

Department of Software Convergence

Data Analysis Capstone Design Project   
Weekly Progress Report

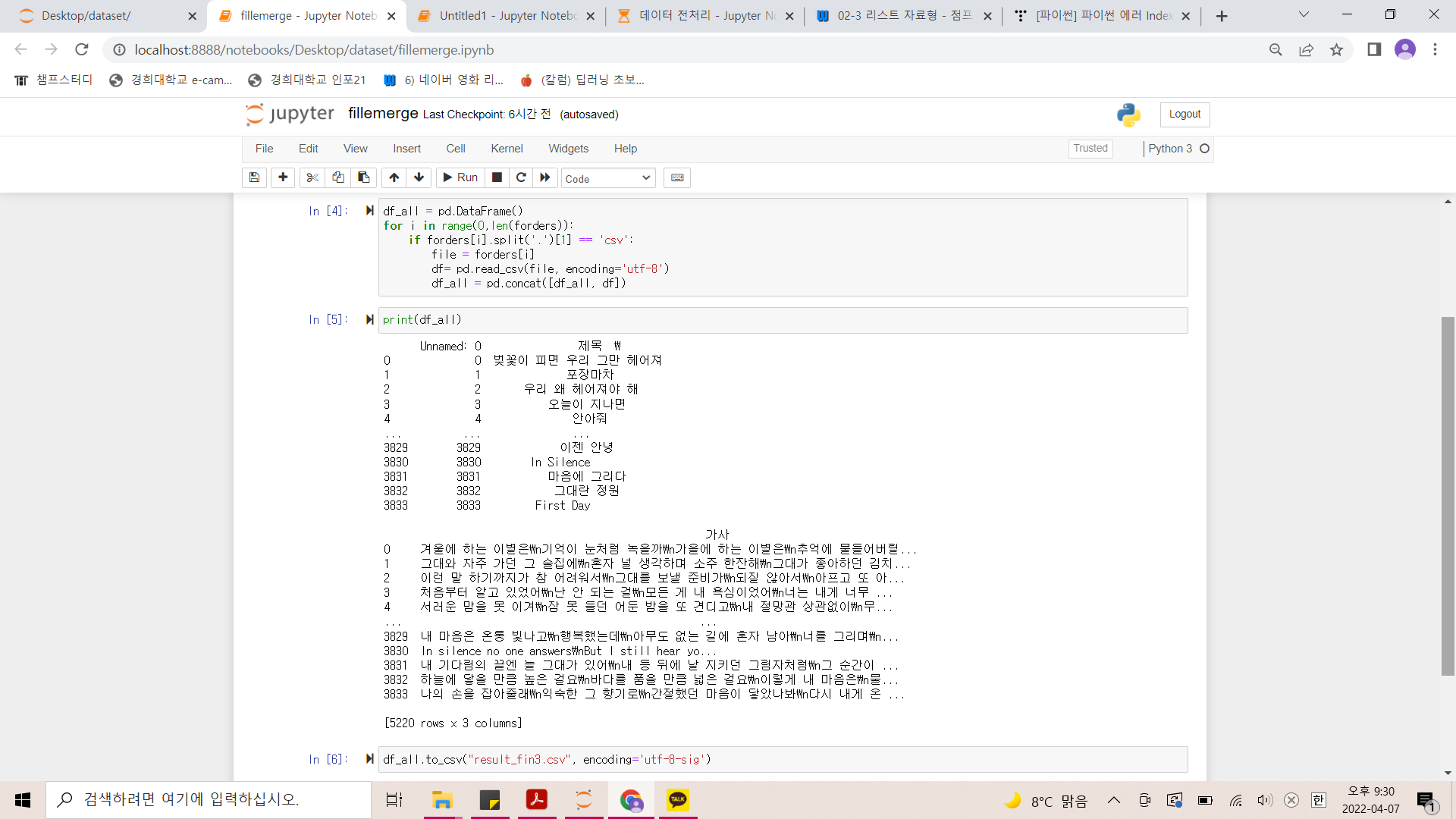
| Project Title | 딥러닝을 활용한 노래 가사 분석 및 추천 시스템 |
| --- | --- |
| Student / ID | 김은비 / 2019100858  유창현 / 2018102121  이은경 / 2019100896  이인석 / 2018110659 |
| Reporting Week | 2022-04-01 ~ 2022-04-07 |
| Project Manager of Reporting Week | 이은경 |
| Faculty Supervisor | 유창현 |

# **Tasks Outlined in Previous Weekly Progress Report** (Provide detailed information on the tasks to be completed in this week)

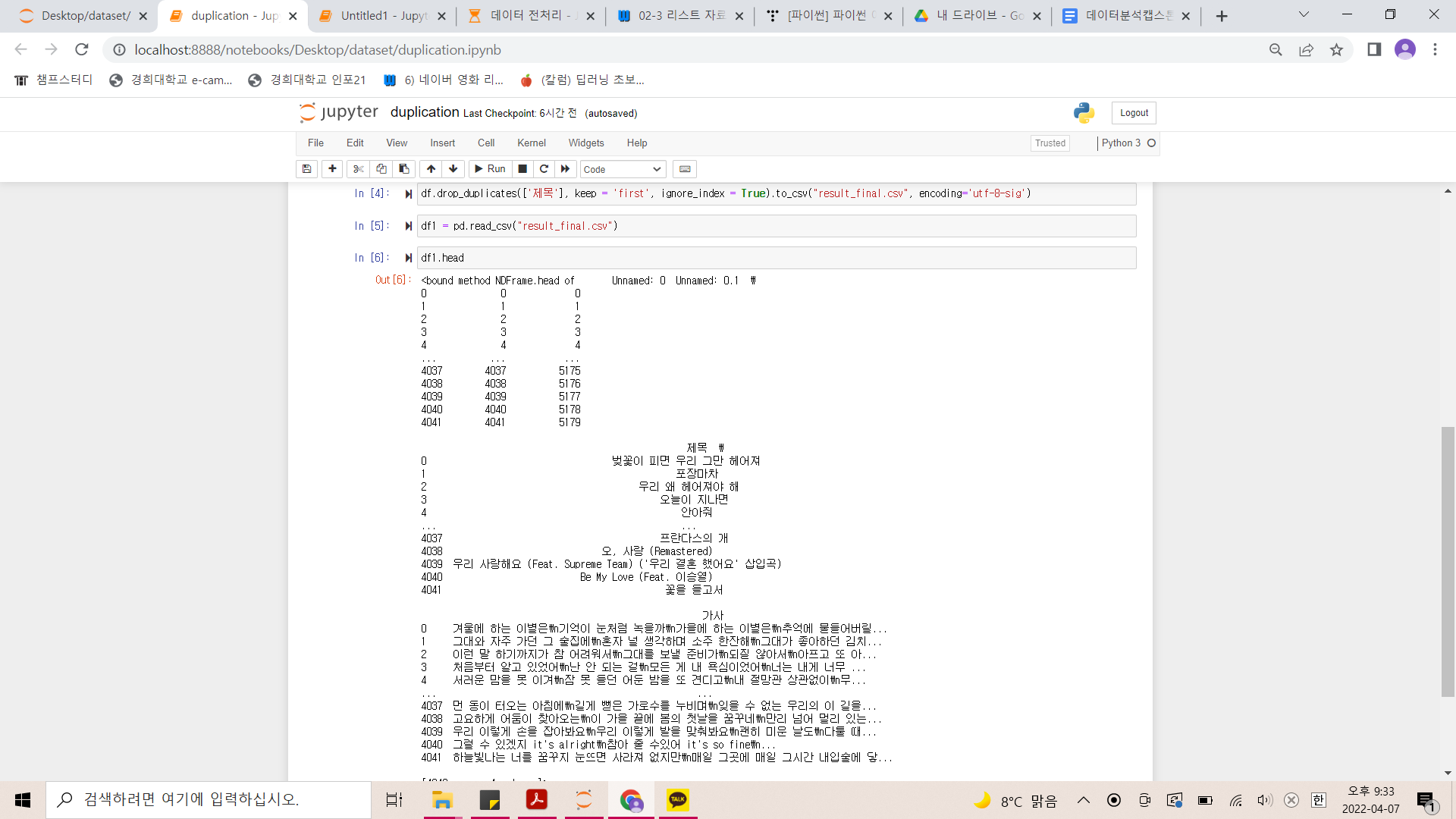
* 조원 4명이 각자 크롤링한 파일 병합하고 중복 제거하여 최종 데이터셋 구축
* 기존에 구축된 한국어 감성사전이 있는지 확인/ 활용가능성 확인
* 수집된 가사 데이터에 대한 전처리(토큰화, pos태깅, 불용어 제거 등)

# **Progress Made in Reporting Week** (Provide detailed information on the progress that you made in the reporting week. Limit your write-up to no more than two page)

* 조원별로 각자 크롤링하여 csv파일로 저장한 데이터를 pandas 라이브러리를 활용하여 하나의 파일로 병합하였고 최종 6,572곡이었다. 그 중 drop\_duplicates 함수를 사용하여 같은 제목을 갖는 중복 데이터를 제거하는 과정과 영어 가사로 이루어진 데이터를 삭제하는 과정을 거쳐서 최종적으로 활용하게 된 데이터는 4,003곡이다.

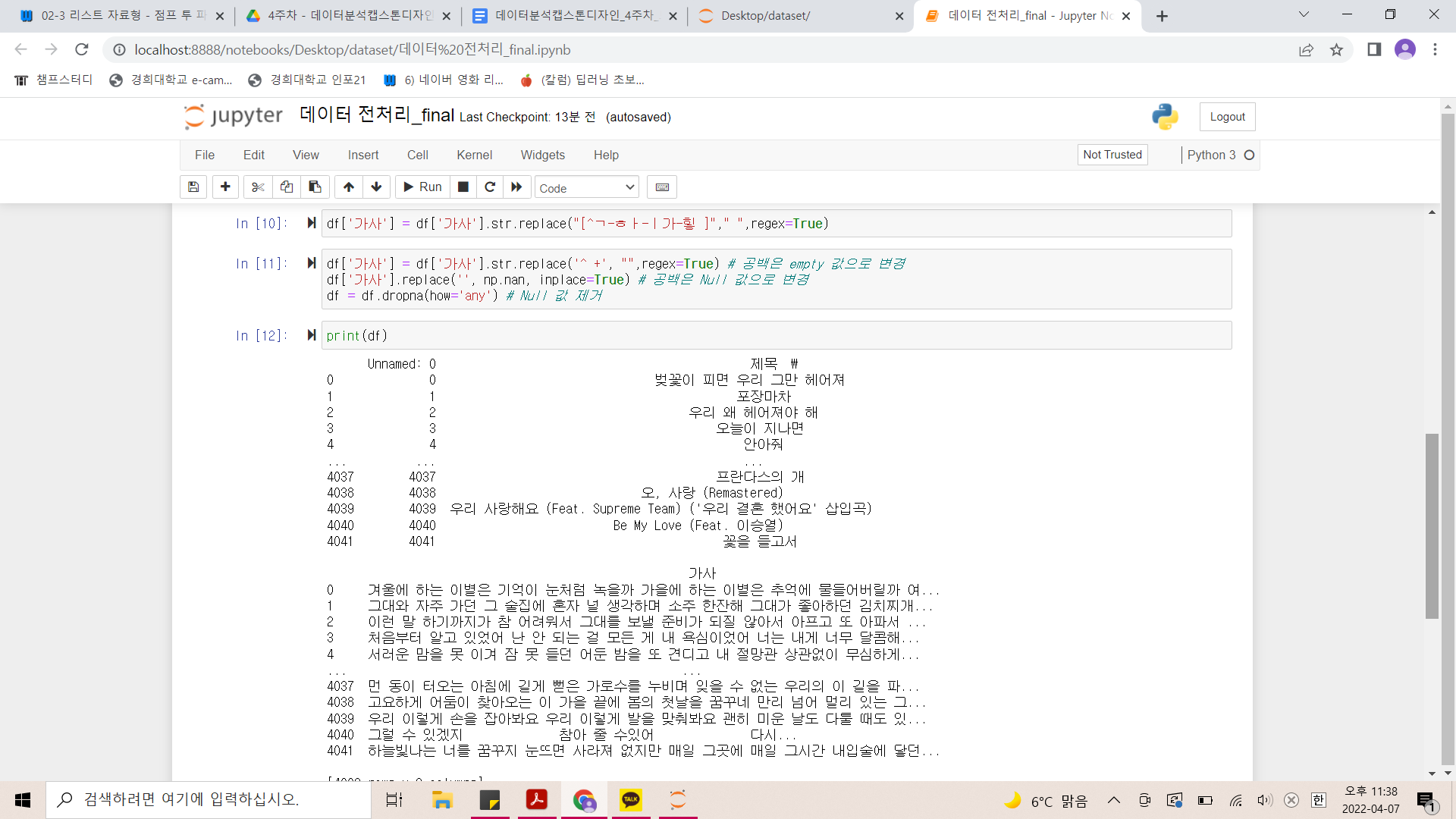


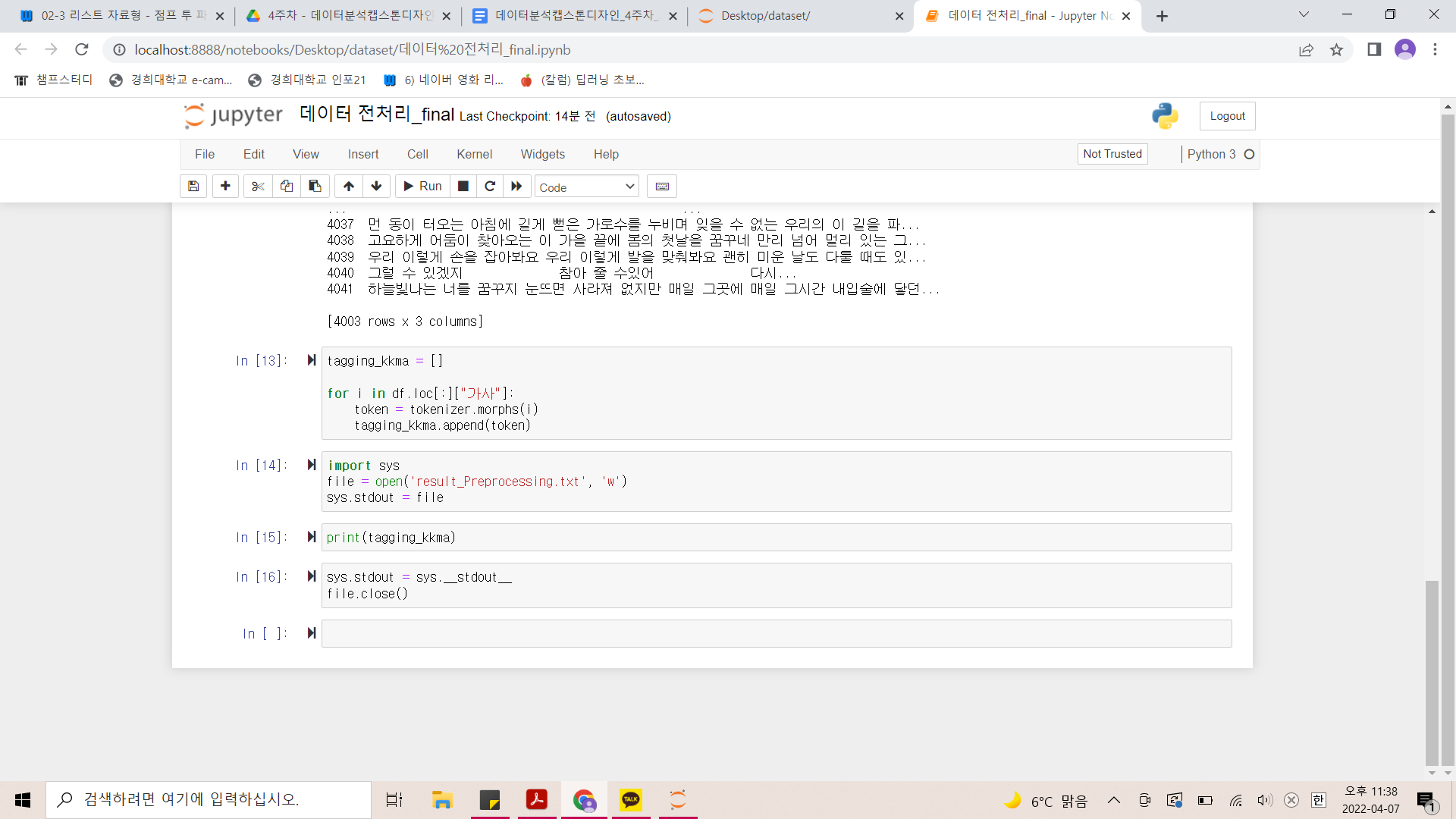
[크롤링한 데이터가 저장된 csv파일 병합 코드]



[csv 파일 내 중복 데이터 제거 코드]

* 데이터 전처리에 사용한 형태소 분석기는 한국어 정보처리를 위한 파이썬 패키지 KoNLPy의 Kkma이다. Kkma(꼬꼬마)는 서울대에서 개발한 한국어 형태소 분석기로 분석 품질이 우수하고 용언의 어간과 어미에 대한 상세한 분석이 가능하며 장점을 가진다. 단점은 고유명사에 대한 형태소 분석이 취약하고 실행 속도가 느리다는 점이다. 하지만, 해당 연구에 사용되는 데이터를 적용하는데는 그리 오랜 시간이 걸리지 않을 것이라 판단했고, 노래 가사에 고유명사가 사용되는 경우가 많지 않기에 Kkma를 활용해 전처리하기로 결정했다.
* Kkma.mporphs를 사용하여 토큰화를 진행하고 Kkma.pos를 통해 각 형태소에 대한 품사를 태깅했다. 그 중 SF(마침표, 물음표, 느낌표), SP(쉼표, 가운뎃점, 콜론, 빗금), SS(따옴표, 괄호표, 줄표), SE(줄임표), SO(붙임표), SW(기타기호), UN(명사추정범주), OL(외국어), OH(한자), ON(숫자)를 불용어로 설정하고 제거를 시도하였다. 하지만 예상과 다른 결과가 도출되었고 그 상세한 설명은 다음 3번에 작성하였다.
* 다른 전처리 방법으로 정규표현식을 사용하여 한국어만을 추출한 후 토큰화하는 방법을 활용하여 최종 전처리를 완료하였다.



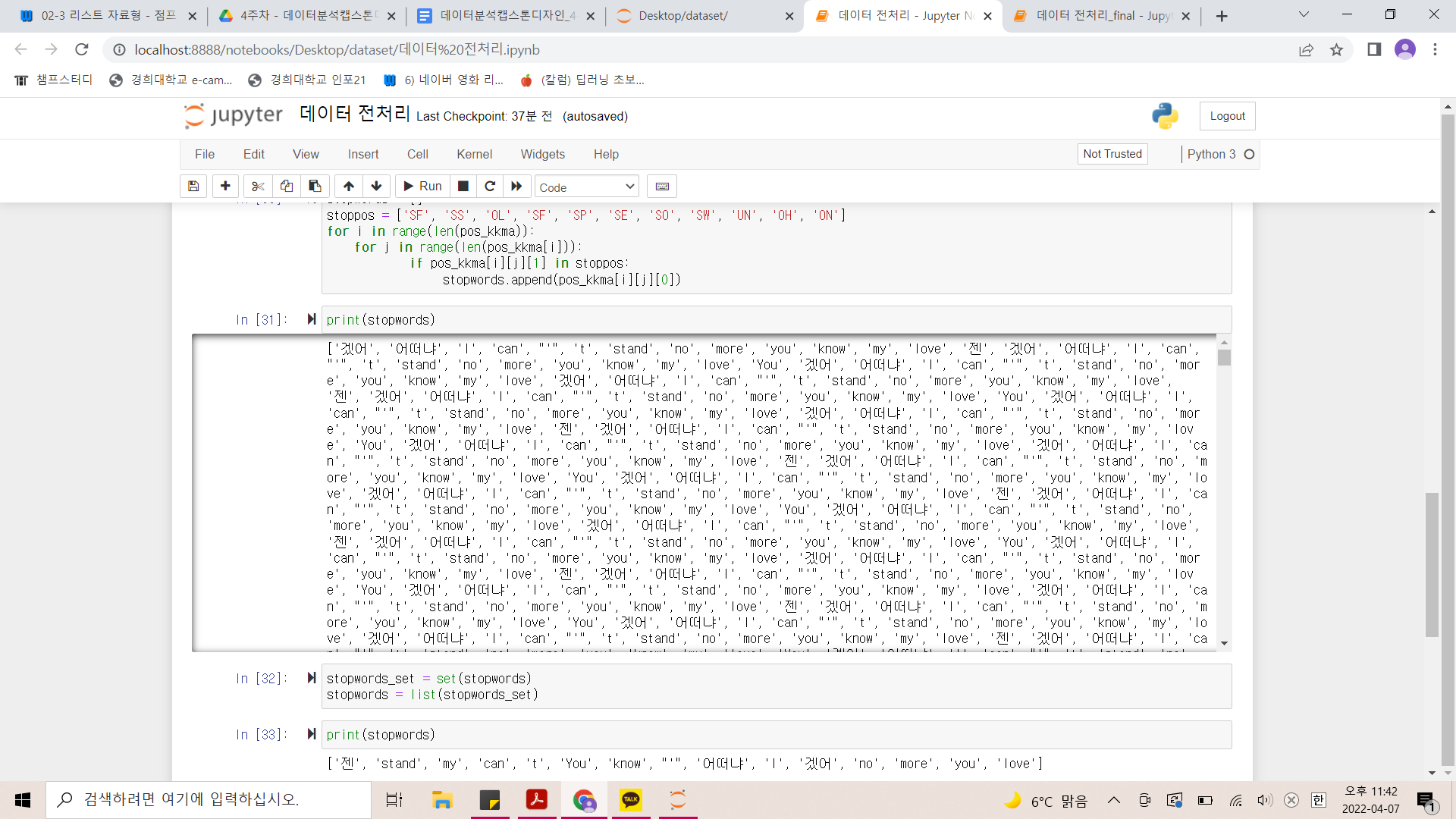


[가사 데이터 전처리 코드]

* 기존에 이미 존재하는 감정사전을 사용할지, 새로 구축할지에 대해 조원들과 논의하였다. 논의 결과 1) 지난주에 정의한 7가지 세부 감정에 맞는 감정사전이 없었고, 2) 노래 가사를 도메인으로 구축한 감정사전은 없어, 전처리한 노래 가사 데이터를 가지고 직접 감정 사전을 구축하기로 결정하였다.

# **Difficulties Encountered in Reporting Week** (Provide detailed information on the difficulties and issues that you encountered in the reporting week. Limit your write-up to no more than one page)

* 기존에 구축된 한국어 감성사전이 존재한다면 그것을 활용하여 보다 효율적인 분석을 진행하려고 하였으나, 우리가 결정한 7가지 감정으로 분류하는데 기존 감성사전을 적용하기는 쉽지 않다는 결론이 났다. 그래서 좀 더 시간이 소요되더라도 word2vec을 사용하여 정의한 7가지 감정과 유사도가 높은 단어들을 선별하여 자체적인 감성사전을 구축하기로 결정하였다.
* 전처리를 위해 kkma를 사용하여 토큰화 및 pos태깅을 진행하였다. 그 후 불용어로 판단되는 품사를 가진 형태소를 제거하는 방식으로 불용어 제거를 수행하려고 하였으나, 토큰화 후 pos태깅을 하는 과정에서 포함되지 않는 형태소들이 있었다. 즉, 품사가 부여되지 않는 형태소들이 있어서 pos태깅을 통해 구축한 불용어 리스트에 포함되는 않는 영어나 기호들이 있어 깔끔한 전처리가 이루지지 않았다. 따라서 pos태깅을 통한 전처리 대신 정규 표현식을 활용하여 한국어만을 추출한 후 kkma로 토큰화를 진행하였다.



이처럼 pos태깅을 통해 불용어를 추출한 결과, ‘I, can, stand, You’과 같은 일부 영어만 포함된 것을 확인할 수 있다.

# **Tasks to Be Completed in Next Week** (Outline the tasks to be completed in the following week)

* Word2vec를 활용하여 이전에 결정한 세부 감정에 대한 유사도 기반 감성사전 구축