DB project #3

20131575 유승재

1. MongoDB 질의문 및 결과

본 프로젝트에서 작성해야 할 질의문은 총 5가지이다. 이는 아래와 같다.

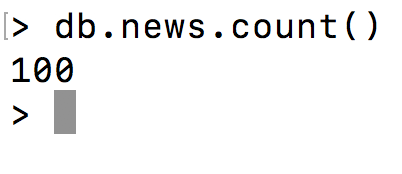
1. 전체 기사의 개수를 구하시오.
2. ‘연합뉴스’ 사에 실린 기사들의 개수를 구하시오.
3. 2016년 6월 1일부터 2016년 6월 12일까지 실린 기사 개수를 구하시오.
4. 전체 기사를 날짜 별로 정렬해서 가장 빠른 날짜와 늦을 날짜의 기사 제목을 출력하시오.
5. ‘연합뉴스’ 사에 실린 기사 혹은 2016년 6월 5일부터 2016년 6월 20일까지 실린 기사의 개수를 구하시오.

다음은 각각에 대한 질의문 과 수행 결과이다.

1. 전체 기사의 개수를 구하시오.

질의문 : db.news.count()

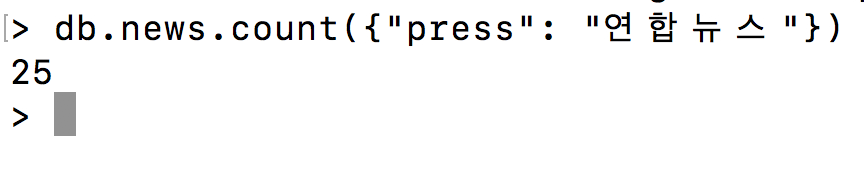
수행결과 :



1. ‘연합뉴스’ 사에 실린 기사들의 개수를 구하시오.

질의문 : db.news.count({"press": "연합뉴스"})

수행결과:



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

질의문 :

db.news.count({

datetime: {

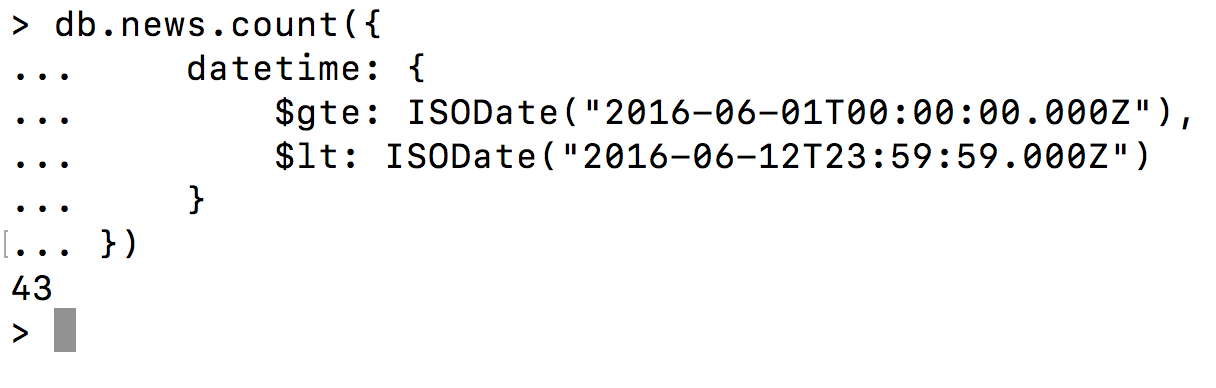
$gte: ISODate("2016-06-01T00:00:00.000Z"),

$lt: ISODate("2016-06-12T23:59:59.000Z")

}

})

수행결과:



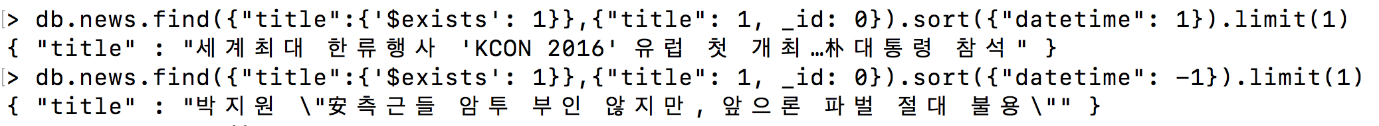
1. 전체 기사를 날짜 별로 정렬해서 가장 빠른 날짜와 가장 늦은 날짜의 기사 제목을 출력하시오.

질의문: (질의문을 두번으로 나누어서 함.)

db.news.find({"title":{'$exists': 1}},{"title": 1, \_id: 0}).sort({"datetime": 1}).limit(1)

db.news.find({"title":{'$exists': 1}},{"title": 1, \_id: 0}).sort({"datetime": -1}).limit(1)

수행결과:



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

질의문 :

db.news.count({

$or: [

{

datetime: {

$lt: ISODate("2016-06-20T23:59:59.000Z"),

$gte: ISODate("2016-06-05T00:00:00.000Z")

}

},

{

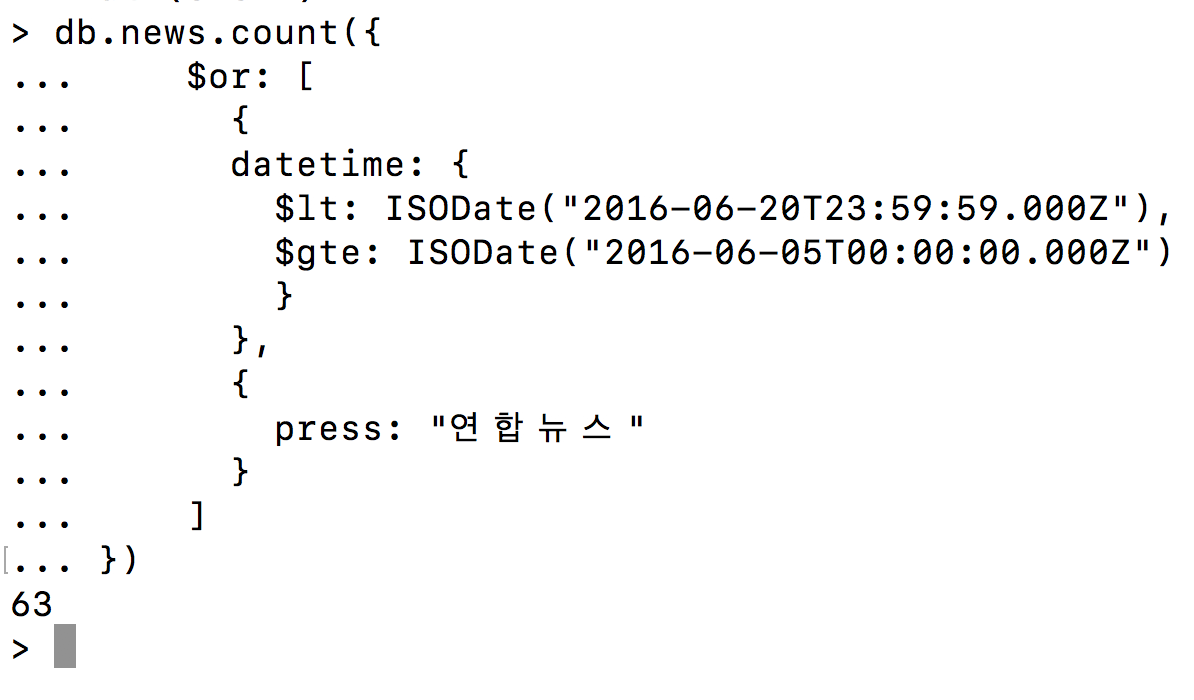
press: "연합뉴스"

}

]

})

수행결과:



1. 텍스트 마이닝을 수행할 때에는 RDB보다 NoSQL DB를 사용하는 것이 좋은가?

텍스트 마이닝을 수행할 때에는 RDB보다 NoSQL DB를 사용하는 것이 좋다. 주로 RDB는 관계형 데이터베이스로 웹에서의 게시판 글 같은 곳에 많이 쓰인다. 즉, 관계형을 저장하기 위해 사용하는 반면, NoSQL인 MongoDB는 빅데이터를 저장하기 위해 많이 쓰인다. 즉, 텍스트 마이닝 같이 많은 데이터베이스를 처리하기에 적합하다. 이는 RDB류인 MySQL은 table-based 데이터베이스인 반면 NoSQL은 document based, key-value based, key-value pairs, graph database이기 때문이다.

NoSQL은 이러한 특징 때문에 다이나믹 스키마가 가능하다. 즉, 다른 존재하는 데이터들을 수정하지 않고 자유롭게 data schema를 바꿀 수 있다. 또한 간단한 쿼리로 데이터를 가져오는 속도가 높다. 또한, 분명한 스키마 정의가 없이 빅데이터가 단 시간에 많이 쌓일 때 RDB보다 안정적이다.

실제로 본인은 MySQL을 샤딩(Sharding)하여 빅데이터를 저장해 보고, 똑같은 데이터를 MongoDB에 저장했을 때, 데이터를 fetch하는 속도가 MongoDB가 9배정도 빨랐다. 또한 요즘 추세는 그보다 더 빠르게 데이터를 불러오기 위하여 엘라스틱서치를 MongoDB와 River Plugin 하여 사용한다.

결론적으로, NoSQL은 수백만 트랜잭션을 처리하는 곳에 더 적합하고 이는 텍스트 마이닝 뿐만 아니라 다양한 빅데이터를 다루기에 적합하다.