

|  |
| --- |
| **Text Indexing Report**  **- inverted index -** |

|  |  |
| --- | --- |
| 과목 | 자연어처리 및 정보검색 |
| 담당교수 | 정재은 교수님 |
| 학과 | 경제학부 |
| 학번 | 20153029 |
| 이름 | 김민지 |
| 제출일 | 2019.03.21 |

**Text Indexing**

**1) 10 Movie scripts**

- Doc01: Beauty and the Beast

- Doc02: Coco

- Doc03: Deadpool

- Doc04: It

- Doc05: La La Land

- Doc06: The Great Gatsby

- Doc07: The Martian

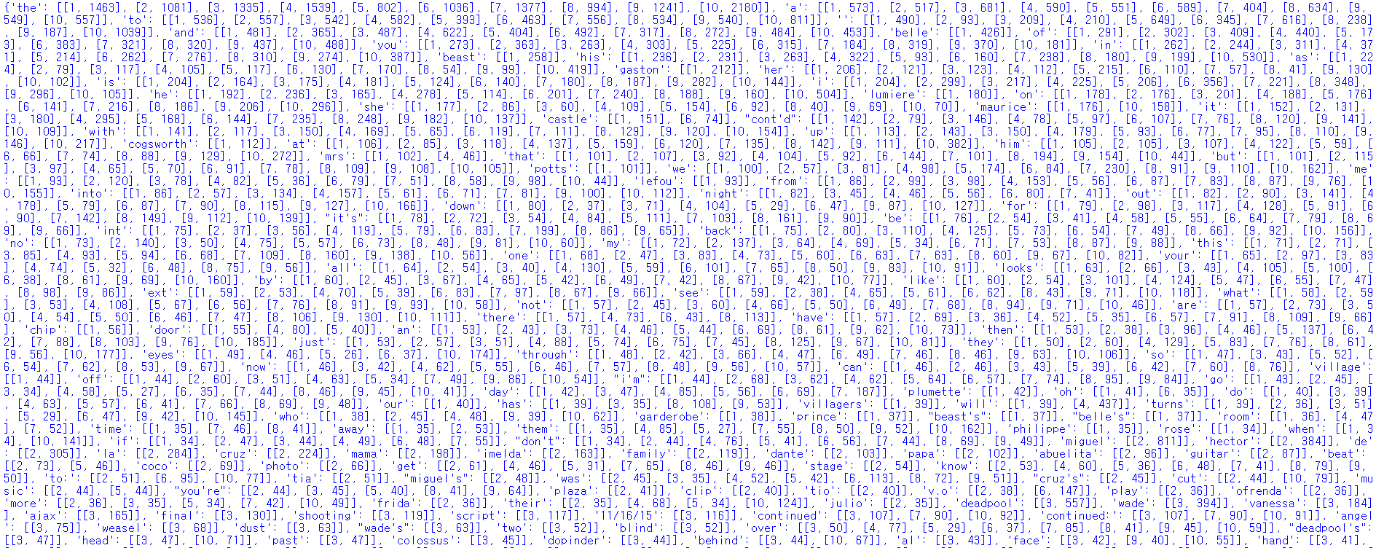
- Doc08: The Secret Life of Walter Mitty

- Doc09: Thor: Ragnarok

- Doc10: War for the Planet of the Apes

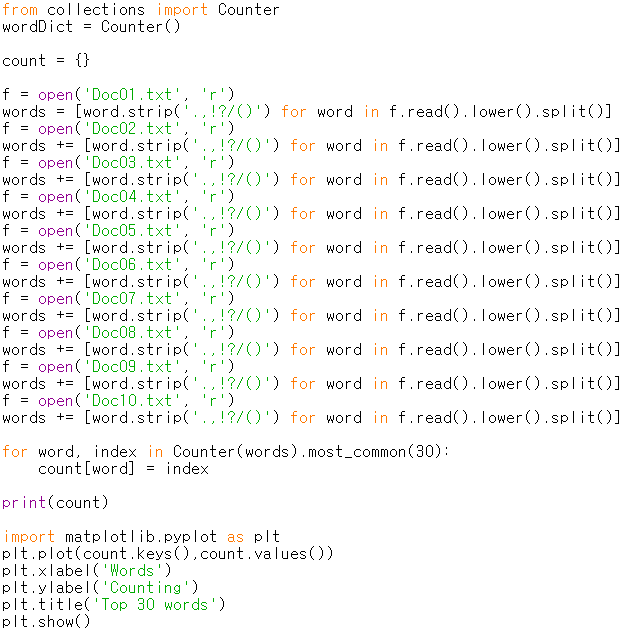
**2) Code for indexing the texts**

****

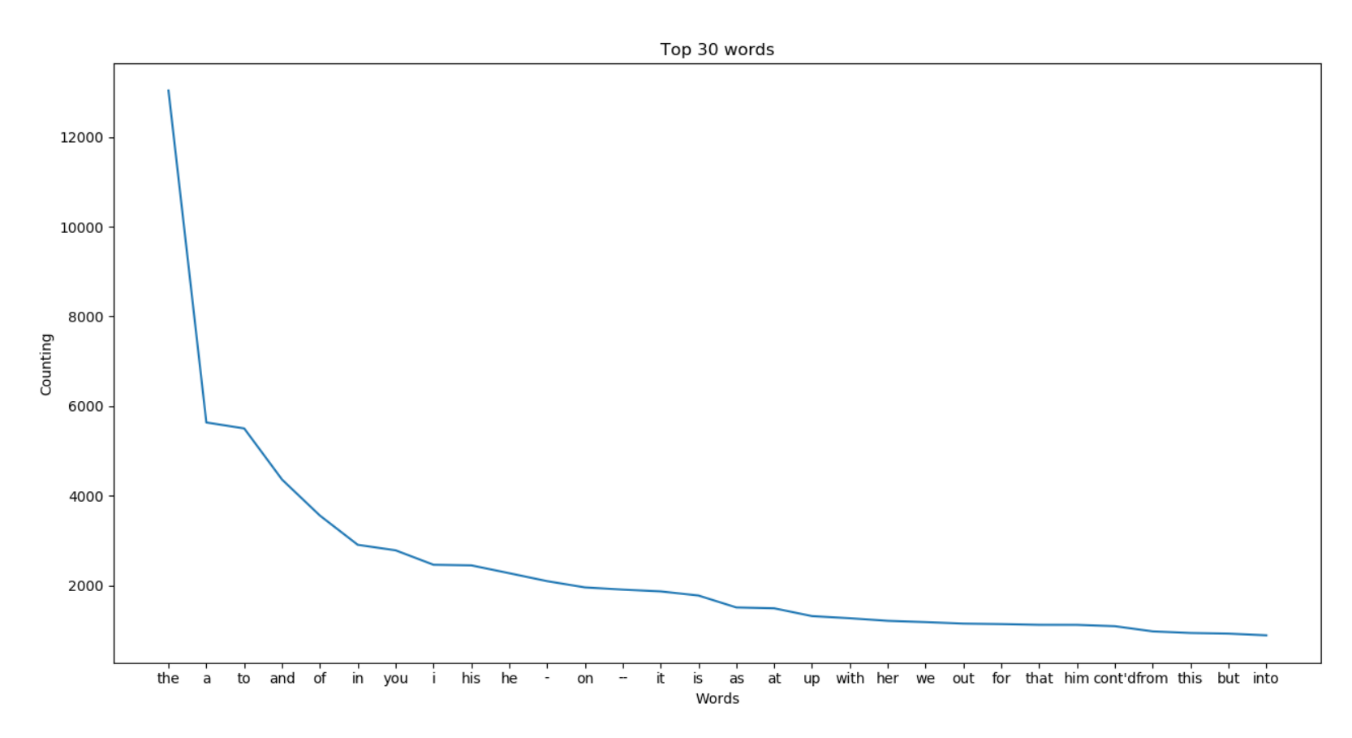


Index 함수를 정의하여 10개 문서들의 단어들을 빈도순으로 100개를 뽑아 {단어 : [[문서번호, 빈도], [문서번호, 빈도]..]}의 형태로 저장되도록 Text Indexing하였다. 빈도순으로 100개가 아닌 모든 단어를 보기 위해서는 most\_common(100)을 most\_common()으로 수정하면 되나, 결과값이 너무 길어 100으로 지정해 주었다.

**3) Charts for showing the word distributions**

****

****

****

모든 단어를 차트로 생성하면 차트가 너무 복잡해지기 때문에,10개 문서들의 단어들을 모두 합한 결과를 빈도수 순으로 정렬하여 상위 30개를 차트로 생성하였다. 이 그래프를 보면 ‘the’라는 단어가 가장 많이 사용되었다는 것을 알 수 있으며, 그 밖의 상위 30개의 단어들이 대부분 관사, 전치사, 대명사라는 것을 알 수 있다.