Data Mining-Project

2018314848

한승희

인공지능융합전공

1. Problem Definition
   1. Topic

코로나 발생 이후 대학생들의 관심사 변경 분석

* 1. Background

2019년 12월 중증급성호흡기증후군 코로나바이러스 2(SARS-CoV-2)로 지정된 한 바이러스가 중국 중부 우한에서 온 환자들에게 원인 불명의 폐렴으로 처음 확인됐다. 코로나바이러스 2019년 1차 보고 이후 전 세계적인 감염병으로 번져 대규모 감염이 발생했다. 현재 ‘코로나 바이러스’는 개인의 사소한 생활습관부터 국가 경제, 외교상황까지 영향을 미치지 않는 곳이 없을 정도로 모든 것을 바꾸어 놓았으며, 전 세계적으로 가장 관심이 많은 주제이다. 그 영향은 우리 대학생들에게도 큰 영향을 미쳤으며, 코로나의 직접적인 영향(확진자수 증가 등)과 간접적인 영향(확진자 수를 줄이기 위해 비대면 강의 실시)이 각 시기에 따라서 학생들의 생각을 어떻게 바꾸어 놓았는지 분석하고자 한다. 현 시기에 이러한 프로젝트를 진행하는 것은 최신 데이터를 이용하기 때문에 데이터의 양과 질이 좋을 것이기 때문에 가치 있는 프로젝트가 될 것으로 기대한다.

* 1. Plan

먼저, 날짜별로 코로나 누적 확진자수, 사망자수, 격리해제 수 등을 기록한 데이터가 필요할 것이다. 이 부분은 어렵지 않게 구할 것으로 생각된다. 또한 엄청나게 많은 수의 대학생들의 생각을 알기 위해서, 학생들이 주로 사용하는 SNS(Facebook, Twitter, Instagram 등등)이나 커뮤니티(에브리타임 등)에서 데이터를 얻은 후의 데이터의 질을 비교하여 더 나은 방법으로 진행할 것이다. 한국어 전처리를 위해 다양한 한국어 전처리 패키지를 사용해서 더욱 성능이 좋은 방법으로 진행할 것이다. 미리 분류된 train set이 없기 때문에 지도학습으로 어떠한 종류의 글인지를 분류하는 것은 대단히 어렵고, 한국어의 특성상 특정 어휘의 등장만으로 그 글의 종류를 결정하는 것은 어렵기 때문에 비지도학습으로 글들을 Clustering하여 분석하는 방법을 사용할 것이다.

* 1. Expected result

각 시기별로 가장 큰 뉴스에 따라 대학생들이 관심있어 하는 주제가 Dynamic하게 바뀔 것으로 생각한다. 예를 들어 신천지 사태 이후로는 대구, 신천지, 각종 비하 표현을 포함한 글이 많이 나타날 것이고, 확진자가 거의 발생하지 않는 구간에서는 오프라인 개강에 대한 논의가 많이 이루어 질 것으로 예상한다.

1. Data Collections
   1. Corona

현재 코로나만큼 뜨거운 이슈를 갖고 있는 주제가 없기 때문에, 이에 관한 데이터를 구하는 것은 어렵지 않았다. 본인은 <https://coronaboard.kr/> 이 사이트에서 정보를 구했으며, 엑셀파일로 저장한 후 이용하였다

* 1. Crowd

데이터는 selenium과 chromedriver를 이용한 웹 크롤링 방식을 이용해 구했으며, 일상의 사진을 공유하는 SNS인 instagram보다 오직 대학생들의 의견만 수집할 수 있고, 익명이 보장되어서 자신의 의견을 좀 더 자유롭게 말할 수 있는 익명 커뮤니티 “에브리타임”에서 정보를 수집했을 때 정보의 질이 좋았다. 웹 크롤링 당시 날짜(6/6)부터 한국에 첫 확진자가 발생한 날짜(1/21)까지 날짜, 게시판, 글 내용, 댓글list의 정보만 추출하였다.

1. Data preprocessing
   1. Bulletin board

성균관대학교 에브리타임에는 수많은 종류의 게시판이 있었다. 그 중에서는 시기에 따라 변화하는 학생들의 의견을 알기 힘든 게시판들이 꽤 있었다. (장터 게시판, 홍보 게시판 등등) 따라서 범위를 자유게시판종류로 한정하고, 이에 속하지 않는 게시판의 글들은 제거해주었다.

* 1. Feature extraction

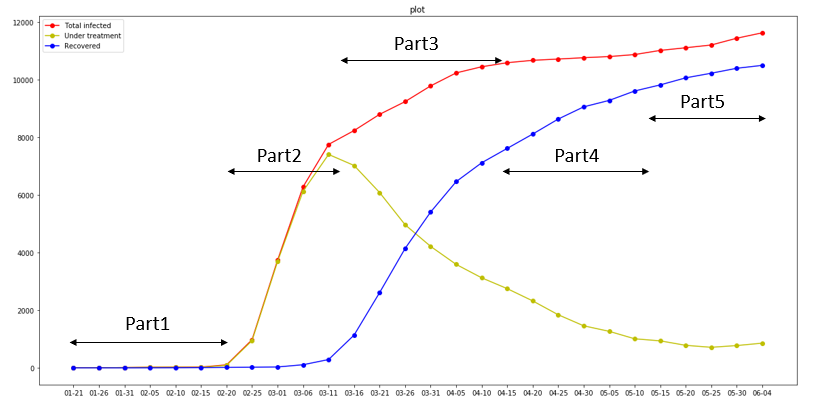
우리가 추출해온 “date” column은 문자열 데이터로 시기별로 파트를 나누는 데에 적합하지 않으며, 그 날의 몇 시에 글이 작성되었는 지는 본 프로젝트에서 필요없는 정보이다. 따라서 date에서 month와 day 데이터를 추출해서 새로운 column으로 추가시키고, 계산의 편리함을 위해 데이터 타입은 정수형으로 지정한다

* 1. Korean preprocessing

sonlp패키지에 있는 normaizer를 통해 공백이나 특수문자 등을 제거해주고, ㅋㅋㅋㅋㅋ과 같은 반복되는 것을 하나로 통일시켜 준다. 또한 인터넷 익명 커뮤니티이다 보니까 띄어쓰기 형식이나 맞춤법 등을 철저히 지키지는 않기 때문에 형태소 분석기로 konlpy의 Okt를 이용하였다. Okt를 이용해 명사만 추출하여 텍스트를 정리하였다.

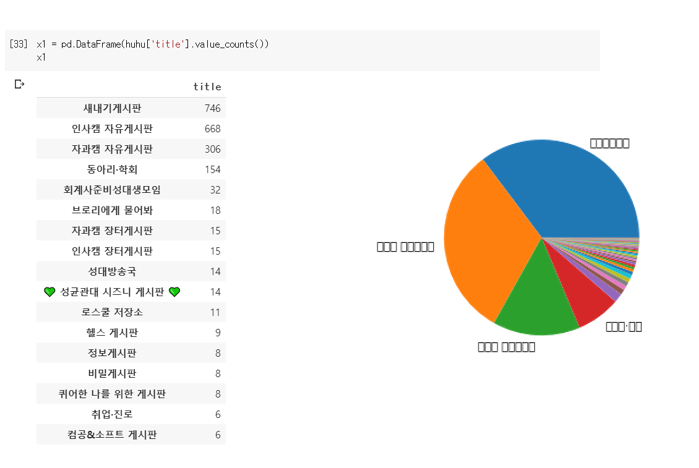
* 1. Part division

코로나 발생이후 초반, 중반, 후반으로 구분하지 않고 확진자 증가율의 변화를 기준으로 삼아 총 5개의 파트로 나누어 시기를 분류하였다.



1. Data analysis
   1. Bulletin board

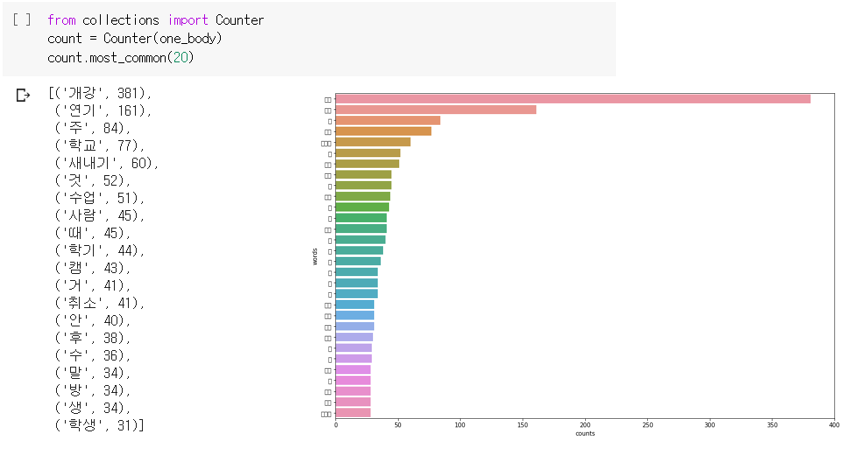
분석을 시작하기에 앞서 게시판 별로 얼마나 많은 글을 쓰는지, 즉 어느 학생들이 다수에 속하는지 등을 알아야 할 필요가 있었다. 이는 pandas문법인 value\_counts로 어렵지 않게 구할 수 있었다.



확인해 본 결과 새내기들의 지분이 가장 큰 것임을 알 수 있었고, 자연과학캠퍼스의 학생들은 커뮤니티를 많이 이용하지 않는 것 같았다. 여론이 형성되는데 새내기의 영향이 가장 클 것이라고 짐작할 수 있다.

* 1. Word count

한국어 전처리가 잘 되었는지, 어떤 단어들이 많이 등장했는지를 확인하기 위해서 collections패키지의 Counter를 통해 상위 20개 단어 빈도수를 확인하였다.

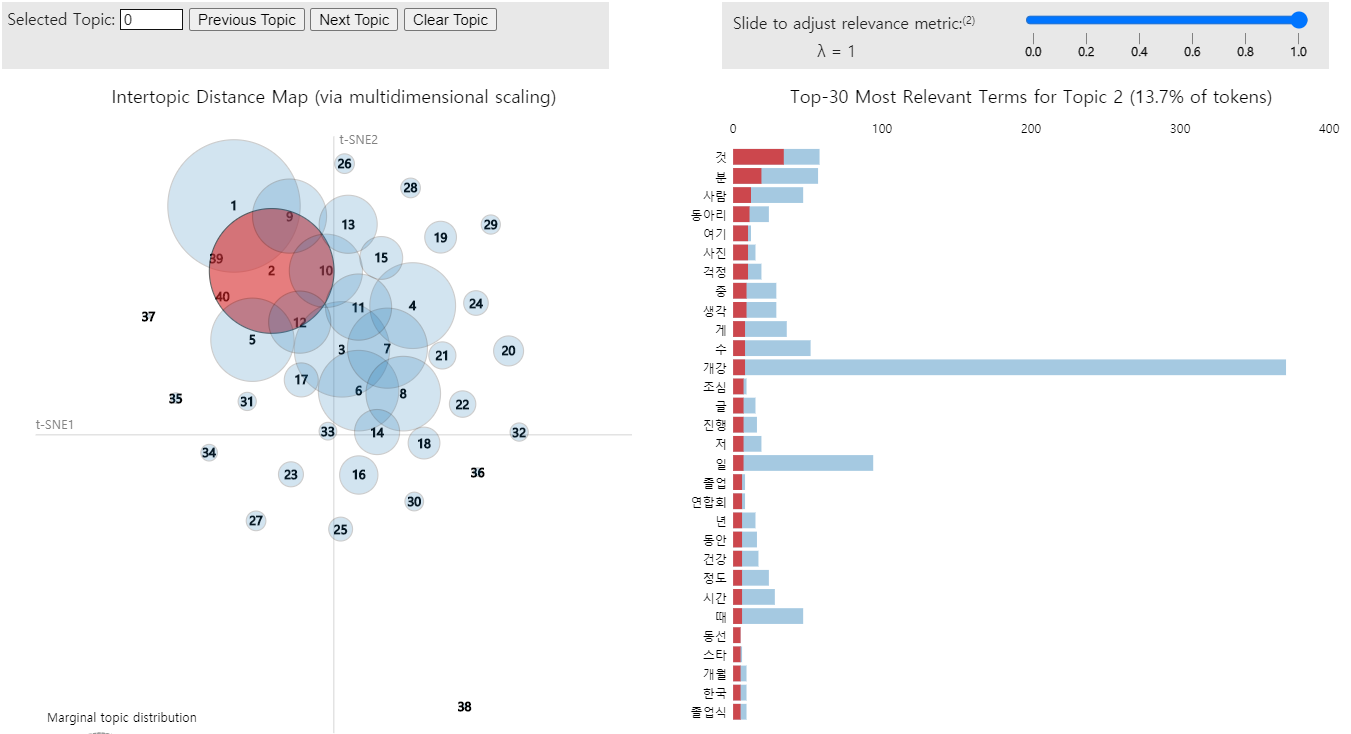


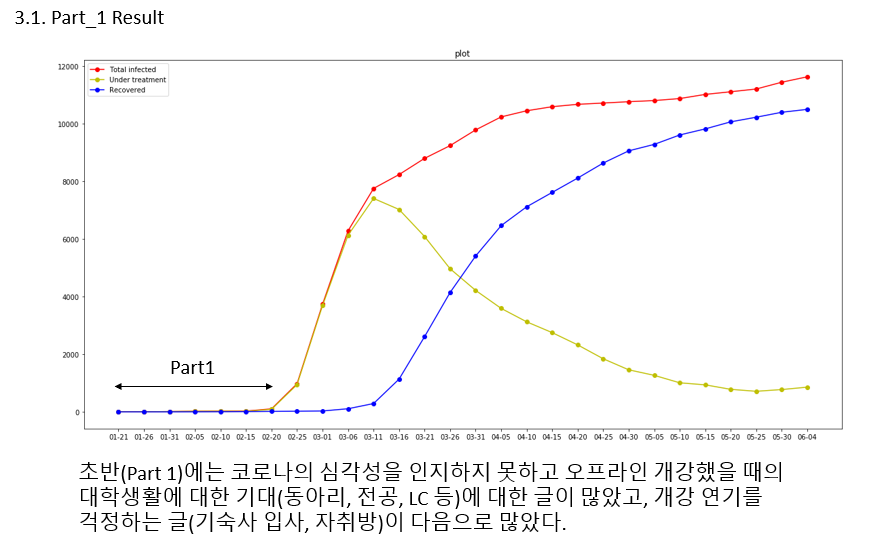
인터넷 글이라 형식이 자유분방해서 잘 분류하지 못할 것 같아 불안한 마음이 있었지만, 필요한 정보를 얻는 데에는 부족함이 없을 정도로 잘 전처리 한 것 같다.

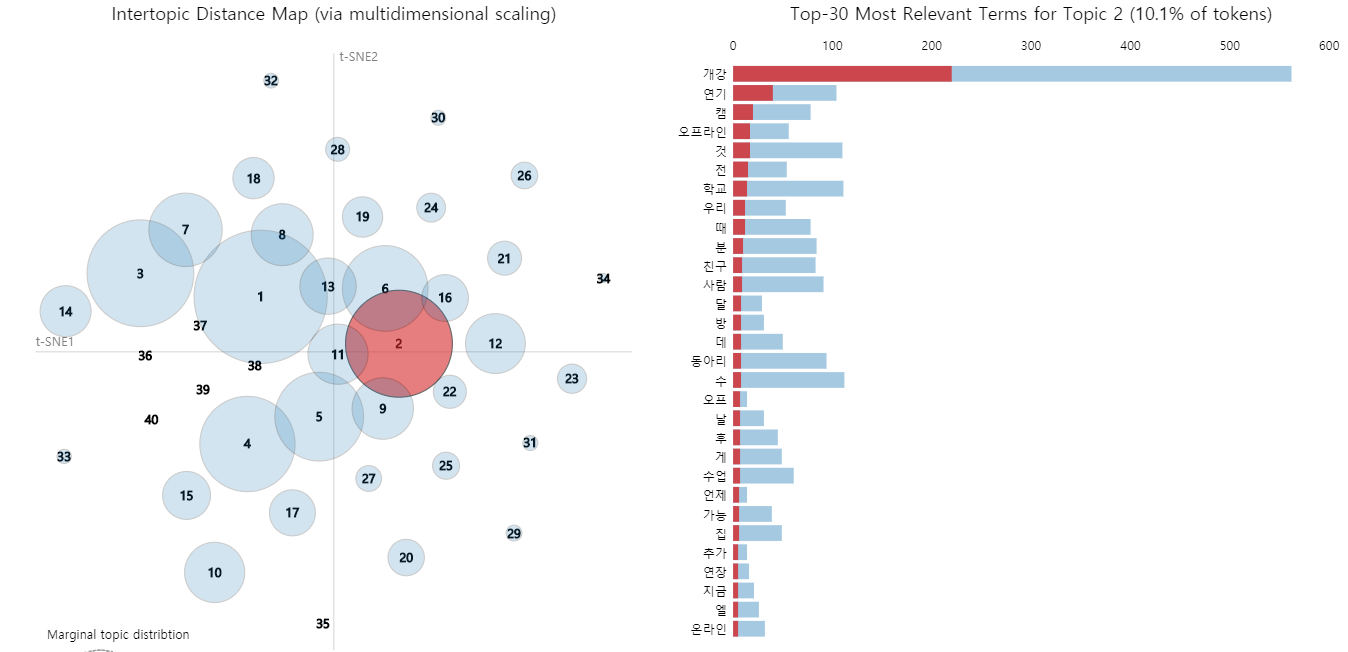
* 1. Clustering

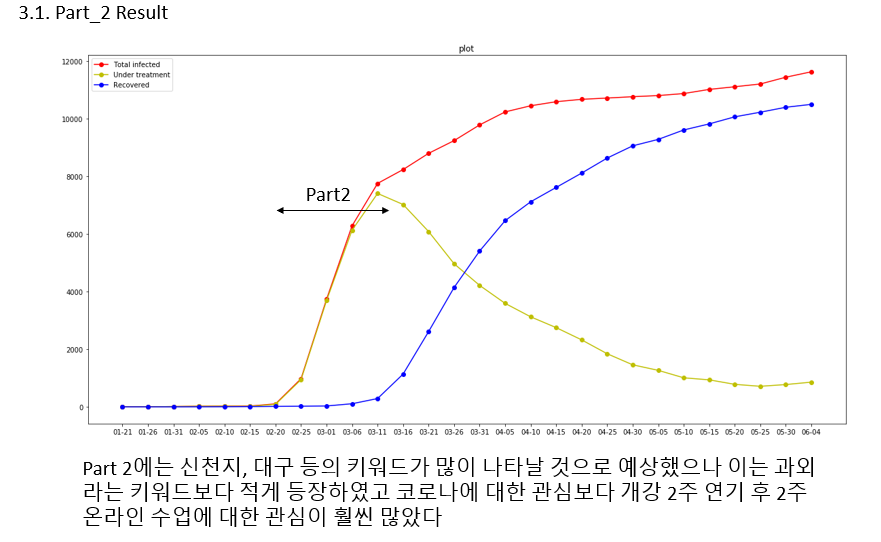
비지도 학습인 K-means clustering을 이용해서 어떤 키워드들이 어떻게 분류해 있는지 분석하였다. 전체 데이터를 이용하지 않고 다섯개의 파트들을 각각 데이터개수에 맞춰 Clustering 하였으며 pyLDAvis를 이용해 시각화 하였다.

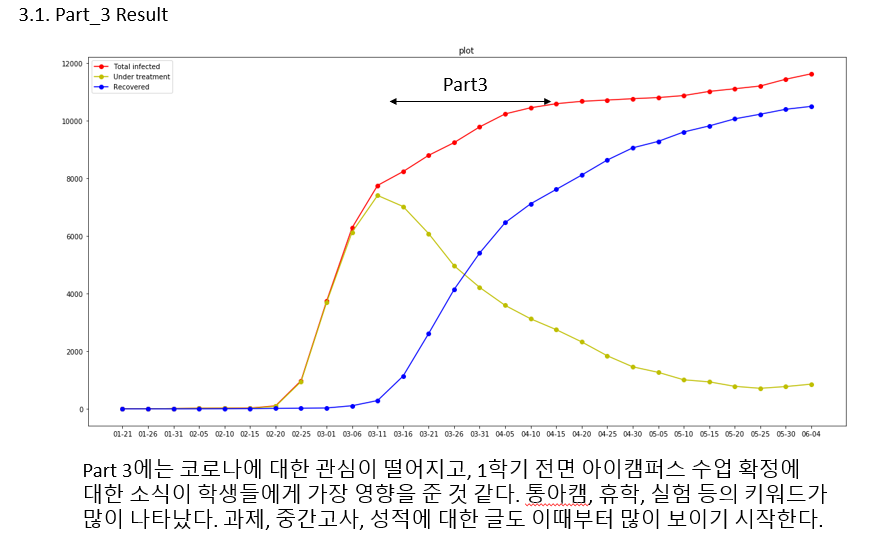
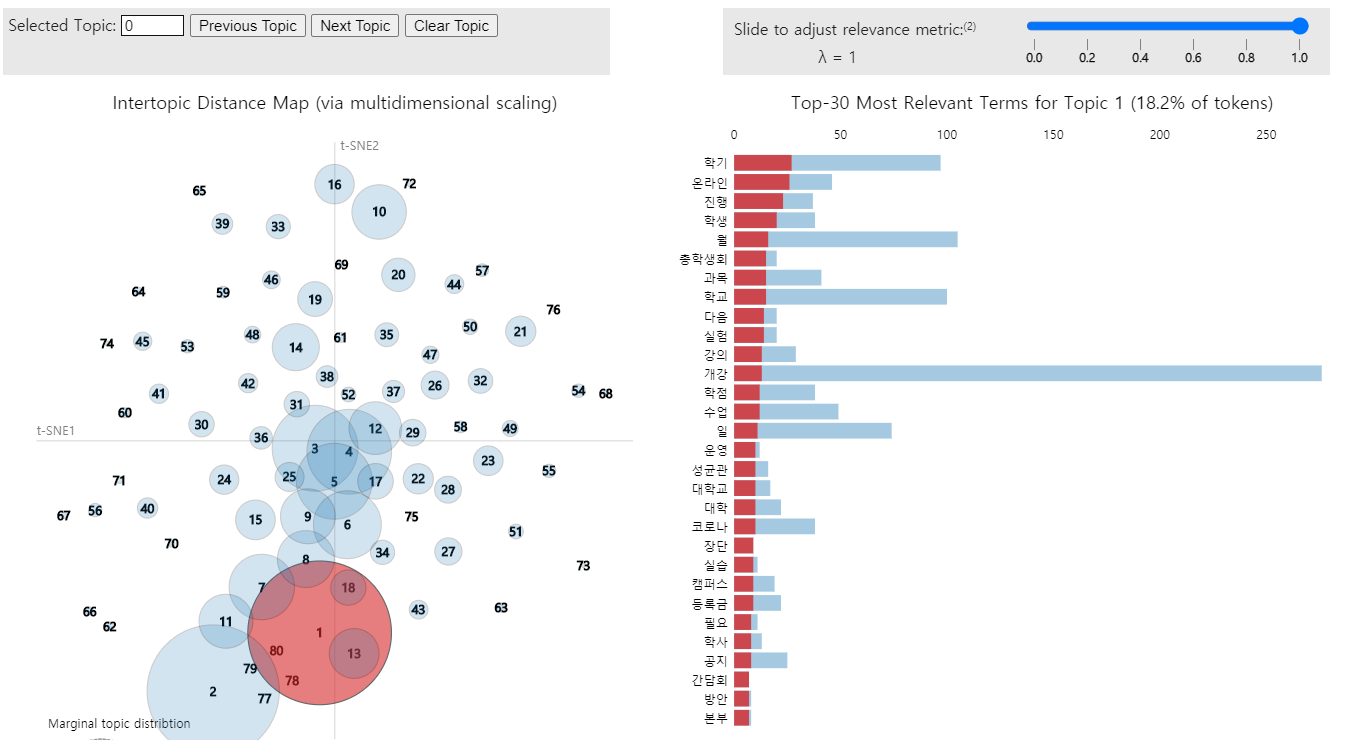
다음은 다섯개의 파트를 순서대로 분석한 결과를 나타낸 것이다.

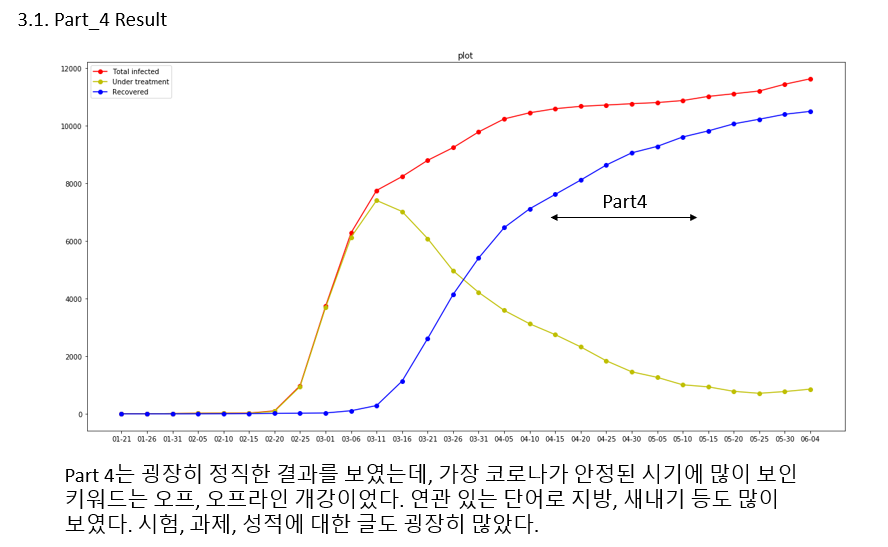
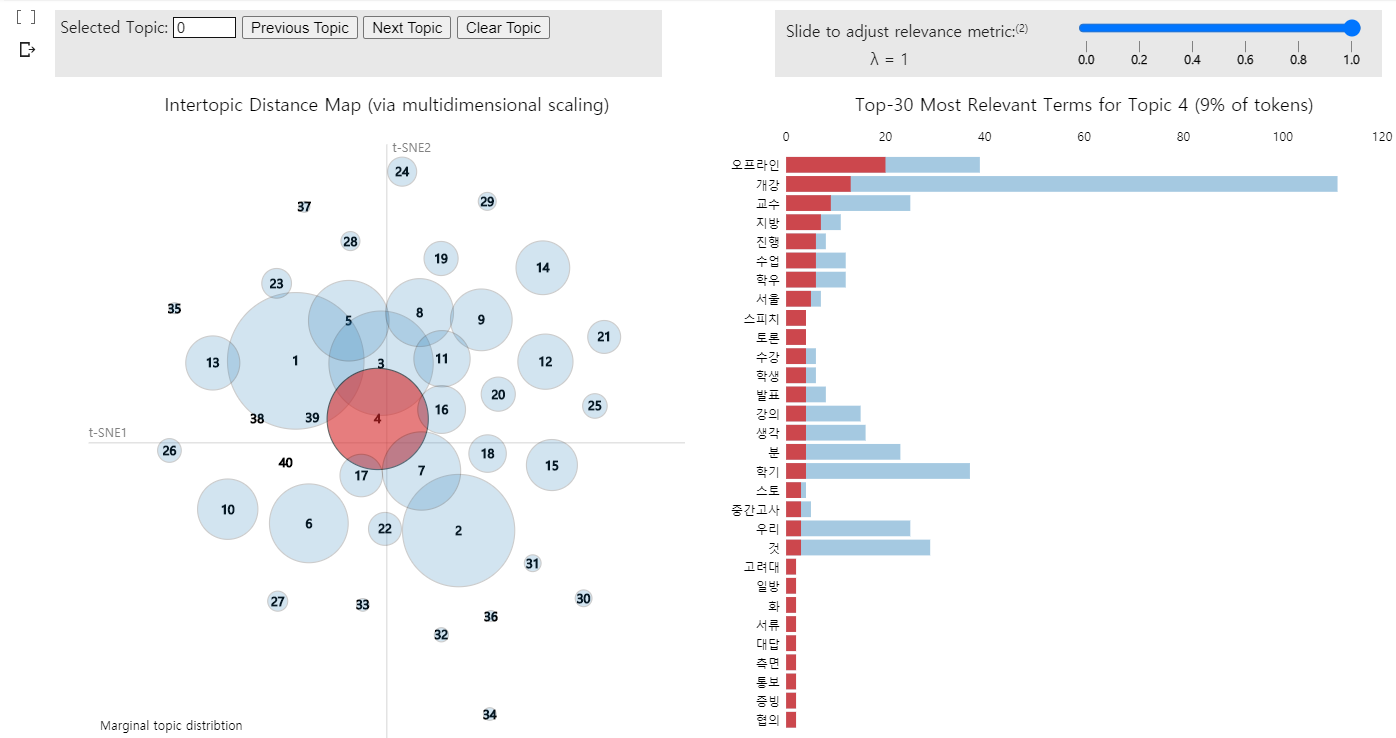


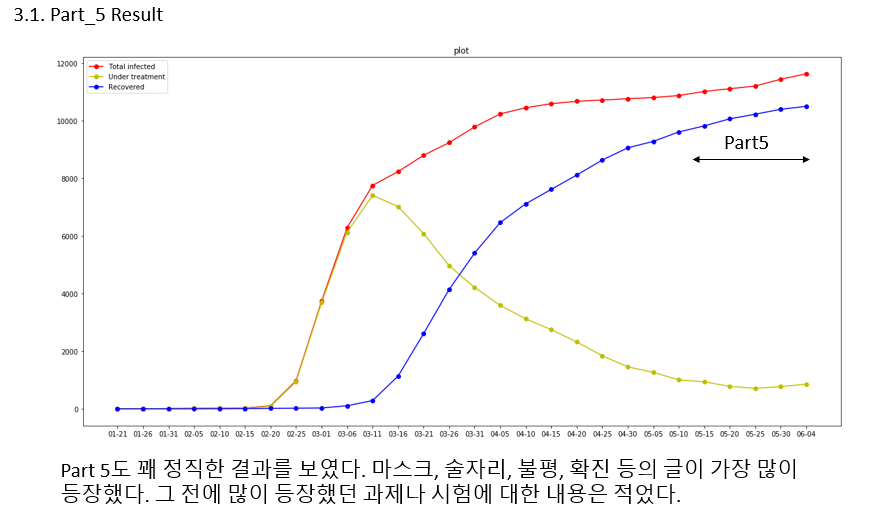
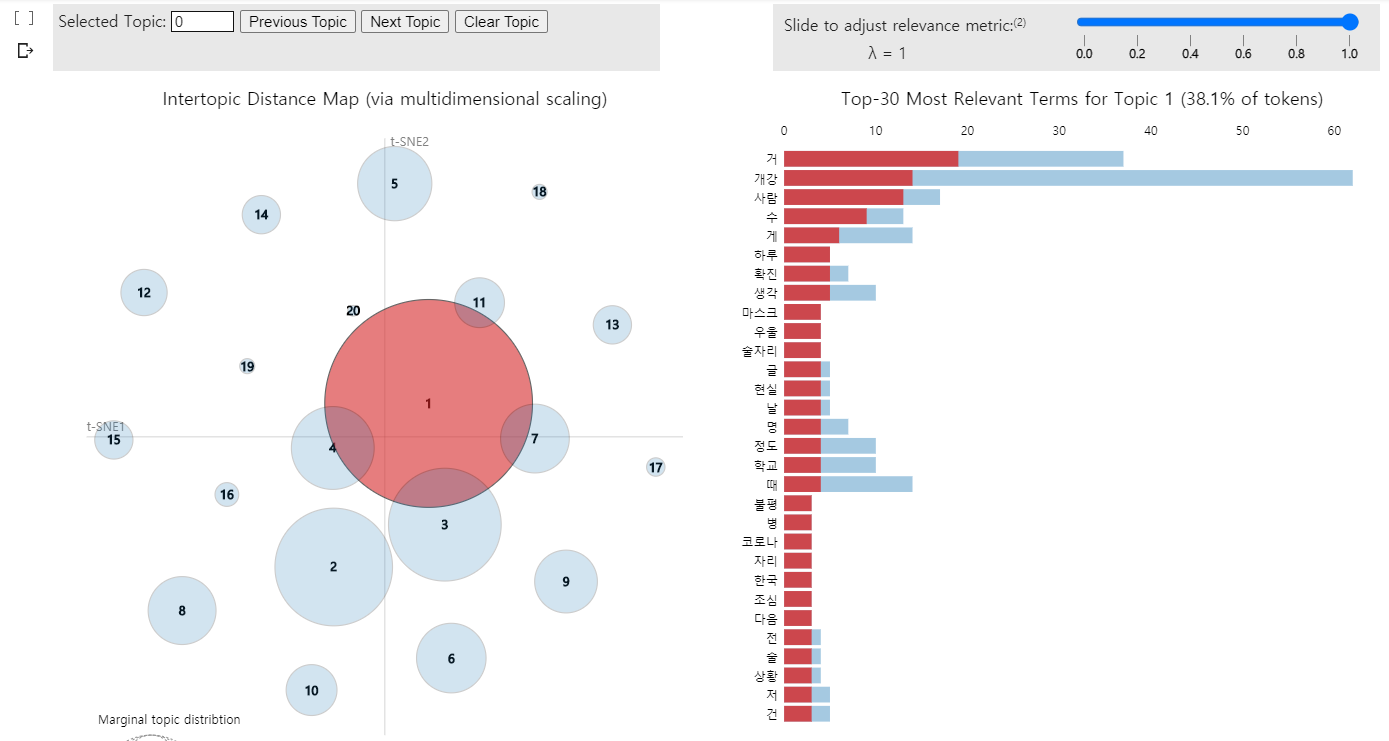












* 1. Words relation analysis

CorpusbasedKeywordExtractor를 통해 단순히 어느 단어가 많이 등장했는 지만 보는 것이 아니라, 어떤 단어가 어떻게 연관이 되어 있는지도 알아보기 위해 연관어 분석을 실행하였다. 주요 키워드는 ‘온라인, 오프라인, 신천지, 과외, 성적, 외국, 과제, 통아(캠)’ 이다



1. Result

먼저, 코로나를 기준으로 파트를 나누었고 가장 많은 관심을 받는 주제는 코로나 관련일 것이라는 내 예상과는 달리 학생들이 코로나, 신천지, 중국 등에는 생각보다 관심이 없었고 시험, 과제, 성적, 중간고사, 성적 산출 방식 등등에 학생들은 더 많은 관심을 보였다. 이는 초반에는 코로나를 위험질병으로 인식하고 두려워했지만 시간이 지날수록 코로나는 먼 나라 이야기가 되었고 학생들에게 좀 더 직접적인 영향을 주는 시험, 과제등에 대한 관심이 커진 것 같다. 즉, 엄청난 이슈이지만 자신들에게 직접적인 영향을 주지 않는 코로나 대신, 사소한 주제이지만 나에게 큰 영향을 미치는 시험, 과제 등에 학생들은 더욱 큰 관심을 보이는 것을 알 수 있었다.

연관어 분석으로 알 수 있는 정보도 많이 있었다. 먼저, 과외라는 키워드를 분석하였을 때 의외의 결과로 대구, 수원이라는 두개의 지역이 나왔는데, 대구상황이 워낙 안좋다 보니까 과외에 대한 변동을 겪은 학생이 많음을 알 수 있었고, 수원지역은 상황이 심각하지는 않지만 자연과학캠퍼스 학생들이 다수 분포해 있고 확진자가 가까운 곳에 나왔기 때문에 마찬가지로 과외를 그만둔 학생이 많음을 알 수 있었다.

또한 많은 키워드 들에서 총학생회, 이루리(총학생회 이름), 설문조사 등에 대한 관련성을 확인할 수 있었는데, 온라인 학기이다 보니까 학교와 소통할 수 있는 길이 많지 않고 그래서 어떠한 이슈가 생길 때마다(등록금, 연기, 온라인 확정, 성적 등등) 총학생회에 건의하거나 설문조사를 하는 사람들이 많음을 알 수 있었다. 이는 cluster로는 발견할 수 없는 정보이다.

마지막으로 비대면 수업에 대한 학생들의 인식도 조금이나마 엿볼 수 있었는데, 통아캠과 연관있는 감정적인 단어가 ‘우울, 짜증’이었다. 이는 사람들을 만나지 못해서 정신적으로 의지할 사람이 없기 때문에 혹은 온라인 수업이다 보니까 교수님들이 학생들의 학업성취정도를 파악하지 못해서 실습이나 수업 등으로 배우는 내용들도 몽땅 과제로 내주기 때문에 과제에 대한 압박 및 스트레스 등 다양한 이유로 통아캠(온라인 학기)에 대한 부정적인 견해가 더욱 큼을 알 수 있었다.