**과목명: 데이터베이스 시스템**

**1 분반**

**박석 교수님**

**<<Project #3>>**

**서강대학교 [컴퓨터 공학과]**

**[20171653]**

**[신나현]**

<목차>

1. **문제 정의**
2. **MongoDB 질의문 및 결과**
3. **RDB vs. NoSQL DB에 대한 비교**
4. **문제 정의**

본 프로젝트에서는 텍스트 마이닝 기법 중 하나인 FP-Growth Algorithm을 이용하여 제공된 뉴스 기사들을 분석하고 뉴스 기사에서 주로 쓰이는 단어들을 알아내는 프로그램을 작성한다. 또한 뉴스 기사 분석을 위해 비정형 데이터를 다루기 쉬운 NoSQL 기반 데이터베이스인 MongoDB를 사용함으로써 NoSQL 데이터베이스의 사용법을 익힐 뿐 아니라 관계형 데이터베이스와 NoSQL 데이터베이스간의 차이점을 인식하는 것을 목적으로 한다.

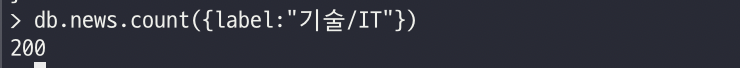
1. **MongoDB 질의문 및 결과**
2. **전체 기사의 개수를 구하시오.**

****

질의문: db.news.count()

결과: 600

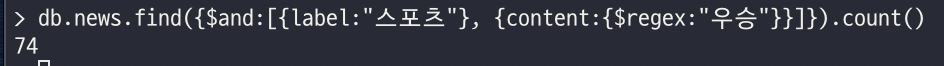
1. **기술/IT면에 실린 기사들의 개수를 구하시오.**



질의문: db.news.count({label:"기술/IT"})

결과: 200

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



질의문: db.news.find({$and:[{label:"스포츠"}, {content:{$regex:"우승"}}]}).count()

결과: 74

1. **RDB vs. NoSQL DB에 대한 비교**

RDB는 관계형 데이터 베이스로, 대표적으로 Mysql, Oracle등이 있다. RDB는 테이블마다 스키마를 정의하여야 하며 SQL 질의문을 통해서 요청을 처리한다. 또한 데이터 타입과 제약을 통해 데이터의 정확성을 보장해 준다. 그러나 데이터를 쿼리하는데 드는 비용이 높다. RDB의 경우 성능을 높이기 위해서는 하드웨어를 고성능으로 교체해 주어야 한다. 또한 전체 문서에 대한 검색은 별도의 도구가 필요하며 테이블에 대한 다양한 가변성이 있는 데이터를 저장하기는 어렵다.

NoSQLdms RDB와 달리 비관계형 DBMS로, 대표적으로 mongoDB, hBase등이 있다. 특히 mongoDB는 문서형 데이터베이스이다. 데이터 형식이 다양하며, 바꾸기 쉬워 대규모의 데이터를 유연하게 처리할 수 있다. 또한 테이블-컬럼과 같은 스키마가 없는 Schema-less구조로, 분산환경에서 단순 검색 및 추가 작업을 위한 키 값을 최적화 하고 지연과 처리율이 우수하다. 이는 대규모 확대가 가능한 수평적인 확장성의 특징을 가지고 있어서 정확성 보다는 데이터 양이 중요한 빅데이터(Big Data)에 많이 사용한다.

이러한 NoSQL DB는 기존의 RDB보다 더 큰 규모, 빠른 응답이 필요한 빅데이터에 적합하며 간단한 데이터 구조와 수평적 확장 덕분에 RDB보다 빠르게 응답이 가능하다. 이는 RDB와 달리 데이터 간의 관계를 정의할 필요가 없어 빠르게 데이터 구조를 설계할 수 있으며 방대한 양을 저장할 수 있다.

텍스트 마이닝이란 비정형 데이터 마이닝의 유형 중 하나로 자연어 처리 기술과 문서 처리 기술을 적용하여 유용한 정보를 추출, 가공하는 것을 목적으로 하는 기술이다. 실생활에서 만들어 지는 대부분의 문서는 형태로 보관되며 제목, 저자, 출판 날짜 등과 같은 구조적인 특징들과 문서의 요약, 내용과 같은 크기가 일정하지 않은 비 구조적 요소들을 포함하기에 반 구조적 데이터로 분류된다.

비구조적인 마이닝 유형인 텍스트 마이닝의 경우는 위와 같은 장점들을 갖는 NoSQL에 유리하다. NoSQL은 데이터간의 관계를 정의할 필요도 없으며 대용량의 데이터를 저장할 수 있어 대규모 데이터를 유연하게 처리할 수 있기 때문이다.