**IT정치와 파이썬 기말고사 과제**

2017111978 이강혁

크롤링 대상: 국민참여 입법센터의 국회 입법 진행 현황

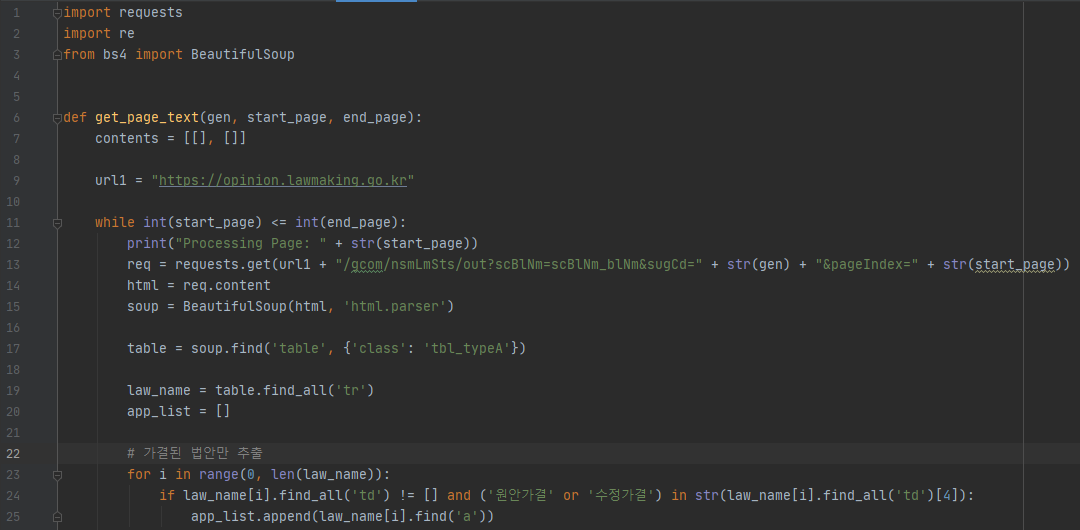
올해 임기가 만료된 제20대 국회에서 지난 4년간 발의한 법안을 통해서 사회적 이슈를 파악할 수 있다고 생각해서 20대 국회 발의 법안을 이번 과제의 대상으로 삼았습니다.

따라서, 이번 과제에서는 20대 국회에서 발의한 법안 중 가결된 법안을 추려내고 형태소 분석기를 활용해서 명사만 추려낸 후 워드 클라우드로 출력하였습니다.

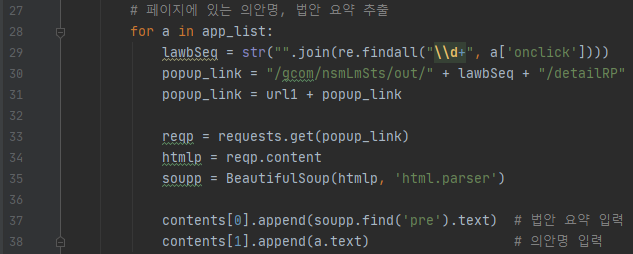
코드 설명



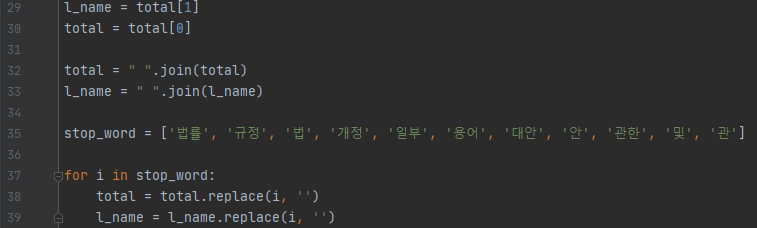
첫 페이지를 읽어와서 전체 페이지 수를 알아냅니다.



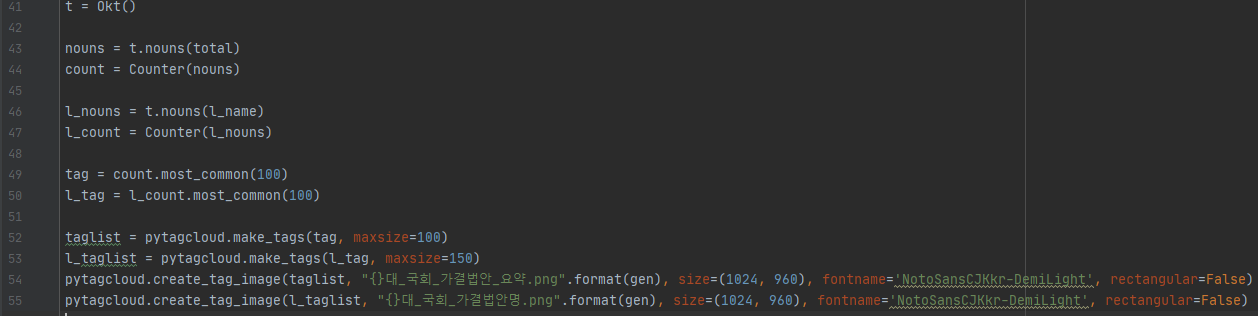
각 페이지에서 법안의 목록이 있는 table을 읽고 ‘a’태그를 통해서 의안명과 입법정보 링크, 의결현황을 알아냅니다. 의결현황에서 ‘원안 가결’과 ‘수정 가결’된 의안만 추출합니다.



추출된 의안에서 의안명을 리스트에 저장하고 입법정보 페이지를 다시 요청하여 ‘제안이유 및 주요내용’ 정보도 다른 리스트에 저장합니다.



모든 페이지에서 같은 작업을 마치면 리스트를 각각 문자열로 만들고 문자열에서 ‘법률’, ‘규정’, ‘법’, 개정’ 등 빈도수는 높지만 중요하지 않은 단어를 제거합니다.



Okt클래스의 nouns 메소드로 명사만 추출합니다. 그리고 각 명사의 빈도수를 세서 상위 100개를 선택해서 워드 클라우드로 저장합니다.

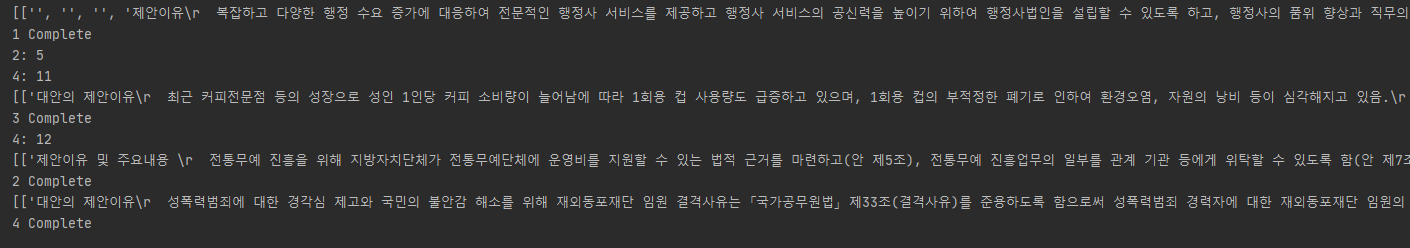


1 의안명 워드 클라우드 2 제안이유 및 주요내용 워드 클라우드

문제점 및 개선방안

가장 큰 문제는 느린 크롤링 속도였습니다. 20대 국회기준 2,415개의 페이지에 일일이 requests를 실행하고 의안별로 입법정보 페이지를 다시 요청하다 보니 한 번 실행하는데 약 2시간의 시간이 걸렸습니다. (jupyter notebook 기준 1페이지당 3초 이상)

보다 빠른 처리를 위해 멀티 프로세싱을 시도하였지만 queue의 크기가 전체 text의 크기보다 고 queue의 빈 자리가 생길 때까지 block 상태를 유지하는 put 메소드의 특성 때문에 각 프로세스가 종료되지 않고 대기하는 상황이 발생했습니다. Queue에 있는 리스트를 처리하는 프로세스를 따로 만들면 해결 가능할 것으로 생각합니다.



다른 문제로 ‘법률’, ‘규정’과 같이 중요하지 않은 단어를 직접 선택해서 제거하였지만 여전히 빈도수는 높지만 중요하지 않은 단어들이 많이 남아있는 것을 볼 수 있었습니다. 따라서, TF-IDF를 계산해서 높은 값을 가진 단어 위주로 추출한다면 유의미한 결과를 얻을 수 있을 것으로 예상합니다.

Github: <https://github.com/GangHyeokLee/Politics-Python-Final-Project.git>