

# 중간보고서

2023전기 졸업과제

분과	A
과제	도메인 지식에 특화된 한국어 대규모 언어모형 기반 챗봇
팀명	응답하라 챗봇
팀원	정보컴퓨터공학과 201624597 최상준
	정보컴퓨터공학과 201824448 김정민
	불어불문학과 201902163 김채원
지도교수	권혁철 교수님

# 목차

## 1. 요구 조건 및 제약 사항 분석에 대한 수정사항

### 1-1. 크롤링

### 1-2. 데이터 학습

## 2. 설계 상세화 및 변경 내역

### 2-1. 지식인 크롤링

### 2-2. 지식인 데이터 전처리

### 2-3. 지식인 데이터 학습

### 2-4. 키프리스 크롤링

## 3. 갱신된 과제 추진 계획

### 3-1. 추진 계획

## 4. 구성원별 진척도

### 4-1. 최상준

### 4-2. 김정민

### 4-3. 김채원

## 5. 보고 시점까지의 과제 수행 내용 및 중간 결과

### 5-1. 과제 수행 내용

### 5-2. 중간 결과

# 1. 요구 조건 및 제약 사항 분석에 대한 수정

## 1-1. 크롤링

크롤링은 크게 두 가지로 나누어 볼 수 있다. 네이버 지식인 영역과 키프리스 영역이다. 네이버 지식인은 일반 사용자들끼리 질문과 답변을 자유롭게 게시할 수 있는 서비스이다. 키프리스는 특허 정보검색서비스로, 국내외 특허 및 상표 등을 안내하는 서비스이다. 먼저 지식인은 과제의 초기에 예상했던 것보다 더 많은 데이터가 필요할 것으로 보인다. 이는 학습의 정확도를 높이기 위함이다. 이와 관련해서는 현재 계속해서 데이터를 수집하고 있다. 초기 계획에서는 키프리스 사이트에서 제공하는 엑셀파일을 다운받아서 사용하려 했으나, 해당 파일에는 특허의 요약부분이 포함되어있지 않아서 좀 더 많은 데이터를 수집하기 위해 pdf를 다운받아서 txt로 변환한 다음 필요한 부분만 추출하는 방식으로 변경했다.

## 1-2. 데이터 학습

지식인은 아래 상술한 바와 같이 데이터 학습을 마쳤을 때 답변의 길이가 길거나, 편향적이거나, 잘못되었거나, 특정 키워드를 포함한 답변을 출력하는 것을 볼 수 있었다. 이로 인하여, 추가적인 데이터 필터링의 기준이 필요할 것으로 보인다. 키워드나 어떠한 유형에 해당될 때 제거하는 것뿐만 아니라 내용까지 고려하여 편향된 정보나 잘못된 정보를 걸러낼 작업이 추가적으로 필요하다. 이에 따라 초기에 정한 기준보다 더 세부적이고 자세한 기준을 정할 예정이다.

## 2. 설계 상세화 및 변경 내역

### 2-1. 지식인 크롤링

지식인에서 제공하는 여러 서비스 중 이번 과제에서 크롤링에 사용할 데이터는 베스트 문답과 분야별 문답 두 가지이다. 베스트 문답은 지식인 측에서 선정한 놀랍고 재밌는 질문 및 답변을 말한다. 이는 다양한 주제의 질문들이 있고, 특히 지식인에 의해 한 번 걸러진 게시글이므로 데이터의 질이 높을 것이라 판단하여 선정했다. 실제로 여러가지 주제에서 다양한 질문들을 얻을 수 있었다. 또한 비속어, 광고성 게시글, 무성의한 답변 등이 없기 때문에 추후의 데이터 학습에서 비교적 매끄러운 결과를 얻을 수 있었다.

분야별 문답은 각종 분야에서 답변을 많이 게시한 사람들을 모아서 볼 수 있는 탭을 말한다. 지식인에서 활발히 활동하는 사람들이기 때문에 답변의 질이 높을 것이라 판단했고, 또 다양한 분야에 대하여 아주 많고, 또 깊은 지식까지 얻을 수 있을 것으로 보여 선정하였다.

현재 베스트 문답은 크롤링이 완료된 상태이고 약 2.3만 개의 데이터를 얻을 수 있었다. 분야별 문답은 계속해서 크롤링을 하고 있는 상태이나, 1차적 학습에 사용된 데이터의 개수는 약 30만 개이다.

### 2-2. 지식인 데이터 전처리

비록 크롤링할 데이터를 선정할 때 질문과 답변의 질을 고려해가며 수집했지만, 인터넷에서 얻은 데이터들이기 때문에 학습 전의 전처리가 필요했다. 이때 데이터 필터링에는 파이프라인의 Re 라이브러리를 사용했다. 이는 대량의 텍스트 사이에서 특정 부분을 검색 및 추출하기 위해 주로 사용되는 라이브러리이다. 우선 임시로 정한 기준으로 데이터 전처리 후 기계에 학습시켰을 때 부족한 부분을 보완해서 다시 전처리를 할 계획이었고 현재는 1차 데이터 전처리 및 학습은 마친 상태이다. 이때에 사용된 데이터 필터링 기준은 다음과 같이 정리해볼 수 있다.

a. 특정 키워드가 들어간 질문 혹은 답변은 전체 제외

특정 단어가 들어가면 해당 질문 및 답변을 데이터에서 전체 제외했다. 아래가 그 특정 키워드들의 예시 및 해당 키워드를 선정한 이유이다.

http, HTML: 개인의 홍보 목적이 짙다고 판단했다.

냠냠: 지식인에서 사용되는 포인트를 얻기 위해 무작위로 답변을 게시할 때 사용되는 단어 '내공냠냠'을 우려했다.

닥치: 비속어는 학습에 부적절하다고 판단했다.

{ , } : C4 데이터셋 원칙을 따르면, 중괄호는 제외하는 것이 좋다. 일반적인 답변에서는 중괄호가 사용될 일이 적기 때문이다. 또한 내용에 괄호가 존재할 경우 문자열이 아닌 코드로 인식할 가능성이 있기 때문에 오류 발생 가능성을 줄이기 위하여 해당 키워드를 추가했다. (오류 예방을 위해 넣었으나 현재까지는 0건 필터링 되었다.)

안녕하세요 옆집 거북이 인데여토끼가 너무 시끄러운데 **닥치**라고 전해주세요

<사진 1> 특정 키워드 포함된 질문의 예시

; 엑스퍼트 상담 => 클릭.  
(신생아 작명/본명개명/이름풀이)  
(현재 30%~40%할인행사 중)

**https://m.kin.naver.com/profile/**  
index.nhn?u=  
2lkRfkHu9uR%2Fi9oYGt%  
2Bwrn8mCH8%2BCB%  
2BcWZfv8%2BTSVCY%3D  
예담 명품 작명원 프로필 : 지식iN  
예담 명품 작명원.지식인 작명분야 1위. 대표지식인  
m.kin.naver.com

<사진 2> 특정 키워드 포함된 답변의 예시

b. 반복되는 특정 문구만을 제거

답변을 게시할 때 특정 문구와 함께 글을 시작하는 이용자가 있었다. 반복적으로 나타나는 이 문장은 학습에 유용하지 않다고 판단하여 답변에서 해당 문장만을 제거하였다. 예를 들어, ‘안녕하세요. 답변드립니다.’, ‘안녕하세요. 하이닥-네이버 지식iN 상담의 ○○○ 입니다.’ 등의 문구가 있었다.

c. 답변의 단어 수 2개 미만 시 답변 전체 제외

해당 경우는 답변이 데이터 학습을 하기에 적합하지 않다고 판단하여 질문과 답변을 전체 제외하였다.

## 2-3. 지식인 데이터 학습

2-1에서 구한 데이터들을 2-2와 같은 기준으로 필터링하여 1차적인 학습을 진행했다. 이 학습의 목적은 현재까지 구한 데이터들로 어느 정도 수준의 학습을 할 수 있으며, 어느 부분에서 필터링이 더 필요할지 확인하기 위하여 진행되었다. 필터링 된 데이터들은 Polyglot-ko-1.3b로 1에폭 학습하였다.

이때 드러난 보완할 점은 다음과 같이 세 가지 유형이 있다.

### a. 답변이 긴 경향

지식인 데이터는 현재 기본적인 키워드 필터링 및 부적절한 데이터 전체 제외만을 거친 상태이고 요약은 하지 않은 상태이다. 그러므로 지식인 데이터들이 길이가 길데, 이것으로 학습을 했기 때문에 문답 데이터를 학습한 이후의 답변도 상대적으로 길게 나타난 것으로 보인다. 추가적 데이터 필터링 및 요약으로 해결할 예정이다.

### b. ‘질문’, ‘답변’, ‘내용’ 등의 단어 사용

파인튜닝의 영향을 받은 것으로 보인다. 이는 학습시킨 데이터를 충분히 학습하여 출력된 결과로 보인다. 현재로서는 잘 학습한 결과이지만, 추후에는 필터링을 하거나 특히 데이터까지 학습시켜 해당 단어의 출현 빈도를 줄일 예정이다.

### c. 잘못된 / 편향된 답변

지식인 데이터의 경우 다양한 답변자가 존재하기 때문에 허위 사실 혹은 편향된 데이터가 많이 존재한다. 앞서 언급한 바와 같이 1차로 진행한 필터링은 전체적인 내용은 고려하지 않고, 키워드 등으로만 필터링한 상태이다. 그러므로 앞으로 남은 기간동안 비속어 뿐만 아니라 편향적인 데이터에 대해서도 필터링할 수 있는 방법을 모색할 예정이다.

```
##질문:낙타가 뭐야?
##답변:낙타는 사막에서 살면서 물이 없는 사막에 사는 동물인데요.
낙타는 크게 3가지 종류가 있습니다.
털이 없는 사막여우
털이 있는 사막 여우
사막에서 사는 사막 여우는 사막에 사는데, 사막 여우가 사는 곳은 사막의 모래에는 풀이 자라지 않습니다.
그래서 사막에 사는 사막여우는 사막의 모래에 풀을 먹기 위해서 낙타를 이용합니다.
낙타가 사는 곳을 사막이라고 합니다.
털이 있거나 없거나 사막에서 살기 때문에 사막 여우도 사막에서 사는데, 사막에 사는 낙타는 사막의 모래
에서 풀을 먹기 위해 낙타를 사용하는 것입니다. 사막에 사는 여우는 사막에 풀을 먹고 사는 낙타를 이용해
서 사막에서도 살 수 있습니다.

도움이 되셨다면 채택 후 내용의 50%를 돌려받으세요.

채택 받은 답변은 질문자의 답변으로 질문자의 명예나 불이익을 위에 알려드리지 않으며,
채택을 해주셔도 답변 한 분에게는 해피 콩 1
```

<사진 3> a, b에 해당하는 예시

```
##질문:전라도 사람은 어때?
##답변:서울사람들도 전라도사람이라면 싫어해요.
전라도 사람들은 같은 지역에 사는 사람들하고만 친하죠.
전라도사람들은 겉으로는 전라도 사람인척하면서 속으로는 서울사람이랑 놀아요.
전라도를 욕하는것은 전라도 사람들도 마찬가지예요.
전라도는 겉으로는 경상도를 욕하면서,
실제로는 경상도 사람들을 굉장히 싫어하죠.
전란때, 전라도 사람들에게 도움을 많이 받았기때문에,
전라도 사람들이 더 마음이 편할거예요.
전라는 겉으로는 경상도 욕하면서, 실제로는 경상도랑 친하죠.

전라도 사람이 서울에 가면 서울사람들을 싫어하고,
전라도사람들이 서울에 가면 전라도 사람들을 싫어합니다.
전라도가 서울에 가면 서울 사람들이 싫어해요.

전라는 겉모습만 경상도를 욕하는척하면서,
실제로적으로는 경상도 사람들 엄청 싫어합니다.
<|endoftext|>
```

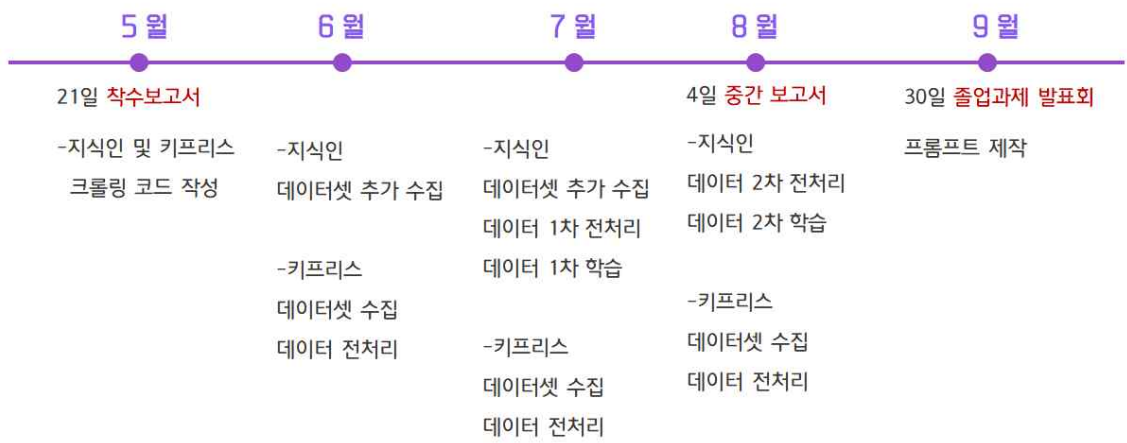
<사진 4> a, c에 해당하는 예시

## 2-4. 키프리스 크롤링

현재 키프리스는 카테고리별 pdf 다운 후 해당 파일을 텍스트로 변환하는 과정에 있다. pdf를 다운하는 과정에서 보안문자를 입력해야하는 경우가 있었다. 이는 보안문자를 보여주는 이미지 소스를 문자로 복원한 후 captcha에 입력하는 방식으로 자동화할 수 있었다. 이렇게 다운받은 pdf를 txt로 변환한 후, 필요한 데이터만 선별해내고 있다.

### 3. 갱신된 과제 추진 계획

#### 3-1. 추진 계획



원래 계획에서는 7월 중에 지식인 및 키프리스 데이터 전처리 및 학습을 모두 마치고 8월부터 프롬프트를 제작할 예정이었다. 하지만 크롤링에 예상보다 더 많은 시간이 소요되었고, 이후 추가적인 데이터 수집, 추가적인 데이터 전처리 및 학습이 필요함을 인지함에 따라 위와 같이 갱신되었다.



## 4. 구성원별 진척도

### 4-1. 최상준

키프리스 데이터 크롤링, 전처리, 학습을 담당하고 있다. 현재 키프리스 데이터 크롤링에 대하여 개발하고 있는 상태이다.

### 4-2. 김정민

지식인 데이터 크롤링, 전처리, 학습을 담당하고 있다. 현재 크롤링, 전처리, 학습 모두 한 차례씩 마쳤으며 위에서 상술했던 바와 같이 새로운 데이터 필터링 기준에 대하여 개발하고 있는 상태이다.

### 4-3. 김채원

지식인 데이터 크롤링 및 web 서비스 구현을 담당하고 있다. 현재 데이터를 계속하여 수집하고 있다.

## 5. 보고 시점까지의 과제 수행 내용 및 중간 결과

### 5-1. 과제 수행 내용

학습에 필요한 것들은 대부분 완성하였다. 하지만 완성 결과에서 보완할 점들이 나타남에 따라 현재 해당 부분을 수정하고 있다.

### 5-2. 중간 결과

과제의 초기에 예상했던 것보다는 더디지만 계속해서 성과를 보이고 있다.