2019 Database System Project #3

Text mining with MongoDB

20131612  
최대운

1. MongoDB 기본 질의
   1. 전체 기사의 개수를 구하시오

스크린샷이(가) 표시된 사진

자동 생성된 설명

* 1. ‘연합뉴스’ 사에 실린 기사들의 개수를 구하시오



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

장치이(가) 표시된 사진

자동 생성된 설명

* 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



1. RDB vs NoSQL DB에 대한 비교

RDB의 특징은 테이블마다 스키마를 정의하고 데이터 타입과 제약을 통해 데이터의 정확성을 보장한다.(ACID 속성) 또한 SQL질의문을 통해 요청을 처리하고 DB를 만들기 위해선 모델링이 우선되어야 한다. RDB의 성능 향상을 위해선 하드웨어를 고성능의 하드웨어로 교체해야한다.(Scale up) 고성능의 하드웨어는 비용이 비싸기 때문에 결과적으로 RDB의 확장성은 좋지않다.

NoSQL DB는 RDB의 확장성을 해결하기 위해 고안된 것으로 분산컴퓨팅을 통해 성능을 향상 시킨다.(Scale Out) 또, NoSQL 데이터베이스는 특정 데이터 모델에 대해 특정 목적에 맞추어 구축되는 데이터베이스로서 유연한 스키마를 갖추고 있다. NoSQL DB는 개발의 용이성, 기능성에 강점이 있고 문서, 그래프, 키 값, 인 메모리, 검색과 같은 다양한 데이터 모델을 사용하고 위와 같은 데이터 모델 액세스 패턴에 대해 최적화 되어 같은 기능을 RDB로 개발 할 때보다 뛰어난 성능을 얻을 수 있다. 다만 RDB에서 사용하는 Join 연산이 필요한 질의는 NoSQL에서는 지원하지 않기 때문에 Join연산은 어플리케이션 수준에서 이루어져야한다. 이번 프로젝트처럼 텍스트 마이닝에 대해선 어플리케이션(파이썬 코드)에서 사용하는 모델 구조 형식(key-value)을 그대로 DB에 저장할 수 있기 때문에 DB 관리가 수월하다. 또한 프로젝트에서 사용된 데이터(뉴스기사 정보)는 작지만 이 후 데이터 양이 더욱 커진다면 확장성이 중요하다. 따라서 이번 프로젝트는 NoSQL DB가 RDB보다 더 적합하다고 할 수 있다.