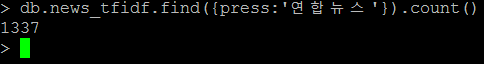
**2017 Database System Project #3**

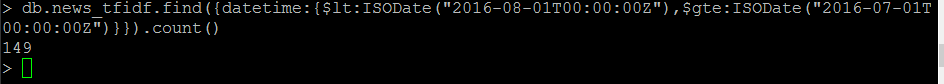
**Text mining with MongoDB**

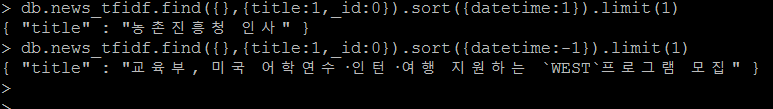
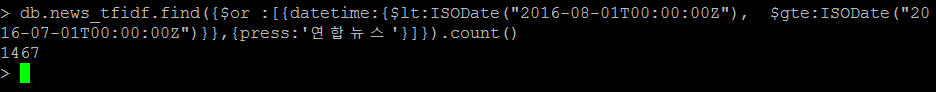
20141575 임형석

**1. MongoDB 질의**

다음의 질의를 작성하고 질의 문과 결과를 보고서에 기입하시오.

1. 전체 기사의 개수를 구하시오.  
   db.news.find().count()  
   
2. ‘연합뉴스’사에 실린 기사들의 개수를 구하시오.  
   db.news\_tfidf.find({press:’연합뉴스’}).count()  
   
3. 2016년 7월 1일부터 2016년 7월 31일까지 7월 한달 동안 실린 기사 개수를 구하시오.  
   db.news\_tfidf.find({datetime:{$lt:ISODate("2016-08-01T00:00:00Z"),$gte:ISODate("2016-07-01T00:00:00Z")}}).count()



1. 전체 기사를 날짜 별로 정렬해서 가장 빠른 날짜와 늦은 날짜의 기사 제목을 출력하시오.  
   
2. ‘연합뉴스’사에 실린 기사 혹은 2016년 7월 1일부터 2016년 7월 31일까지 7월 한 달 동안 실린 기사의 개수를 구하시오.  
   

**2. RDB vs NOSQL DB에 대한 비교**

텍스트 마이닝을 수행할 때에는 RDB보다 NOSQL DB를 사용하는 것이 좋은가?

이에 대한 답을 하고 이유를 기술하시오.

텍스트 마이닝을 수행할 때에는 NOSQL DB가 RDB를 사용하는 것보다 더 좋다.

RDB는 관계형 데이터 베이스로, 관계를 기반으로 한 DB 개체의 내부 관계가 아닌  
 외부의 관계를 적용한 것이다. 즉 table간의 관계가 맺어지는 DB형태로 대표적으로  
 oracle, Mysql등이 있다.

NOSQL은 RDB와 다르게 Table-Column과 같은 스키마 개념이 없어서 데이터 관계와 정해진 규격이 없어서 단순 검색 및 저장 작업에 매우 최적화된 저장 공간이다.

BigData의 발전에 있어서 hadoop과 함께 발전해 나가고 있는 새로운 DB의 유형이다.

대표적으로 이번 프로젝트에서 사용한 MongoDB가 있다.

RDB는 정형 데이터 처리에 유리하고 대용량 데이터 처리 시에는 성능이 저하 된다. 그에 반면에 NOSQL은 비정형,반정형 데이터 처리에 유리하고 대용량 처리 시 성능이 좋다. 그 뿐만 아니라 RDB는 주로 고가 스토리지와 서버를 이용하는 반면에 NOSQL은 Hadoop 과 같이 분산 시스템을 이용하여 스토리지를 제공한다. 예를 들어 100GB짜리 서버 10개를 연결하여 1TB 서버하나처럼 동작하게 하는 가상의 공간을 제공하고, 추가적으로 cluster에 서버를 계속 연결해서 용량을 점차적으로 RDB에 비해 더 저렴한 비용으로 확장시켜나갈 수 있다. 따라서 NOSQL이 RDB에 비해 대용량 데이터를 처리하는데에 더 유리하다고 한다.

텍스트마이닝은 이번 프로젝트에서 사용한 기사와 같은 비정형 데이터 마이닝의 유형 중에 하나다. 주로 텍스트 마이닝은 비정형 및 반정형 데이터에 대해 자연어 처리 기술이나 문서 처리 기술을 적용해 유용한 정보를 추출, 가공하여 적합한 정보를 획득하고 의미를 정제하고 범주화하는 것을 목적으로 사용하기 때문에 RDB를 이용하게 된다면 데이터 사이에서 원하는 정보를 추출하는데에는 제약이 있기 때문에 스키마 개념이 없는NoSQL이 더 적당하다. 또한 NoSQL이 단순 검색 및 저장 작업에 매우 최적화된 저장 공간이고 indexing기능이 정말 뛰어나므로 주로 정보를 추출 및 가공을 하는 텍스트 마이닝에 적합하다. 예를 들면 이번 프로젝트에서 stop words가 아닌 단어들을 추출해 count를 세거나 형태소 분석등 어려운 연산보다는 간단하게 정보를 추출하는 또한 텍스트같이 순식간에 용량이 늘어 날 수 있고대용량 데이터를 저장하기에는 RDB보다도 NoSQL이 비용적인 면에서나 확장성면에서나 더 저렴하고 사용하기에 적합하다.