



Lawyer ChatBot

박희진 개인 PROJECT





목차

01

주제 선정 배경

02

데이터 수집 및 전처리

03

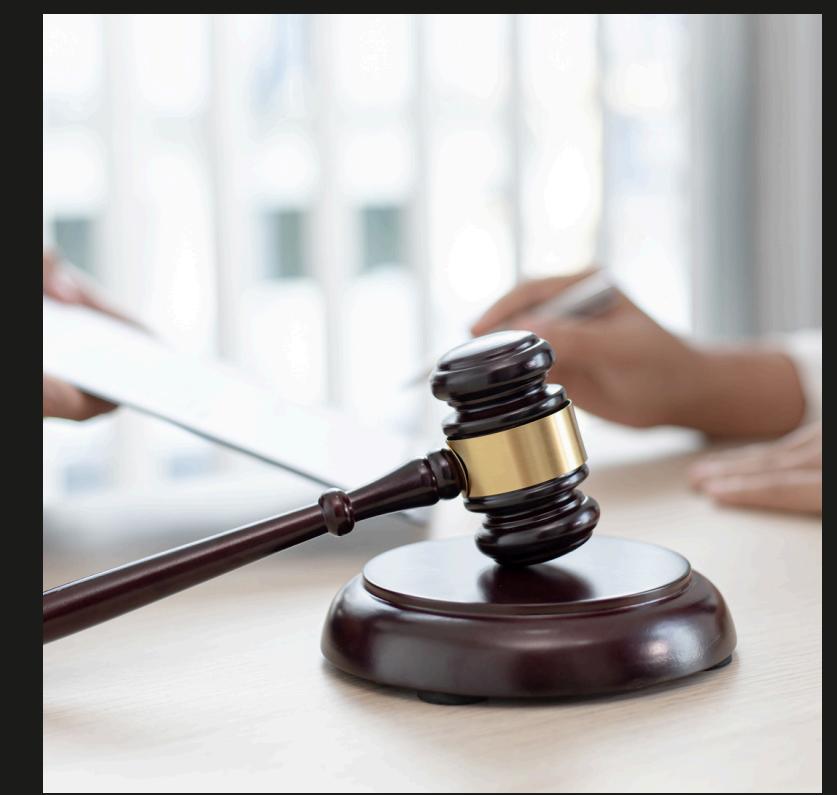
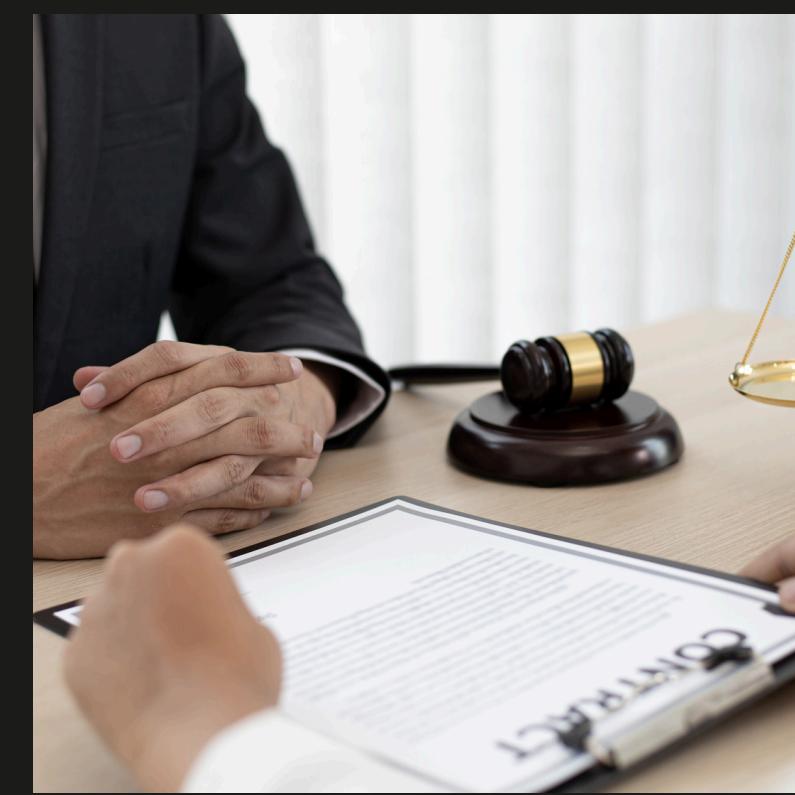
모델링

04

웹 구현

05

계획





주제 선정 배경

01

인터넷을 통한 법률 상담의 증가

인터넷을 통해 법률 상담을 받고자 하는 사람들이 많아 비스를 제공하는 것은 시대적인 요구에 부합하다고 판

02

법률 지식의 접근성 향상

일반 대중이 법률 지식에 접근하기 어려운 경우가 많아 문제에 대한 정보를 얻을 수 있도록 지원할 수 있다.

03

비용 절감과 효율성 향상

변호사에게 상담을 받거나 법률 서류를 작성하기 위해 용을 절감하고 빠르고 효율적으로 법률적인 문제를

the L 박준제 변호사 법률상담

#일산변호사#파주변호사#형사전문변호사#성공사례다수...

박준제 변호사 238명 | 30분 전



제14회 변호사시험 정보 공유 with 메가로이어스

#변호사시험 #변시 #제14회변시 #제14회변호사시험 #...

메가로이어스 변호사 786명 | 30분 전



변호사 직접상담방

#법률상담 #변호사 #민사 #가사 #형사 #이혼 #상속 #...

변호사 332명



법무법인윤승

법률상담 수원변호사 안산변호사 성범죄 음주운전 형사전문...

법무법인윤승 584명 | 30분 전



교통사고 무료 상담방(합의, 과실, 대인)

#교통사고#산재#운전자#보험#렌트카#음주운전#스포츠...

교통사고 무료 상담방 584명 | 30분 전



우리변호사

#법률상담 #민사 #손해배상 #이혼 #형사변호사 #고소 ...

우리변호사 205명



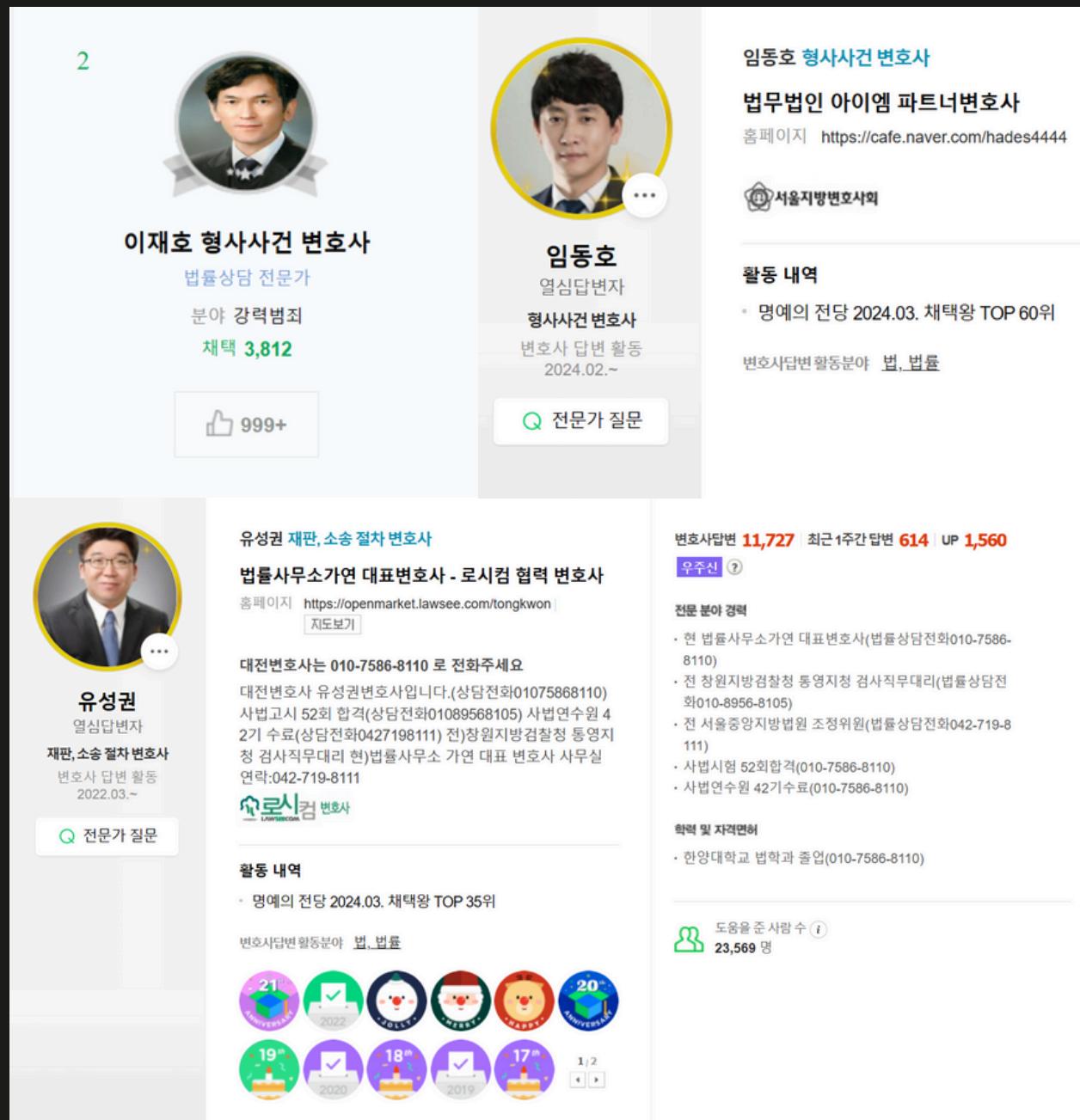


데이터 수집 & 전처리



데이터 수집 : 크롤링

약 10명의 변호사 지식인 답변 데이터 수집



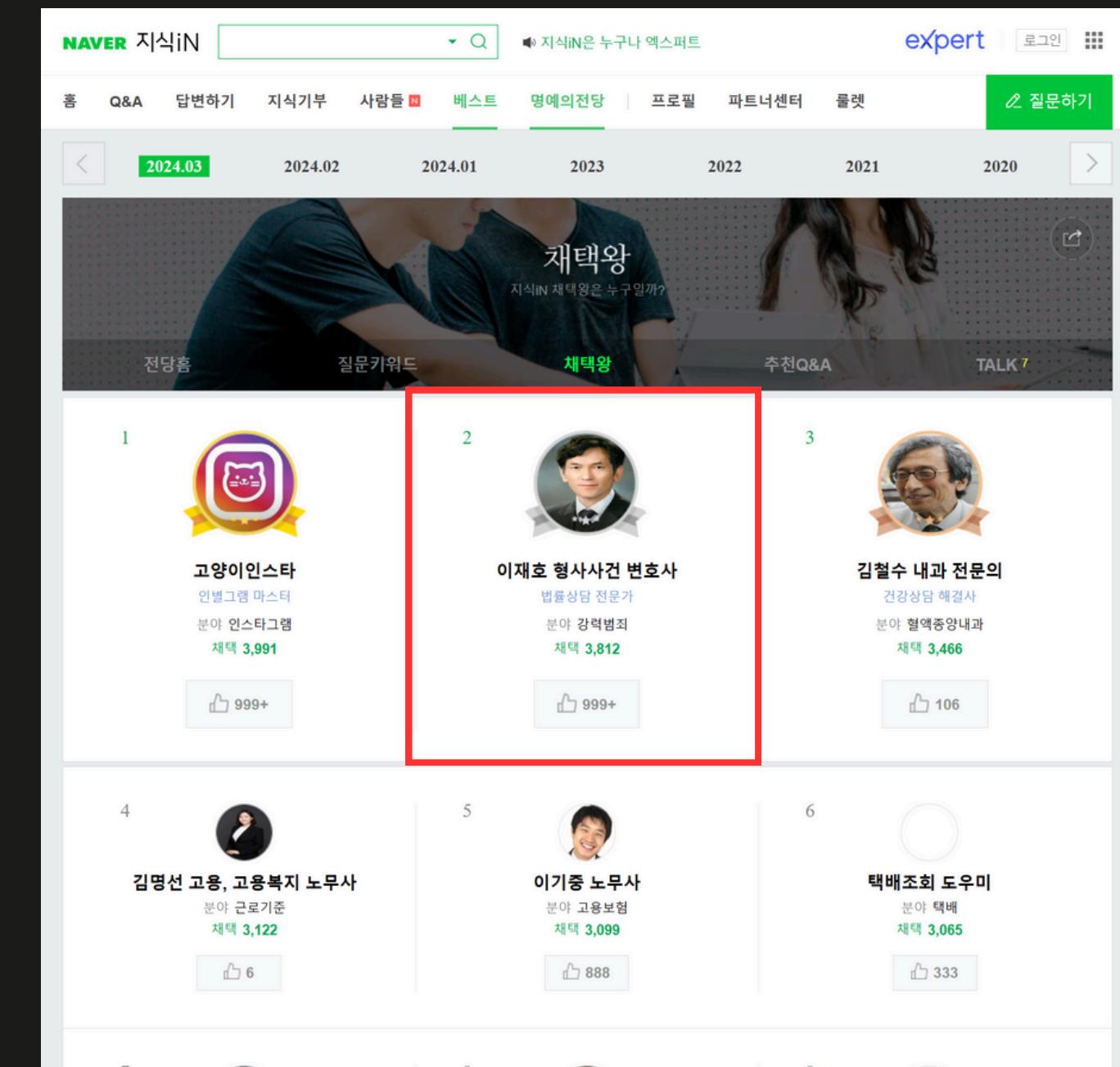
2

이재호 형사사건 변호사
법률상담 전문가
분야 강력범죄
채택 3,812
999+

임동호 형사사건 변호사
법무법인 아이엠 파트너변호사
홈페이지 <https://cafe.naver.com/hades4444>
서울지방변호사회

활동 내역
· 명예의 전당 2024.03. 채택왕 TOP 60위
변호사답변 활동분야 범, 범률

3
유성권 재판, 소송 절차 변호사
법률사무소가연 대표변호사 - 로시컴 협력 변호사
홈페이지 <https://openmarket.lawsee.com/tongkwan>
지도보기
대전변호사는 010-7586-8110 로 전화주세요
대전변호사는 유성권변호사입니다.(상담전화01075868110)
사법고시 52회 합격(상담전화01089568105) 사법연수원 4
2기 수료(상담전화0427198111) 전)창원지방검찰청 통영지
청 검사직무대리 현)법률사무소 가연 대표 변호사 사무실
연락:042-719-8111
로시컴 변호사
999+



NAVER 지식IN

2024.03 2024.02 2024.01 2023 2022 2021 2020

채택왕

질문하기

1 고양이인스타
인생그램 마스터
분야 인스타그램
채택 3,991
999+

2 이재호 형사사건 변호사
법률상담 전문가
분야 강력범죄
채택 3,812
999+

3 김철수 내과 전문의
건강상담 해결사
분야 혈액증양내과
채택 3,466
106

4 김명선 고용, 고용복지 노무사
분야 근로기준
채택 3,122
6

5 이기중 노무사
분야 고용보험
채택 3,099
888

6 택배조회 도우미
분야 택배
채택 3,065
333

데이터 수집 : 크롤링

예상 데이터 수 : 약 13만개

이재호 형사 사건 변호사 (서울)
법무법인 공신 대표 변호사
홈페이지 <http://www.byunhosa.kr> [지도보기]
상담예약 010 8759 8282
서울지방변호사회
활동 내역
· 명예의 전당 2024.03. 채택왕 TOP 2위
변호사답변 활동분야 범_법률
2023 분야별 지식인 전문가 분야
2022 분야별 지식인
2021 분야별 지식인
2020 분야별 지식인
2019 분야별 지식인
2018 분야별 지식인
답변 보기 (128,436)
전문가 답변 (128,538) **더보기**
오픈사전 보기 (0)
Q 응주운전 벌금 600만원 2024.04.29.
A 안녕하세요. 서울지방변호사회 소속으로 지식IN 법률상담을 진행하고 있는 이재호 변호사입니다. 벌금에 대한 분할납부는 '제3형' 등에...
Q ㅋㅋㅋㅋㅋ 라고만 댓글 달아도 처벌받나요? 2024.04.29.
A 안녕하세요. 서울지방변호사회 소속으로 지식IN 법률상담을 진행하고 있는 이재호 변호사입니다. 구체적인 사안에 따라 달리 판단될 수...
Q 소년재판 처분 2024.04.29.
A 안녕하세요. 서울지방변호사회 소속으로 지식IN 법률상담을 진행하고 있는 이재호 변호사입니다. 비행사실, 과거비행력, 반성하는 태도,...
Q 여자회장실 호기심 무단출입 관련 질문 2024.04.29.
A 안녕하세요. 서울지방변호사회 소속으로 지식IN 법률상담을 진행하고 있는 이재호 변호사입니다. 작성된 내용만 보면 사건화 가능성은 낮...

전문가 답변수 128,538 | 어제 한 답변 52 | 오늘 한 답변 57 | 전문가 답변 활동 2015.11.~
전체질문
제목 분야 작성
스토킹처벌법 형벌, 형집행 2024.04.28
청주 성범죄자 성범죄 2024.04.28
스토킹죄로 신고 가능 할까요? 공갈, 협박 2024.04.28
성범죄 관련 강력범죄 2024.04.28
폭행 합의금 폭행 2024.04.28
실스로 여자랑 부딪힘 성범죄 2024.04.28
실수로 여자랑 부딪힘 성범죄 2024.04.28
어깨빵 합의금 여쭈어봅니다. 공개 2024.04.28
공연음란죄 질문 성범죄 2024.04.28
통매음 고소 재판, 소송 절차 2024.04.27
고소 당하나요? 성범죄 2024.04.27
여자회장실 호기심 무단출입 관련.. 성범죄 2024.04.27
형사사건변호사수입료 얼마정도 하나요? 성범죄 2024.04.27
성폭행무고죄 맞고소 당하기 쉬워요? 성범죄 2024.04.27
장애인성폭행 상대는 제가 장애인인거 몰랏... 성범죄 2024.04.27
동성성폭행 조용히 해결하고 싶은데 성범죄 2024.04.27
여자회장실 호기심으로 출입 관련 성범죄 2024.04.27
여자회장실 호기심으로 무단침입 관련 성범죄 2024.04.27
강제추행구공판 실형 위기입니다 성범죄 2024.04.27
학폭생기부기재 몇호부터 기재될까요?? 성범죄 2024.04.27

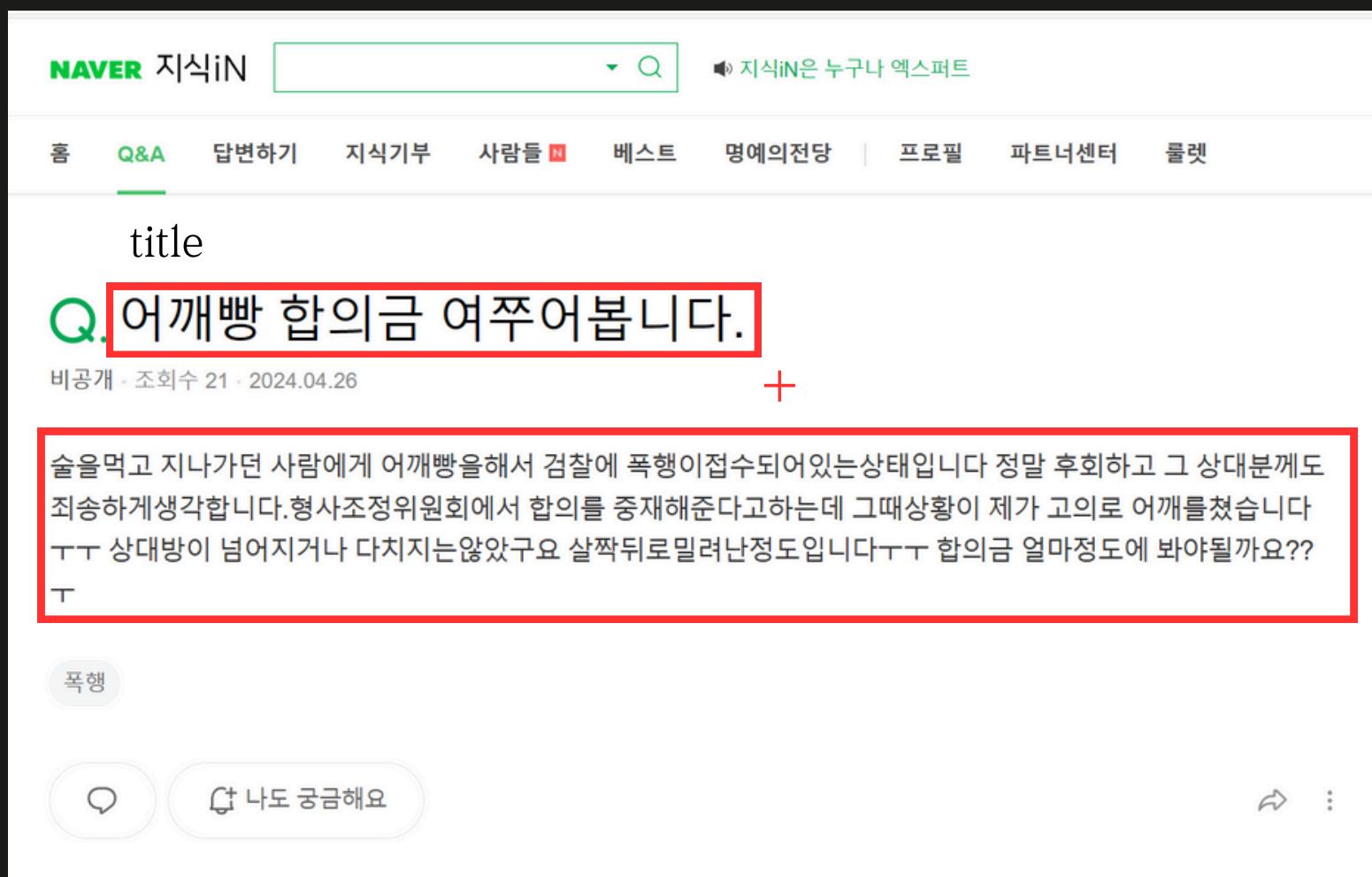
분야 추출

링크 추출

데이터 수집 : 크롤링

추출한 링크로 접속 후 크롤링

question



NAVER 지식iN

title

Q **어깨빵 합의금 여쭈어봅니다.**

비공개 · 조회수 21 · 2024.04.26

술을먹고 지나가던 사람에게 어깨빵을해서 검찰에 폭행이접수되어있는상태입니다 정말 후회하고 그 상대분께도 죄송하게생각합니다.형사조정위원회에서 합의를 중재해준다고하는데 그때상황이 제가 고의로 어깨를쳤습니다 ㅜㅜ 상대방이 넘어지거나 다치지는않았구요 살짝뒤로밀려난정도입니다ㅜㅜ 합의금 얼마정도에 봄야될까요??

폭행

나도 궁금해요

answer



이재호

변호사 🔥 열심답변자

법무법인 공신 · 2023 전문가 분야 지식인

본인 입력 포함 정보 ⓘ

안녕하세요. 서울지방변호사회 소속으로 지식iN 법률상담을 진행하고 있는 이재호 변호사입니다.

합의금은 양 당사자의 합의에 의해 정해지는 금액이기 때문에 정해져있는 최대나 적정선이 없습니다.

합의는 가해자와 피해자의 의사의 합치로써 합의조건은 다양하고 당사자 중 일방의 제시안을 다른 일방이 수용할 때 성립되므로 어떤 금액을 제시하였을 때 피해자가 거절하거나, 피해자가 제시한 금액을 질문자님이 수용하지 못하면 합의는 성립되지 않습니다.

상담예약 010 8759 8282

데이터 수집 : 크롤링

크롤링 함수 코드



해당 변호사의 답변만 크롤링

```
def crawl_data(url, name):  
    page_num = 1 # 크롤링할 페이지 번호  
    # page_num 0/ 1인 경우에 컬럼만 적힌 csv 파일을 생성  
    if page_num == 1:  
        df = pd.DataFrame(columns=['field', 'question'])  
        file_path = './data/' + name + '.csv'  
        df.to_csv(file_path, mode='w', index=False)  
  
    while True:  
        try:  
            page_url = url + f'&page={page_num}'  
            html = urlopen(page_url)  
            bs = BeautifulSoup(html, 'html.parser')  
  
            q_tbody = bs.find('tbody').find_all('tr')  
  
            # 필드 추출  
            fields = [q.find('td', class_='field').text for q in q_tbody]  
  
            # 제목 추출  
            titles = [q.find('a').text for q in q_tbody]  
  
            # 링크 추출  
            links = [title.find('a').get('href') for q in q_tbody for title in q.find_all('td', class_='title')]  
  
            # 질문 추출  
            questions = []  
  
            # 답변 추출  
            answers = []  
  
            # 링크 안에 들어가서 질문과 답변 추출  
            for link in links:  
                try:  
                    html = urlopen('https://kin.naver.com/' + link)  
                    bs = BeautifulSoup(html, 'html.parser')  
                    question_elements = bs.find('div', class_='questionDetail')  
                    question = question_elements.text.strip()  
                    cleaned_question = re.sub(r'[\n\t\r\x0a]', '', question)  
                    questions.append(cleaned_question)  
  
                    answer_elements = bs.find_all('div', id=re.compile('^answer_\d+$'))  
                    for answer_element in answer_elements:  
                        author = answer_element.find('strong', class_='name').text.strip()  
                        if author == name:  
                            answer = answer_element.find_all('p', class_='se-text-paragraph')  
                            answer_texts = [p.text.strip().replace('\u200b', '') for p in answer]  
                            combined_answer = "\n".join(answer_texts)  
                            answers.append(combined_answer)  
                except Exception as e:  
                    print("Error questions for link:", link, e)  
                    questions.append(None) # 예외가 발생한 경우 None 추가  
                    answers.append(None)  
  
            # 최종 questions ( 제목 + 질문 )  
            questions = [title + '.' + question if question else '' for title, question in zip(titles, questions)]  
  
        url_lee = 'https://kin.naver.com/userinfo/expert/answerList.naver?u=dEgr29uBZ5z0sTY%2FIic1U8r4rT%2BfX97EC6sMbz2os5Q%3D'  
        crawl_data(url_lee, '이재호')
```

간단한 전처리

데이터 전처리

데이터 전처리 함수



```
def remove_repeated_punctuation(sentence):  
    # 한글, 영어, 숫자, 일부 구두점만 포함되도록 정규표현식 수정  
    cleaned_sentence = re.sub(r'[^ㄱ-ㅎㅏ-ㅣ가-힣a-zA-Z0-9\.\\\'!?\s]', '', sentence)  
    # 반복되는 'ㅠ', 'ㅎ', 'ㅋ'를 지움  
    cleaned_sentence = re.sub(r'(ㅠ)\1+', r'\1', cleaned_sentence)  
    cleaned_sentence = re.sub(r'(ㅎ)\1+', r'\1', cleaned_sentence)  
    cleaned_sentence = re.sub(r'(ㅋ)\1+', r'\1', cleaned_sentence)  
    cleaned_sentence = re.sub(r'(ㅜ)\1+', r'\1', cleaned_sentence)  
    # 반복되는 구두점을 지움  
    cleaned_sentence = re.sub(r'([^\w\s])\1+', r'\1', cleaned_sentence)  
    return cleaned_sentence  
  
def predata_to_csv(df, repeat_sentences, mode = 'a'):  
    # 결측치 제거  
    df.dropna(inplace = True)  
    # 한 열이라도 빈문자열인 행 제거  
    df = df.drop(index=df[df.apply(lambda row: row.str.strip().eq('')).any(axis=1)].index)  
    # 중복되는 질문 제거  
    df.drop_duplicates(subset='question', keep='last', inplace=True)  
    # 중복되는 인사를 제거  
    df['answer'] = df['answer'].str.replace('안녕하세요.', '')  
    for r in repeat_sentences:  
        df['answer'] = df['answer'].str.replace(re.escape(r), '', regex=True)  
    # 반복 구두점 제거  
    df['answer'] = df['answer'].apply(remove_repeated_punctuation)  
    df['question'] = df['question'].apply(remove_repeated_punctuation)  
    print(df)  
    if mode == 'w':  
        df.to_csv('rawal_data.csv', mode = 'w', index=False)  
    else:  
        df.to_csv('rawal_data.csv', mode = 'a', index = False, columns = None)
```

```
leeDF = pd.read_csv('./data/이재호.csv')  
leement = ['상담예약 010 8759 8282', '서울지방변호사회 소속으로 지식IN 법률상담을 진행하고 있는 이재호 변호사입니다.',  
predata_to_csv(leeDF,leement)
```

```
ohDF = pd.read_csv('./data/오지영.csv')  
ohment = ['서울지방변호사회 소속으로 지식IN 법률상담을 진행하고 있는 오지영 변호사입니다.', '힘들고 복잡한 법률 문제, 혼자 고민하지 마시고 언제든지 전화주세요 법률상담 전화번호 010-7494-2459']  
predata_to_csv(ohDF, ohment, 'a')
```

데이터 취합

```
uDF = pd.read_csv('./data/유성권.csv')
ument = ['대전모욕죄변호사 유성권 변호사입니다',
          '대전변호사 유성권 변호사 사시52회 입니다',
          '대전상속변호사 유성권 변호사 사시52회입니다',
          '대전합의서변호사 유성권입니다',
          '대전변호사 유성권 입니다',
          '대전이혼전문변호사 유성권입니다',
          '대전변호사 유성권 변호사 사시52회입니다',
          '대전변호사 유성권 변호사입니다',
          '개인회생변호사 유성권 변호사입니다',
          '로톡네이버 지식IN 상담변호사 유성권 입니다',
          '로시컴-네이버 지식IN 상담변호사 유성권 입니다.',
          '대전이혼변호사 유성권 변호사입니다',
          '대전상해죄변호사 유성권 변호사 사시52회입니다',
          '대전변호사 유성권 변호사 사시52회입니다',
          '대전전세사기변호사 유성권 변호사입니다',
          '대전경찰조사 변호사 유성권 변호사입니다']
predata_to_csv(uDF, ument, 'a')
```

최종 데이터 수 : 71744개

```
<class 'pandas.core.frame.DataFrame'>
Index: 71744 entries, 0 to 321817
Data columns (total 3 columns):
 #   Column   Non-Null Count  Dtype  
---  --  
 0   field    71744 non-null   object 
 1   question  71744 non-null   object 
 2   answer   71744 non-null   object 
dtypes: object(3)
memory usage: 2.2+ MB
```

데이터 확인 : filed, question, answer

field	question	answer
0 재판, 소송 절차	남해상속상담 부탁드립니다.본인이 부동산을 상속 받는 게 부모님 유언이라고 우기고 ...	질문자님의 공동상속인이 유언이라 말하는 동영상이 녹음유언의 조건에 부합는지 ...
1 신용, 파산	안성개인회생전문 사무실 찾고 있어요.상담 좀 받고 싶은데., 채무가 많아요. 저도...	안성개인회생전문 변호사사무실 찾는 질문에 답변 드립니다. 개인회생의 성공은 곧...
2 신용, 파산	평택시개인회생전문 변호사 추천부탁드려요. 대출금을 상환하지 못해 현재 개인회생 신청...	평택시개인회생전문 변호사 찾는 질문에 답변 드립니다. 개인회생은 법적으로 채무...
3 재판, 소송 절차	구미학교폭력변호사님 저희 아들 이야기입니다. 구미학교폭력변호사님 부모로써 대응할것입니다...	질문자님, 글 작성해주신 교내 학교폭력 관련하여 구미학교폭력변호사가 직접 답변...
4 재판, 소송 절차	동생이랑 제가 나이차이가 많이납니다.그리고 동생이 어렸을 때부터 대학 들어갈 때까지...	질문자님, 글 작성해주신 유류분 반환 관련하여 대구상속로펌에서 직접 설명 드리...
5 재판, 소송 절차	당진교통사고변호사 도움 좀 받고 싶습니다 당진교통사고변호사 도움 좀 받고 싶습니다빨...	대인 대물사고를 일으켰으나 책임을 져야 한다는 것에 무서워 현장을 도주하였다면...
6 재판, 소송 절차	아동학대 방임 관련 사안입니다.태안형사전문변호사 찾고 있는데 잘하시는 분 있나요?이...	최근 각종 뉴스와 기사를 통해 아동학대 방임에 대한 문제가 커지고 있다는 점 ...
7 재판, 소송 절차	전여자친구폭행 변호사 도움 전여자친구폭행으로 신고당했습니다.경찰에서 조사받으러 오라...	폭행 사건에 연루되어 변호사의 도움이 필요한 상황 같습니다. 폭행죄는 타인에...
8 재판, 소송 절차	부천마약변호사 제발 도와주세요 부천마약변호사 도움받고 싶어요호기심에 딱 한번 텔레그...	마약 사건에 연루되어 부천마약변호사의 도움이 필요한 상황 같습니다. 마약사범...
9 재판, 소송 절차	경찰조사변호사 도움이 절실한 상황입니다. 경찰조사변호사 도움이 절실한 상황입니다억울...	아동에 대한 정서적 학대행위의 법정형은 5년 이하의 징역 또는 5천만원 이하의...
10 재판, 소송 절차	카촬죄 기소유예 받을 수 있을까요? 카촬죄 기소유예 받을 수 있을까요?제가 회식하고...	카촬죄의 형량은 높은 편입니다. 징역은 최대 7년까지 가능하며, 벌금은 최대 ...
11 재판, 소송 절차	함양변호사사무실 상담 받을 수 있을까요? 함양변호사사무실 방문하여 상담 받을 수 있...	질문자님의 경우 상속포기가 아닌 한정승인을 진행하시는 게 더 맞는 선택지라고...
12 재판, 소송 절차	창원마약상담 처음 걸린건데 위험할까요 저 진짜 호기심에 처음으로 먹어봤다가 경찰에서...	무조건 법률 대리인의 도움이 필요합니다. 마약은 초범도 5년 이하의 징역 또는...
13 재판, 소송 절차	부산공사대금변호사 복잡한 상황이라 질문. 저희는 하도급업체이고 도급업체와 공사대금 ...	질문자께서 말씀하신 대로 '계약서 대로' 문제없이 진행되었다면 상대가 소송을 걸...
14 재판, 소송 절차	성균관대역변호사 찾고있습니다 성균관대역변호사 찾고 있는데요임대인이 전세보증금을 안돌...	전세보증금을 반환해 주지 않는다면 전세보증금반환소송을 진행해 보시기 바랍니다....
15 재판, 소송 절차	화서역변호사님 세입자월세미납 어떻게 해. 화서역변호사 찾고 있어요세입자가 2달치 월...	세입자가 두달치의 월세를 밀렸다면 명도소송을 할 수 있습니다. 만약, 본인 ...
16 재판, 소송 절차	수원역법무법인 아동학대 고소 상담 가능한. 저는 어린이집 교사이고 아이의 학부모에게...	아동에 대한 정서적 학대행위의 법정형은 5년 이하의 징역 또는 5천만원 이하의...
17 재판, 소송 절차	매교역변호사 빌려준돈 소송 문의드립니다 안녕하세요. 전 여자친구에게 돈을 빌려주었는...	대여금 반환 청구 소송은 비용과 시간 또한 많이 소요됩니다. 그렇기 때문에 비...
18 재판, 소송 절차	서울민사변호사 폭행으로 인한 손해배상청. 서울민사변호사님, 제가 폭행으로 인한 손해...	우선 형사와 민사는 별개의 절차라는 점 인지하고 계셔야 합니다. 만일 형사소...
19 재판, 소송 절차	강남변호사사무실 음주운전 초범의 선처 안녕하세요. 음주운전 초범인데 강남변호사사무실...	음주운전 초범 관련 문의는 강남변호사사무실에 꾸준히 들어오고 있습니다. 음주...
20 재판, 소송 절차	의정부개인파산전문 변호사 선임 꼭 해야. 의정부개인파산 신청하려고 합니다. 알아보니 ...	의정부개인파산 질문에 답변 드립니다. 개인파산제도는 자신의 모든 재산으로도 채...
21 신용, 파산	남양주개인회생 대출받은게 감당이 안됩니다. 남양주개인회생 가까운 로펌 찾고있습니다.빠...	남양주개인회생 로펌 찾는 질문에 답변 드립니다. 개인회생 제도는 신청인의 지...

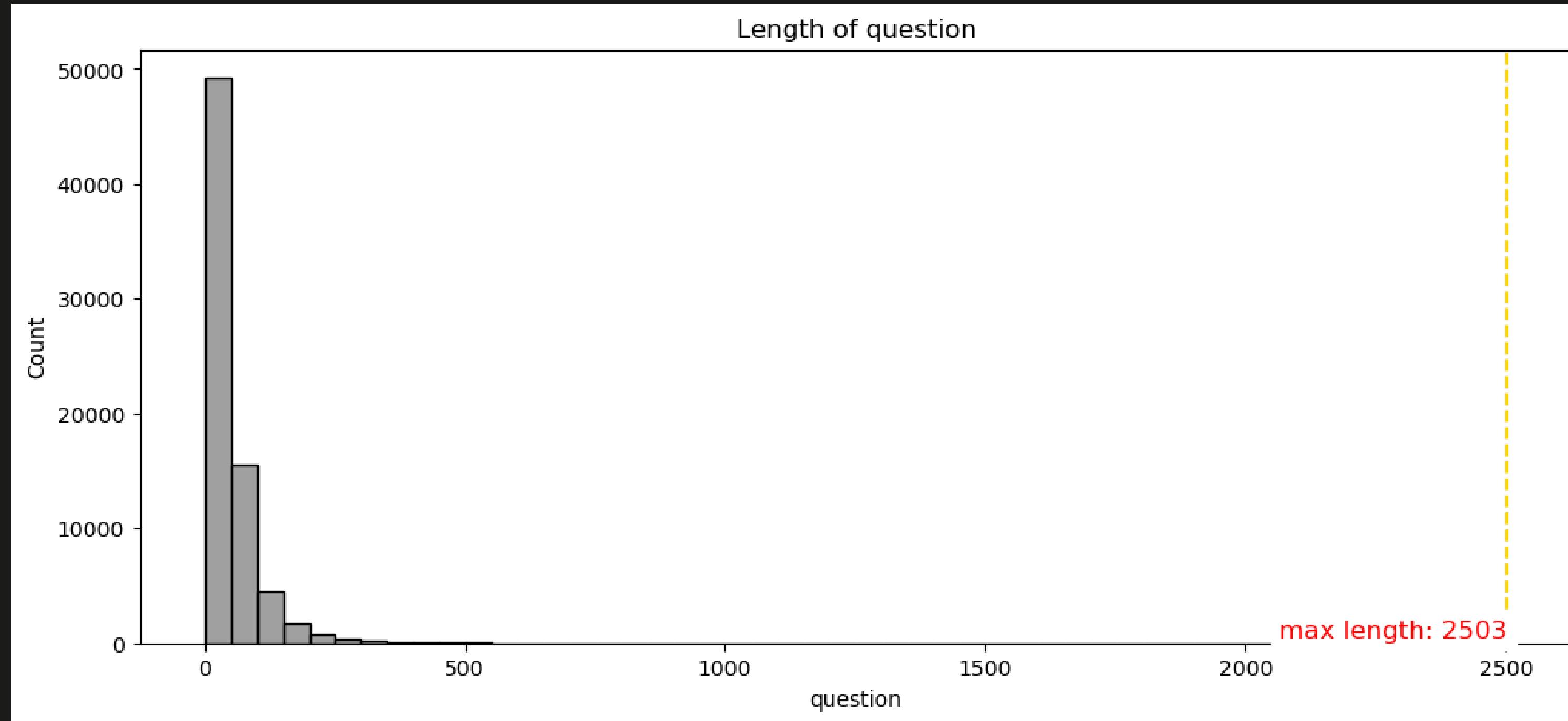


Modeling



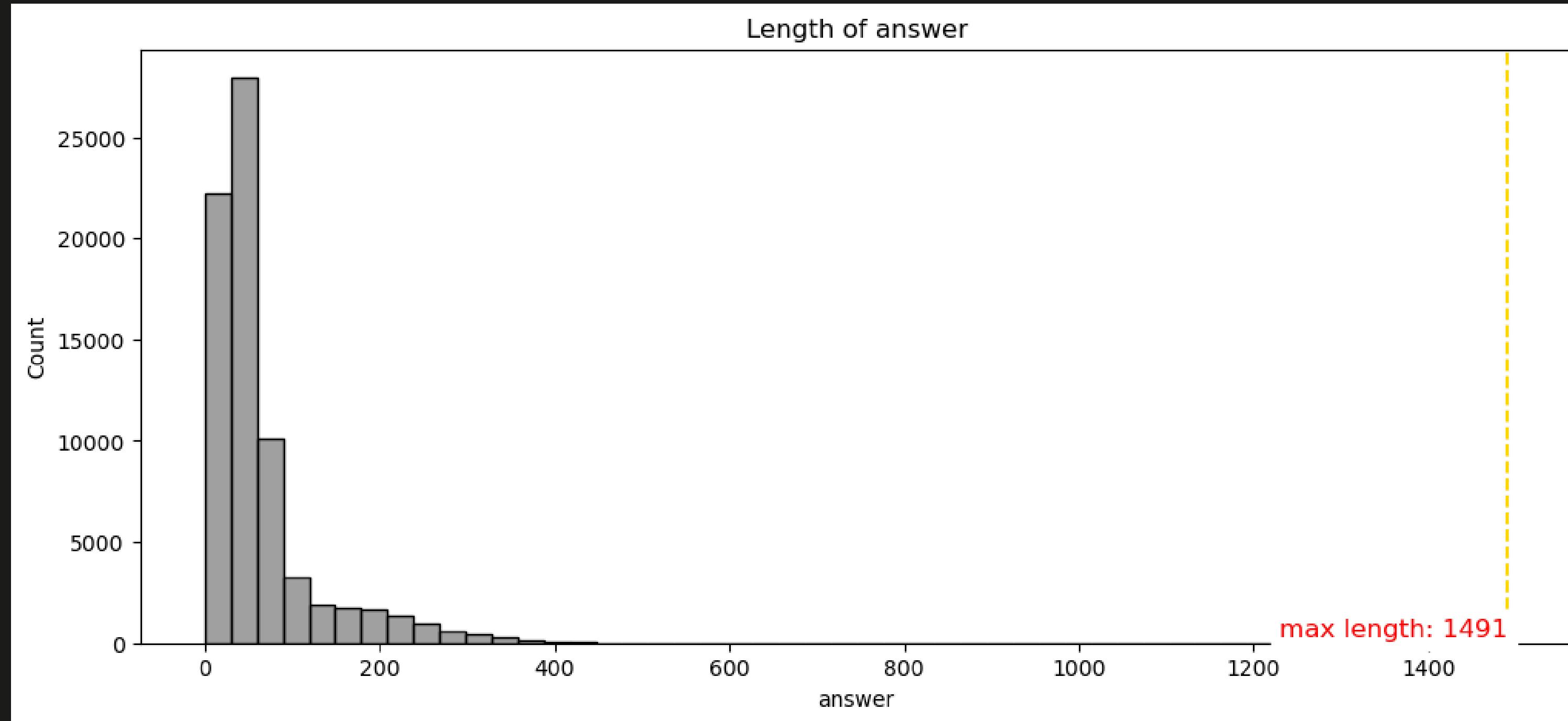
데이터 분포 확인 : question

기준 : 토큰 개수



데이터 분포 확인 : answer

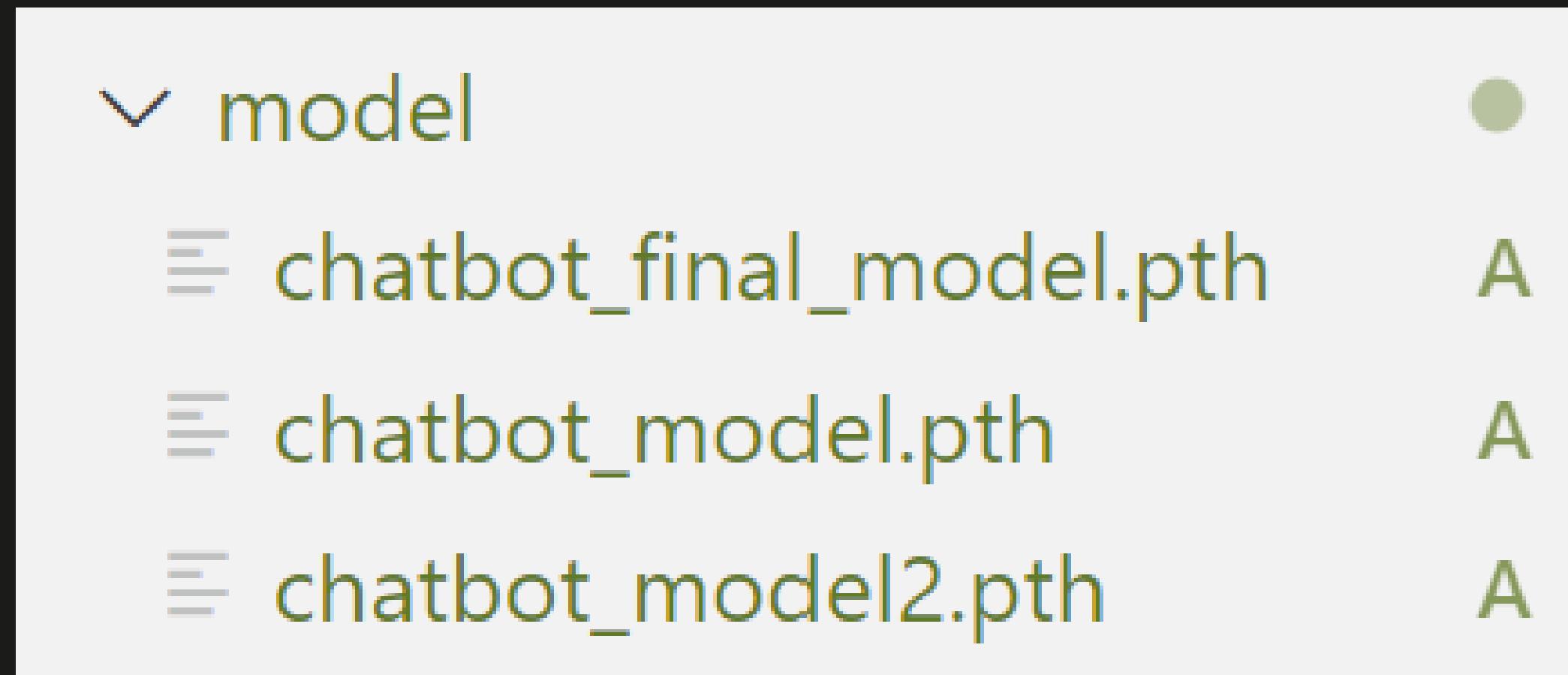
기준 : 토큰 개수



GPT2 Fine Tuning

"skt/kogpt2-base-v2"

이틀에 걸쳐...



Predict

Transformer Model

```
OutOfMemoryError: CUDA out of memory. Tried to allocate 764.00 MiB. GPU 0 has a total capacity of 11.99 GiB of
which 0 bytes is free. Of the allocated memory 10.95 GiB is allocated by PyTorch, and 115.14 MiB is reserved by
PyTorch but unallocated. If reserved but unallocated memory is large try setting
PYTORCH_CUDA_ALLOC_CONF=expandable_segments:True to avoid fragmentation. See documentation for Memory Management
(https://pytorch.org/docs/stable/notes/cuda.html#environment-variables)
```

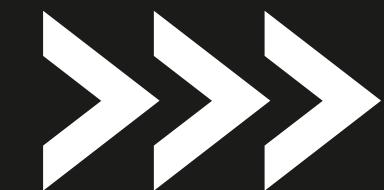


Web 구현

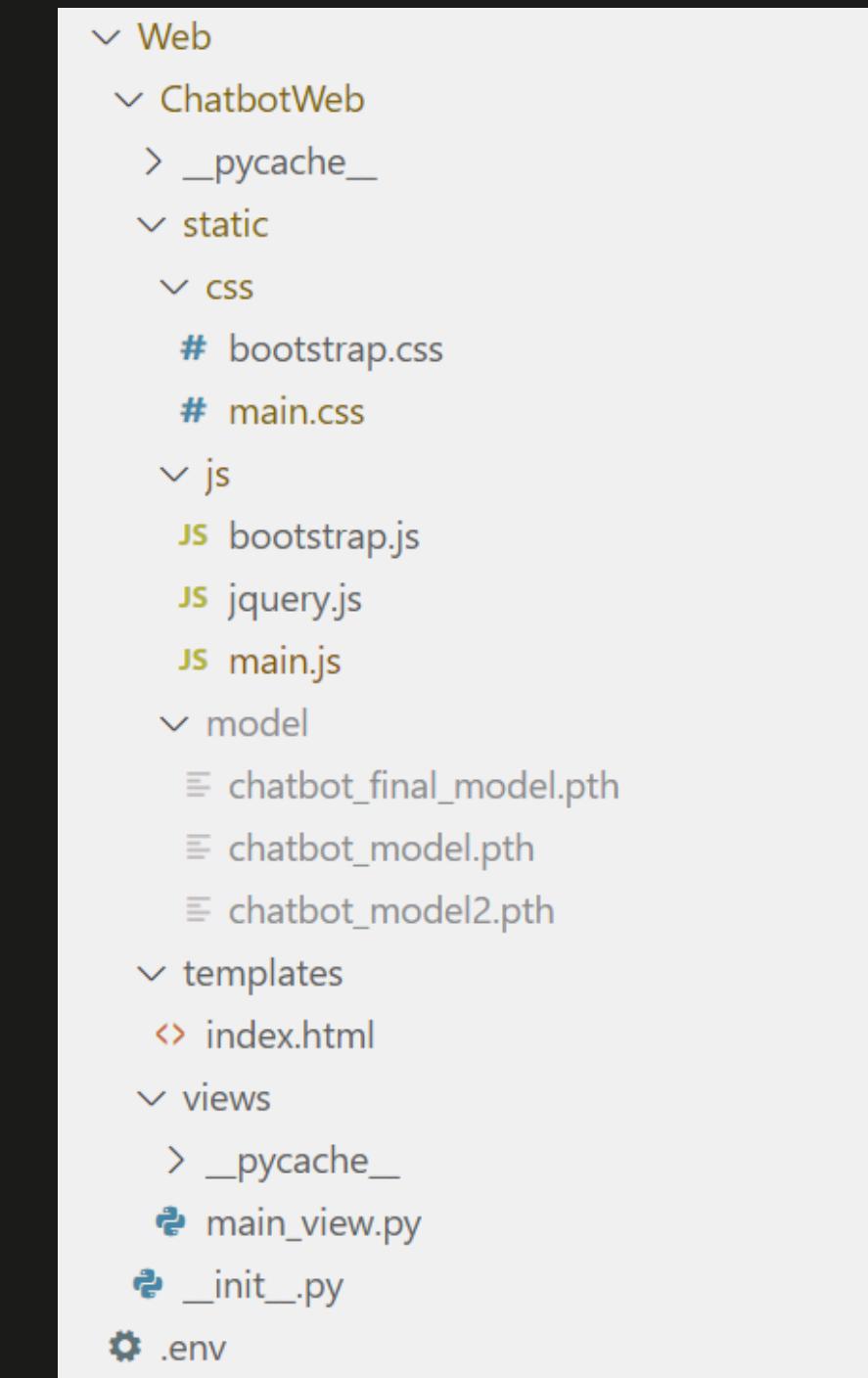


Web 구현

bootstrap 이용



내 프로젝트에 맞게 변형





앞으로의 계획



모델링

01

모델 경량화

02

offset 주는 모델로 모델링

03

2가지 방법 시도

1. 요약 모델 -> Transformer 2단계에 걸쳐 모델링
2. 과감히 1%의 데이터를 버리고 온전한 데이터로 학습



1. 요약 데이터 생성 : 허깅페이스 요약 모델 이용

"lcw99/t5-base-korean-text-summary"

Q. 원문

마트절도죄.해결가능할까요 ㅠ 마트절도죄 문제인데 도움 주시면 감사하겠습니다제가 마트에서 물건을 몇개 훔쳤습니다 ㅠ가격 환산하면 10만원도 안되긴 하는데.사장이 저를 신고해서 경찰조사 앞두고 있어요.마트절도죄 처벌 받을까요? 저 어떻게해야하나요?

요약문

마트에서 물건을 몇개 훔쳤는데 사장이 신고해서 경찰조사를 앞두고 있는데 마트절도죄 처벌 받을까요?

A. 원문

마트절도죄 관련하여 변호사의 도움이 필요한 상황 같습니다. 절도는 형법 제329조에 따라 타인의 재물을 절취했을 때 성립되는 범죄입니다. 짙 문자님의 경우 마트에서 계산하지 않고 물건을 가져가셨다면 당연히 절도가 될 수밖에 없습니다. 물건의 가치가 크지 않더라도 절도는 결국 똑같기 때문에 처벌을 피할 수 있다거나 금방 선처를 받을 수 있다고 안일하게 생각해서는 안됩니다. 따라서 변호사와 상담부터 진행해보심이 좋을 것 같습니다. 감사합니다.

요약문

마트절도죄는 형법 제329조에 따라 타인의 재물을 절취했을 때 성립되는 범죄로 마트에서 계산하지 않고 물건을 가져가셨다면 당연히 절도가 될 수밖에 없다

1. 요약 데이터 생성 : 허깅페이스 요약 모델 이용

"lcw99/t5-base-korean-text-summary"

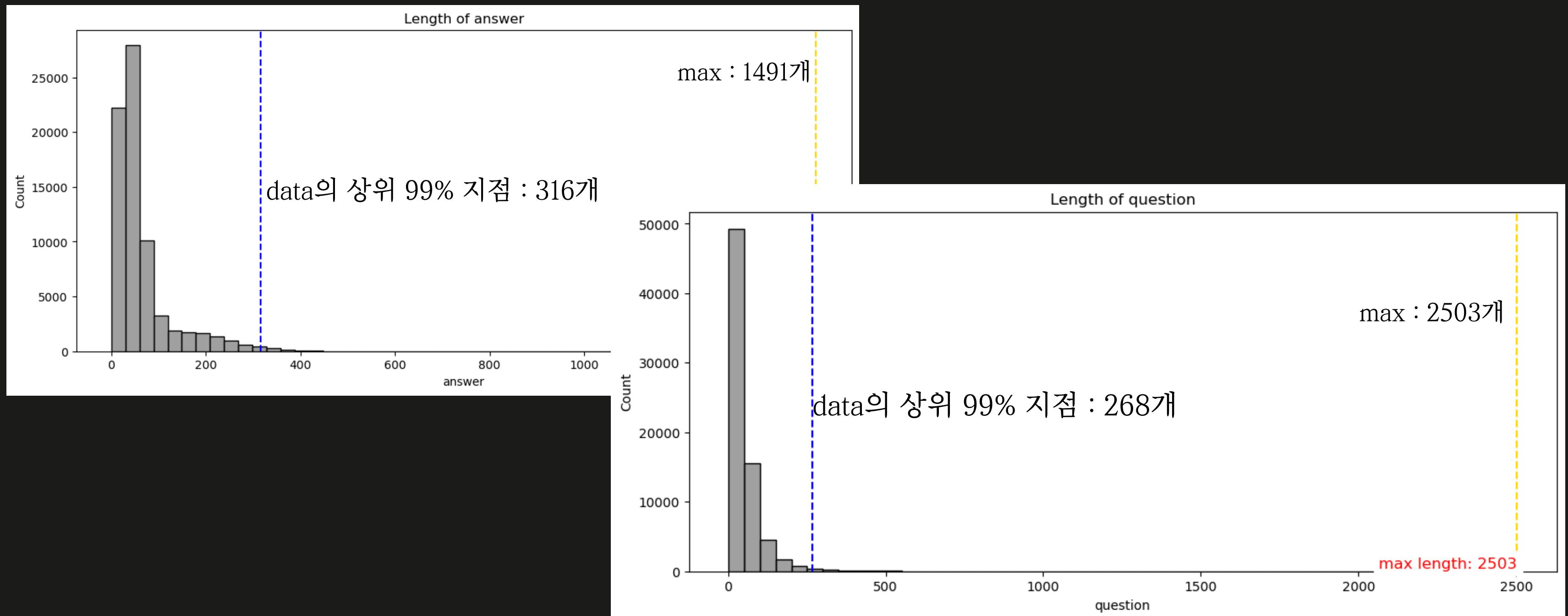
```
1 # question, answer 을 입력으로 받아서, question, answer를 요약한 df를 생성
2
3 def summarize_text(text):
4     inputs = ["summarize: " + text]
5     inputs = tokenizer(inputs, max_length=max_input_length, truncation=True, return_tensors="pt")
6     output = model.generate(**inputs, num_beams=8, do_sample=True, min_length=10, max_length=100)
7     decoded_output = tokenizer.batch_decode(output, skip_special_tokens=True)[0]
8     predicted_title = nltk.sent_tokenize(decoded_output.strip())[0]
9     return predicted_title
10
11 df['question'] = df['question'].apply(summarize_text)
12 df['answer'] = df['answer'].apply(summarize_text)
13
14 df.to_csv('./data/summarized_data.csv', index=False)
```

⌚ 41m 3.7s

2. 데이터 제거 후 모델 학습 : 1%의 긴 문장 제거

기준 : 토큰 개수

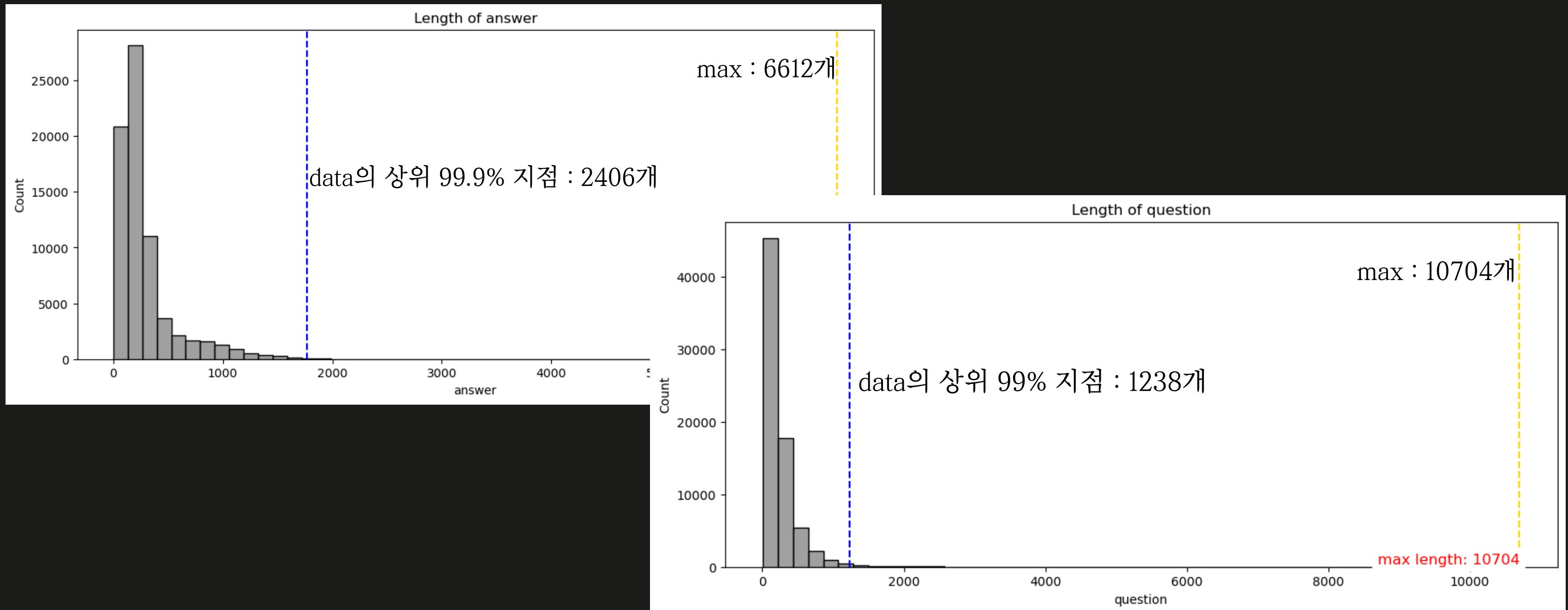
전체 데이터의 최대 2%의 데이터 손실 : 약 1400개



2. 데이터 제거 후 모델 학습 : 1%의 긴 문장 제거

기준 : 문장길이

전체 데이터의 최대 1.1%의 데이터 손실 : 약 770개



데이터 손실 불가피하다.

Question 중 가장 긴 문장 요약

원문 : 힘듭니다. 안녕하세요 . 4년전에 잊었던일입니다 그당시 저는 24살이였고 사촌 여동생은 저랑 9살차이인 15살 중 3이였습니다 우리는 서로 사랑을 하게되었고 합의하에 관계를 맺었습니다 그리고 시간이 흘러 사촌이라는것때문에 서로 불안하고 앞이막막해져 싸우는것도 심해지고 아래도돼나 싶고 그래서 저도 독하게 헤어지게되었습니다 절대 강제로 성관계를 하진 않았습니다 아직도 솔직하게 왜그랬나 싶고 서로 사랑한게 밉습니다 사촌여동생과 저는 피를 나눈 것은 아닙니다 그당시 삼촌이 딸과 아들이잇는 숙모와 재혼을 했던것입니다 삼촌 이낳은 딸이 아닙니다 하지만 현재 지금은 서로 맞지않아 다시 이혼을 하는 상황이고요. 만약 사촌여동생이 저를 고소를 하게된다면 처벌을 받나요 몇달 전에 통화를 하면서 제가혹시나하는 마음에 통화녹음을 한 상황이고요 내용은 그때 오빠랑 사귀면서 후회했냐 그때 서로 생리안해서 엄청 힘들었지않냐 이런내용들과 다른 및 및 내용들입니다 헤어진이유도 인정한 부분이고 전부 인정 했던 통화 녹음입니다 만약 고소를 당하게된다면 증거가되나요 이거말고도 헤어진후 자주통화를 하였고 몇몇 친하게지내는 녹음 파일도잇습니다 이런것들이 증거가 돼나요? 부탁드립니다 변호사님



요약문 : 사촌 여동생과 저는 피를 나눈것은 아니지만 삼촌이 딸과 아들이 있는 숙모와 재혼을 해서 서로 맞지 않아 다시 이혼을 하는 상황이고 만약 사촌 여동생이 저를 고소를 하게 된다면 처벌을 받나요

모델링



긴 문장의 입력 데이터
(Question)

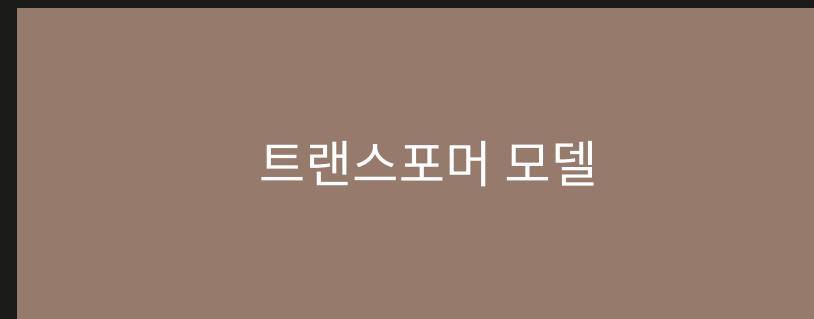


2단계

요약된 입력 데이터



예측 데이터(Answer)



Web 구현

01

DB 연결 :

question & answer 저장

02

모델과 연결: predict

입력값을 요약 모델에 넣은 후 트랜스포머 모델에 넣어 출력값 도출

웹에 모델을 제대로 연결하기 위해서 Java Script 이해 필요



기능 추가 & 성능 향상

01

멀티턴

이전의 대화를 기억하고 대답할 수 있도록 like GPT

02

질문 상황에 적합한 법률 제시

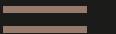
법령 데이터 크롤링 후, Seq2Seq 이용

03

계속해서 데이터 수집하며 모델 성능 향상

크롤링 과정에서 데이터 손실 최소화





Thank You : 시연
