**2019학년도 1학기 데이터베이스 Project 3 보고서**

**담당교수: 서강대학교 컴퓨터공학과 박석**

**1. MongoDB 질의문 및 결과**

아래는 본 프로젝트에서 작성해야 할 5가지 질의문을 작성하고 이를 실행한 결과를 사진과 함께 정리한 것이다.

|  |  |  |
| --- | --- | --- |
| 1) | 전체 기사의 개수를 구하시오.     |  | | --- | | db.news.count() | |
| 2) | ‘연합뉴스’ 사에 실린 기사들의 개수를 구하시오.     |  | | --- | | db.news.count({"press": "연합뉴스"}) | |
| 3) | 2016년 6월 1일부터 2016년 6월 12일까지 실린 기사 개수를 구하시오.     |  | | --- | | db.news.count({"datetime": {"$gte": ISODate("2016-06-01T00:00:00Z"), "$lt": ISODate("2016-06-13T00:00:00Z")}}) | |
| 4) | 전체 기사를 날짜 별로 정렬해서 가장 빠른 날짜와 늦은 날짜의 기사 제목을 출력하시오.     |  | | --- | | db.news.aggregate([{$sort: {"datetime": 1}}, {$group: { \_id: null, first\_date\_title: {$first: "$title"}, last\_date\_title:{$last: "$title"}}}]) | |
| 5) | ‘연합뉴스’ 사에 실린 기사 혹은 2016년 6월 5일부터 2016년 6월 20일까지 실린 기사의 개수를 구하시오.     |  | | --- | | db.news.count({$or: [{"press": "연합뉴스"}, {"datetime": {"$gte": ISODate("2016-06-05T00:00:00Z"), "$lt": ISODate("2016-06-21T00:00:00Z")}}]}) | |

**2. RDB와 NoSQL DB에 대한 비교**

텍스트 마이닝을 수행할 때 RDB(Relational Database, 관계형 데이터베이스)보다 NoSQL DB를 사용하는 것이 상대적으로 더 효율이 좋다. RDB는 수업 시간에 배운 바와 같이 Relation에 해당하는 테이블을 정의하고, 여러 테이블 간의 관계를 정의함으로써 schema를 형성한다. 그래서 웹사이트의 게시판 또는 Excel과 같은 프로그램을 사용하여 상품 정보를 정리하는 등 명확한 규칙 또는 제약이 존재하거나 데이터의 잦은 수정이 필요한 분야에서 주로 사용된다. 그러나 NoSQL DB는 key와 그에 대응되는 value가 pair로 묶인 여러 데이터를 document로 저장한다. 그래서 RDB와 달리 document 간의 관계가 존재하지 않아서 다른 데이터에 제약을 받지 않고 특정한 데이터 schema를 자유롭게 변경할 수 있으며(dynamic schema), 수정이 거의 없이 단순히 데이터를 검색하거나 저장 또는 삭제하기 위한 작업에 적합하다. 또한 단순한 Query로 데이터를 가져오는 속도가 빠른데, 이는 RDB에서 필요한 데이터를 탐색할 때 데이터와 관계를 맺는 다른 데이터를 JOIN 하거나 관계의 제약 사항을 고려할 필요가 적기 때문이다. 그래서 여러 transaction을 안정적으로 처리하기에 적절하다.

텍스트 마이닝은 많은 데이터를 처리하는 것을 요구하기에 빅 데이터 분야와 매우 밀접한 관련이 있으며, 위에서 서술한 NoSQL DB의 특징은 텍스트 마이닝처럼 빅 데이터를 처리하는 데 유리하다. 프로젝트에서 진행한 바와 같이 형태소를 분석하거나, 자주 등장하는 데이터를 추출하고 strong 연관 규칙을 생성하는 작업들은 텍스트 마이닝과 관련이 되므로 NoSQL DB에 속하는 MongoDB를 사용한 것이다.