과목 명: 데이터베이스시스템

담당 교수 명: 박 석

<<Project 3>>

서강대학교컴퓨터공학과

**[20121635]**

**[**장종석**]**

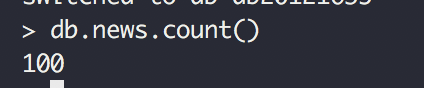
2 요구사항

본 프로젝트에서는 제공된 뉴스기사에 대해 형태소 분석을 통해 문서를 키워드의 집합으로 분할한 후 Apriori 알고리즘을 수행한다. 다음은 MongoDB에 저장된 뉴스기사에 대한 예제 화면이다. 아래 예제 화면을 참고하여 문서의 구조를 분석하고, 2.1 ~ 2.4까지의 작업을 진행 한다.

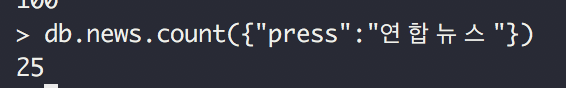
2.1. MongoDB 기본 질의

다음의 질의를 작성하고 질의문과 결과를 보고서에 기입하시오.

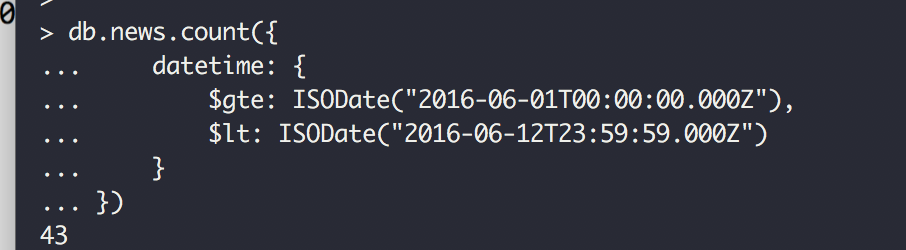
1) 전체 기사의 개수를 구하시오.



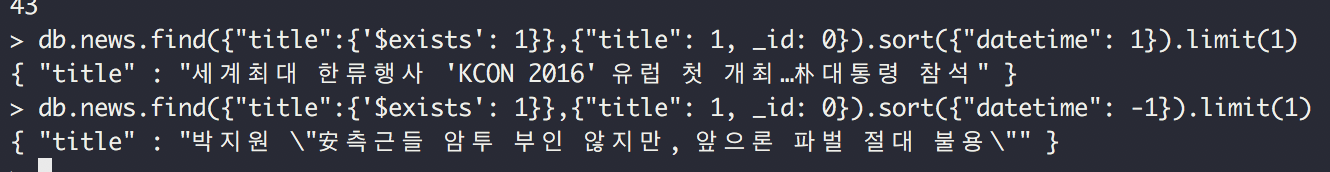
2) ‘연합뉴스’ 사에 실린 기사들의 개수를 구하시오.



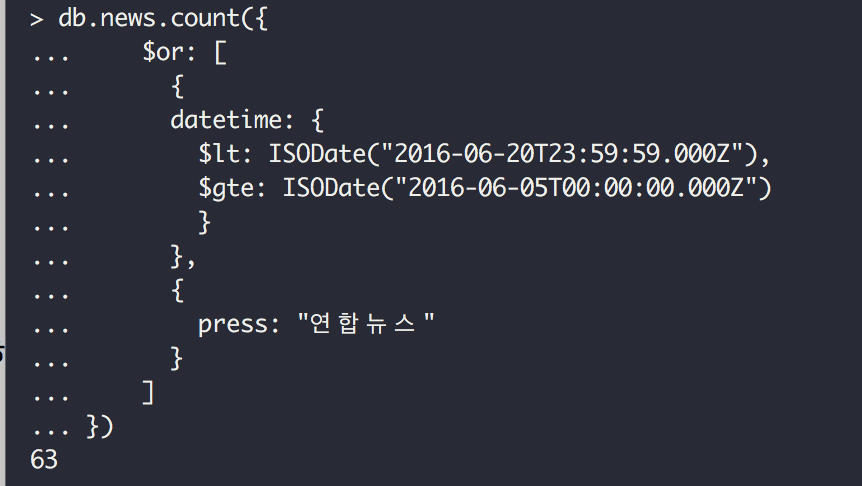
3) 2016년 6월 1일부터 2016년 6월 12일까지 실린 기사 개수를 구하시오.



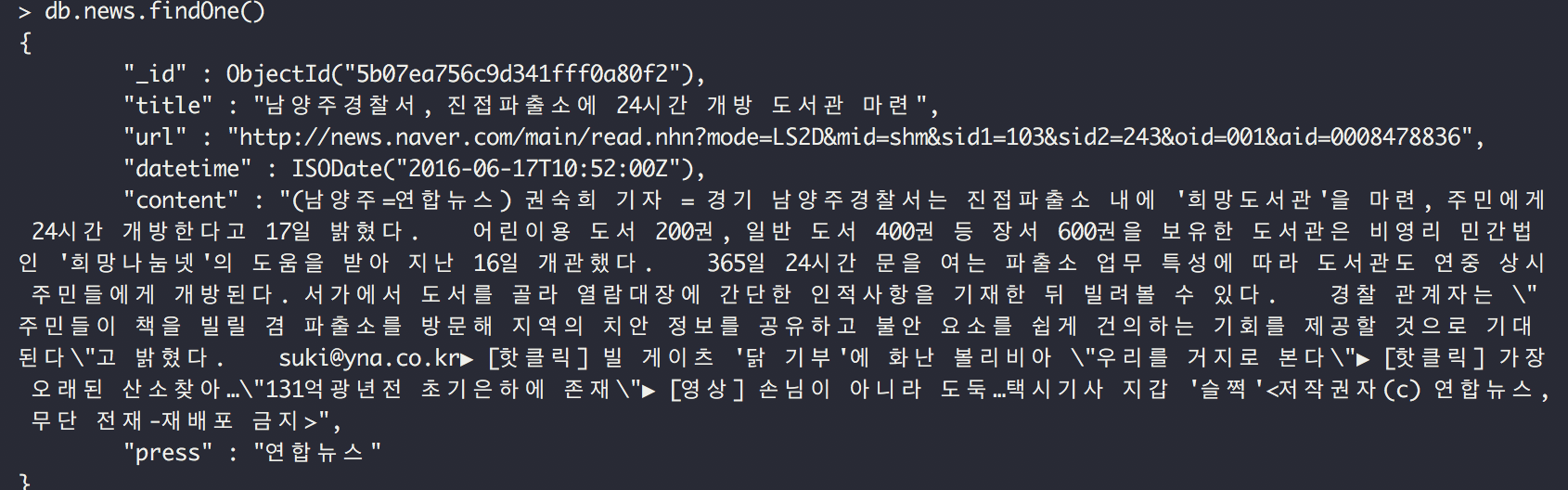
4) 전체 기사를 날짜 별로 정렬해서 가장 빠른 날짜와 늦은 날짜의 기사 제목을 출력하시오.



5) ‘연합뉴스’ 사에 실린 기사 혹은 2016년 6월 5일부터 2016년 6월 20일까지 실린 기사의 개수를 구하시오.



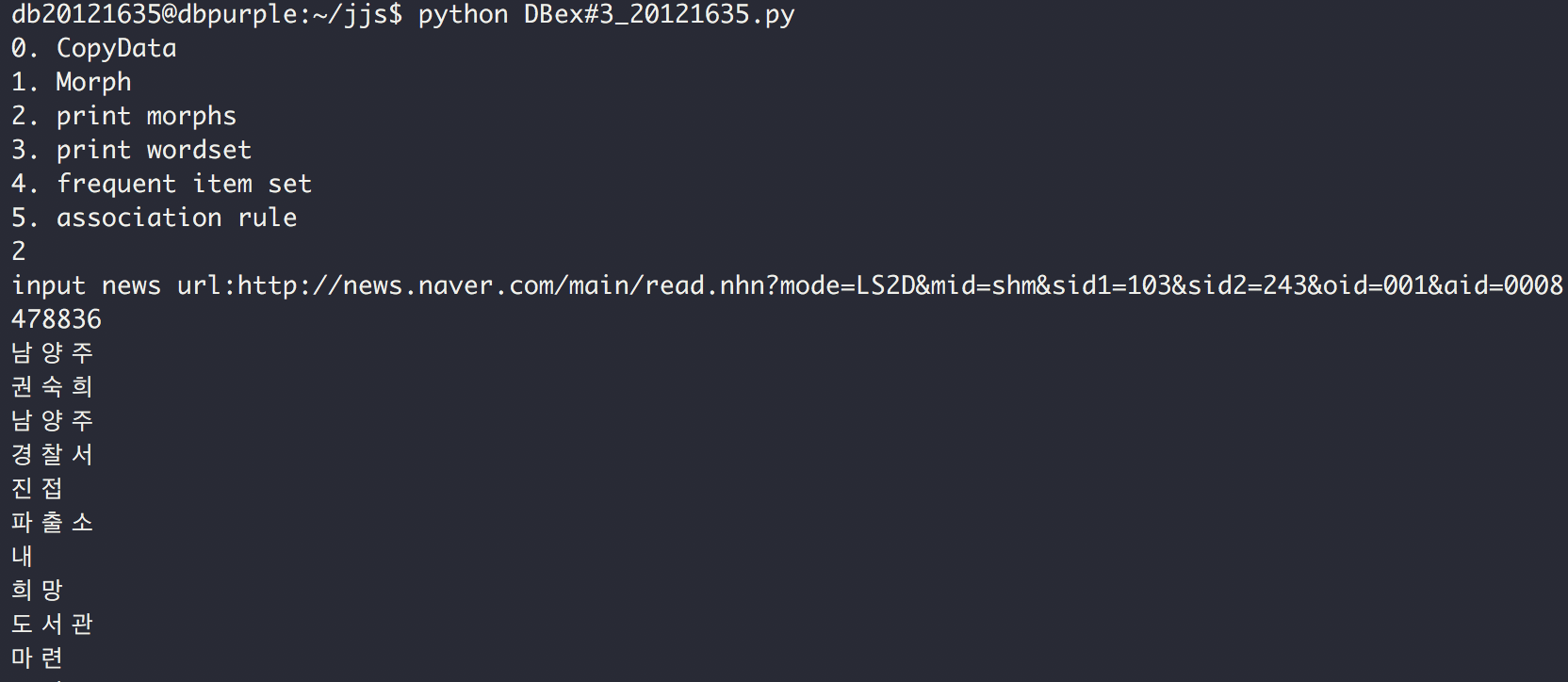
2.2. 뉴스 기사 전처리 과정



위에 첩부한 기사를 바탕으로 아래 질문들을 수행하였다.

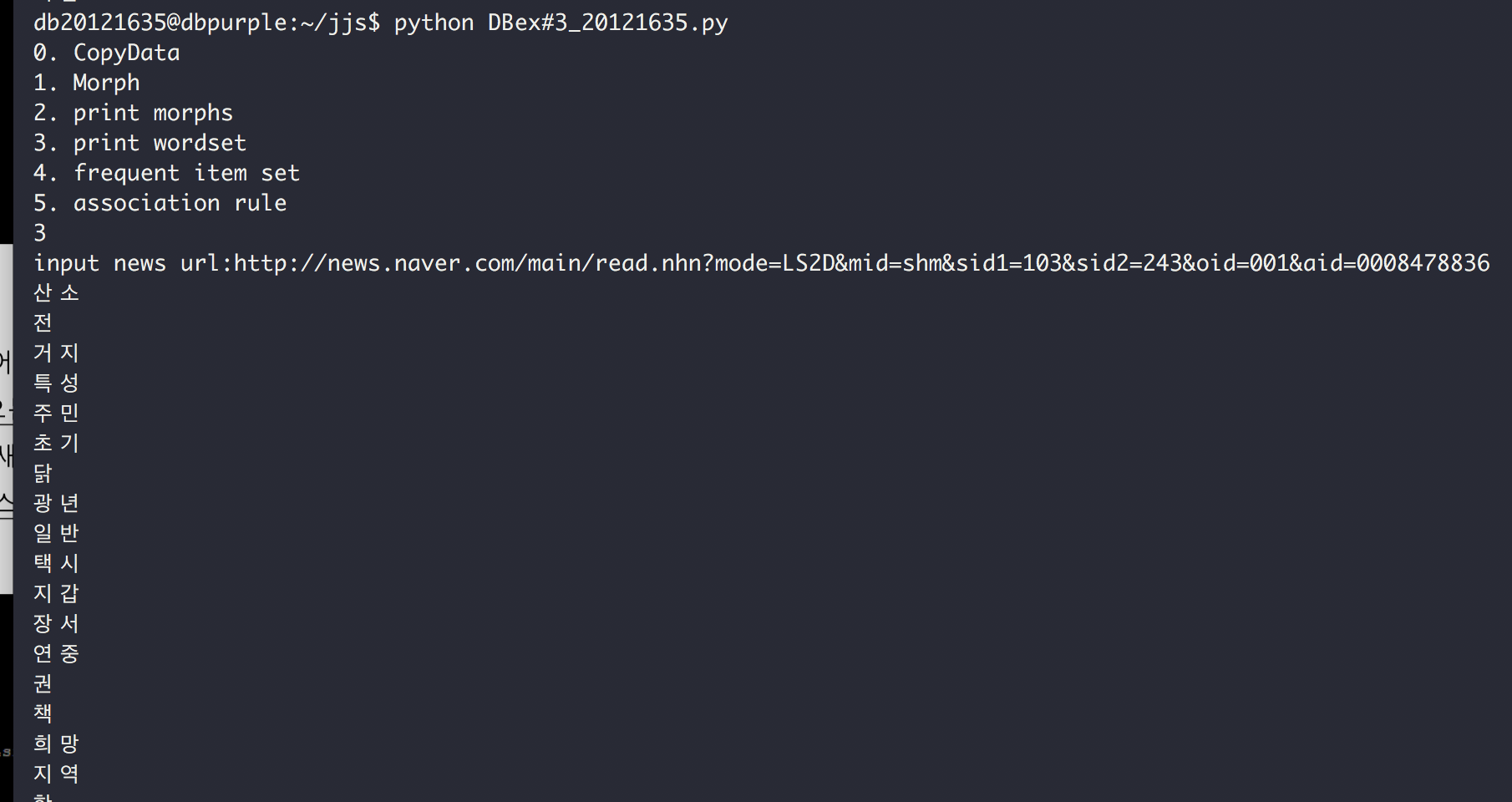
(1)모든 뉴스 기사에 대해 제공된 형태소 분석 소스 코드와 불용 어 리스트 파일을 이용해 텍스트 분석에 불필요한 단어(불용어)를 제거하고, 형태소 열이 추 가된 상태로 데이터베이스를 update한다.

(2)사용자로부터 뉴스 기사의 url을 입력 받아 해당하는 뉴스 기사의 형태소들을 출력 해준다.



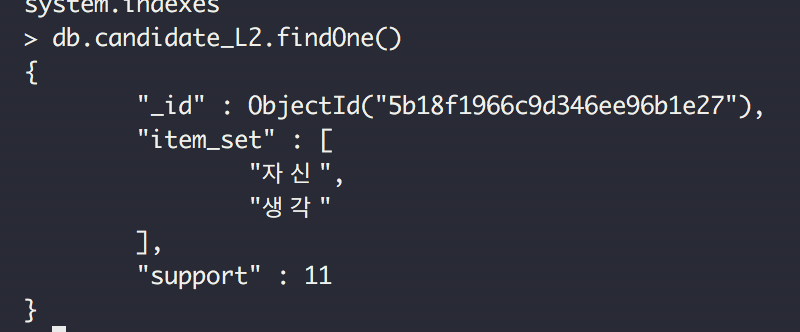
(3)모든 뉴스 기사에 대해 각 기사에 나오는 단어들을 확인하고, 그 단어들을 집합으로 만들어 새로운 collection에 저장한다.

(4)사용자로부터 뉴스 기사의 url를 입력 받아 해당하는 뉴스 기사의 Word set을 출력 해준다.

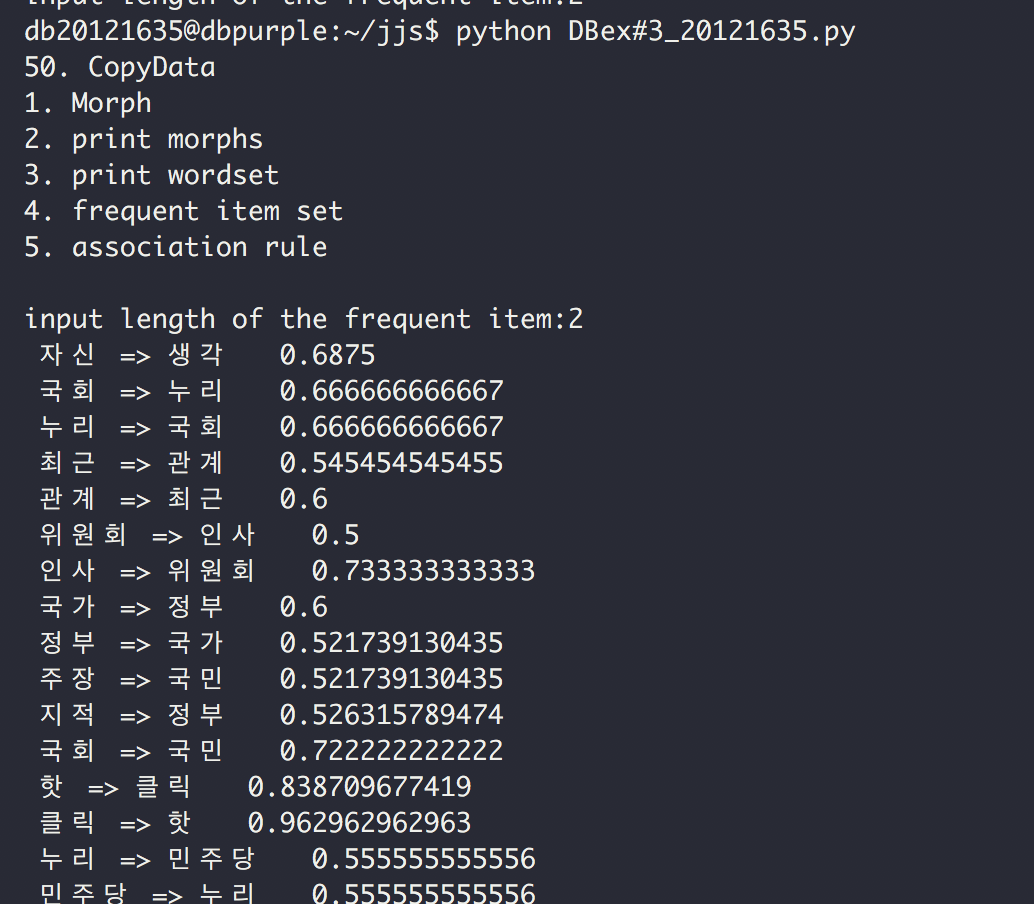


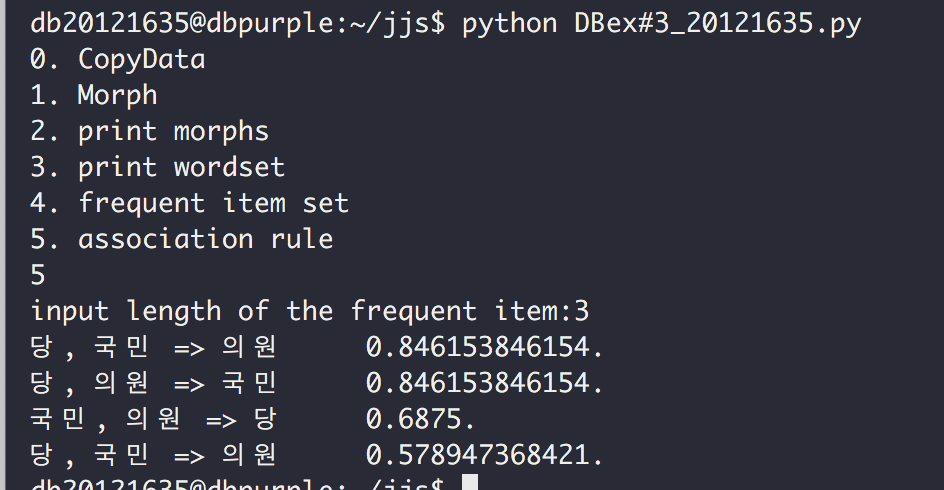
(5)frequent 1-itemset, frequent 2-itemset,frequent 3-itemset을 형성하고 DB에 저장하는 프로그램을 작성하라.





(6)본 프로젝트에서는 frequent n-th item set의 n을 입력 받았을 때 frequent n-th itemset에서 의 strong 연관 규칙을 모두 출력하는 프로그램을 작성하여라.





RDB vs. NoSQL DB에 대한 비교

텍스트 마이닝을 수행할 때에는 RDB보다 NoSQL DB를 사용하는 것이 좋은가? 이에 대한 답을 하고 이유를 기술하시오(진행된 프로젝트와 관련 지어 기술).

텍스트 마이닝을 수행하기 위해서 RDB 보다 NoSQL DB를 사용하는 것이 좋다. 텍스트 마이닝의 경우 비정형화된 데이터를 처리하는 기술인데, 비정형화된 데이터를 처리하기 위해서는 관계형 데이터베이스인 RDB보다, NoSQL을 사용 하는 것이 적절한데, 어떤 데이터가 존재할지 모르는 상황에서, 릴레이션 관계를 정의하거나, 엔티티나, 어트리뷰트를 정의하는 것은 적절한 방법이 아니며, NoSQL DB의 특징을 살리는 것이 좋다.

NoSQL DB의 경우, 빅데이터를 저장하기 위해 사용되며, SNS와 같은 비정형 데이터를 처리하기에 적합하여, FacdBook과 같은 대용량 SNS 서비스에서는 NoSQL을 활용하고 있다.

물론, NoSQL이 장점만 가지고 있는 것은 아니다. Search 시에 Fullscan을 해야하므로 다소 비용적인 측면이 많이 발생을하며, 이를 해결하기 위한 방법으로 다방면의 Index를 구축하여 해결할 수 있는데, 이 역시 비용이 많이 발생하게 된다.

최근 각광받는 NoSQL DB의 경우 Mongo DB, DynamoDB 등이 있으며, 최근 다양한 분야에서 사용되고 있다.

아래는 두가지 DB에 대해 차이점을 잘 나타내고 자료이다.

