**소프트웨어 설계 및 실험 보고서 (5조)**

**앱 명 : 혹시나도...?**

**(https://github.com/kkorona/HoxyNado)**

**201624476 박상운**

**201624440 김연재**

**201424545 지승원**

**제출 일자 : 2018/06/25**

**개요**

**앱 명**: 혹시나도(HoxyNado)

**분 류**: 웹 수집 및 검색 어플리케이션

**요 약**: 마이피누(mypnu.net) 반짝이 게시판에 올라온 글의 요소를 분석하여 DB에 태그를 추가하여 저장하고, 사용자의 신상정보 및 추가 입력 정보와 일치하는 요소들이 있는 글을 검색하여 보여주는 웹 기반 어플리케이션입니다.

**상 세**: 마이피누 반짝이 게시판에서 제목만 보고 본인에 대한 내용이 없다고 판단하는 경우가 많이 있습니다. 혹은 시간이 없어서 대충 훑어보거나 못 보고 넘어가는 경우도 있습니다. ‘혹시나도’는 이런 문제를 해소하고자 만든 마이피누 반짝이 게시판의 글들을 수집하여 제목과 내용을 분석해 사용자에 대한 언급의 유무를 판단하는 어플리케이션입니다.

‘혹시나도’는 마이피누 반짝이 게시판의 게시글을 크롤러를 이용하여 수집하고, 수집한 글에서 작성자가 찾는 대상의 특성들과 글을 올린 시간, 해당 글의 URL 등의 정보를 추출하여 태그를 달아서 DB에 저장합니다.

‘혹시나도’가 제공하는 웹사이트에서 사용자가 이름, 학번, 성별 등 기본적인 신상정보와 학과, 수강하는 과목 등의 추가적인 키워드를 입력하여 제출하면, 입력된 데이터를 기반으로 DB를 탐색하여 사용자 정보와 공통되는 키워드를 가지는 게시글을 검색합니다. 검색되어진 게시글은 마이피누 반짝이 게시판에 올린 시간 순으로, 즉 최신 순으로 정렬되어 사용자에게 제공됩니다. 이때 DB에 저장한 URL을 활용하여 출력된 게시물의 제목을 클릭하면 반짝이 게시판으로 넘어가서 댓글을 달 수 있습니다.

**프로그램 기능**

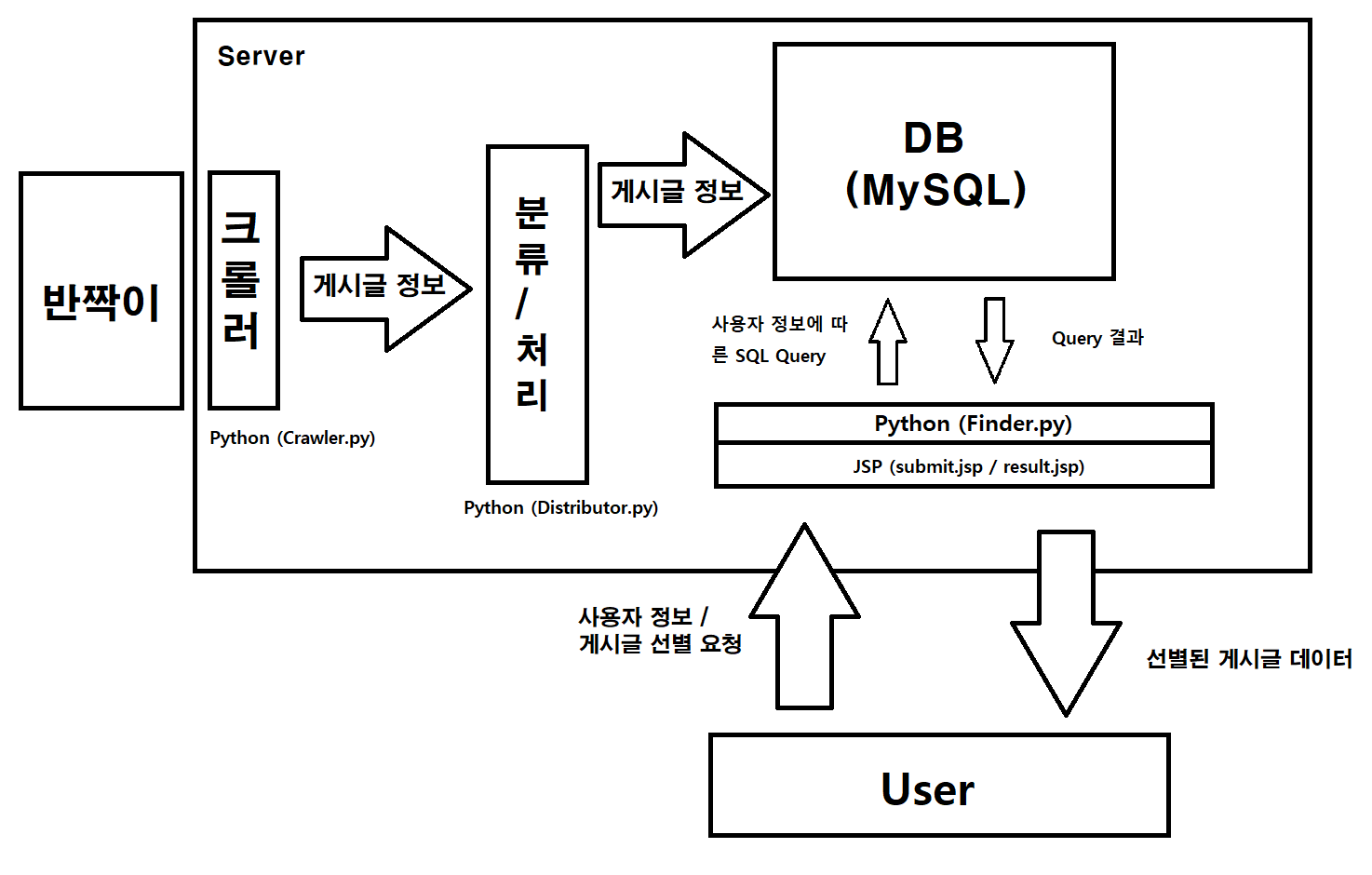
**1. 게시글 수집 및 분류**

웹 크롤링을 통하여 마이피누 반짝이 게시판의 글을 수집합니다. 수집한 게시글은 탐색 대상의 초성, 학번, 학과 등의 다양한 조건으로 분류되어 DB에 저장됩니다. 이를 통해서 각각의 게시글이 가지고 있는 특성 등을 알 수 있으며, 특정 조건을 가지고 있는 게시물을 쉽게 찾아낼 수 있습니다.

**2. 사용자 정보 기반 게시글 탐색 기능**

웹을 통해서 전달받은 사용자 정보는 서버에 전달되어 DB검색을 위한 키워드로 사용됩니다. 검색 결과는 게시글 ID / 게시글 제목 / 게시글 날짜의 정보 묶음으로 추출되어 사용자에게 제공됩니다.

**파이프라인 구조도**



1. 백엔드 전부 Python (DB 접근 : PyMySQL, Article Collector로 Script Data 작성)
2. 불편하게 직접 만들지 말고 PhoneGap 쓰기
3. 게시글 수정에는 투자하지 말고 Indent 사용

**파이프라인 구조 상세 설명 – 크롤링 모듈**

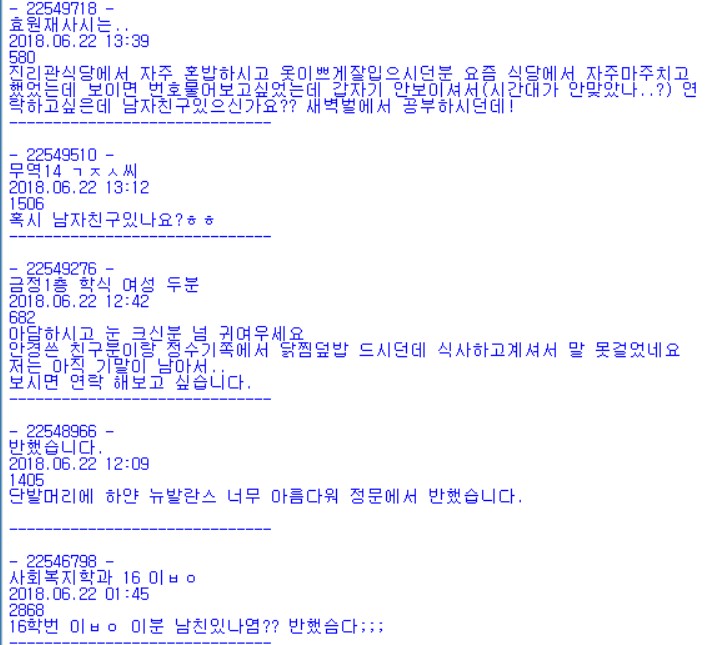
1. **게시판 및 게시물 크롤링**

<https://github.com/kkorona/HoxyNado/blob/master/crawler.py>

사용 언어: Python

사용한 프레임워크: codecs(파일 R/W), requests(웹 상호작용), BeautifulSoup4(HTML 파싱 자료를 Python 데이터화)

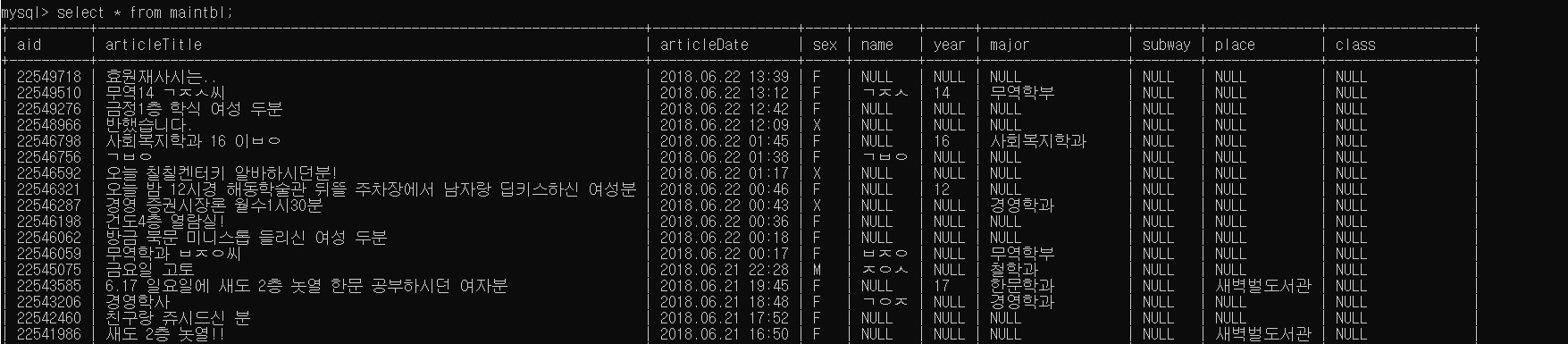
코드 개요 : requests 모듈을 이용하여 마이피누 사이트의 HTML 소스코드를 요청합니다. 이때 사이트에 처음 접속하면 게시판 페이지 대신 경고문 페이지가 뜨는 것을 방지하기 위해서 해당 조건을 결정하는 쿠키 변수인 **‘guidance\_sun\_d41d8cd98f00b204e9800998ecf8427e’**의 값을 true로 변경해줍니다. 이후 HTML 코드를 받아온 후에 BeautifulSoup로 파싱합니다. 파싱한 데이터에 대해서 게시글 ID, 제목, 내용, 날짜 등을 추출하여 분류 처리 모듈로 넘겨줍니다.



위 화면은 크롤링한 데이터들의 결과물인 result를 print한 결과물 중 일부이며 글의 제목과 내용들이 분류처리 모듈에서 분류가 될 것입니다.

**파이프라인 구조 상세 설명 – 분류처리모듈**

1. **크롤링 된 데이터 분류 처리**



우선 크롤링하여 나온 데이터들을 입력받습니다. 입력한 데이터를 분류 조건리스트에 저장된 내용들을 하나씩 반복하면서 분류를 진행합니다. 모든 분류조건들은 ‘ : ’ 으로 구분을 하고 해당 분류조건에서 줄이 바뀌기 전까지는 같은 소분류 조건으로 판단합니다. 만약 조건에 하나라도 부합한다면 같은 소분류조건의 처음에 나오는 조건을 메인으로 하여 element에 저장하면서 다음 분류조건으로 넘어갑니다. 성별의 경우 조건에 부합하지 못하여 element가 초기 값인 None가 그대로 유지 될 경우 ‘X’라는 값이 들어갑니다. 이렇게 분류되어 나온 값들은 처음에 분류되어 나올 값들을 저장할 튜플에 하나씩 저장이 되어 집니다. 이 튜플은 분류를 하지 않는 게시물ID(aid), 제목, 작성시간 및 날짜를 기본적으로 저장하며 분류되어서 나오는 값들을 하나씩 추가하여 최종적으로 10개의 값을 가지게 됩니다.

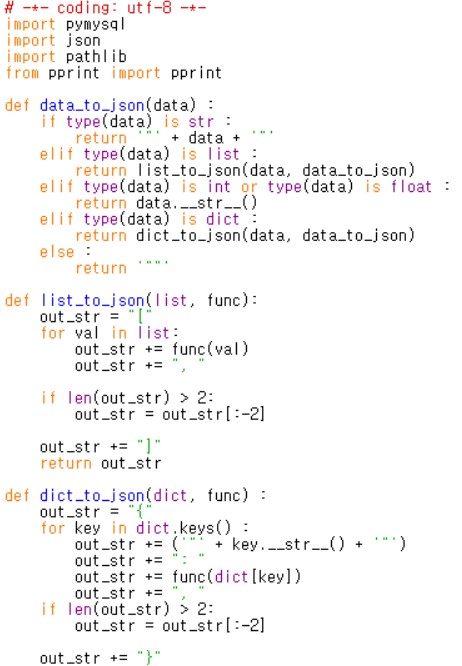
분류가 완료 되면 처음에 편의를 위해 만들어 놓은 conn을 사용하여 서버와 연결을 합니다. 그리고 마찬가지로 편의를 위해 만들어 놓은 sql을 사용하여 총 10개의 컬럼을 가진 maintbl에 문자열로 임시 저장용으로 만든 튜플의 값들을 저장합니다. 최종적으로 이러한 변경사항들을 commit하여 저장을 하고 서버와의 접속을 종료합니다.

만약 분류처리모듈을 모듈이 아닌 단독 파일로 실행 시 크롤링된 데이터가 없기 때문에 임의로 값을 추가하여야 합니다.

위 출력 화면은 Mysql의 서버에 있는 DB에서 분류된 값들이 저장되는 테이블인 maintbl을 출력하여 10개의 컬럼들의 값들이 출력이 된 화면입니다. 크롤링된 데이터는 위와 같이 분류 처리되어 최종적으로 DB에 저장이 되어 집니다.

**파이프라인에 따른 각 모듈 설명 – 검색모듈**

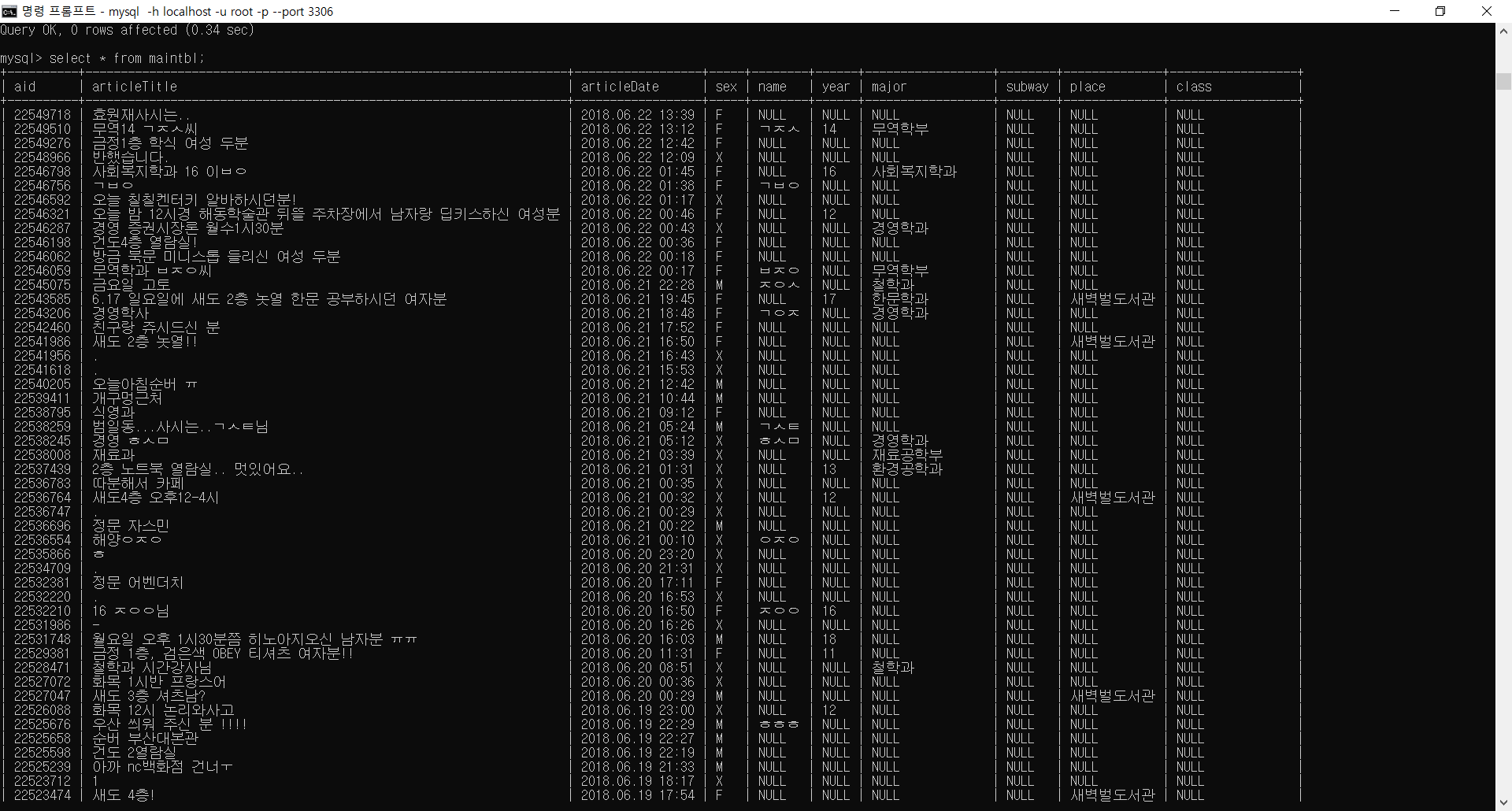
1. **DB에서 사용자로부터 받은 데이터를 검색**



...

**파이프라인에 따른 각 모듈 설명 – DB**

1. **크롤링한 데이터를 저장**

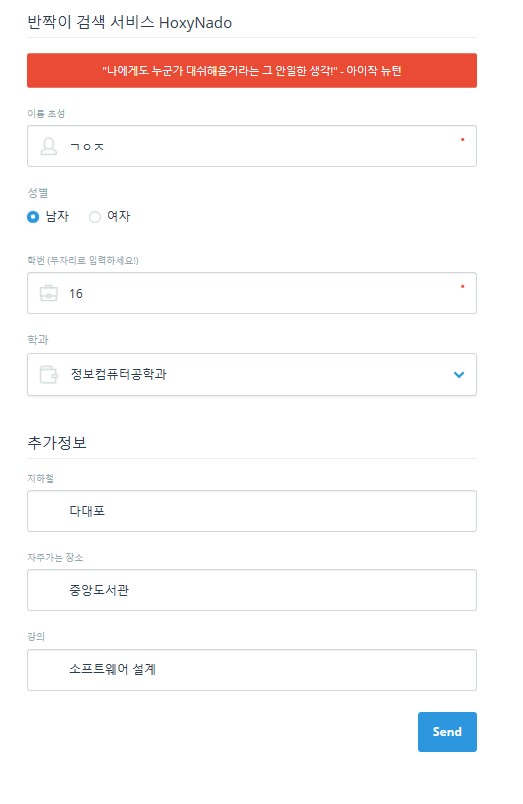
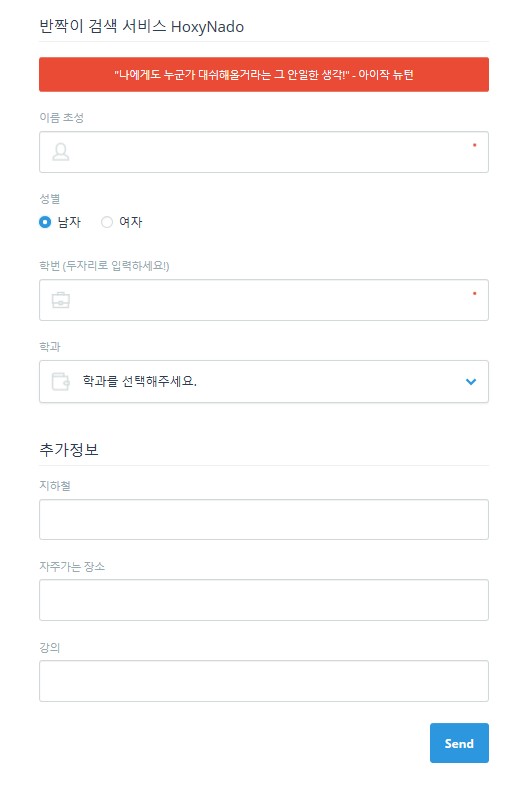


DB에 저장되는 내용들은 게시물에서 가져오는 aid(게시물 URL), 게시물 제목, 게시물 작성 날짜 및 시간뿐만 아니라 sex, name, year, major, subway, place, class 라는 총 7가지의 특성들을 저장합니다. 그리고 DB는 MYSQL을 사용하여 서버를 구축하고 DB를 구성합니다. Python을 사용하기 때문에 파일들은 pymysql을 통해 DB가 있는 서버로 접속을 합니다.

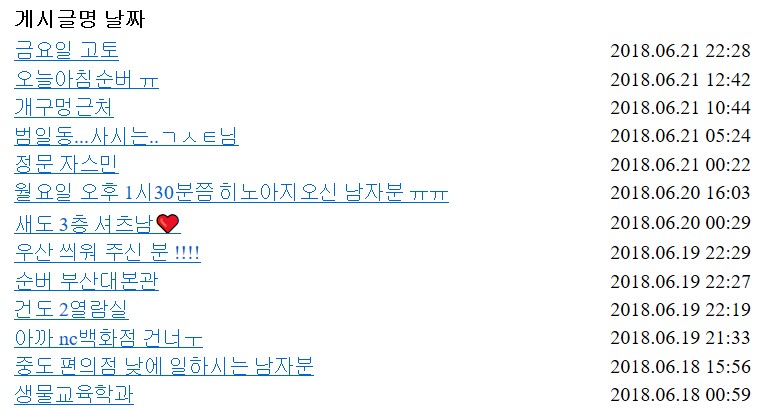
분류처리 모듈의 경우 pymysql로 서버-DB에 접속을 하여 분류 처리된 데이터들을 DB에 저장을 하고 검색 모듈의 경우 사용자가 어플리케이션에서 입력한 정보를 받아서 기존에 DB에 저장된 내용에서 검색하여 확인 후 일치하는 조건을 가진 게시물의 데이터를 DB에서 가져갑니다.

**실제 실행 예시**

실행 화면1 – 사용자가 정보 입력하는 화면



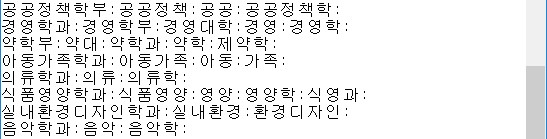
실행 화면2 – 사용자가 입력한 정보를 토대로 검색한 결과가 출력되는 화면



**개선해야하는 점**

1. 분류 조건 및 순서(?)

분류 조건 때문에 의도하지 않은 데이터가 분류되어 나올 수 있습니다. 현재 분류 조건들을 살펴보면 다음과 같습니다.



위 화면은 분류 조건 중 major의 일부분입니다. 학과 이름의 경우 줄임말로 사용하는 경우가 많아서 줄임말 또한 분류조건에 포함이 되어 있습니다. 이러한 줄임말들 중에서 실생활에서 학과이름 줄임말 이외에도 일상적으로 사용되는 언어들 또한 있습니다. 예를 들면 ‘공공시설’에서 공공이 분류되어 공공정책학부가 나올 수 있고 ‘음악을 들으면서 가는 여성분’ 이라는 글에서 음악에서 분류되어 음악학과로 분류되어 질 수 있습니다. 학과 이외에도 이름 초성에서 ㅋㅋㅋ라던가 ㅎㅎㅎ같이 웃음으로 주로 쓰는 초성들 또한 마찬가지로 이름으로 분류되어 질 수 있습니다. ㅋㅋㅋ같은 경우 사람이름으로 쓰이는 경우가 없다고 할 수 있지만 외국인 이름의 한글이 ㅋㅋㅋ가 될 수 있으며 ㅎㅎㅎ또한 사람이름에서 사용되는 초성이기 때문입니다. 이러한 가능성 때문에 예외적으로 해당 조건만 삭제 할 수 없다고 생각합니다. 이러한 부분은 더 많은 논의를 해서 해결을 해 나가야할 문제점 중 하나입니다.

이와 같은 문제와 연계되어 나타나는 또 다른 개선사항이 있습니다. 현 분류처리 모듈의 경우 조건에 부합하는 데이터가 있다면 그 자리에서 해당 조건의 분류를 종료합니다. 제일 처음 부합되는 데이터가 올바른 값이라면 전혀 문제가 없지만 글쓴이가 의도한 조건이 두 번째 혹은 그 이후에 나오는 값에 있다면 문제가 됩니다. 예를 들면 ‘어떤 남성분과 같이 가던 여성분’ 이라는 내용에서 글쓴이는 여자를 찾고 있으나 앞의 남성에서 이 게시물은 남자를 찾는 여자로 분류가 되어 해당 게시물은 어플리케이션을 사용하는 남성들에게 링크가 가서 남자와 남자가 연결되는 슬픈 일이 발생합니다. 이 문제는 대부분 분류조건에서 해결된다면 남녀 구별을 제외한 나머지의 경우는 해결이 가능할 것으로 예상되어집니다. 그러나 남녀 같은 경우에는 좀 더 자세한 분류과정이 필요할 것입니다.

**반성 및 느낀점 (내용 수정 안했음)**

