"eta_secret_board.csv" 는 2020년 10월 5일 이전까지의 everytime 사이트의 비밀 게시판 글과 댓글을 수집한 파일이다. 이 데이터에 대해서 아래 task를 수행하여 결과를 정리하시오.

1. Data를 정리하여 tokenization을 수행하시오. 영어가 아닌 한글 데이터이기 때문에 필요한 전처리가 다를 것이다. KoNLP와 같은 패키지를 사용하여 전처리를 해보자. 어떠한 전처리가 필요해서 수행했는지 어떤 방법을 사용했는지 설명하여라.

참고: https://lsjsj92.tistory.com/216

2. 정리된 data를 이용해서 데이터 탐색을 수행하고 그 결과를 설명하여라.

예) 글 하나에 댓글이 평균 몇 개씩 달리는지, 학기 중과 방학 중에 글의 수가 어떻게 다른지, 글이 많이 작성되는 시간대가 있는지 등.

3. 전체 구간을 학기(+방학)별로 나누어 tf-idf를 계산하여라. 각 학기별로 tf/idf가 높게 나타나는 단어를 정리하여 설명하여라.

... ,2020년 봄학기 2020년 하계방학 2020년 가을학기,

4. Topic Modeling을 수행하여라. 결과를 바탕으로 비밀게시판에는 주로 어떠한 주제들이 언급되는지 설명하여라.

5. (bonus 문제: optional) 시기별(분기별, 월별, 주차별)로 감성분석을 수행하고 그 결과를 분석해보아라.

참고: 군산대학교 한국어 감성 사전: https://github.com/park1200656/KnuSentiLex