**2020 DATABASE SYSTEM PROJECT #3**

**Text mining with MongoDB**

**학번 : 20171640**

**이름 : 박수진**

1. **MongoDB 질의문 및 결과**

**1) 전체 기사의 개수를 구하시오.**

db.news.find().count()



**2) 기술/IT 면에 실린 기사들의 개수를 구하시오.**

db.news.find({label: “기술/IT”}).count()



**3) 스포츠 기사 중 ‘우승’ 단어를 포함하는 기사들의 개수를 구하시오.**

db.news.find({$and: [{label: “스포츠”}, {$where: “this.content.match(‘우승’)”}]}).count()



1. **RDB vs. NoSQL DB에 대한 비교**

**텍스트 마이닝을 수행할 때는 RDB보다 NoSQL DB를 사용하는 것이 좋은가?**

RDB는 Relational Database의 약자로, key와 value들의 간단한 relation들을 테이블화 시킨 데이터베이스이며 관계를 기반으로 한 DB 개체의 내부 관계가 아닌 외부의 관계를 적용한 것이다. 예를 들어 oracle, MySQL 등 label 간의 관계가 존재하는 DB 형태가 RDB이다.

NoSQL은 Not only SQL의 약자로, SQL 계열 쿼리 언어를 사용할 수 있을 뿐만 아니라 table-column과 같은 스키마 개념이 없고 트랜잭션을 지원하지 않는다. 스키마 개념이 없기 때문에 데이터 관계와 정해진 규격이 없다. NoSQL은 덜 제한적인 일관성 모델을 이용하는 데이터의 저장 및 검색을 위한 매커니즘을 제공한다. 단순 검색 및 추가 작업에 최적화되어 있고 indexing 기능이 매우 뛰어나며 처리 속도나 효율이 좋다. 대표적 예시로 프로젝트에서 사용한 MongoDB가 있다.

텍스트 마이닝 수행 시 NoSQL DB를 사용하는 것이 효율적이다. 텍스트 마이닝은 비정형 데이터 마이닝 중 하나이다. 이번 프로젝트에서 사용한 기사를 예로 들 수 있다. 텍스트 마이닝은 이러한 비정형 혹은 반정형 데이터들에 대해 문서 처리 기술 등을 적용하여 원하는 정보를 추출 및 가공 후 의미를 정제하여 범주화하는 것을 목적으로 한다. 따라서 RDB를 이용하여 텍스트 마이닝을 수행하면 데이터에서 원하는 정보만 추출하기 어렵기 때문에 schema 개념이 없는 NoSQL이 적절하다. 또한 텍스트 마이닝에서는 테이블 간의 조인을 수행할 일이 없고, 대용량 데이터를 처리해야 하기 때문에 RDB보다 더 많은 데이터를 저장 가능한 NoSQL이 더 효율적이다.

|  |  |
| --- | --- |
| **MongoDB** | **SQL** |
| Collection | Table/View |
| Document | Record/Tuple |
| PK : \_id filed | PK : Any Attributes |
| Dynamic Schema | Fixed Schema |
| Index | Index |
| Embedded Document | Joins |
| Shard | Partition |