**1.프로젝트 명 : The Reason**

**2. 프로젝트 목표 / 프로젝트 주제를 정한 이유**올해 한이음 멘토링 과정을 통해 개발했고, 한이음 공모전에 출전한 ‘자기소개서 기반 모의면접 인공지능 챗봇 서비스’ 은 사실상 인공지능이 반영된 부분이 없었다.

주제는 참신하다는 평가를 많이 받았지만, 인공지능이라는 타이틀에 걸맞는 기술을 적용하지 못해 다소 아쉬웠던 프로젝트다.

공모전 기간내에는 적용하지 못 했지만, 평소 개발 기간동안 시간이 된다면 꼭 적용해 보고 싶었던 주제를 위해 현재 DataCrawling 과목을 수강신청 했다. 다행이도 정확하게 수강신청 당시 원했던 주제를 다룬 수업이라 배운 내용들을 평소 적용해 보고 싶었던 부분에 적용해보려 했다.

따라서 기존의 MBTI성향에 따라 성향점수를 분류 해 자기소개서가 나타내는 MBTI성향에 따라 Modeling 되던 방식에서, 비지도 학습을 통해 Clustering 된 분류방식으로 Topic Modeling하여 알고리즘을 수정할 예정이며, 분류되는 개수를 8개부터 2개까지 줄여가며 가장 뚜렷하게 구별되는 개수로 분류했다.

**3. 프로젝트 수행을 위해 학습한 내용**

**Bag Of Words**

* 단어들의 순서는 고려하지 않고, 단어들의 출현 빈도(Frequency) 에만 집중하는 텍스트 데이터의 수치화 표현방법이다.

**Konlpy**

* 한글 자연어 처리를 위해 제공되는 Python Library로 한글의 형태소 분석을 위한 라이브러리이다.

**Unsupervised Learning(비지도 학습)**

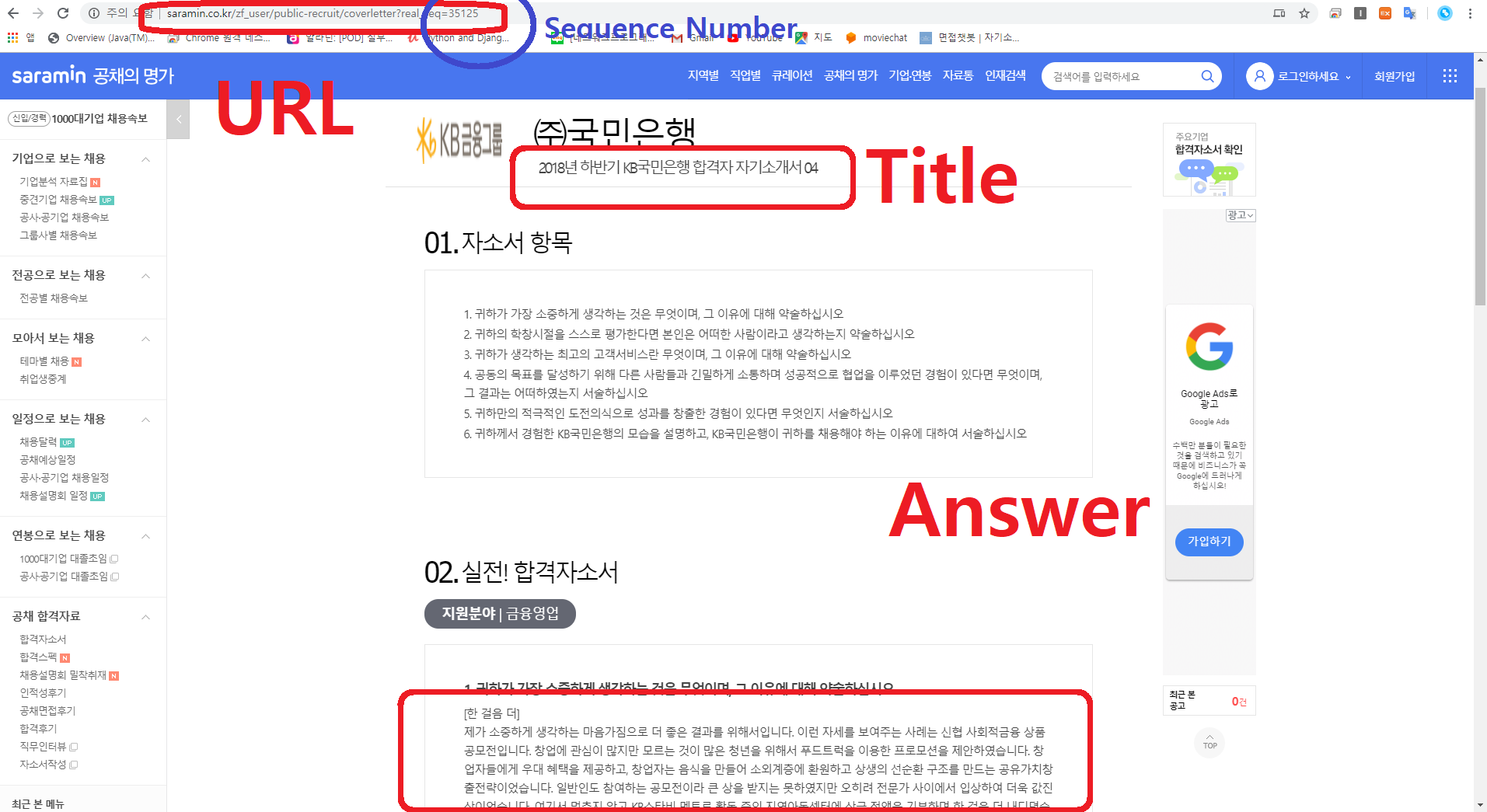
* 기계 학습의 일종으로, 데이터가 어떻게 구성되었는지를 알아낼 때 주로 사용하며, 입력 값에 대한 목표치가 제공되지 않고, Clustering(군집화) 를 하는데 유리하다.

**DataCrawling**

* 많은 량의 데이터를 분류하는 프로젝트기 때문에, 다량의 자기소개서를 읽어 오기 위해 빠질 수 없는 기술이다.

**4. 데이터 수집(Crawler.py)**

구인구직 사이트에 게시되어 있는 합격 자기소개서의 내용을 html 코드에서 **제목**에 해당하는 부분과 **답변**에 해당하는 부분을 Crawling해 csv 파일로 저장한다.



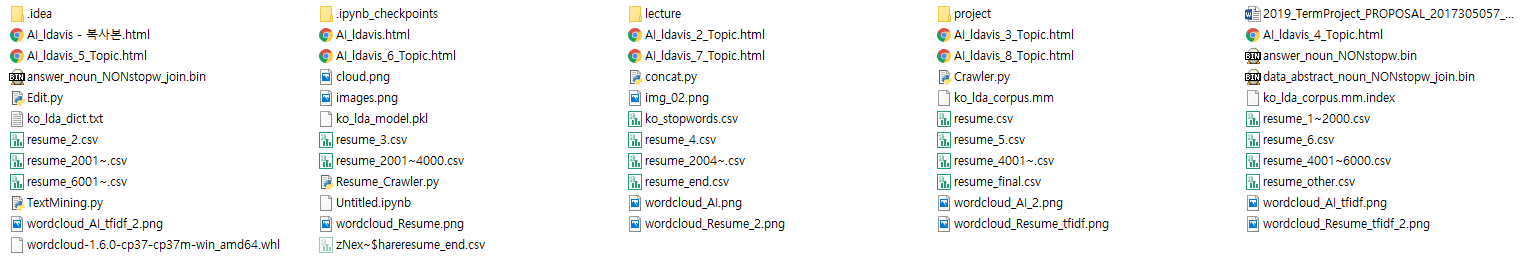
URL은 Sequence Number를 1부터 Count()를 사용해 마지막 글인 36000번째 Index 까지 크롤링했다. 총 작성되었던 글의 개수는 약 36000개 이상인 듯 한데, 실제로 명시된 자기소개서의 개수가 7300개이고, 크롤링 된 자기소개서의 개수가 6600인 것을 보면 삭제되거나, 비공개 처리된 글들이 있는듯 하다.

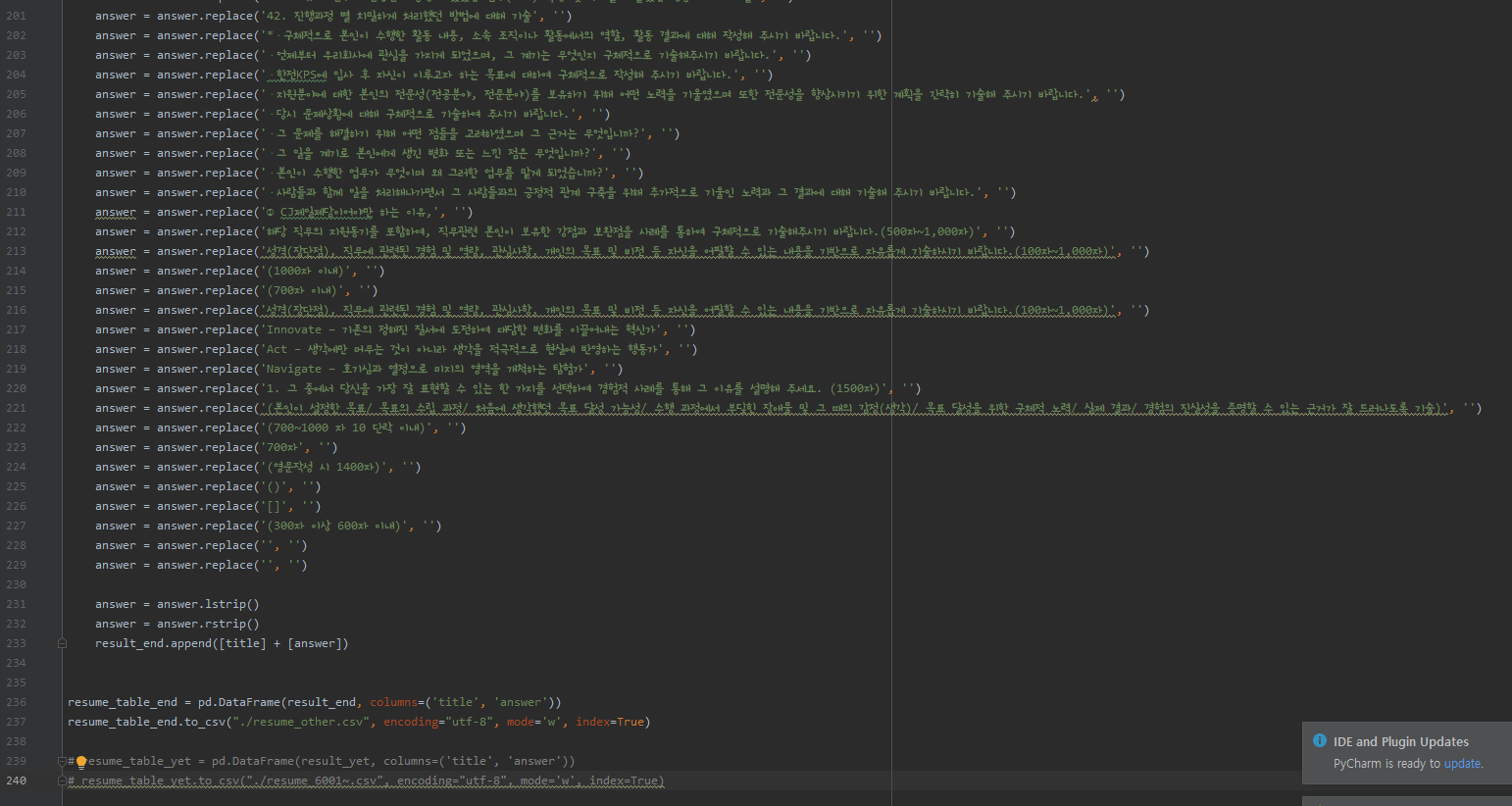
Csv 파일에 저장되는 DataFrame의 열은 Title와 Answer 로 총 2열이기 때문에 복잡하지 않으나, 자기소개서의 답변 내용이 평균적으로 5000자 정도 되기 때문에 꽤나 많은 시간이 소요 되었다.(약 6 ~9시간)

**5. 데이터 준비 및 탐색 작업**

빅데이터라는 학문은 인내의 학문이라는 교수님의 말씀대로 Crawling 하는 과정에서 우선적으로 전처리를 한 후에 저장되도록 했지만, 자고 일어나 저장된 데이터를 보니 역시나 불필요한 데이터가 많았다.

불필요한 데이터를 위해 필요없는 문장들과 기호를 replace 함수를 활용해 대체하다보니 230 Line이 넘었다.

****

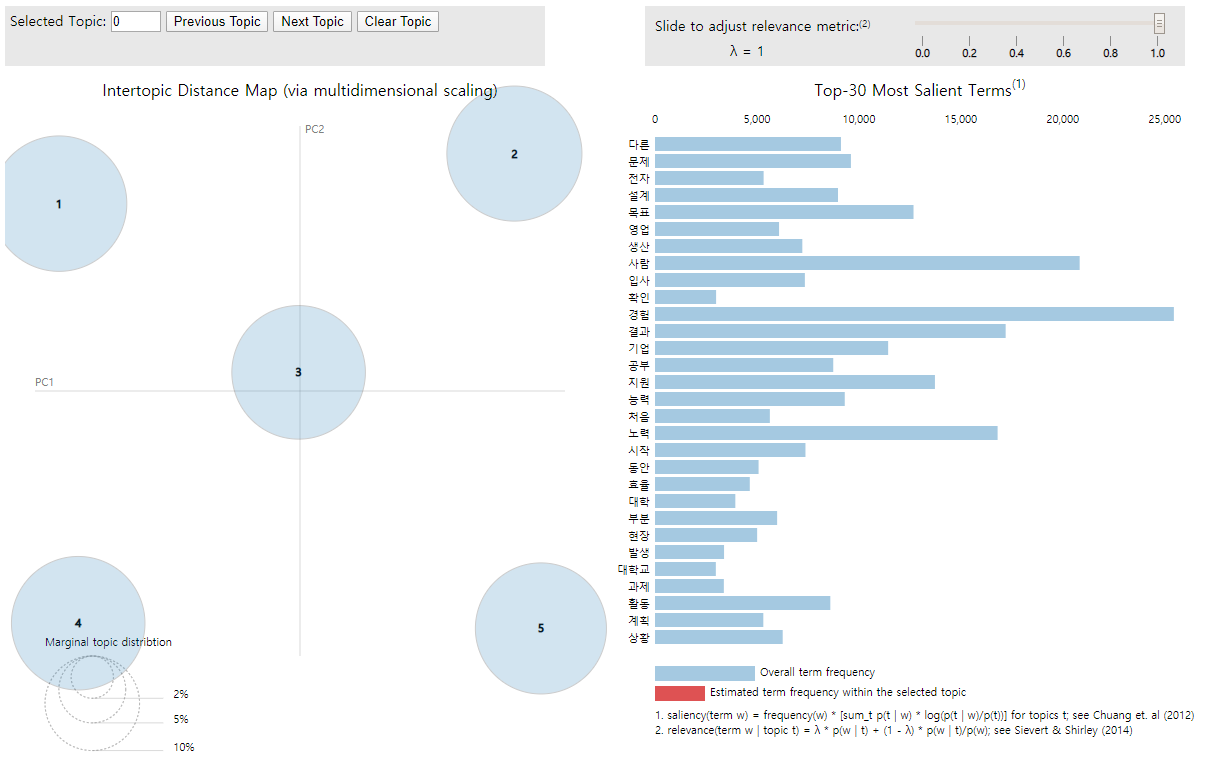
****

**6. 분석 모델**

|  |  |
| --- | --- |
| **8 Topic** | **7 Topic** |
|  |  |
| **6 Topic** | **4 Topic** |
|  |  |
| **5 Topic** | |
|  | |
| **3 Topic** | |
|  | |
| **2 Topic** | |
|  | |

**7. 결과**

뚜렷하게 구별되는 군집화 모델 중 가장 다양하게 분류된 **5 Topic**으로 Topic Modeling을 진행하기로 결정했다.





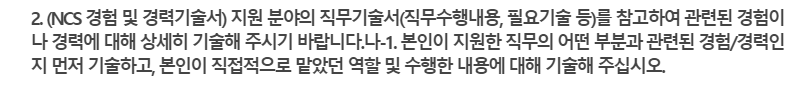
**8. 프로젝트를 진행하는 동안 어려웠던 점 & 발생했던 문제 + 해결 방법**

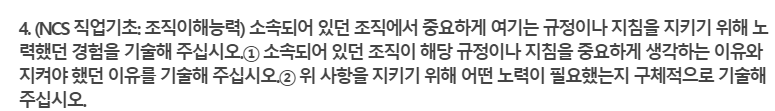
자기소개서의 Text가 너무 많아 Crawling하는데 많은 시간이 소요되었다.

* 비교적 사용자가 적은 새벽 시간을 이용함

자기소개서의 질문에 대한 부가설명과 제한 조건이 모두 달라 원하는 정보만을 Crawling하는 것이 사실상 불가능하다는 생각이 든다.

* 인내심 발휘한 전처리(Preprocessing)





저장된 자기소개서의 전처리를 위해 csv 파일을 Load해두었을 경우, Text의 양이 너무 많아 Memory에 부담이 되어 전처리를 하는데 제한이 되었음

* 2000개 단위로 분리 한 후 전처리하고, 다시 Concat했음