2018 Database System Project #3

Text mining with MongoDB

20161584

민현홍

4.1.1 MongoDB 질의문 및 결과

2.1. MongoDB 기본 질의

다음의 질의를 작성하고 질의문과 결과를 보고서에 기입하시오.

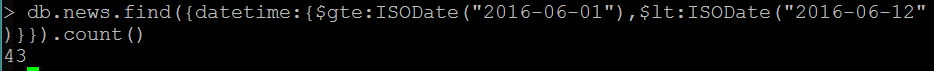
1. 전체 기사의 개수를 구하시오.

C:\Users\sskim\AppData\Local\Microsoft\Windows\INetCache\Content.Word\1-1.png

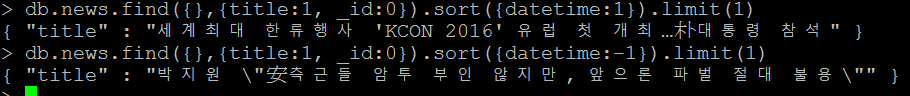
1. ‘연합뉴스’ 사에 실린 기사들의 개수를 구하시오.

C:\Users\sskim\AppData\Local\Microsoft\Windows\INetCache\Content.Word\1-2.png

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



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



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

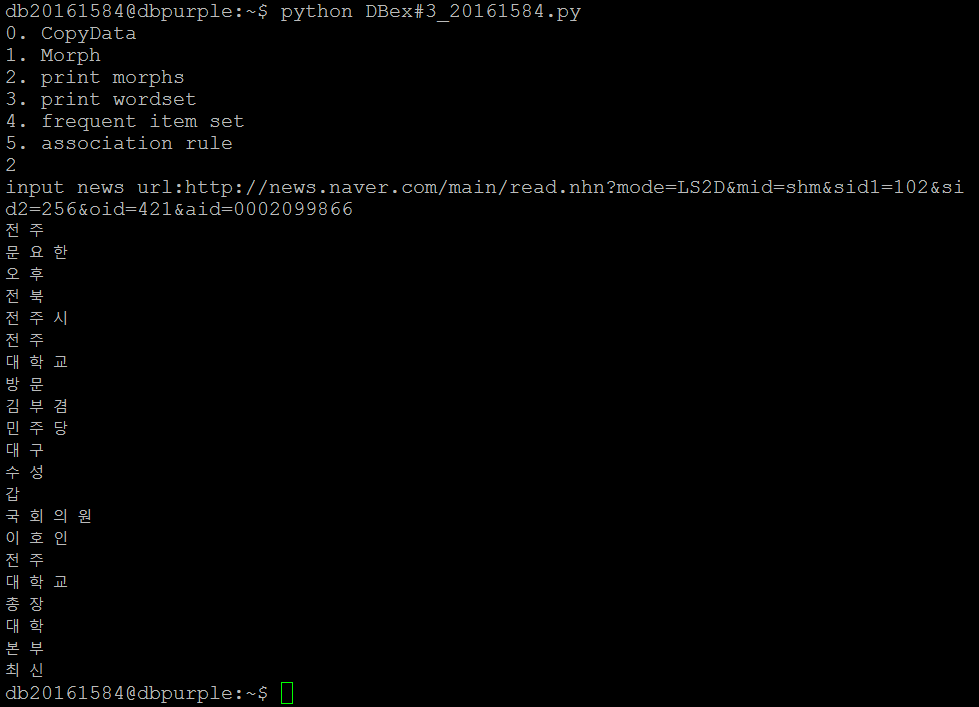
C:\Users\sskim\AppData\Local\Microsoft\Windows\INetCache\Content.Word\1-5.png

2.2. 뉴스 기사 전처리 과정

2.2.1. 형태소 분석 및 불용어 처리

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

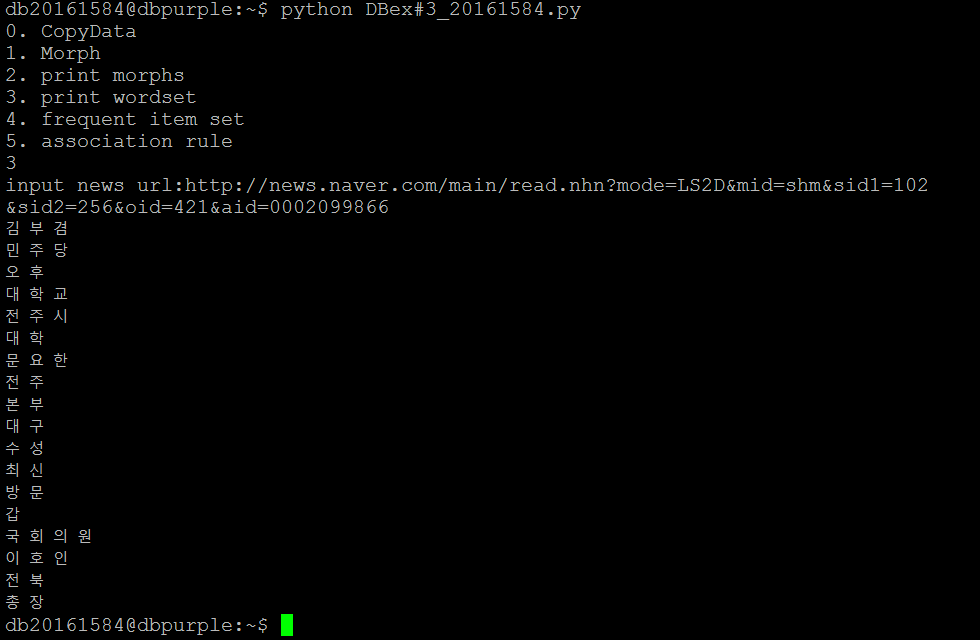
(2) 사용자로부터 뉴스 기사의 url을 입력 받아 해당하는 뉴스 기사의 형태소들을 출력 해준다.



2.2.2 한 기사 내의 형태소 집합 구하기

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

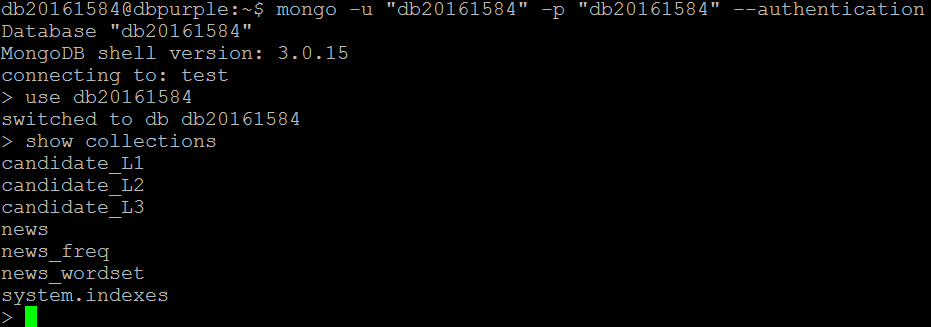
(4) 사용자로부터 뉴스 기사의 url를 입력 받아 해당하는 뉴스 기사의 Word set을 출력 해준다.

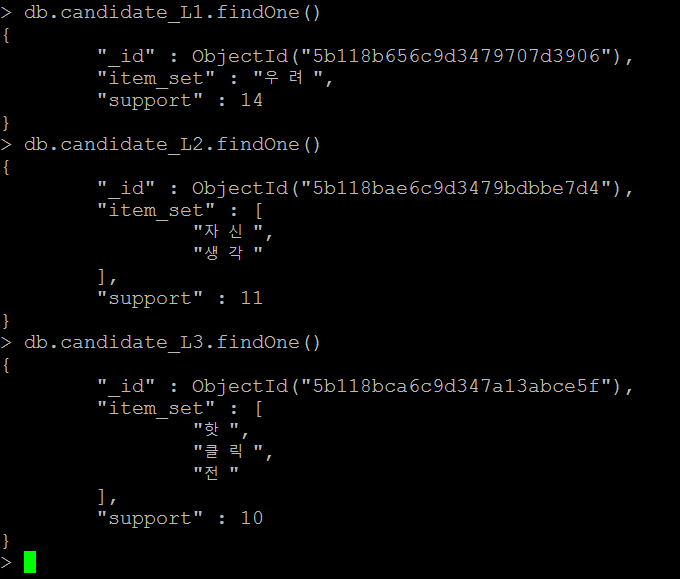


2.3 Apriori 알고리즘 구현

2.3.1 min sup을 만족 시키는 frequent itemset 생성

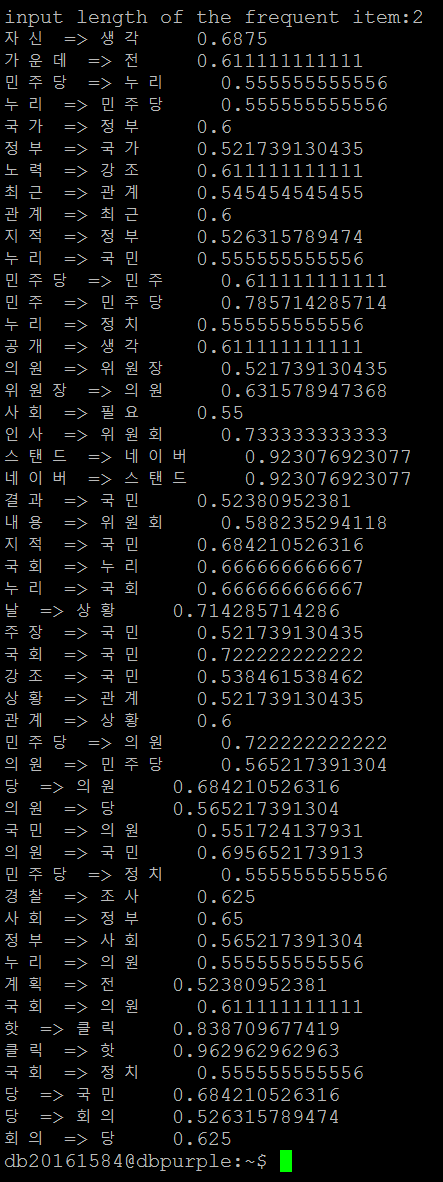
(5) frequent 1-itemset, frequent 2-itemset, frequent 3-itemset을 형성하고 DB에 저장하는 프로그램을 작성하라.

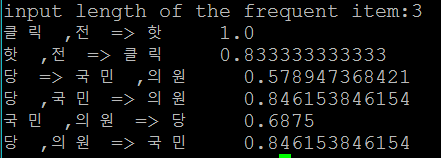




2.4.2 strong 연관 규칙을 생성

(6) 본 프로젝트에서는 frequent n-th item set의 n을 입력 받았을 때 frequent n-th itemset에서 의 strong 연관 규칙을 모두 출력하는 프로그램을 작성하여라. (min\_conf = 50%)





4.1.2 RDB vs. NoSQL DB에 대한 비교

* 텍스트 마이닝을 수행할 때에는 RDB보다 NoSQL DB를 사용하는 것이 좋은가?

텍스트 마이닝이란 비정형 데이터 마이닝 유형 중 하나로써 자연어 처리 기술과 문서 처리기술을 적용하여 반정형/비정형 텍스트 데이터를 정형화하고, 특징을 추출하기 위한 기술과 추출된 특징으로부터 의미 있는 정보를 발견할 수 있도록 하는 텍스트 마이닝 기술이다. 텍스트 마이닝의 주요 기술로는 자연어 처리(파싱, 형태소 분석, 품사 태깅, 관계 추출, 의미 추출), 언어모델링(언어 감지, 규칙기반 개체명 · 상용어 인식), 기계 학습 알고리즘(반복훈련을 통하여 습득한 정보 사용능력을 개선), 마이닝 기술(각종 통계적 기법을 활용한 정보 분류 및 분석 기술 · 기법) 등이 있다.

NoSQL은 은 SQL 계열 쿼리 언어를 사용할 수 있다는 사실을 강조한다는 면에서 "Not only SQL"로 불리기도 한다. NoSQL 데이터베이스는 전통적인 관계형 데이터베이스 보다 덜 제한적인 일관성 모델을 이용하는 데이터의 저장 및 검색을 위한 매커니즘을 제공한다. NoSQL 데이터베이스는 단순 검색 및 추가작업에 있어서 매우 최적화된 키 값 저장 기법을 사용하여 응답속도나 처리효율 등에 있어서 매우 뛰어난 성능을 나타낸다.

RDB란 관계형 데이터베이스라고 하며, 관계형 모델에 의한 데이터베이스로서 수학에서의 관계 개념을 응용한 것이다. 관계형 데이터 모델이란 데이터를 구성하는데 필요한 방법 중 하나로 모든 데이터를 2차원의 테이블 형태로 표현해 준다. RDB는 데이터의 독립성이 강하다. 또한, 최종 사용자와 프로그래머간의 의사 소통을 원활하게 하는 구조적 단순함이 RDB의 장점이다.

따라서 비구조적인 요소들을 포함하는 텍스트 마이닝의 경우 NoSQL에 유리하다. 왜냐하면 NoSQL은 RDB와 달리 따로 데이터 간의 relation을 정의 하지 않고, RDB보다 더 많은 대용량의 데이터를 저장 가능하고, 테이블의 schema가 유동적이기 때문이다.

즉, 이번 프로젝트와 같은 텍스트 마이닝의 경우 NoSQL이 RDB보다 적합하다.