정보검색

컴퓨터학부 2019117865

김나형

1. DocId, Title, content 파싱

<https://hiseon.me/python/python-string-parse/> 을 참고하여

텍스트이(가) 표시된 사진

자동 생성된 설명

Parse 설치

파일을 읽어오기 위해 주어진 full\_corpus.hwp를 corpus.txt로 저장

한 줄씩 읽어와서 doc Id와 title, content를 분리하는 파싱 작업부터 시작하였다.

텍스트이(가) 표시된 사진

자동 생성된 설명

와 같은 에러가 발생. 주어진 한글 파일을 그대로 복사 붙여넣기 하였더니 </title> 뒤에 띄어쓰기 등이 있는 경우 주어진 full\_corpus 파일에 오류가 발생하여 해당 파일 파싱을 위해 조금 수정하였다.

<title>5. 하인리히 뵐<title> 등의 타이틀도 <title>5. 하인리히 뵐</title>로 수정하였다.

<title>27.야세르 아라파트</title>와 같이 점(.) 뒤 space가 없는 경우도 수정해주었다.

텍스트이(가) 표시된 사진

자동 생성된 설명

출력할 경우 docId, title, content 모두 잘 파싱되어 하나의 딕셔너리 배열에 저장되 출력되는 것을 확인할 수 있다.

1. 형태소 분석기

고려사항 : 조사와 같은 frequent terms는 가치가 없기 때문에 DF 고려할 필요 없어서 명사만을 형태소 분석에서 명사 추출

KoNLPy 라이브러리 사용

<https://konlpy.org/ko/latest/#start> 공식 문서를 따라가며 환경 세티

텍스트이(가) 표시된 사진

자동 생성된 설명

라이브러리 설치

https://mr-doosun.tistory.com/22

Td-idf를 weight로 사용. Wt,d = tf \* idf