데이터베이스시스템 프로젝트3

2016\*\*\*\*

1. MongoDB 기본 질의
2. 전체 기사의 개수

db.news.count()



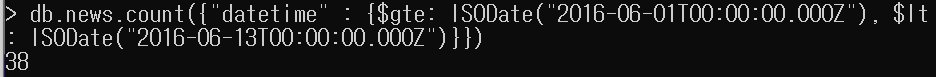
1. 연합뉴스 기사의 개수

db.news.count({press: "연합뉴스"})



1. 2016년 6월 1일부터 6월 12일까지 기사 개수

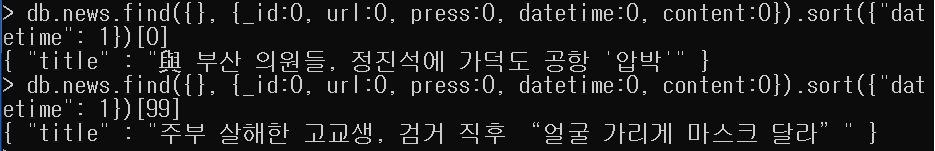
db.news.count({"datetime" : {$gte: ISODate("2016-06-01T00:00:00.000Z"), $lt: ISODate("2016-06-13T00:00:00.000Z")}})



1. 날짜 오름차순으로 정렬하고 첫번째 기사 제목과 마지막 기사 제목 출력

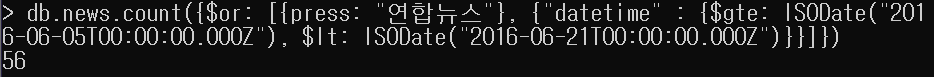
db.news.find({}, {\_id:0, url:0, press:0, datetime:0, content:0}).sort({"datetime": 1})[0]

db.news.find({}, {\_id:0, url:0, press:0, datetime:0, content:0}).sort({"datetime": 1})[99]



1. 연합뉴스 기사이거나 2016년 6월 5일부터 20일 사이에 나온 기사의 개수

db.news.count({$or: [{press: "연합뉴스"}, {"datetime" : {$gte: ISODate("2016-06-05T00:00:00.000Z"), $lt: ISODate("2016-06-21T00:00:00.000Z")}}]})



1. RDB와 NoSQL 비교

질문 : 텍스트 마이닝을 수행할 때에는 RDB보다 NoSQL DB를 사용하는 것이 좋은가?

답:

결론적으로, 텍스트마이닝과 같이 복잡한 대규모의 데이터를 처리할 때에는 RDB보다 NoSQL의 사용이 효과적이다. NoSQL은 RDB의 복잡한 관계형 데이터 구조와 그로 인해 발생하는 유지보수에 들어가는 시간, 저장공간 관리 등의 한계를 보완하기 위해 등장하였다. 그 대표적인 예로, NoSQL은 RDB와 달리 데이터 간의 관계를 정의하지 않는다. 따라서 각 데이터의 table 끼리 연산이 불가능한 대신 직관적으로 데이터에 접근할 수 있다. 또한, 하나의 고성능 서버에서 수행되는 RDBMS과 달리 NoSQL은 클라우드 분산처리 방식을 이용해 데이터를 관리한다. 여러 대의 클라우드에 데이터를 복제하고 분산시켜 저장하기 때문에 한 서버에서 장애가 발생하여 데이터의 손실이 일어나더라도 서비스가 중단되지 않고 데이터의 손실 역시 최소화할 수 있다. 마지막으로, NoSQL은 RDB와 달리 스키마가 유동적이다. RDB는 한 번 지정된 스키마의 규칙에 따라 데이터를 저장해야 한다. 하지만 NoSQL에는 스키마에 제한받지 않고 하나의 \_ID에 자유로운 필드 추가 및 삭제를 할 수 있어 데이터를 저장하고 수정하는 데 효율적이다.

실제로 본 프로젝트에서 NoSQL 형태인 MongoDB를 이용해 텍스트마이닝을 수행했을 때, 수백 여개의 말뭉치 데이터를 직관적이고 효율적으로 처리할 수 있었다. 그리고 프로젝트 수행 도중 백여 개의 데이터 중 오류가 발생했을 때 오류가 발생한 해당 데이터만 고치면 해결되어서 데이터의 용량이 커졌을 때 발생할 지 모르는 오류에 RDB 보다 유연하게 대처할 수 있을 것이라는 생각을 했다. RDBMS를 이용해서 수백 개가 아닌 수백만개의 데이터를 다루는 텍스트마이닝 프로젝트였다면 오류가 발생했을 때 모든 데이터의 관계 구조를 추적하여 오류의 원인을 찾아내고 디버깅하는 데 시간과 노력이 많이 소요되었을 것이다.