관련이 없는 문장들이 나온 것을 확인



2.3 모델 성능 테스트

Word2Vec VS TF-IDF

TF-IDF를 이용해 조문 데이터 1만개 학습 후 input 값에 해당하는 조문이 어떻게 나오는지 테스트

```
laws_data_list_cut = laws_data_list[0:10000]
```

```
def tokenizer(raw, pos = ['Noun', 'Verb', 'Adjective']):
    okt = Okt()
    # 길이가 1 이하인 토큰은 제외, 위에서 지정한 (Okt 사전에 따른) 토큰들만 특징으로 삼기
    return [word for word, tag in okt.pos(raw) if len(word) > 1 and tag in pos]
```

```
tfidfvectoriser=TfidfVectorizer(tokenizer= tokenizer, ngram_range=(1,2), min_df=2, max_features = 20000)
```

```
tfidfvectoriser.fit(laws_data_list_cut) # 벡터라이저가 단어들을 학습
```

```
C:\Users\skygg\anaconda3\envs\nlp_test\lib\site-packages\sklearn\feature_extraction\text.py:516: UserWarning
ne'
  warnings.warn(
TfidfVectorizer(max_features=20000, min_df=2, ngram_range=(1, 2),
                tokenizer=<function tokenizer at 0x00000195D67278B0>)
```

```
tfidfvectoriser.vocabulary_ # 벡터라이저가 학습한 단어사전을 출력
```

```
...
```

```
sorted(tfidfvectoriser.vocabulary_.items()) # 단어사전을 정렬
```

```
...
```

```
tfidfvectoriser.idf_.shape
```

```
(3753,)
```

```
tfidf_vectors=tfidfvectoriser.transform(laws_data_list_cut)
print(tfidf_vectors)
```

```
...
```

```
features = tfidfvectoriser.get_feature_names()
features
```

```
...
```

```
test_data = '소방 대원 관련'
```

```
srch = [ t for t in tokenizer(test_data) if t in features]
```

```
srch
```

```
['소방', '관련']
```

2.3 모델 성능 테스트

Word2Vec VS **TF-IDF**

소방 대원 관련 이라는 input 값과 관련된 법조문이 결과로 나온 것을 확인

제16조119항공대의 업무 119항공대는 다음 각 호의 업무를 수행한다 인명구조 및 응급환자의 이송의사가 동승한 응급환자의 병원 간 이송을 포함한다 화재 진압 장기이식환자 및 장기의 이송 항공 수색 및 구조 활동 공중 소방 지휘통제 및 소방에 필요한 인력장비 등의 운반 방역 또는 방재 업무의 지원 그 밖에 재난관리를 위하여 필요한 업무 /score : 0.18811915099699872
제3조119구조대에서 갖추어야 할 장비의 기준 119구조구급에 관한 법률 시행령이하 영이라 한다 제5조에 따른 119구조대이하 구조대라 한다 중 특별시광역시특별자치시도특별자치도이하 시도라 한다 소방본부 및 소방서119안전센터를 포함한다에 설치하는 구조대에서 법 제8조제3항에 따라 갖추어야 하는 장비의 기본적인 사항은 소방력 기준에 관한 규칙 및 소방장비관리법 시행규칙에 따른다 소방청에 설치하는 구조대에서 법 제8조제3항에 따라 갖추어야 하는 장비의 기본적인 사항은 제1항을 준용한다 제1항과 제2항에서 규정한 사항 외에 구조대가 갖추어야 하는 장비에 관하여 필요한 사항은 소방청장이 정한다 /score : 0.16741547136497015
제2조정의 이 법에서 사용하는 용어의 뜻은 다음과 같다 구조란 화재 재난재해 및 테러 그 밖의 위급한 상황이하 위급상황이라 한다에서 외부의 도움을 필요로 하는 사람이하 요구조자라 한다의 생명 신체 및 재산을 보호하기 위하여 수행하는 모든 활동을 말한다 119구조대란 탐색 및 구조활동에 필요한 장비를 갖추고 소방공무원으로 편성된 단위조직을 말한다 구급이란 응급환자에 대하여 행하는 상담 응급처치 및 이송 등의 활동을 말한다 119구급대란 구급활동에 필요한 장비를 갖추고 소방공무원으로 편성된 단위조직을 말한다 응급환자란 응급의료에 관한 법률 제2조제1호의 응급환자를 말한다 응급처치란 응급의료에 관한 법률 제2조제3호의 응급처치를 말한다 구급차등이란 응급의료에 관한 법률 제2조제6호의 구급차등을 말한다 지도의사란 응급의료에 관한 법률 제52조의 지도의사를 말한다 119항공대란 항공기 구조구급 장비 및 119항공대원으로 구성된 단위조직을 말한다1 119항공대원이란 구조구급을 위한 119항공대에 근무하는 조종사 정비사 항공교통관제사 운항관리사 119구조구급대원을 말한다1 119구조견이란 위급상황에서 소방기본법 제4조에 따른 소방활동의 보조를 목적으로 소방기관에서 운용하는 개를 말한다1 119구조견대란 위급상황에서 119구조견을 활용하여 소방기본법 제4조에 따른 소방활동을 수행하는 소방공무원으로 편성된 단위조직을 말한다 /score : 0.13408781941566572

2.3 모델 성능 테스트

Doc2Vec VS SBERT

Doc2Vec를 이용해 조문 데이터 50,000개 학습 후 input 값에 해당하는 조문이 어떻게 나오는지 테스트

```
tagged_data = [TaggedDocument(words=okt.morphs(_d), tags=[str(i)]) for i, _d in enumerate(laws_data_list)]
```

```
# 모델 설계
max_epochs = 100
model = Doc2Vec(vector_size=20,
                alpha=0.025,
                min_alpha=0.00025,
                min_count=1,
                dm=1)
```

```
# doc2vec이 단어 사전을 만들어냄
model.build_vocab(tagged_data)
```

```
for epoch in range(max_epochs):
    model.train(tagged_data,total_examples=model.corpus_count,epochs=model.epochs)
    # decrease the learning rate
    model.alpha -= 0.0002
    # fix the learning rate, no decay
    model.min_alpha = model.alpha

model.save("d2v.model")
```

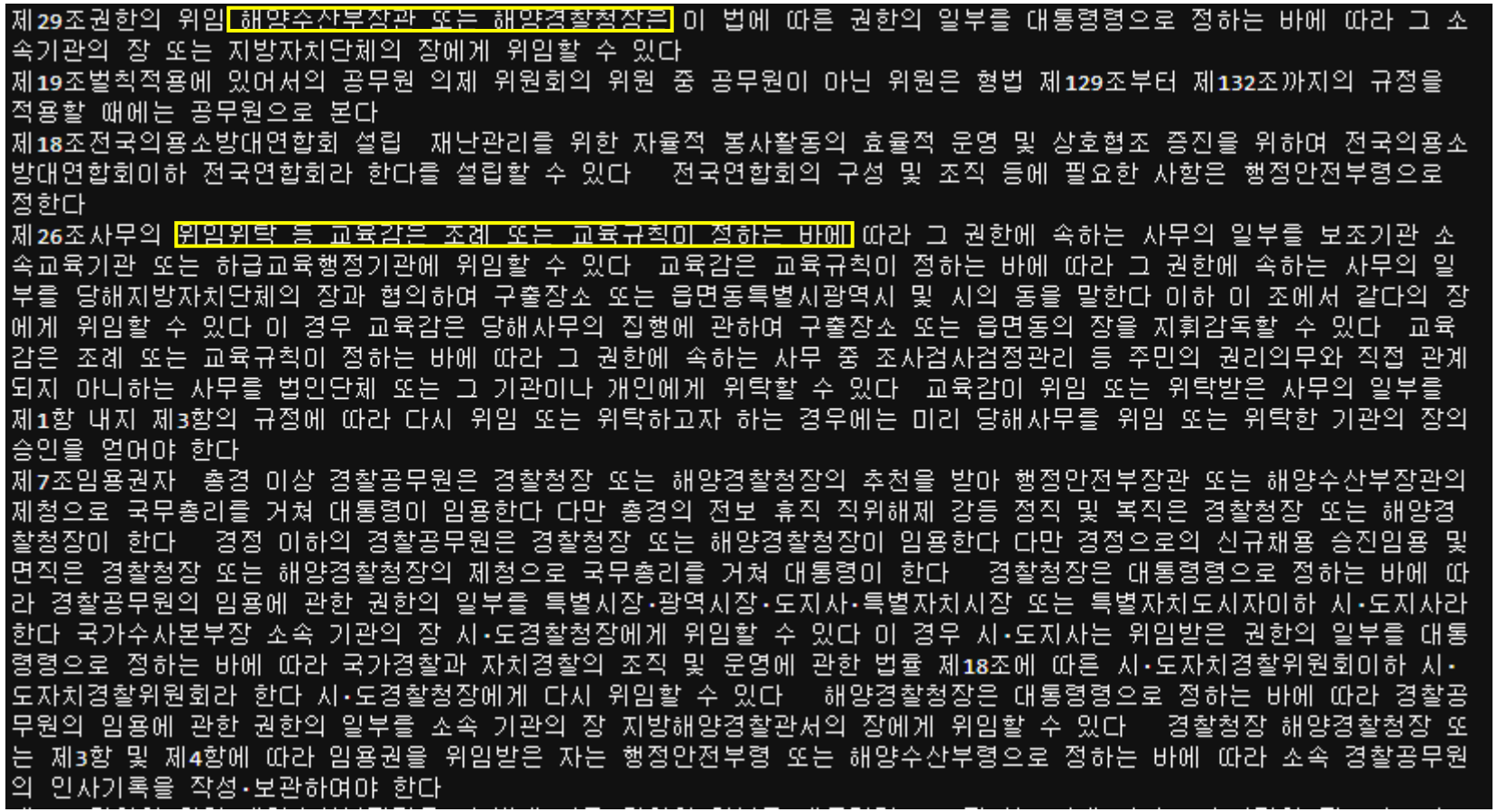
```
load_model= Doc2Vec.load("d2v.model")

test_string = "소방 대원 관련"
tokened_test_string = okt.nouns(test_string) # input을 형태소 분석을 실시
tokened_test_string
```

```
['소방', '대원', '관련']
```

2.3 모델 성능 테스트

Doc2Vec VS SBERT

소방 대원 관련과는 관련 없는 문장들이 나오는 것을 확인


제29조권한의 위임 해양수산부장관 또는 해양경찰청장은 이 법에 따른 권한의 일부를 대통령령으로 정하는 바에 따라 그 소속기관의 장 또는 지방자치단체의 장에게 위임할 수 있다
제19조벌칙적용에 있어서의 공무원 의제 위원회의 위원 중 공무원이 아닌 위원은 형법 제129조부터 제132조까지의 규정을 적용할 때에는 공무원으로 본다
제18조전국의용소방대연합회 설립 재난관리를 위한 자율적 봉사활동의 효율적 운영 및 상호협조 증진을 위하여 전국의용소방대연합회이하 전국연합회라 한다를 설립할 수 있다 전국연합회의 구성 및 조직 등에 필요한 사항은 행정안전부령으로 정한다
제26조사무의 위임위탁 등 교육감은 조례 또는 교육규칙이 정하는 바에 따라 그 권한에 속하는 사무의 일부를 보조기관 소속교육기관 또는 하급교육행정기관에 위임할 수 있다 교육감은 교육규칙이 정하는 바에 따라 그 권한에 속하는 사무의 일부를 당해지방자치단체의 장과 협의하여 구출장소 또는 읍면동특별시광역시 및 시의 동을 말한다 이하 이 조에서 같다의 장에게 위임할 수 있다 이 경우 교육감은 당해사무의 집행에 관하여 구출장소 또는 읍면동의 장을 지휘감독할 수 있다 교육감은 조례 또는 교육규칙이 정하는 바에 따라 그 권한에 속하는 사무 중 조사검사검정관리 등 주민의 권리의무와 직접 관계되지 아니하는 사무를 법인단체 또는 그 기관이나 개인에게 위탁할 수 있다 교육감이 위임 또는 위탁받은 사무의 일부를 제1항 내지 제3항의 규정에 따라 다시 위임 또는 위탁하고자 하는 경우에는 미리 당해사무를 위임 또는 위탁한 기관의 장의 승인을 얻어야 한다
제7조임용권자 총경 이상 경찰공무원은 경찰청장 또는 해양경찰청장의 추천을 받아 행정안전부장관 또는 해양수산부장관의 제청으로 국무총리를 거쳐 대통령이 임용한다 다만 총경의 전보 휴직 직위해제 강등 정직 및 복직은 경찰청장 또는 해양경찰청장이 한다 경정 이하의 경찰공무원은 경찰청장 또는 해양경찰청장이 임용한다 다만 경정으로의 신규채용 승진임용 및 면직은 경찰청장 또는 해양경찰청장의 제청으로 국무총리를 거쳐 대통령이 한다 경찰청장은 대통령령으로 정하는 바에 따라 경찰공무원의 임용에 관한 권한의 일부를 특별시장·광역시장·도지사·특별자치시장 또는 특별자치도시자이하 시·도지사라 한다 국가수사본부장 소속 기관의 장 시·도경찰청장에게 위임할 수 있다 이 경우 시·도지사는 위임받은 권한의 일부를 대통령령으로 정하는 바에 따라 국가경찰과 자치경찰의 조직 및 운영에 관한 법률 제18조에 따른 시·도자치경찰위원회이하 시·도자치경찰위원회라 한다 시·도경찰청장에게 다시 위임할 수 있다 해양경찰청장은 대통령령으로 정하는 바에 따라 경찰공무원의 임용에 관한 권한의 일부를 소속 기관의 장 지방해양경찰관서의 장에게 위임할 수 있다 경찰청장 해양경찰청장 또는 제3항 및 제4항에 따라 임용권을 위임받은 자는 행정안전부령 또는 해양수산부령으로 정하는 바에 따라 소속 경찰공무원의 인사기록을 작성·보관하여야 한다

2.3 모델 성능 테스트

Doc2Vec VS SBERT

SBERT를 이용해 조문 데이터 50,000개 학습 후 input 값에 해당하는 조문이 어떻게 나오는지 테스트

```
input_text_data = "소방 대원 관련"
```

```
q_list = []
q_list.append(input_text_data)
```

```
from sentence_transformers import SentenceTransformer, util
import numpy as np

embedder = SentenceTransformer("jhgan/ko-sroberta-multitask")

corpus = json_data_chan_list

corpus_embeddings = embedder.encode(corpus, convert_to_tensor=True)

queries = q_list
top_k = 5
for query in queries:
 query_embedding = embedder.encode(query, convert_to_tensor=True)
 cos_scores = util.pytorch_cos_sim(query_embedding, corpus_embeddings)[0]
 cos_scores = cos_scores.cpu()

 top_results = np.argpartition(-cos_scores, range(top_k))[0:top_k]

 print("\n\n======================\n\n")
 print("Query:", query)
 print("\nTop 5 most similar sentences in corpus:")

 for idx in top_results[0:top_k]:
  print(corpus[idx].strip(), "(Score: %.4f)" % (cos_scores[idx]))
```

2.3 모델 성능 테스트

Doc2Vec VS **SBERT**

소방 대원 관련이란 input값을 넣었을 때 Doc2Vec보다 관련 있는 문장들이 나오는 것을 확인

```
Query: 소방 대원 관련

Top 5 most similar sentences in corpus:
제20조임무 의무소방원의 임무는 다음과 같다  화재 등에 있어서 현장활동의 보조 가 화재 등 재난재해사고현장에서의 질서유지 등 진압업무의 보조와 구조구급활동의 지원 나 소방용수시설의 확보 다 현장 지휘관의 보좌 라 상황관리의 보조 마 그밖에 현장활동에 필요한 사항의 지원 소방행정의 지원 가 문서수발 등 소방행정의 보조 나 통신 및 전산 업무의 보조 다 119안전센터에서의 소내근무의 보조 라 소방용수시설 유지관리의 지원 마 소방순찰 및 예방활동의 지원 바 차량운전의 지원 소방관서의 경비  의무소방원은 제1항의 규정에 의한 임무를 수행함에 있어서 소방청장이 정하는 근무수칙을 준수하고 그 임무를 성실히 수행하여야 한다 (Score: 0.7658)
제8조소방활동구역의 출입자 법 제23조제1항에서 대통령령으로 정하는 사람이란 다음 각 호의 사람을 말한다  소방활동구역 안에 있는 소방대상물의 소유자관리자 또는 점유자 전기가스수도통신교통의 업무에 종사하는 사람으로서 원활한 소방활동을 위하여 필요한 사람 의사간호사 그 밖의 구조구급업무에 종사하는 사람 취재인력 등 보도업무에 종사하는 사람 수사업무에 종사하는 사람 그 밖에 소방대장이 소방활동을 위하여 출입을 허가한 사람 (Score: 0.7528)
제24조소방활동 종사 명령  소방본부장 소방서장 또는 소방대장은 화재 재난재해 그 밖의 위급한 상황이 발생한 현장에서 소방활동을 위하여 필요할 때에는 그 관할구역에 사는 사람 또는 그 현장에 있는 사람으로 하여금 사람을 구출하는 일 또는 불을 끄거나 불이 번지지 아니하도록 하는 일을 하게 할 수 있다 이 경우 소방본부장 소방서장 또는 소방대장은 소방활동에 필요한 보호장구를 지급하는 등 안전을 위한 조치를 하여야 한다  삭제  제1항에 따른 명령에 따라 소방활동에 종사한 사람은 시도지사로부터 소방활동의 비용을 지급받을 수 있다 다만 다음 각 호의 어느 하나에 해당하는 사람의 경우에는 그러하지 아니하다 소방대상물에 화재 재난재해 그 밖의 위급한 상황이 발생한 경우 그 관계인 고의 또는 과실로 화재 또는 구조구급 활동이 필요한 상황을 발생시킨 사람 화재 또는 구조구급 현장에서 물건을 가져간 사람 (Score: 0.7480)
제8조민간 소방인력의 운용 등  소방본부장 또는 소방서장은 재난 현장의 활동인력을 확보하기 위하여 필요한 경우 관할 지역의 의용소방대원 퇴직 소방공무원 및 소방 관련 학과 학생을 시기별시간대별로 구분하여 소방대원으로 편성운영할 수 있다  소방본부장 또는 소방서장은 소방관서와 응원출동협정이 체결된 자체소방대위험물안전관리법제19조에 따라 설치된 자체소방대를 말한다를 소방출동대로 편성하여 화재 현장에 출동하게 할 수 있다  제1항 및 제2항에 따라 민간 소방인력을 소방대원으로 운영할 경우 그 인건비 등 운영비용에 관한 사항은 지방자치단체의 조례로 정한다 (Score: 0.7444)
제20조관계인의 소방활동 관계인은 소방대상물에 화재 재난재해 그 밖의 위급한 상황이 발생한 경우에는 소방대가 현장에 도착할 때까지 경보를 울리거나 대피를 유도하는 등의 방법으로 사람을 구출하는 조치 또는 불을 끄거나 불이 번지지 아니하도록 필요한 조치를 하여야 한다 (Score: 0.7394)
```

2.4 최종 모델 선정

TF-IDF + SBERT

- 단어인식만 사용하기에는 문맥적인 유사도가 조금 부족
- 문장인식만 사용하기에는 제시된 데이터의 내용과 조금 동떨어지는 경우가 생김

결론

TF-IDF 로 최대한 관련 단어가 있는 문장을 걸러내고
그 결과물을 SBERT를 이용해 학습시켜서 최대한 관련있는 문장으로 반영하는 것으로 결정

2.5 SBERT 한국어 모델 성능 테스트

sentence-transformers/xlm-r-100langs-bert-base-nli-stsb-mean-tokens

한글이 포함된 다언어 변환 모델

```
Query: 소방 대원 관련

Top 5 most similar sentences in corpus:
제8조소방안전교부세의 대상사업 등  영 제10조의4제2항에 따른 소방안전교부세를 사용할 수 있는 대상사업은 다음 각 호의 사업으로 구분하며 그 세부사항은 행정안전부장관이 정한다  소방 인력 운용 소방공무원 인건비
중점사업 중요하고 시급한 소방시설소방장비를 포함한다 이하 같다 및 안전시설의 확충 소방안전관리 및 안전관리 강화 사업 재량사업 중점사업에 해당하지 아니하는 일반적인 소방시설 및 안전시설의 확충 소방안전관리
및 안전관리 강화 사업  시도지사는 세출예산을 편성할 때에 소방안전교부세를 재원으로 함을 표시하여야 한다 (Score: 0.7173)
제14조한국소방산업기술원의 설립  소방청장은 소방산업의 진흥발전을 효율적으로 지원하기 위하여 한국소방산업기술원이하 기술원이라 한다을 설립할 수 있다  기술원은 법인으로 한다  기술원은 다음 각 호의 사업을
행한다  소방산업의 육성과 소방산업 기술진흥을 위한 정책제도의 조사연구 소방산업의 기반조성 및 창업지원 소방산업 전문인력의 양성 지원 소방산업 발전을 위한 소방장비 보급의 확대와 마케팅 지원 소방산업의 발전
을 위한 국제협력 및 해외진출의 지원 소방사업자의 품질관리능력과 전문성 향상에 필요한 사업 소방장비의 품질 확보 품질 인증 및 신기술신제품에 관한 인증 업무 소방산업에 관한 데이터베이스의 구축운영 출판 기술
강습 및 홍보 소방용 기계기구 소방시설 및 위험물 안전에 관한 조사연구기술개발 및 지원1 위험물안전관리법 제8조제1항 후단에 따른 탱크안전성능시험1 이 법 또는 다른 소방 관계 법령에 규정된 사업으로서 소방청장
시도지사 또는 소방기관의 장이 위탁하거나 대행하게 하는 사업1 그 밖에 기술원의 설립 목적을 달성하는데 필요한 사업  기술원에 관하여 이 법에서 규정한 것을 제외하고는 민법의 재단법인에 관한 규정을 준용한다
소방청장은 기술원의 시설 및 운영에 필요한 경비를 예산의 범위에서 출연하거나 지원할 수 있다 (Score: 0.7161)
제3조119구조대에서 갖추어야 할 장비의 기준  119구조구급에 관한 법률 시행령이하 영이라 한다 제5조에 따른 119구조대이하 구조대라 한다 중 특별시광역시특별자치시도특별자치도이하 시도라 한다 소방본부 및 소방서1
19안전센터를 포함한다에 설치하는 구조대에서 법 제8조제3항에 따라 갖추어야 하는 장비의 기본적인 사항은 소방력 기준에 관한 규칙 및 소방장비관리법 시행규칙에 따른다  소방청에 설치하는 구조대에서 법 제8조제3
항에 따라 갖추어야 하는 장비의 기본적인 사항은 제1항을 준용한다  제1항과 제2항에서 규정한 사항 외에 구조대가 갖추어야 하는 장비에 관하여 필요한 사항은 소방청장이 정한다 (Score: 0.6941)
제7조소방의 날 제정과 운영 등  국민의 안전의식과 화재에 대한 경각심을 높이고 안전문화를 정착시키기 위하여 매년 11월 9일을 소방의 날로 정하여 기념행사를 한다  소방의 날 행사에 관하여 필요한 사항은 소방청장
또는 시도지사가 따로 정하여 시행할 수 있다  소방청장은 다음 각 호에 해당하는 사람을 명예직 소방대원으로 위촉할 수 있다  의사상자 등 예우 및 지원에 관한 법률 제2조에 따른 의사상자로서 같은 법 제3조제3호
또는 제4호에 해당하는 사람 소방행정 발전에 공로가 있다고 인정되는 사람 (Score: 0.6820)
제4조기관장의 책임 제2조에 따른 공공기관의 장이하 기관장이라 한다은 다음 각 호의 사항에 대한 감독책임을 진다 소방시설 피난시설 및 방화시설의 설치·유지 및 관리에 관한 사항 소방계획의 수립·시행에 관한 사항
소방 관련 훈련 및 교육에 관한 사항 그 밖의 소방안전관리 업무에 관한 사항 (Score: 0.6752)
제25조강제처분 등  소방본부장 소방서장 또는 소방대장은 사람을 구출하거나 불이 번지는 것을 막기 위하여 필요할 때에는 화재가 발생하거나 불이 번질 우려가 있는 소방대상물 및 토지를 일시적으로 사용하거나 그 사
용의 제한 또는 소방활동에 필요한 처분을 할 수 있다  소방본부장 소방서장 또는 소방대장은 사람을 구출하거나 불이 번지는 것을 막기 위하여 긴급하다고 인정할 때에는 제1항에 따른 소방대상물 또는 토지 외의 소방
대상물과 토지에 대하여 제1항에 따른 처분을 할 수 있다  소방본부장 소방서장 또는 소방대장은 소방활동을 위하여 긴급하게 출동할 때에는 소방자동차의 통행과 소방활동에 방해가 되는 주차 또는 정차된 차량 및 물건
등을 제거하거나 이동시킬 수 있다  소방본부장 소방서장 또는 소방대장은 제3항에 따른 소방활동에 방해가 되는 주차 또는 정차된 차량의 제거나 이동을 위하여 관할 지방자치단체 등 관련 기관에 견인차량과 인력 등에
대한 지원을 요청할 수 있고 요청을 받은 관련 기관의 장은 정당한 사유가 없으면 이에 협조하여야 한다  시도지사는 제4항에 따라 견인차량과 인력 등을 지원한 자에게 시도의 조례로 정하는 바에 따라 비용을 지급할
수 있다 (Score: 0.6726)
```

연관된 문장이 나오지 않음. 더 조사해 보니 저품질 임베딩 생성으로 더 이상 사용되지 않는 모델로 확인

2.5 SBERT 한국어 모델 성능 테스트

distiluse-base-multilingual-cased-v1

한글이 포함된 다언어 변환 모델

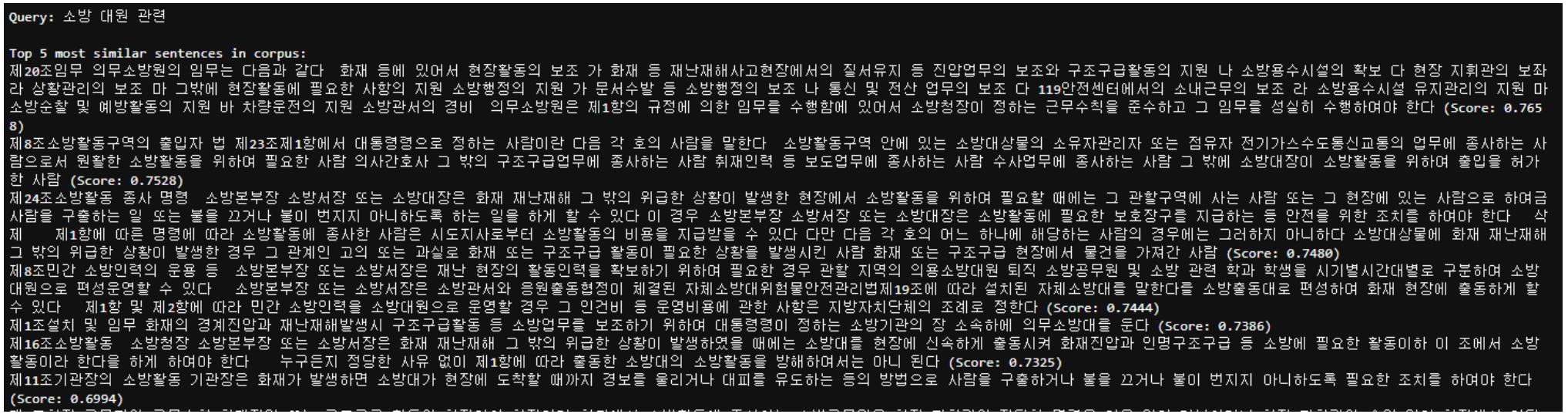
```
Query: 소방 대원 관련

Top 5 most similar sentences in corpus:
제3조직무 소방청은 소방에 관한 사무를 관장한다 (Score: 0.5126)
제15조소방공사감리자의 지정신고 등  법 제17조제2항에 따라 특정소방대상물의 관계인은 공사감리자를 지정한 경우에는 해당 소방시설공사의 착공 전까지 별지 제21호서식의 소방공사감리자 지정신고서에 다음 각 호의 서류전자문서를 포함한다를 첨부하여 소방본부장 또는 소방서장에게 제출해야 한다 다만 전자정부법 제36조제1항에 따른 행정정보의 공동이용을 통하여 첨부서류에 대한 정보를 확인할 수 있는 경우에는 그 확인으로 첨부서류를 갈음할 수 있다  소방공사감리업 등록증 사본 1부 및 등록수첩 사본 1부 해당 소방시설공사를 감리하는 소속 감리원의 감리원 등급을 증명하는 서류전자문서를 포함한다 각 1부 별지 제22호서식의 소방공사감리계획서 1부 법 제21조의3제2항에 따라 체결한 소방시설설계 계약서 사본화재예방 소방시설 설치유지 및 안전관리에 관한 법률 시행규칙 제4조제2항에 따라 건축허가등의 동의요구서에 소방시설설계 계약서가 첨부되지 않았거나 첨부된 서류 중 소방시설설계 계약서가 변경된 경우에만 첨부한다 1부 및 소방공사감리 계약서 사본 1부  특정소방대상물의 관계인은 공사감리자가 변경된 경우에는 법 제17조제2항 후단에 따라 변경일부터 30일 이내에 별지 제23호서식의 소방공사감리자 변경신고서전자문서로 된 소방공사감리자 변경신고서를 포함한다에 제1항 각 호의 서류전자문서를 포함한다를 첨부하여 소방본부장 또는 소방서장에게 제출하여야 한다 다만 전자정부법 제36조제1항에 따른 행정정보의 공동이용을 통하여 첨부서류에 대한 정보를 확인할 수 있는 경우에는 그 확인으로 첨부서류를 갈음할 수 있다  소방본부장 또는 소방서장은 제1항 및 제2항에 따라 공사감리자의 지정신고 또는 변경신고를 받은 경우에는 2일 이내에 처리하고 그 결과를 신고인에게 통보해야 한다 (Score: 0.5029)
제30조시험위원  소방청장은 법 제26조제2항에 따라 관리사시험의 출제 및 채점을 위하여 다음 각 호의 어느 하나에 해당하는 사람 중에서 시험위원을 임명하거나 위촉하여야 한다  소방 관련 분야의 박사학위를 가진 사람 대학에서 소방안전 관련 학과 조교수 이상으로 2년 이상 재직한 사람 소방위 이상의 소방공무원 소방시설관리사 소방기술사  제1항에 따른 시험위원의 수는 다음 각 호의 구분에 따른다  출제위원 시험 과목별 3명 채점위원 시험 과목별 5명 이내제2차시험의 경우로 한정한다  제1항에 따라 시험위원으로 임명되거나 위촉된 사람은 소방청장이 정하는 시험문제 등의 출제 시 유의사항 및 서약서 등에 따른 준수사항을 성실히 이행하여야 한다  제1항에 따라 임명되거나 위촉된 시험위원과 시험감독 업무에 종사하는 사람에게는 예산의 범위에서 수당과 여비를 지급할 수 있다 (Score: 0.4894)
제26조피난 명령  소방본부장 소방서장 또는 소방대장은 화재 재난재해 그 밖의 위급한 상황이 발생하여 사람의 생명을 위험하게 할 것으로 인정할 때에는 일정한 구역을 지정하여 그 구역에 있는 사람에게 그 구역 밖으로 피난할 것을 명할 수 있다  소방본부장 소방서장 또는 소방대장은 제1항에 따른 명령을 할 때 필요하면 관할 경찰서장 또는 자치경찰단장에게 협조를 요청할 수 있다 (Score: 0.4856)
제21조공무원의 파견 소방병원은 그 업무를 수행하기 위하여 필요하면 소방청장을 거쳐 관계 행정기관의 장에게 그 소속 공무원의 파견을 요청할 수 있다 (Score: 0.4812)
제5조소방사업자의 신고  영 제20조제2항에 따른 소방사업자의 신고서는 별지 제7호서식에 따른다  소방청장은 소방사업자 신고를 받았을 때에는 별지 제8호서식에 따른 신고확인서를 발급하고 별지 제9호서식에 따른 신고대장에 필요한 사항을 기록하여야 한다  영 제20조제1항에 따라 소방사업의 세부 분야와 그 분야별 신고에 필요한 사항은 소방청장이 정하여 고시한다 (Score: 0.4750)
제8조소방시설업자가 보관하여야 하는 관계 서류 법 제8조제4항에서 행정안전부령으로 정하는 관계 서류란 다음 각 호의 구분에 따른 해당 서류전자문서를 포함한다를 말한다  소방시설설계업 별지 제10호서식의 소방시설설계기록부 및 소방시설 설계도서 소방시설공사업 별지 제11호서식의 소방시설공사 기록부 소방공사감리업 별지 제12호서식의 소방공사 감리기록부 별지 제13호서식의 소방공사 감리일지 및 소방시설의 완공 당시 설계도서 (Score: 0.4736)
```

소방청의 역할에 관련한 법률이 주로 나옴

2.5 SBERT 한국어 모델 성능 테스트

jhgan/ko-sroberta-multitask

한국어 중심의 문장 변환 모델

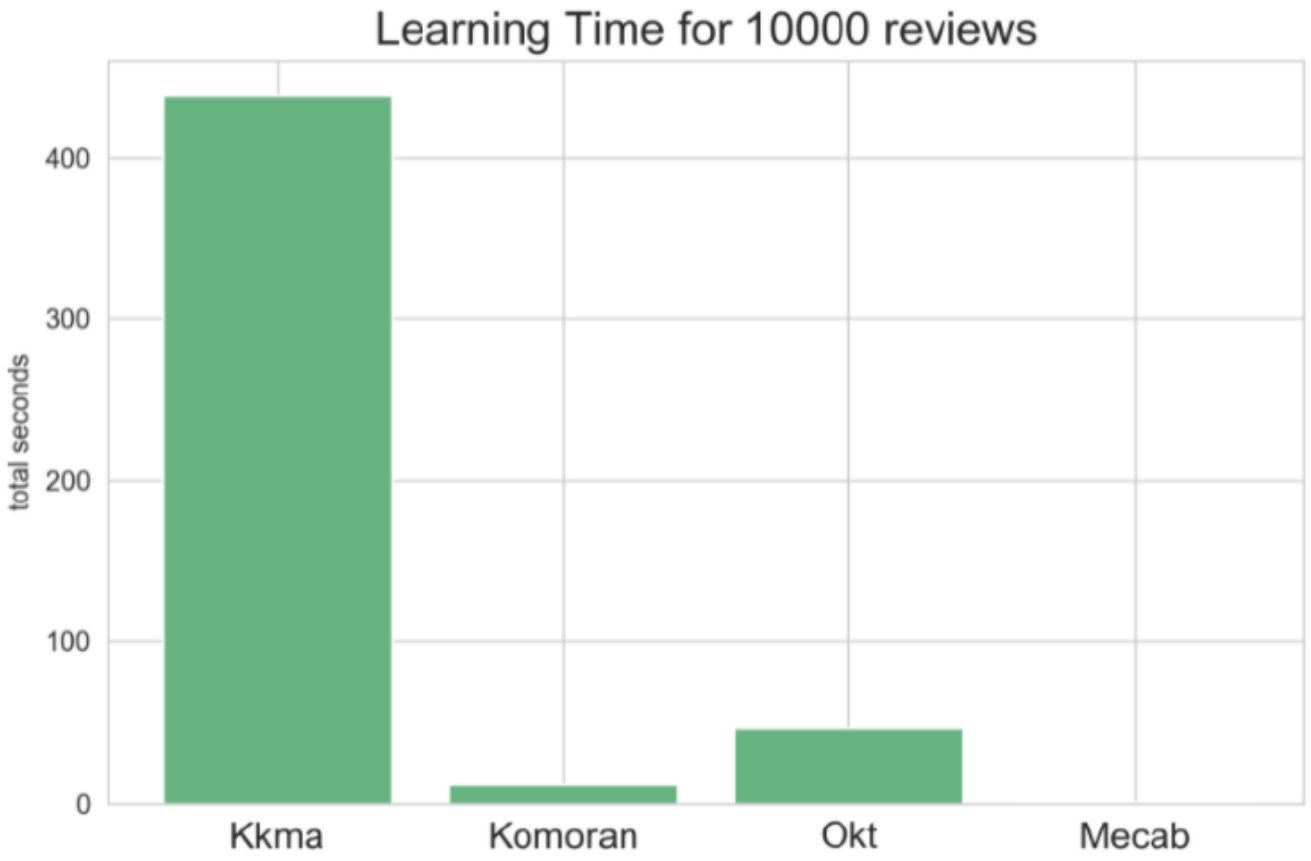


소방 대원이 어떤 일을 하는 사람인지에 대한 법률 조문이 유사도 0.7658로 나왔고
관련 조문 역시 소방대원과 관련된 내용이 주였다.

=> 이 모델을 사용하기로 결정

2.6 형태소 분석기 비교 및 선정

형태소 분석기 속도 비교



Mecab(4.77ms)> Komoran(27.6ms) > Okt(340ms) > Kkma(1.33s)

2.6 모델 및 형태소 분석기 비교 및 선정

Mecab, Okt, Komoran 성능 비교

Mecab

```
%%time
print(mecab.pos(text))
```

```
[('정의', 'NNG'), ('이', 'MM'), ('법', 'NNG'),
CPU times: user 4.18 ms, sys: 964 µs, total: 5
Wall time: 4.77 ms
```

Komoran

```
%%time
print(kom.pos(text))
```

```
[('정', 'NNP'), ('의', 'JKG'), ('이', 'MM'),
CPU times: user 43.2 ms, sys: 69 µs, total:
Wall time: 27.6 ms
```

Okt

```
%%time
print(okt.pos(text))
```

```
[('정의', 'Noun'), ('이', 'Noun'), ('법', 'Noun'),
CPU times: user 630 ms, sys: 7.7 ms, total: 638 ms
Wall time: 340 ms
```

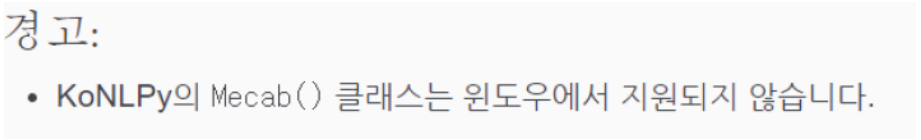
코모란은 단어를 너무 세부적으로 쪼개 내어 원하는 형태와는 맞지 않는다고 판단

2.6 모델 및 형태소 분석기 비교 및 선정

Okt, Mecab 편의성 비교

Okt
- Konlpy를 이용하여 사용가능
- 이용자 사전을 추가가 용이함 : txt 파일에 단어만 추가

Mecab
- 따로 설치 필요
- 이용자 사전을 추가 까다로움 : 넣으려는 데이터의 품사까지 지정해야 함



결론
프로젝트의 진행사항을 봤을 때 좀 더 사용하기 용이한 **Okt를 사용하기로 결정**

Part 3

모델 구축 및 학습

1.라이브러리 및 데이터 로드
2.TF-IDF 단어 학습 및 저장
3.1차 문장 추출
4.SBERT 학습
5.결과 반영

3.1 라이브러리 및 데이터 로드

라이브러리

```
import pandas as pd
import json
import re
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.metrics.pairwise import linear_kernel
import pickle
from konlpy.tag import Okt
from sentence_transformers import SentenceTransformer, util
import numpy as np
okt=Okt() #형태소 분해 객체 생성
```

Data load

```
file_path = r'D:\Study\Python\Jupyter\law_json_data\clean_laws_jo_total.json'
with open(file_path, 'r', encoding="UTF-8") as jsonfile:
    json_data = json.load(jsonfile)
```

```
laws_data_list = json_data["laws"]
len(laws_data_list)
```

```
145279
```

3.2 TF-IDF 단어 학습 및 저장

TF-IDF 단어 학습

```
def tokenizer(raw, pos = ['Noun', 'Verb', 'Adjective']):
    okt = Okt()
    # 길이가 1 이하인 토큰은 제외, 위에서 지정한 (Okt 사전에 따른) 토큰들만 특징으로 삼기
    return [word for word, tag in okt.pos(raw) if len(word) > 1 and tag in pos]
```

```
tfidfvectoriser=TfidfVectorizer(tokenizer= tokenizer, ngram_range=(1,2), min_df=2, max_features = 20000)
```

```
tfidf = tfidfvectoriser.fit(laws_data_list) # 벡터라이저가 단어들을 학습
```

```
C:\Users\admin\anaconda3\envs\tfenv\lib\site-packages\sklearn\feature_extraction\text.py:516: UserWarning: The parameter
'token_pattern' will not be used since 'tokenizer' is not None'
  warnings.warn(
```

모델 학습 및 모델 저장

```
#학습한 TF-IDF 저장
pickle.dump(tfidf, open("tfidf_fit.pickle", "wb"))
```

```
# TF-IDF 불러오기
with open("tfidf_fit.pickle", 'rb') as f:
    load_tfidf_fit = pickle.load(f)
```

3.2 TF-IDF 단어 학습 및 저장

단어 학습 확인

```
load_tfidf_fit.vocabulary_ # 벡터라이저가 학습한 단어사전을 출력
```

```
{'관련': 2041,
 '하여': 18013,
 '피해': 17531,
 '입은': 13141,
 '명예': 5906,
 '회복': 19825,
 '시켜': 10041,
 '인권': 12831,
 '신장': 10293,
 '국민': 2669,
 '이바지': 12517,
 '목적': 5961,
 '한다': 18446,
 '관련 하여': 2091,
 '피해 입은': 17540,
 '명예 회복': 5908,
 '이바지 목적': 12518,
 '목적 한다': 5980,
 '사용': 8116,
 '하는': 17660,
 '용어': 11587,
 '정의': 14519,
 '다음': 4008,
 '같다': 277,
 '계엄': 1382,
 '사령부': 7853,
 '합동': 18821,
 '수사': 9594,
```

3.2 TF-IDF 단어 학습 및 저장

Transform

```
tfidf_vectors=load_tfidf_fit.transform(laws_data_list)
print(tfidf_vectors)
```

```
  (0, 19825)    0.25288694964574965
  (0, 18446)    0.049079615037825666
  (0, 18013)    0.0672464710950464
  (0, 17540)    0.32549827566182826
  (0, 17531)    0.2079158148691132
  (0, 13141)    0.2650892898155746
  (0, 12831)    0.2795052776297303
  (0, 12518)    0.25550132649681034
  (0, 12517)    0.24206215057113548
  (0, 10293)    0.3424528883912471
  (0, 10041)    0.28061404000166396
  (0, 5980)     0.17842200460512908
```

저장 후 feature 추출

```
#TF-IDF transform저장
pickle.dump(tfidf_vectors, open("tfidf_vectors.pickle", "wb"))
```

```
# TF-IDF transform불러오기
with open("tfidf_vectors.pickle", 'rb') as f:
    load_tfidf_vectors = pickle.load(f)
```

```
features = load_tfidf_fit.get_feature_names()
```

3.3 1차 문장 추출

Input Text Tokenize

```
input_text_data = '소방 대원 관련'
```

```
srch = [ t for t in tokenizer(input_text_data) if t in features]
srch
```

```
['소방', '대원', '관련']
```

```
srch_vector = load_tfidf_fit.transform([input_text_data])
```

```
cosine_similar =  linear_kernel(srch_vector, load_tfidf_vectors).flatten()
```

코사인 유사도를 이용해 1차 유사 문장 추출

```
sim_rank_idx = cosine_similar.argsort()[::-1]
```

```
#tfidf result
tf_idf_result_index = [] #실 데이터 인덱스
tf_idf_sentences = []# 결과 조문
for i in sim_rank_idx:
    if cosine_similar[i] > 0.13:
        print('{} /score : {}'.format(laws_data_list[i],cosine_similar[i]))
        tf_idf_result_index.append(i)
        tf_idf_sentences.append(laws_data_list[i])
```

3.3 1차 문장 추출

1차 문장 추출 결과

제4조청장 청장은 소방총감으로 보한다 /score : 0.4990716556881685
제3조계급 구분 소방공무원의 계급은 다음과 같이 구분한다 소방총감 소방정감 소방감 소방준감 소방정 소방령 소방경 소방위 소방장 소방교 소방사 /score : 0.4713956894485098
제8조소방력의 기준 등 소방기관이 소방업무를 수행하는 데에 필요한 인력과 장비 등이하 소방력이라 한다에 관한 기준은 행정안전부령으로 정한다 시도지사는 제1항에 따른 소방력의 기준에 따라 관할구역의 소방력을 확충하기 위하여 필요한 계획을 수립하여 시행하여야 한다 소방자동차 등 소방장비의 분류표준화와 그 관리 등에 필요한 사항은 따로 법률에서 정한다 /score : 0.4355005993421697
제2조정의 이 규칙에서 사용하는 용어의 뜻은 다음과 같다 소방기관이란 소방장비 인력 등을 동원하여 소방업무를 수행하는 소방서119안전센터119구조대119구급대119구조구급센터119항공대소방정대119지역대119종합상황실소방체험관을 말한다 소방장비란 소방장비관리법 제2조제1호에 따른 소방장비를 말한다 /score : 0.4183677172250224
제6조소방정책사업비계정의 세입과 세출 소방정책사업비계정의 세입은 다음 각 호와 같다 소방안전교부세 중 제5조제1항제1호를 제외한 금액에서 소방분야에 해당하는 금액 제7조에 따른 시도 일반회계로부터의 전입금 지역자원시설세 중 대통령령으로 정하는 금액 소방사무 관련 국고보조금과 다른 특별회계 및 기금으로부터의 전입금 소방사무 관련 법령 및 조례 위반자로부터 징수한 과태료 과징금 및 이행강제금 소방사무 관련 법령 및 조례 이행에 따른 각종 수수료 수입 그 밖의 수입금 소방정책사업비계정의 세출은 다음 각 호와 같다 소방사무 수행에 필요한 경비 소방시설 확충에 필요한 경비 그 밖에 소방특별회계의 설치 목적에 부합하여 시도지사가 필요하다고 인정하는 사업 관련 경비 /score : 0.406530335468857
제9조인증대상 소방장비 법 제12조제1항에서 대통령령으로 정하는 소방장비란 다음 각 호의 소방장비를 말한다 소방펌프차 소방고가차 방화복 그 밖에 소방청장이 정하여 고시하는 소방장비 /score : 0.4028055136973048
제5조차장 차장은 소방정감으로 보한다 /score : 0.3953615490194425
제36조복제 등 민방위 대원은 교육훈련 중이나 임무 수행 중에는 행정안전부령으로 정하는 민방위 대원 복장을 착용하거나 표지장識을 달 수 있다 /score : 0.3953167296168951
제2조정의 이 법에서 사용하는 용어의 뜻은 다음과 같다 소방장비란 소방업무를 효과적으로 수행하기 위하여 필요한 기동장비화재진압장비구조장비구급장비보호장비정보통신장비측정장비 및 보조장비를 말한다 소방업무란 소방기본법 제3조제1항에 따른 업무를 말한다 소방기관이란 중앙소방학교중앙119구조본부소방본부소방서지방소방학교119안전센터119구조대119구급대119구조구급센터항공구조구급대소방정대119지역대 및 소방체험관 등 소방업무를 수행하는 기관을 말한다 관리란 소방장비의 안전성을 확보하고 효율적으로 활용하기 위하여 소방장비의 구매를 위한 기획에서부터 불용의 결정과 폐기양도까지 전 주기에 걸쳐 언제든지 본래의 성능을 발휘하도록 하는 점검정비 및 그 밖의 모든 행위를 말한다 운용이란 소방장비를 그 기능 및 목적에 맞도록 안전하게 사용하는 것을 말한다 내용연수年란 소방장비의 운용에 지장이 없는 상태에서 소방업무를 원활하게 수행할 수 있을 것으로 예측한 소방장비의 경제적 사용연수를 말한다 소방장비운용자란 소방장비를 직접 운용하는 소방공무원 의무소방원 및 의용소방대원을 말한다 /score : 0.39377903613824683

3.4 SBERT 학습

TF-IDF 결과에 해당하는 문장 토대로 SBERT에 학습

```
sbert_result_index = []
embedder = SentenceTransformer("jhgan/ko-sroberta-multitask")

# TF-IDF 결과 조문을 Corpus로
corpus = tf_idf_sentences
corpus_embeddings = embedder.encode(corpus, convert_to_tensor=True)

q_list = []
q_list.append(input_text_data)
queries = q_list

top_k = 20
for query in queries:
    query_embedding = embedder.encode(query, convert_to_tensor=True)
    cos_scores = util.pytorch_cos_sim(query_embedding, corpus_embeddings)[0]
    cos_scores = cos_scores.cpu()

    top_results = np.argpartition(-cos_scores, range(top_k))[0:top_k]

    print("\n\n======================\n\n")
    print("Query:", query)
    print("\nTop 5 most similar sentences in corpus:")

    for idx in top_results[0:top_k]:
        print(corpus[idx].strip(), "(Score: %.4f)" % (cos_scores[idx]))
        sbert_result_index.append(tf_idf_result_index[idx])
```

3.5 결과 반영

결과를 실제 법령 데이터와 연결 – 법령명 목록 반영

```
result_law_dic = {} #EX {"우체국보험특별회계법 시행규칙":[123513,12345,234,123]}
for i in sbert_result_index:
    index_num = int(i)
    #print(laws_data_list_cut[index_num])
    data_detail = tatal_laws_data.iloc[[index_num],:]
    # print(tatal_laws_data.iloc[index_num]["법령명"])

    law_name = tatal_laws_data.iloc[index_num]["법령명"]

    print('법령명 : ',tatal_laws_data.iloc[index_num]["법령명"])
    if law_name in result_law_dic:
        value_list = result_law_dic[law_name]
        value_list.append(index_num)

    else:
        result_law_dic[law_name] = [index_num]
```

```
법령명 :  의무소방대설치법 시행령
법령명 :  소방기본법 시행령
법령명 :  소방기본법
법령명 :  소방력 기준에 관한 규칙
법령명 :  의무소방대설치법
법령명 :  소방기본법
법령명 :  공공기관의 소방안전관리에 관한 규정
법령명 :  소방공무원 복무규정
법령명 :  소방기본법
법령명 :  소방공무원 복무규정
법령명 :  화재예방, 소방시설 설치·유지 및 안전관리에 관한 법률
법령명 :  소방청과 그 소속기관 직제
법령명 :  119구조·구급에 관한 법률 시행령
법령명 :  화재예방, 소방시설 설치·유지 및 안전관리에 관한 법률
법령명 :  소방기본법
법령명 :  소방청과 그 소속기관 직제
법령명 :  소방력 기준에 관한 규칙
법령명 :  소방장비관리법
법령명 :  소방공무원 보건안전 및 복지 기본법
법령명 :  소방력 기준에 관한 규칙
```

3.5 결과반영

결과를 실제 법령 데이터와 연결 – 해당 법령명 법조문 반영

```
choice_law_name_value = result_law_dic['의무소방대설치법 시행령']

for i in choice_law_name_value:

    #print(laws_data_list_cut[index_num])
    data_detail = tatal_laws_data.iloc[[i],:]
    print('법령명 : ',tatal_laws_data.iloc[i]["법령명"])
    print('법령MST : ',tatal_laws_data.iloc[i]["법령MST"])
    print('법령ID : ', tatal_laws_data.iloc[i]["법령ID"])
    print('시행일자 : ', tatal_laws_data.iloc[i]["시행일자"])
    print('공포번호 : ', tatal_laws_data.iloc[i]["공포번호"])
    print('법령구분명 : ', tatal_laws_data.iloc[i]["법령구분명"])
    print('조문번호 : ', tatal_laws_data.iloc[i]["조문번호"])
    print('조문 : ', tatal_laws_data.iloc[i]["조문내용"])
    print("")
```

```
법령명 :  의무소방대설치법 시행령
법령MST :  222757
법령ID :  9250
시행일자 :  2021.2.11
공포번호 :  제31147호
법령구분명 :  대통령령
조문번호 :  20
조문 :  제20조(임무)
 ①의무소방원의 임무는 다음과 같다. <개정 2003.5.29, 2006.6.30>
                                1. 화재 등에 있어서 현장활동의 보조 가. 화재 등 재난·재해사고현장에서의 질서유지 등 진압업무의 보조와 구조·구급활동의 지원
                                        나. 소방용수시설의 확보
                                        다. 현장 지휘관의 보좌
                                        라. 상황관리의 보조
                                        마. 그밖에 현장활동에 필요한 사항의 지원
                                       2. 소방행정의 지원 가. 문서수발 등 소방행정의 보조
                                        나. 통신 및 전산 업무의 보조
                                        다. 119안전센터에서의 소내근무의 보조
                                        라. 소방용수시설 유지관리의 지원
                                        마. 소방순찰 및 예방활동의 지원
                                        바. 차량운전의 지원
                                       3. 소방관서의 경비
                                         ②의무소방원은 제1항의 규정에 의한 임무를 수행함에 있어서 소방청장이 정하는 근무수칙을 준수하고 그 임무를 성실히 수행
하여야 한다. <개정 2004.5.24, 2014.11.19, 2017.7.26>
```

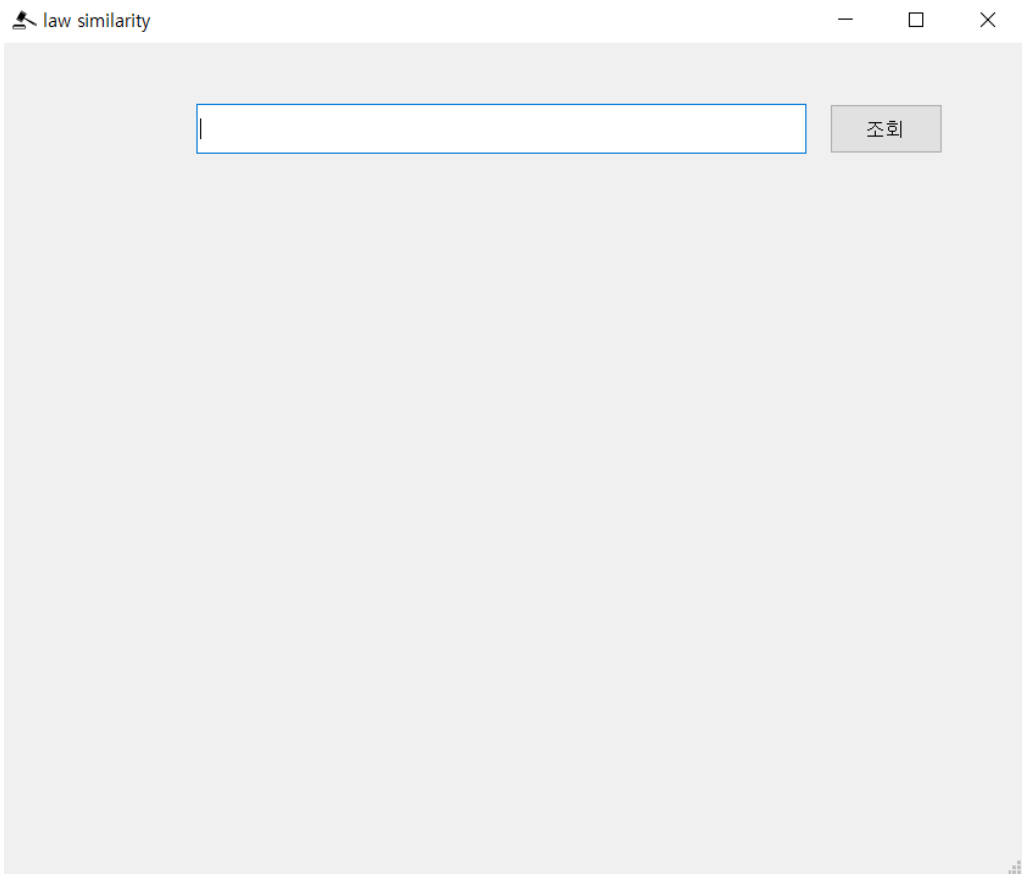
Part 4

시각화 및 결과

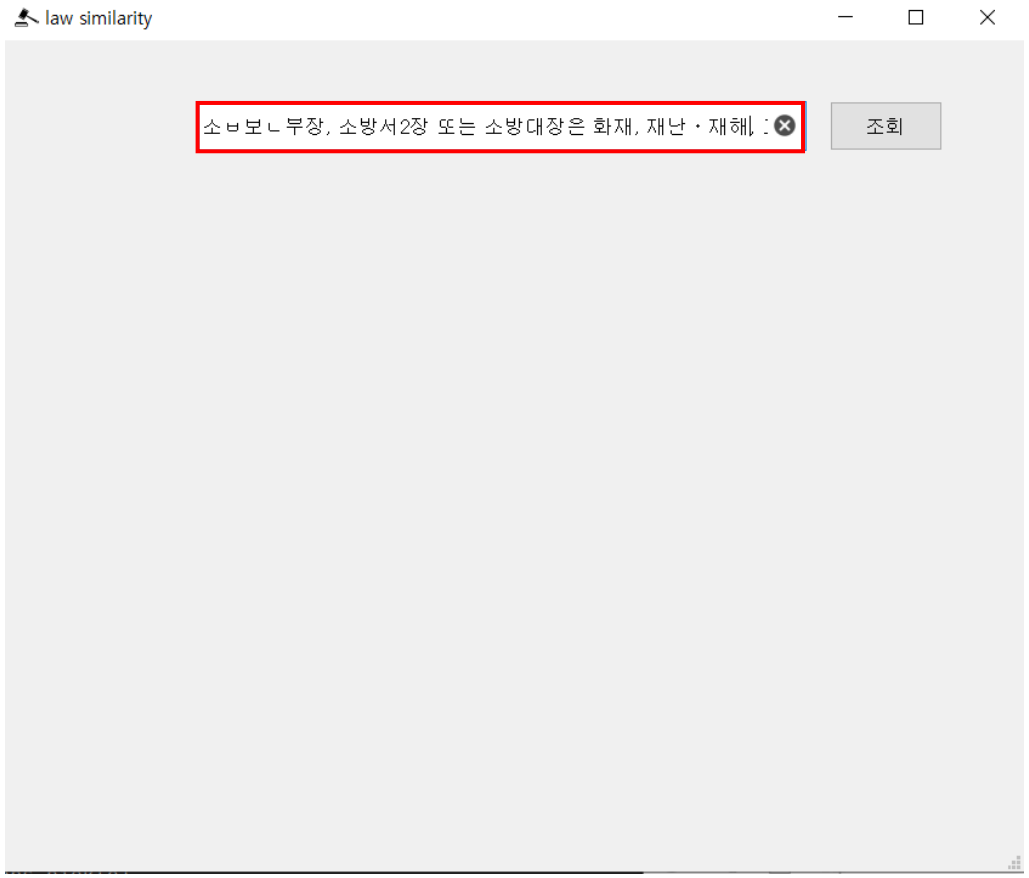
1.PyQt 시각화 및 결과

4.1 PyQt 시각화 및 결과

첫 화면

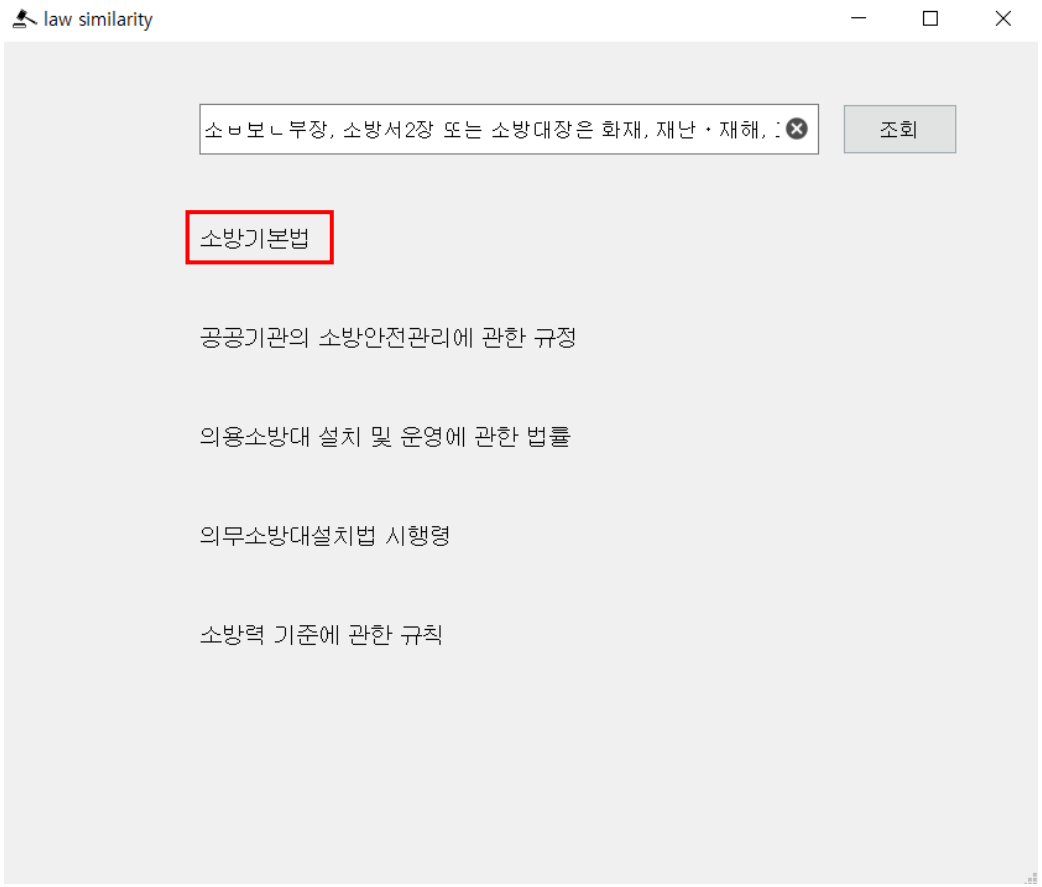


검색 정보 입력
오타가 들어가는 경우도 확인

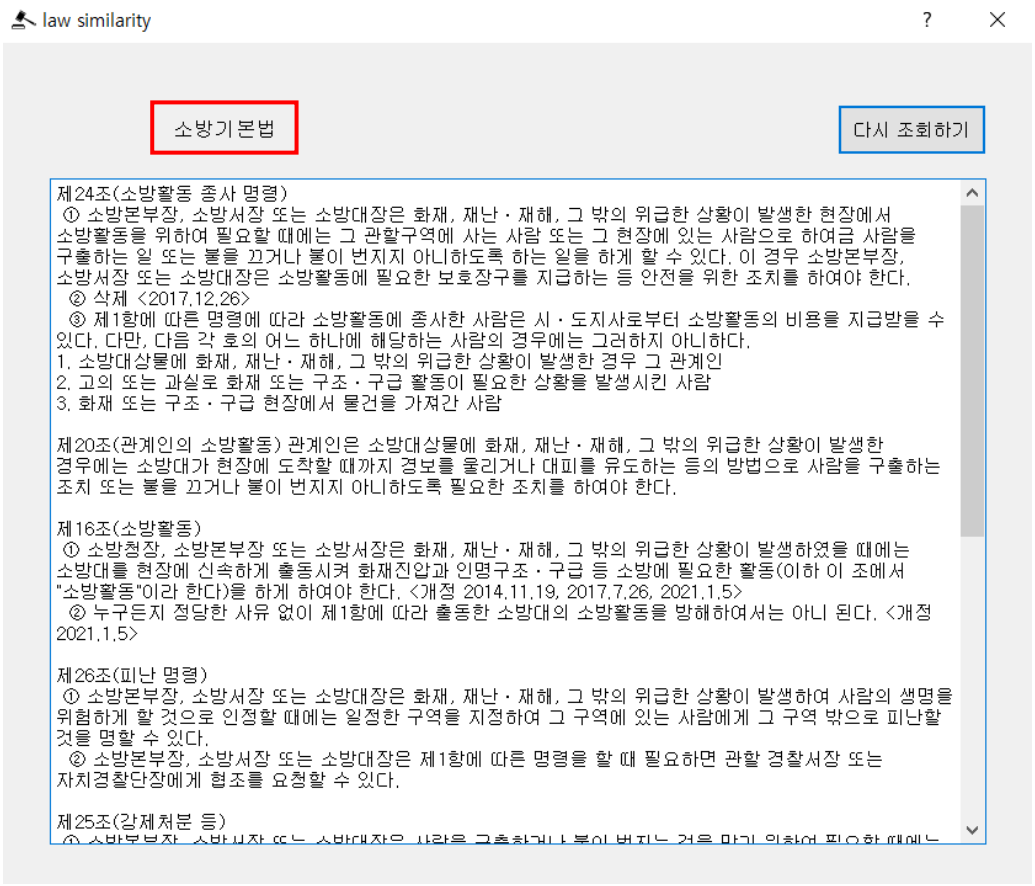


4.1 PyQt 시각화 및 결과

검색 정보에 관련된 법령명 목록 출력

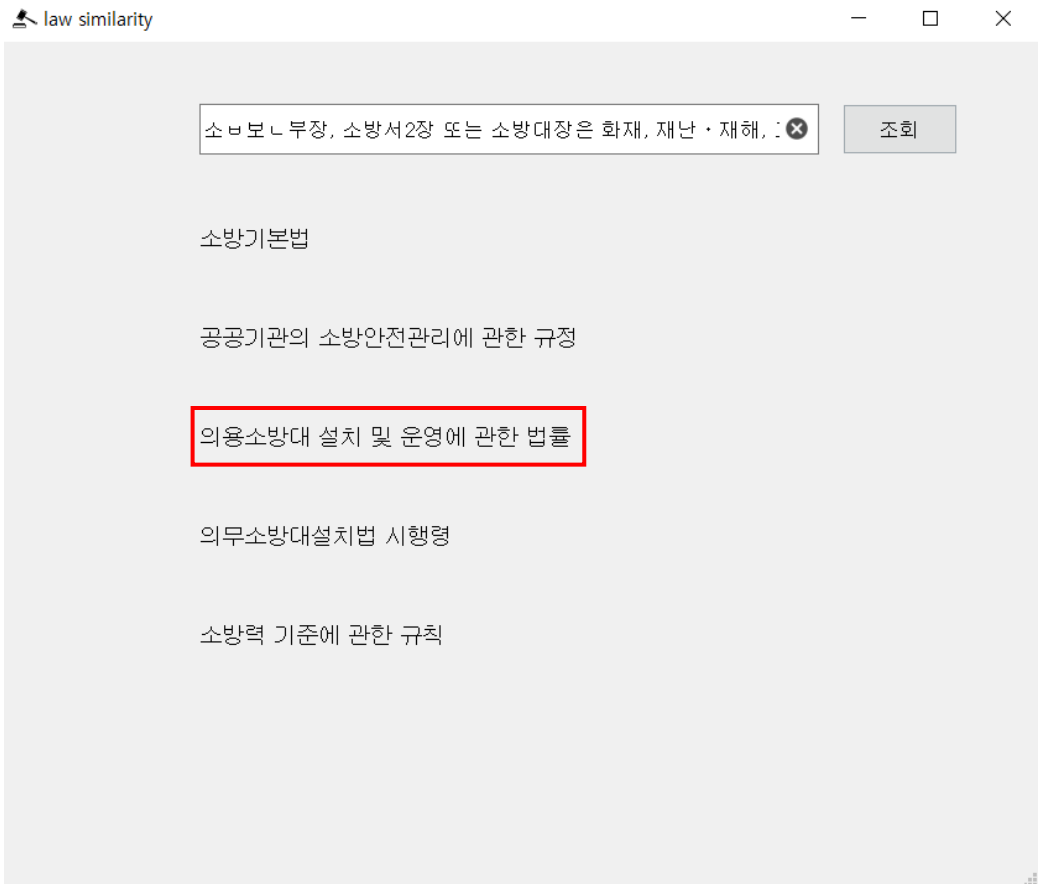


법령명 클릭 -> 관련 조문 출력

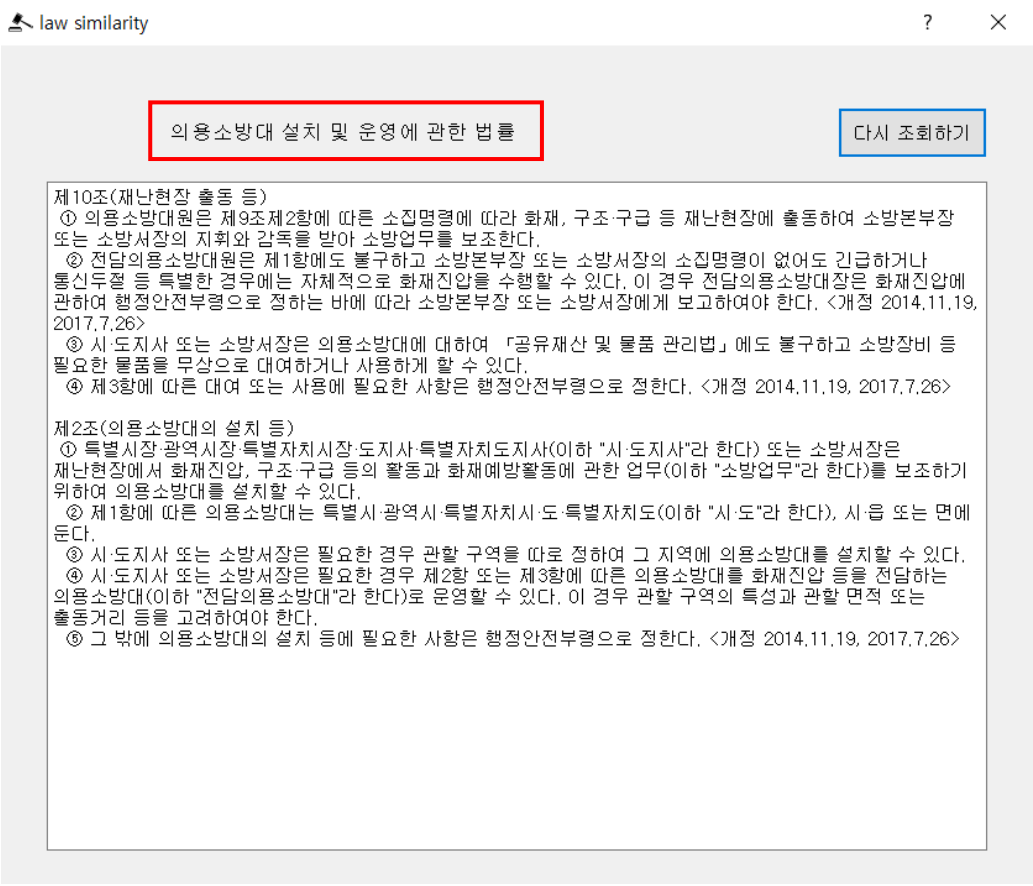


4.1 PyQt 시각화 및 결과

검색 정보에 관련된 법령명 목록 출력



법령명 클릭 -> 관련 조문 출력



Part 5

개선사항 및 소감

1.개선 사항
2.소감

5.1 개선 사항

TF-IDF 학습 시 형태소 분석기 선정 부분

시간관계상 적용하기 힘들다고 판단한 Mecab을 사용해 보지 못한 점
프로젝트 마무리 후에 간단한 테스트를 시도 => Okt 보다 확실히 빠른 속도를 자랑함

Okt를 이용한 학습 시간 : 4시간 이상
Mecab을 이용한 학습 시간 : 약 2시간

SBERT 한국어 모델 선정, 학습 부분

여러 한국어 모델을 테스트 해보지 못한 점
법률에 맞은 파인튜닝을 해보지 못한 점

PyQt 구현 부분

구현을 프로젝트 기간 후반에 정해서 구현에 시간이 부족했던 점
예외 처리 부족한 점

5.2 소감

김찬희

자연어처리를 두번째 하기 때문에 어느정도 자신이 있었지만 깊이 파고 들어가보니 논리적으로 이유를 찾아야 하는 복잡한 분야라는 생각이 많이 들었습니다. 하지만 그만큼 결과를 추출하는 흥미로웠으며 앞으로도 관련분야를 공부하고 싶다는 생각이 많이 들었습니다. 모두 고생하셨습니다!

박유정

모델을 여러 개 적용해 보면서 어떤 것이 더 나은지 명확한 기준이 없어서 어려웠지만, 직접 구현하고 비교해 보면서 방향을 찾아갈 수 있었습니다. 이 과정에서 해보기 전까지는 모르는 것이며 정답이란 없다는 걸 깨달았습니다. 무엇보다 자연어는 성능 개선을 위해 전처리가 중요하다는 것 도요. 짧은 기간 동안 팀원들과 고생하며 핵심적으로 많은 것들을 배울 수 있어서 좋았습니다

정새하

처음 시작할 때 갈피를 잡지 못해 힘들었지만 자연어 처리에 대한 시각을 넓힐 수 있는 기회여서 좋았습니다. 여러 모델을 비교해 보면서 내가 이걸 왜 써야하는 지를 생각하며 스스로 방향을 잡아 나갈 수 있는 능력이 많이 필요하다는 걸 느꼈습니다. 팀원들 다들 고생 많으셨습니다.

정한슬

생소한 법령 데이터를 이용한 NLP 프로젝트여서 어려웠으나 자연어 처리는 실생활뿐만 아니라 전문적인 영역에서도 필요로 하는 것을 알게 되었습니다. 또한 새로운 시각화 방법인 PyQt에 대해 알게 되고 공부할 수 있어서 좋았습니다.

Part 6

참고 자료

1.참고 논문
2.데이터 수집 및 전처리 관련
3.모델 구축 관련
4.시각화 및 기타

6.1 참고 논문

인공지능 기법을 활용한 법률안 예측 모델 연구
http://journal.dcs.or.kr/xml/25182/25182.pdf

법률정보시스템을 위한 텍스트마이닝 적용방안
https://www.koreascience.or.kr/article/JAKO202012854885152.pdf

딥러닝 알고리즘을 이용한 유사 판례 매칭 데이터셋 구축 방안 연구
https://www.koreascience.or.kr/article/CFKO202130060864862.pdf

KorNLI and KorSTS: New Benchmark Datasets for Korean Natural Language Understanding
https://arxiv.org/abs/2004.03289

[논문 리뷰] KorNLI and KorSTS: New Benchmark Datasets for Korean Natural Language Understanding
https://misconstructed.tistory.com/28

Universal Sentence Encoder(2018)
https://dodonam.tistory.com/204

법률정보시스템을 위한 텍스트 마이닝 적용 방안
https://www.koreascience.or.kr/article/JAKO202012854885152.pdf

KorNLI and KorSTS: New Benchmark Datasets for Korean Natural Language Understanding
https://arxiv.org/abs/2004.03289

6.2 데이터 수집 전처리 관련

데이터

국가법령정보 공동활용 사이트
https://open.law.go.kr/LSO/main.do

AI 허브
https://aihub.or.kr/aihub-data/natural-language/about

크롤러

[빅데이터] 웹 크롤링 : BeautifulSoup(1) find, xml 파싱, 태그 속성값 크롤링
https://heannim-world.tistory.com/43

Python에서 XML 파서 만들기
https://www.delftstack.com/ko/howto/python/python-xml-parser/

파이썬 xml 태그값 가져오기 및 파싱
https://lee-mandu.tistory.com/519?category=838684

네이버 영화 리뷰 키워드분석 (4) 전처리 시작
https://haystar.tistory.com/11?category=962597

6.3 모델 구축 관련

[Python] 한국어 형태소 분석기 체험 및 비교(Okt, Mecab, Komoran, Kkma) –
https://soohee410.github.io/compare_tagger

Python #eunjeon, mecab 모듈 설치
https://sputnik-kr.tistory.com/173?category=997532

추천시스템 분석 입문하기(유사도 학습)
https://www.youtube.com/watch?v=g2-z0saMteA&list=PL9mhQYIlKEhdkOVTZWJJIy8rv6rQaZNNc&index=2

사이킷런 TFIDF 와 코사인유사도 로 문서 유사도 구하기
https://sikaleo.tistory.com/m/62

딥러닝으로 동네생활 게시글 필터링하기
https://medium.com/daangn/%EB%94%A5%EB%9F%AC%EB%8B%9D%EC%9C%BC%EB%A1%9C-%EB%8F%99%EB%84%A4%EC%83%9D%ED%99%9C-%EA%B2%8C%EC%8B%9C%EA%B8%80-%ED%95%84%ED%84%B0%EB%A7%81%ED%95%98%EA%B8%B0-263cfe4bc58d

Hugging face : AI 커뮤니티 SBERT Model Search
https://huggingface.co/

6.3 모델 구축 관련

딥 러닝을 이용한 자연어 처리 입문 BERT/SBERT
https://wikidocs.net/156176

문서 벡터를 이용한 추천 시스템
https://wikidocs.net/102705

Word2Vector using Gensim
https://medium.com/analytics-vidhya/word2vector-using-gensim-e055d35f1cb4

Calculating Document Similarities using BERT, word2vec, and other models
https://towardsdatascience.com/calculating-document-similarities-using-bert-and-other-models-b2c1a29c9630

doc2Vec Example
https://reliable-poultry-5ba.notion.site/doc2Vec-39f290edb3484ef1b7bdc8dbf9012b7e

네이버 영화 리뷰 키워드 분석
https://haystar.tistory.com/11?category=962597

6.3 모델 구축 관련

NLP - 11. 워드투벡터(Word2Vec)
https://bkshin.tistory.com/entry/NLP-11-Word2Vec

딥 러닝을 이용한 자연어 처리 입문 TF-IDF
https://wikidocs.net/31698

[python]Doc2Vec -3) Doc2vec tokenizing
https://m.blog.naver.com/gpdlswkd17/221494617376

[elasticsearch] doc2vec으로 korquad 데이터 유사도 분석하기 : 문단 수준의 문서를 검색
https://skagh.tistory.com/32

Pretrained Models — Sentence-Transformers documentation
https://www.sbert.net/docs/pretrained_models.html

Word2Vec 과 Doc2Vec
https://dailyheumsi.tistory.com/165

GENSIM
https://radimrehurek.com/gensim/models/doc2vec.html

6.3 모델 구축 관련

Beauty Domain-Specific Pre-trained Language Model 개발하기
http://blog.hwahae.co.kr/all/tech/tech-tech/5876/

[PYTORCH / HUGGINGFACE] CUSTOM DATASET으로 BERTTOKENIZER 학습하기
https://cryptosalamander.tistory.com/139

버트(BERT) 파인튜닝 간단하게 해보자.
http://freesearch.pe.kr/archives/4963

[Basic NLP] sentence-transformers 라이브러리를 활용한 SBERT 학습 방법
https://velog.io/@jaehyeong/Basic-NLP-sentence-transformers-%EB%9D%BC%EC%9D%B4%EB%B8%8C%EB%9F%AC%EB%A6%AC%EB%A5%BC-%ED%99%9C%EC%9A%A9%ED%95%9C-SBERT-%ED%95%99%EC%8A%B5-%EB%B0%A9%EB%B2%95

Kopora
https://ko-nlp.github.io/Korpora/ko-docs/

6.4 시각화 및 기타

시각화

구글 번역기 프로그램

https://codetorial.net/pyqt5/examples/translator.html

PyQt

https://ehclub.net/category/Python/QT

기타

Paperswithcode : 코드와 논문을 matching해서 볼 수 있는 사이트

https://paperswithcode.com/

Slack 사용법

https://infrajp.tistory.com/1?category=805802

정규표현식 의미를 확인 할 수 있는 사이트

https://regexr.com/

순서도 작성

https://app.diagrams.net/

THANK YOU!