|  |
| --- |
| 영화제 각본상 수상작 감성분석 & Wordcloud |

2020. 03.09.

**이은성**

**목차**

**1. 프로젝트 개요**

1.1 프로젝트 기획 배경

1.2 구성원 및 역할

1.3 프로젝트 추진 일정

**2. 프로젝트 소개**

2.1 데이터 수집 및 불러오기

2.2 proto-typing

**3. 프로젝트 결과**

3.1 결과 출력

**4. 후기**

4.1 문제점 시사 및 향후 개선 사항

**1. 프로젝트 개요**

1.1 프로젝트 기획 배경

본 교육과정을 통해 학습한 R machine learning algorithm을 활용하여 유명 영화제 각본상 수상작 스크립트 감성분석 및 wordcloud를 출력하여 수상작 텍스트의 특징을 파악하고자 이 프로젝트를 기획했다.

1.2 구성원 및 역할

|  |  |
| --- | --- |
| 이름 | 구현 부분 |
| 이은성 | 프로젝트 기획 및 수행 |

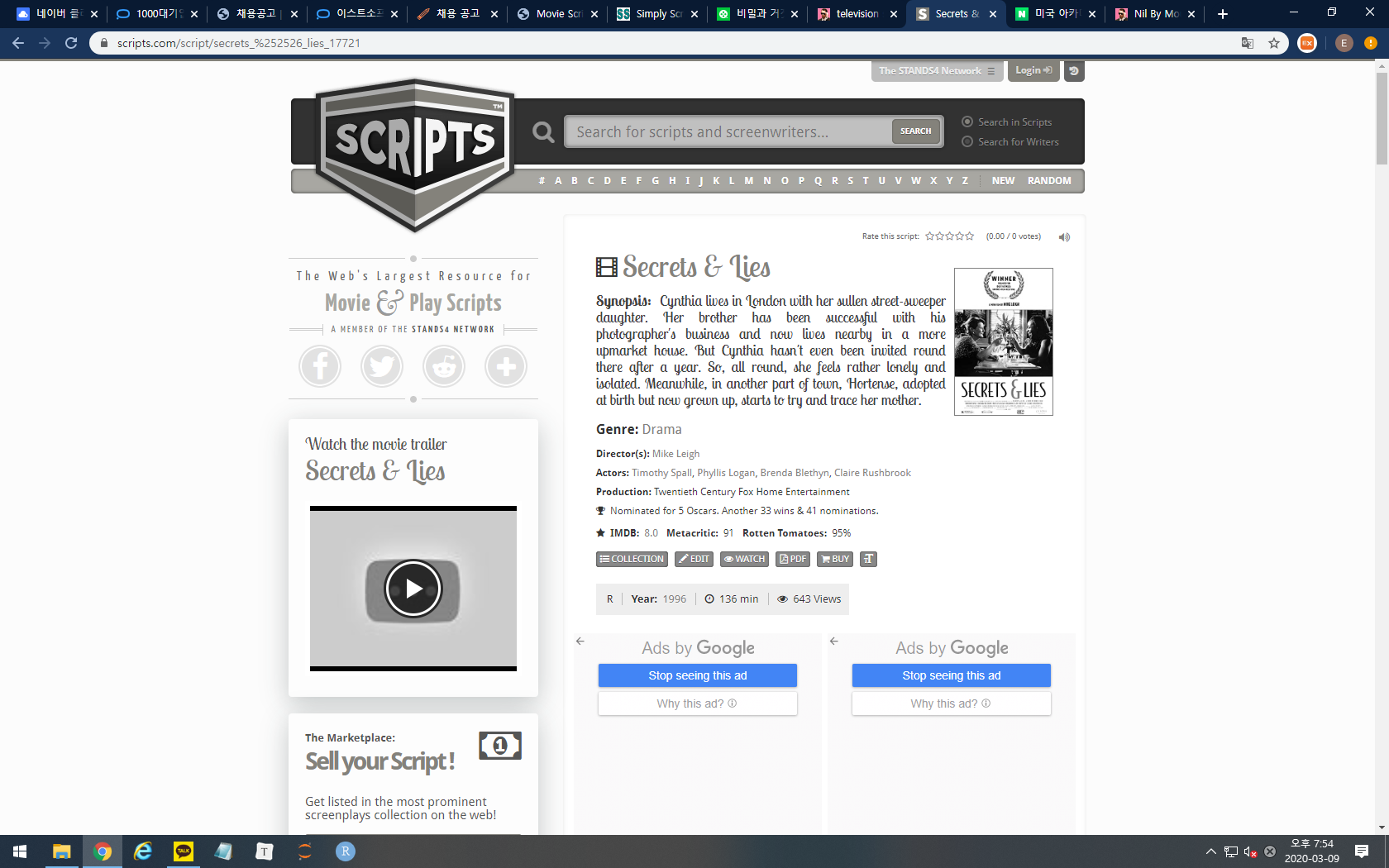
1.3 프로젝트 추진 일정

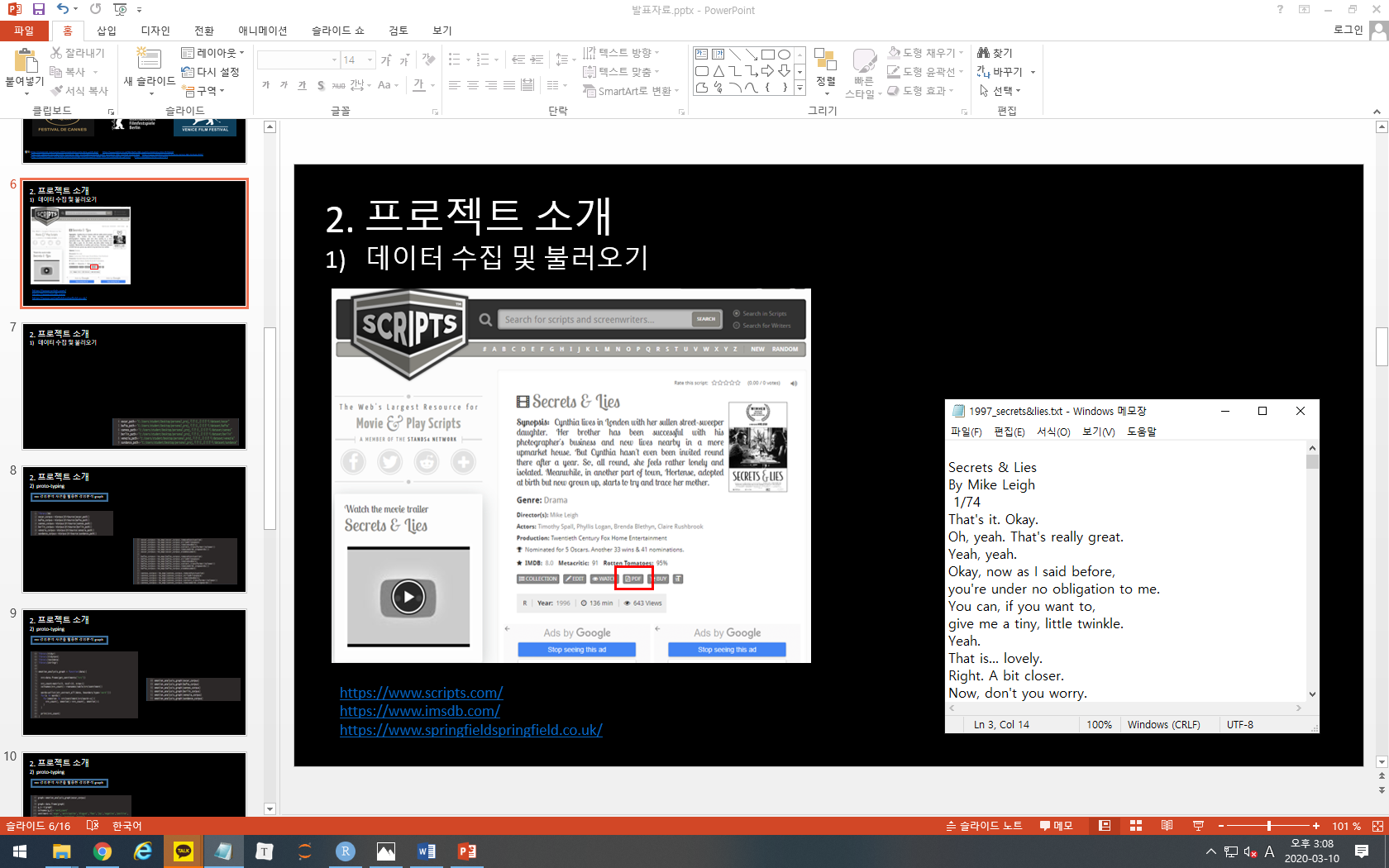
|  |  |  |  |
| --- | --- | --- | --- |
| 구분 | 기간 | 활동 | 비고 |
| 사전  기획 | 02/21(금) ~ 02/24(월) | 프로젝트 기획 및 주제 선정 |  |
| Proj  수행/  완료 | 02/25(화) ~ 03/08(일) | 프로젝트 수행 |  |
| 03/09(월) ~ 03/11(수) | 보고서 작성 및 발표 |  |

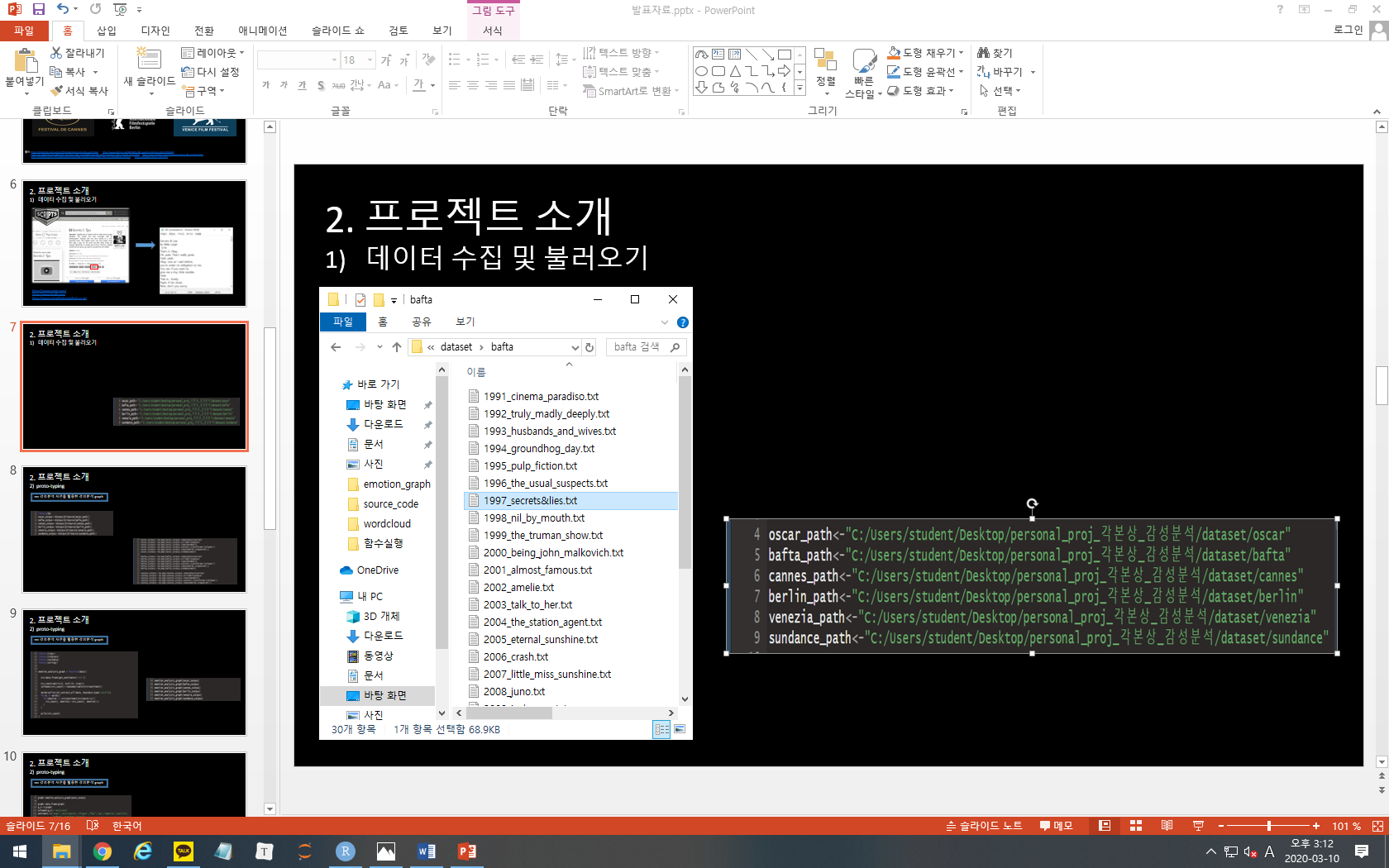
**2. 프로젝트 소개**

2.1 데이터 수집 및 불러오기

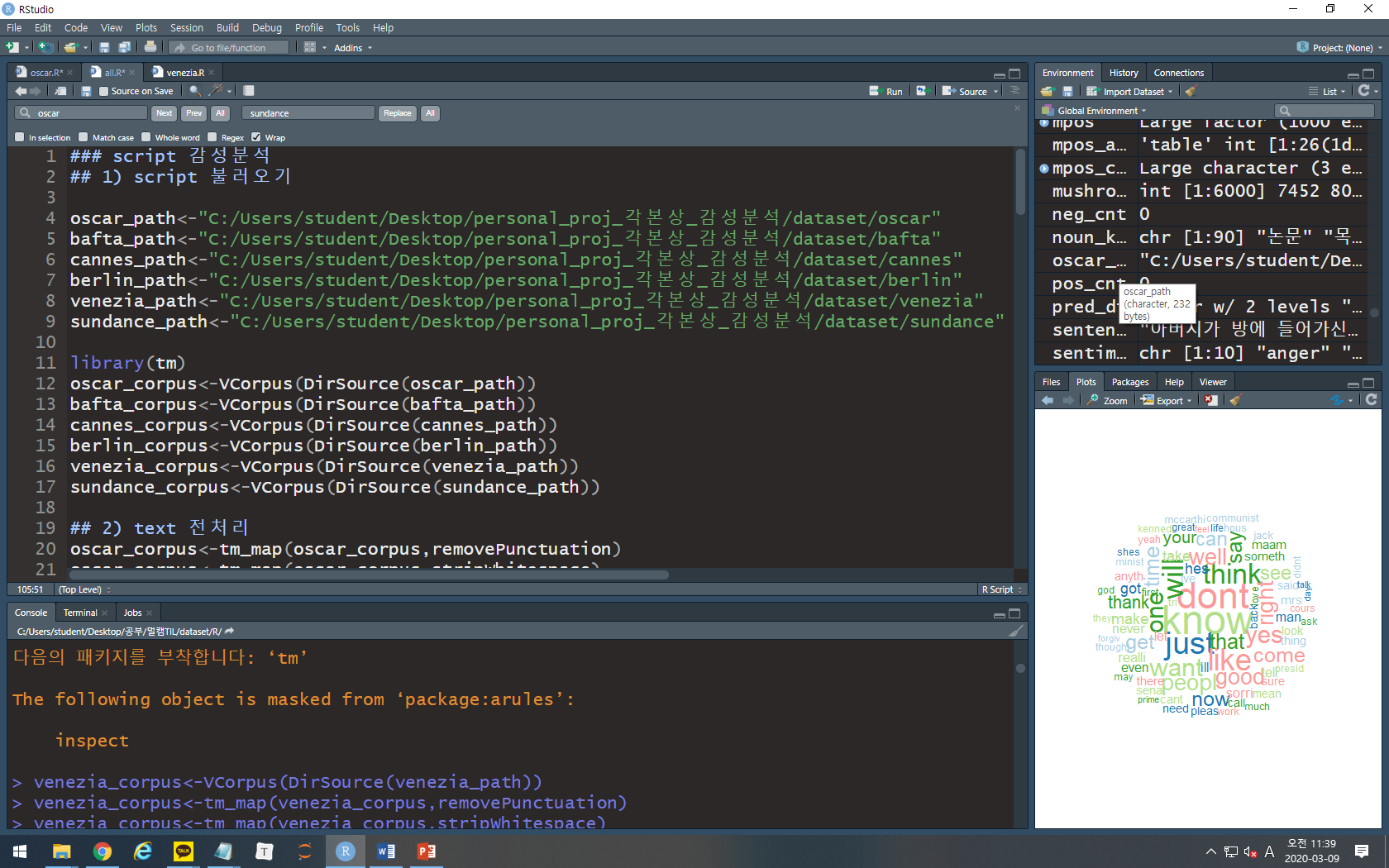
1. 6개 영화제 (OSCAR, BAFTA, Sundance, Cannes, Berlin, Venezia) 의 최근 10년간 각본상 수상작 스크립트를 스크립트 공유 웹 페이지를 통해 수집하고, txt 형식으로 저장한다.







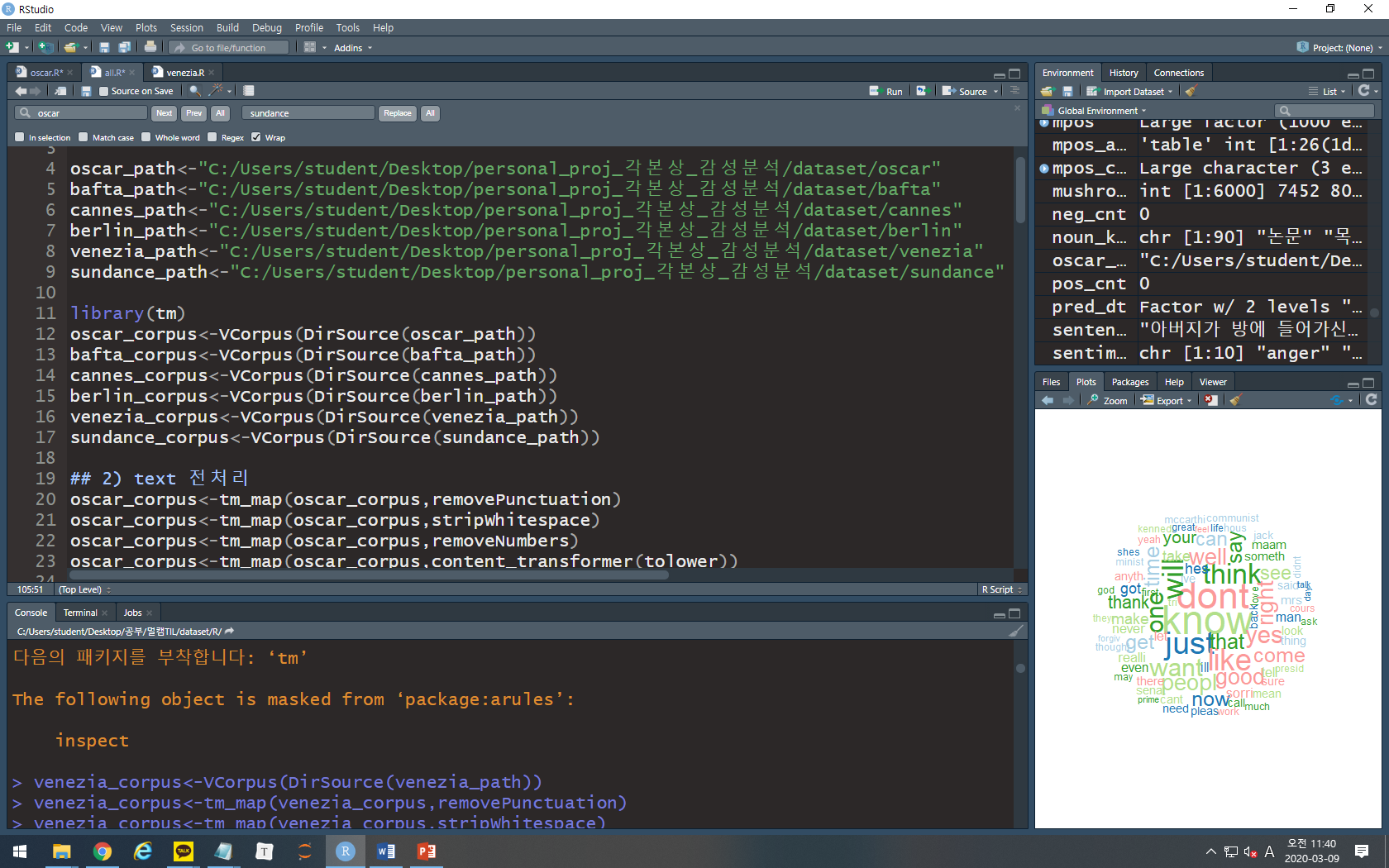
1. Rstudio 상에서 수집한 스크립트 txt 파일을 각 영화제 별로 불러오기 한다.



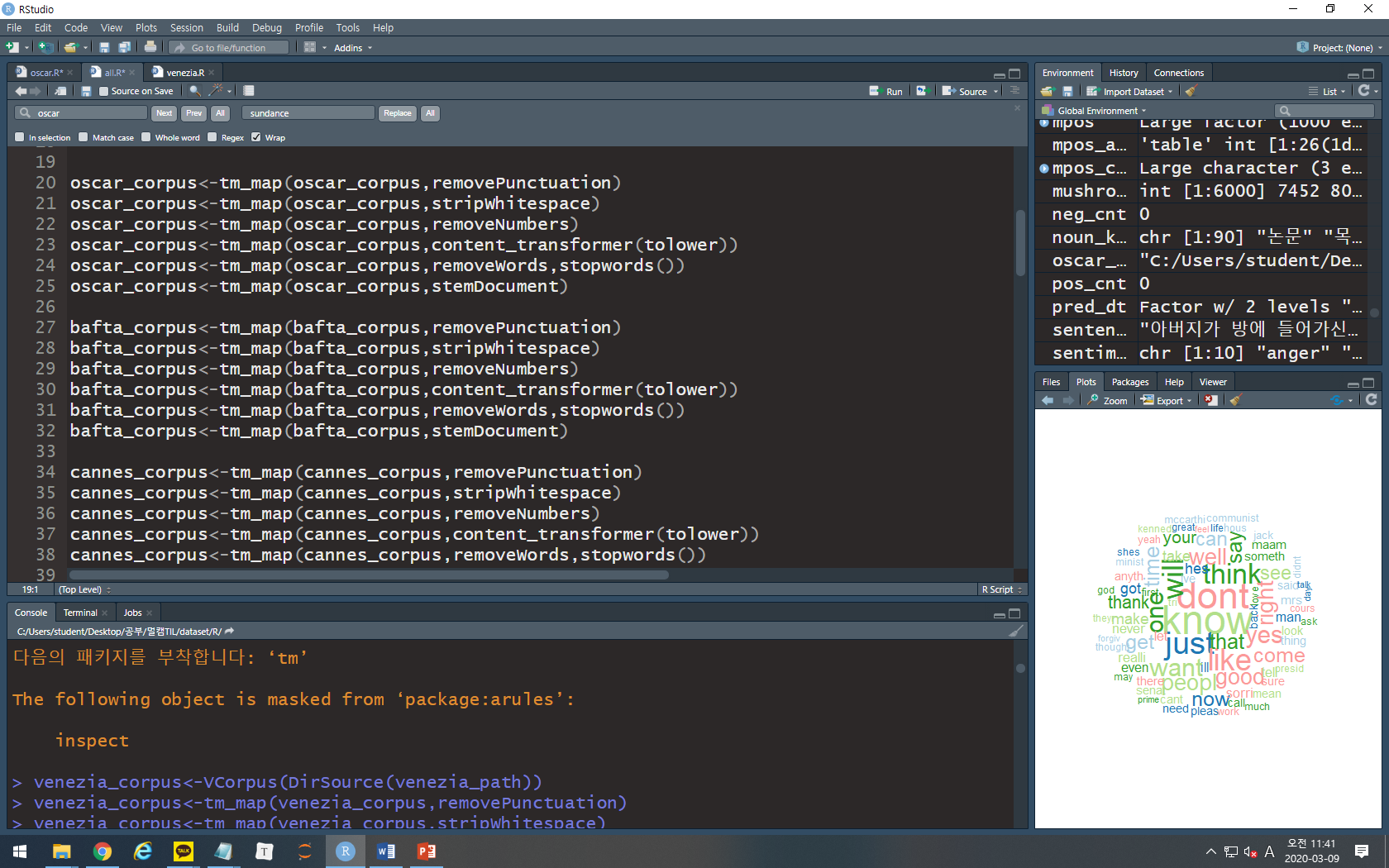
2.2 proto-typing

**<nrc 감성분석 사전을 활용한 감성분석graph>**

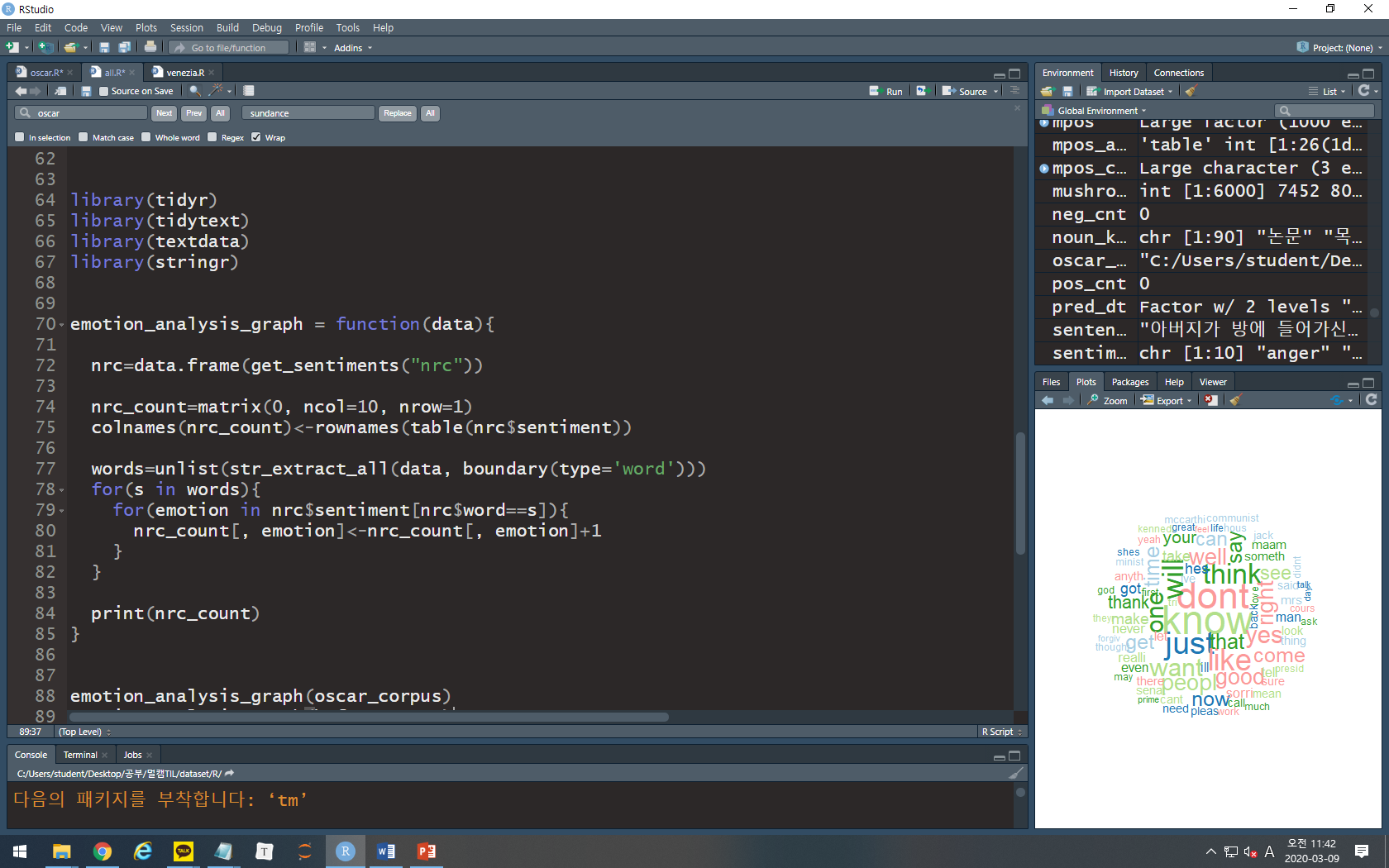
1. tm package를 활용하여 각 영화제 별 script의 corpus를 생성한다.



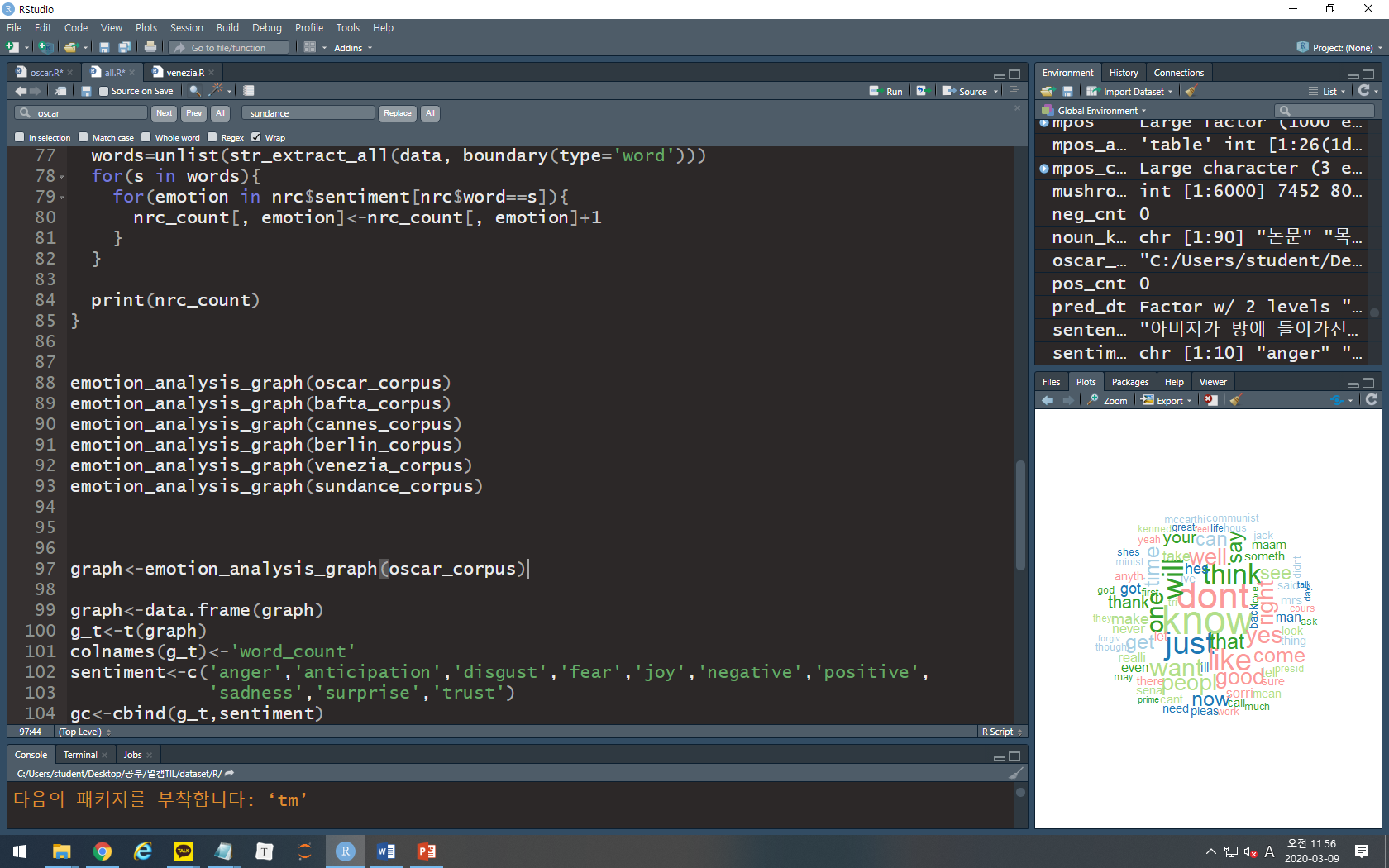
1. 감성분석 graph 함수 실행 전, 생성된 corpus를 전처리한다.



1. nrc 감성분석 사전을 활용한 감성분석 graph 함수를 생성한다.

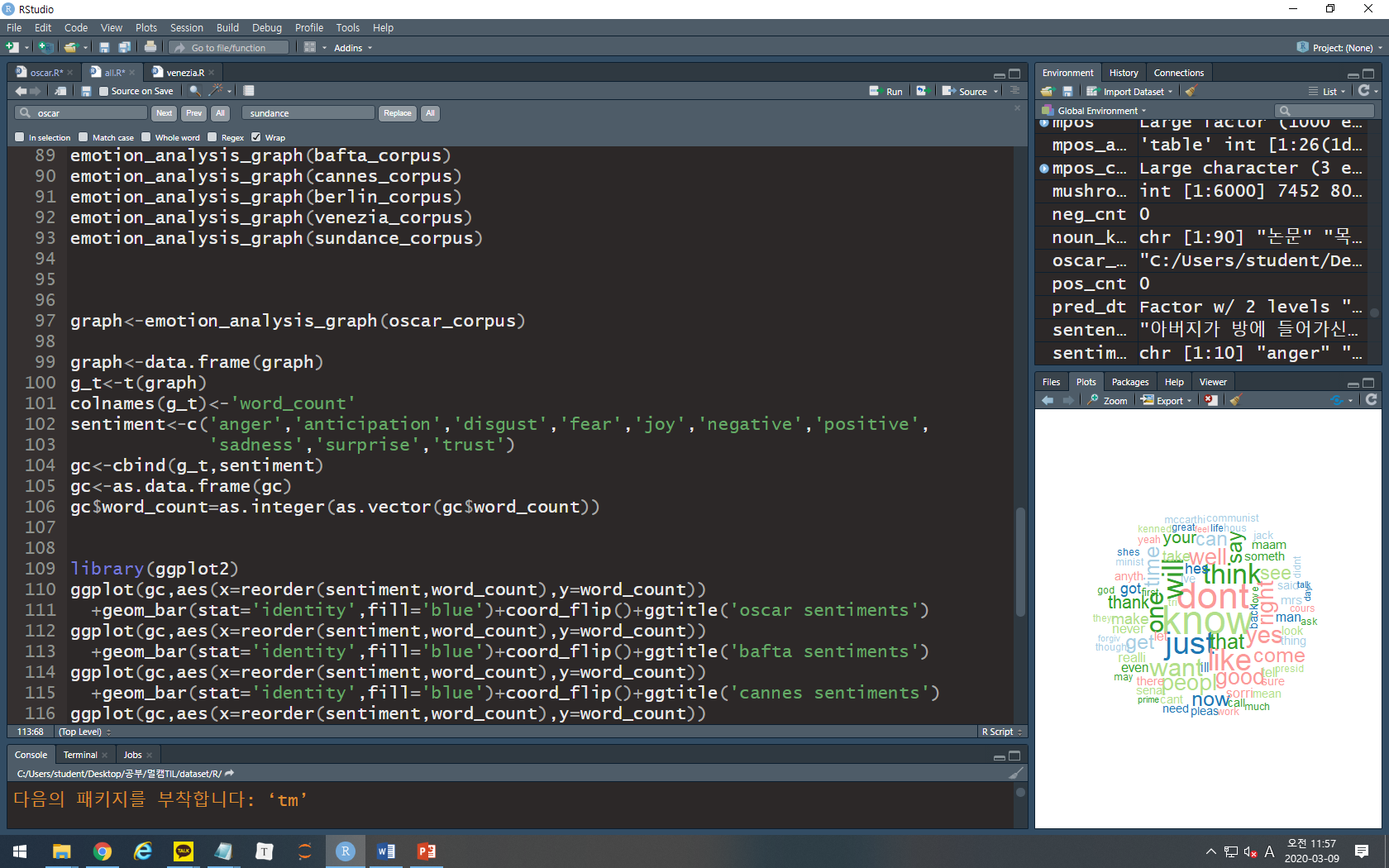


1. 감성분석 graph 함수를 각 corpus 별로 실행한다.

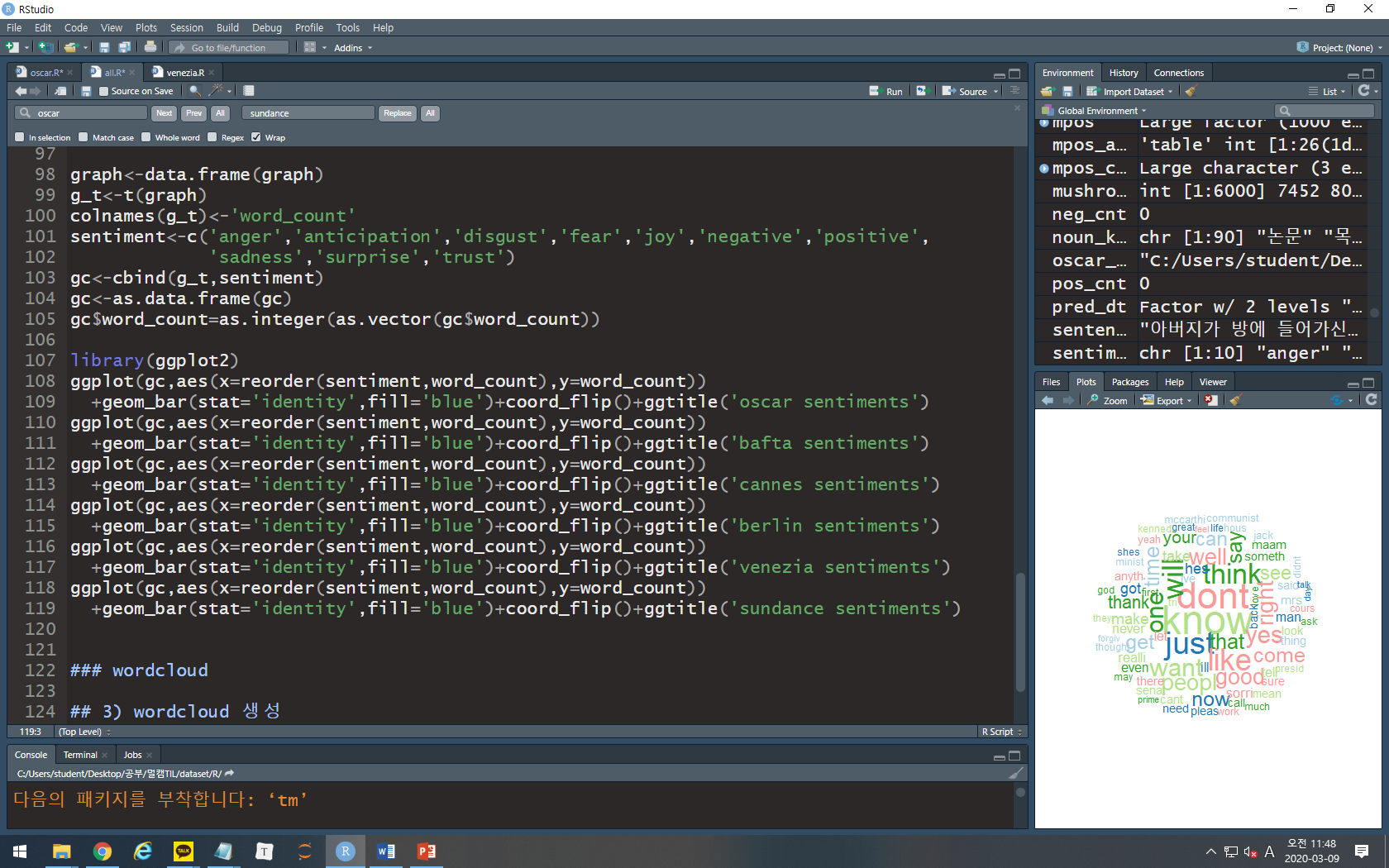


1. graph 출력을 용이하게 하기 위해 함수 결과값을 전치한 후, nrc 감성사전의 10가지 sentiment를 column name으로 하는 column 추가한다.

또한, 출력된 word count 값을 integer로 바꿔준다.

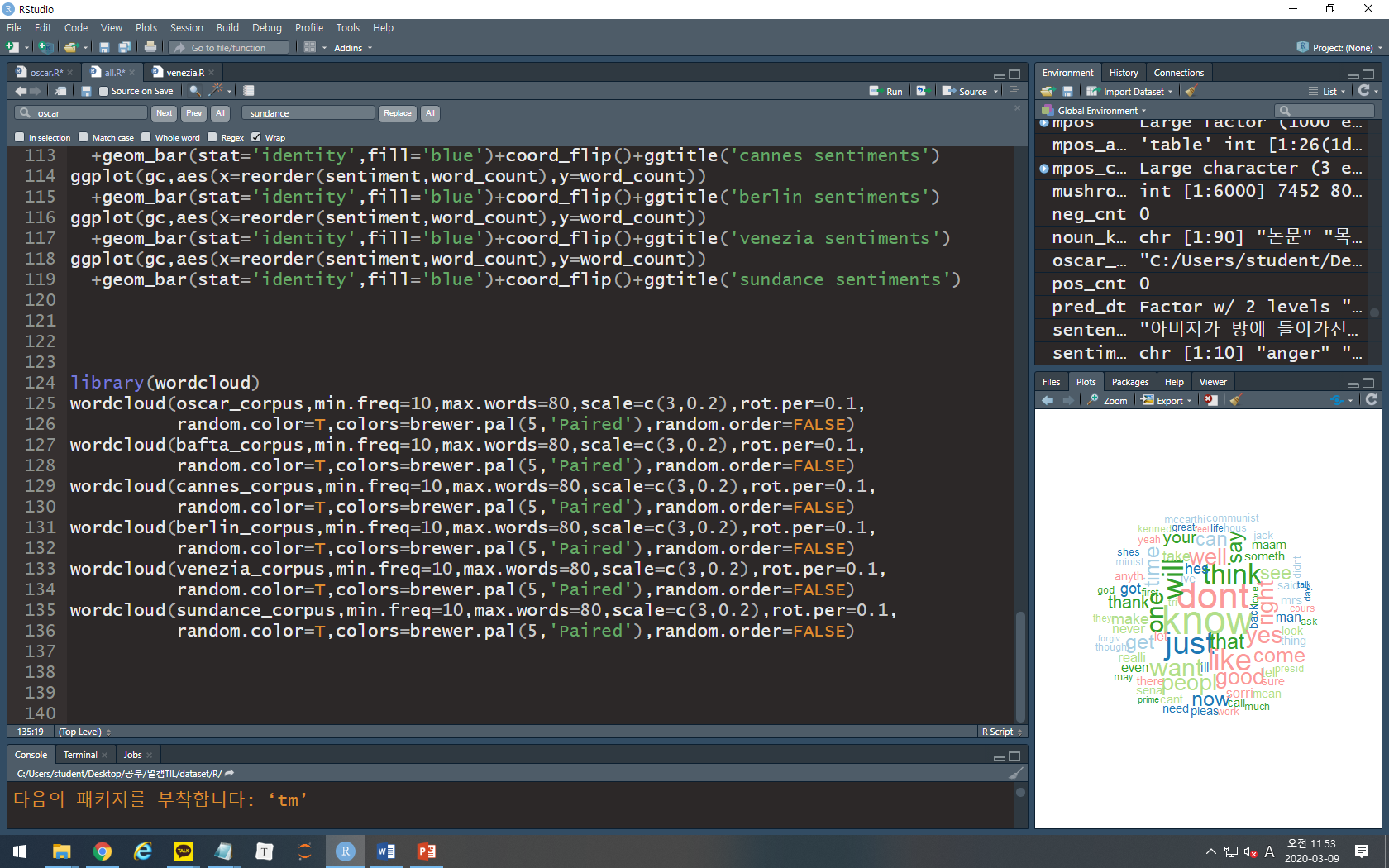


1. ggplot2 package를 활용하여 많이 count된 감성 순으로 각 영화제 별 graph를 출력한다.



**<wordcloud>**

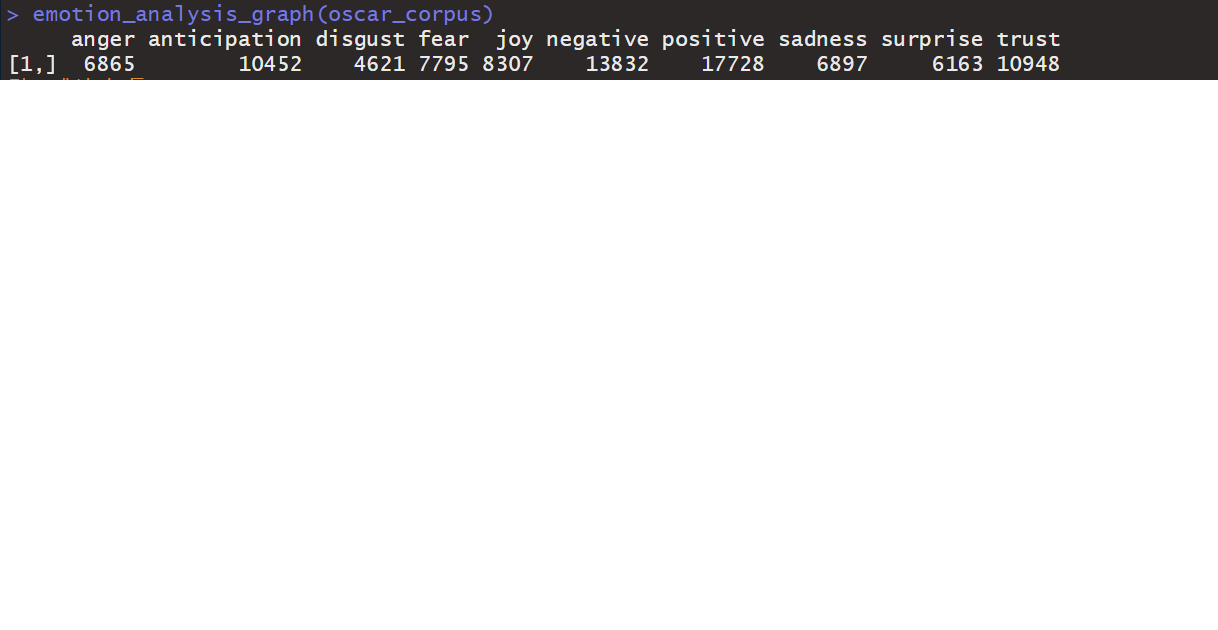
wordcloud package를 활용하여 각 영화제 작품상 script 별 wordcloud를 출력한다.

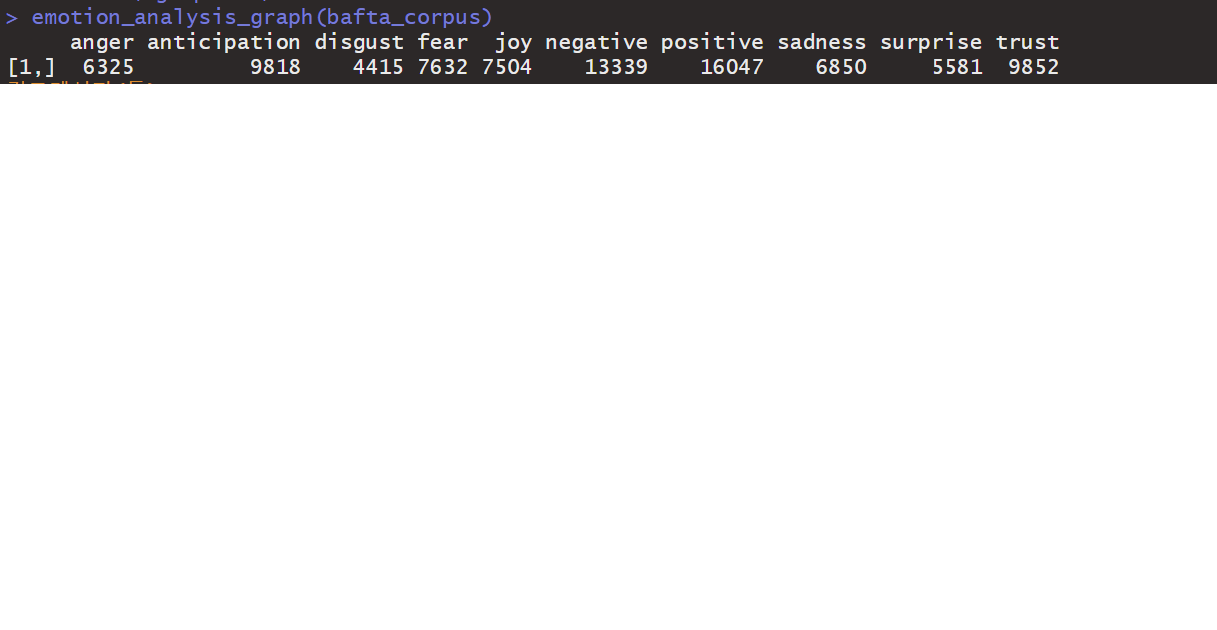


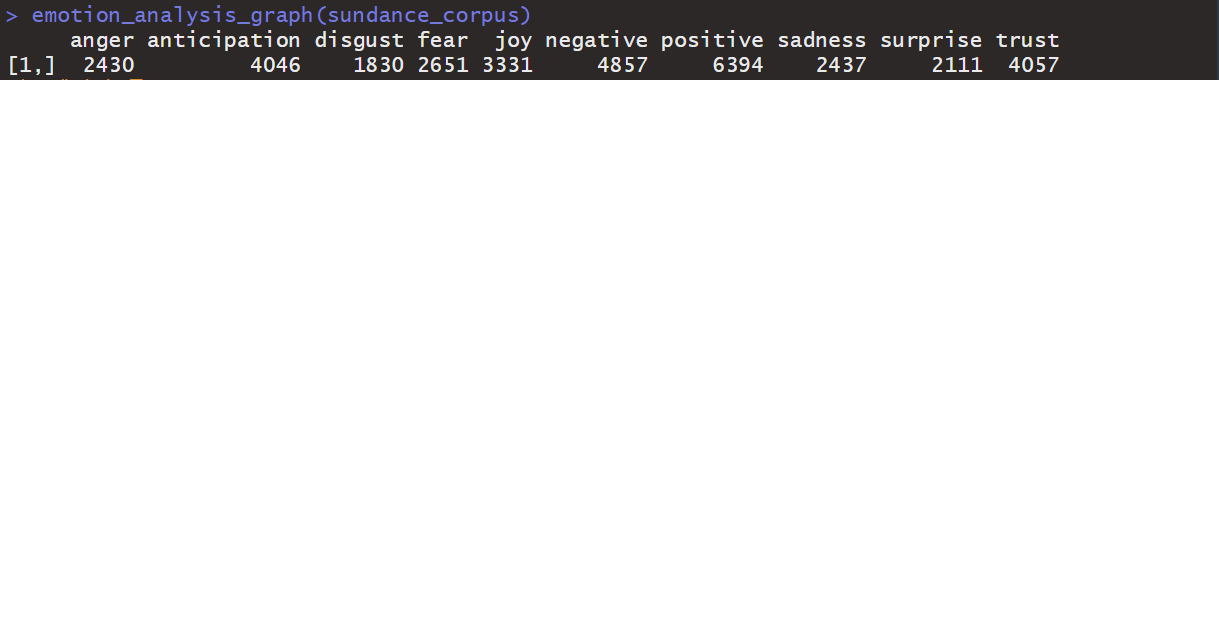
**3. 프로젝트 결과**

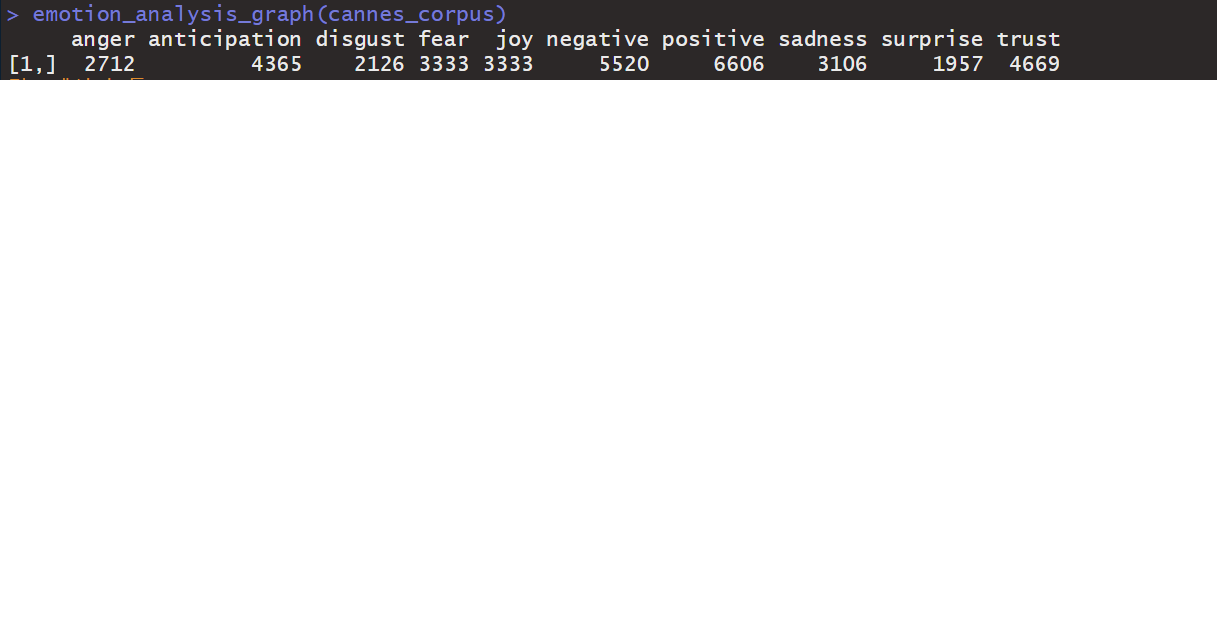
3.1 결과 출력

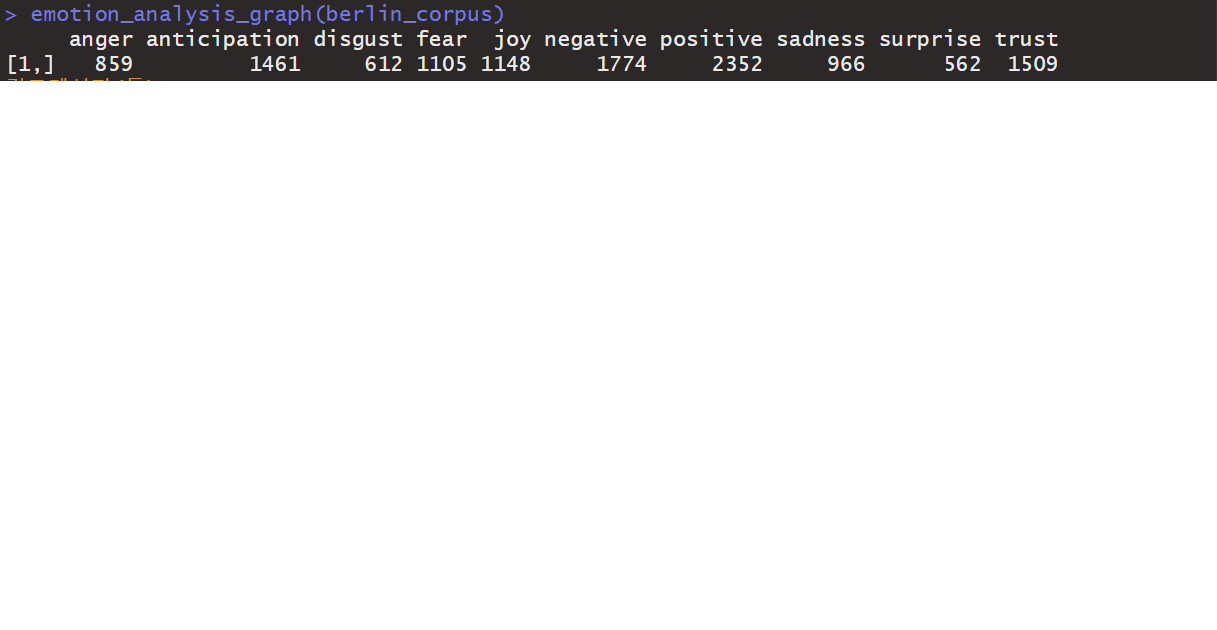
**<nrc 감성분석 사전을 활용한 감성분석 graph 함수 실행>**

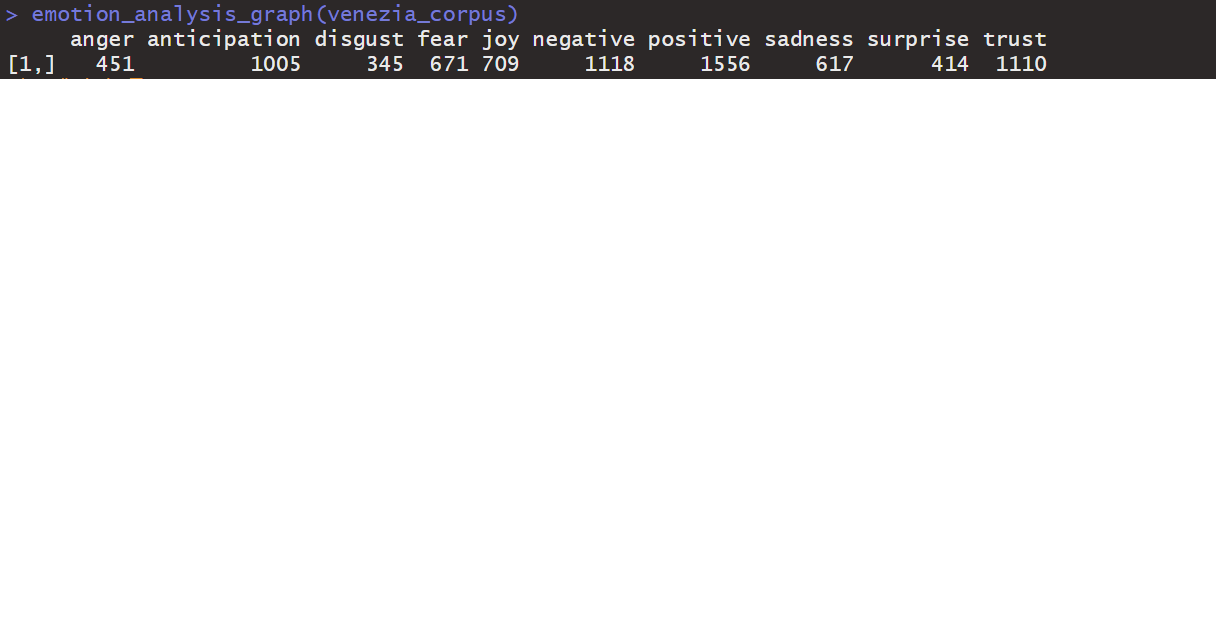












**<nrc 감성분석 사전을 활용한 감성분석 graph>**

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |

**<각 영화제 script 별 wordcloud 출력>**

|  |  |
| --- | --- |
|  |  |
| OSCAR wordcloud | BAFTA wordcloud |
|  |  |
| Sundance wordcloud | Cannes wordcloud |
|  |  |
| Berlin wordcloud | Venezia wordcloud |

**4. 개발 후기**

4.1 문제점 시사 및 향후 개선 사항

본 프로젝트의 결과물을 통해 6개의 영화제 각본상 수상작에 내포된 공통의 감성을 nrc 감성분석 사전의 10개의 감성에 따라 분류 및 비교할 수 있었다. 그러나 연도별 각본상 수상작이 없거나, 영화제 별로 수상작이 겹치는 등 여러 변동 사항이 함께 반영된 결과물이라는 점이 아쉽다. 또한 영화제마다 거의 동일한 감성이 우위에 있었고, 특정 단어들만 wordcloud를 통해 출력되었다. 이는 R만을 이용한 감성분석으로는 좀 더 정확한 감성분석에 어려움이 있고, 영화 스크립트의 특성상 회화체에 많이 사용되는 특정 단어들이 많이 사용되었다는 개선점을 시사한다.

따라서 프로젝트 취지에 맞는 결과물을 얻기 위해서는, Python 등 다른 프로그래밍 언어 혹은 정교한 결과물을 얻을 수 있는 algorithm을 보충할 필요가 있다.