**데이터베이스시스템 Project #3**

**학과: 컴퓨터공학과**

**학번: 20171609**

**이름: 김민식**

**4.1.1 MongoDB 기본질의**

**다음의 질의를 작성하고 질의문과 결과를 보고서에 기입하시오.**

**1) 전체 트윗의 개수를 구하시오.**

db.tweet\_tfidf.find().count()



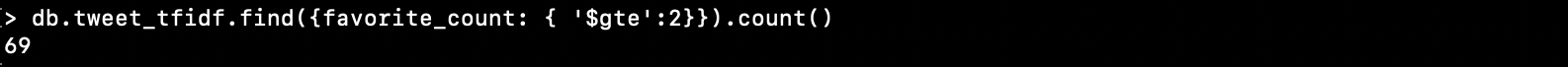
**2) retweet된 트윗의 개수를 구하시오.**

db.tweet\_tfidf.find({retweeted: true}).count()



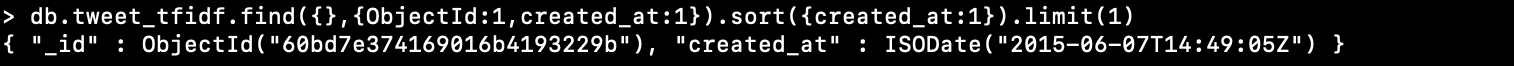
**3) 좋아요(favorite\_count)가 2 이상인 트윗의 개수를 구하시오.**

db.tweet\_tfidf.find({favorite\_count: { '$gte':2}}).count()



**4) 전체 트윗을 날짜 별로 정렬해서 가장 빠른 날짜의 트윗을 출력하시오. (ObjectID와 날짜만 출 력할 것.)**

db.tweet\_tfidf.find({},{ObjectId:1,created\_at:1}).sort({created\_at:1}).limit(1)



**5) 2015년 6월 1일부터 2015년 6월 30일까지 6월 한 달 동안 실린 기사의 개수를 구하시오. (6월 30일에 올라온 트윗도 포함.)**

db.tweet\_tfidf.find({created\_at: {$gte:ISODate("2015-06-01"), $lte:ISODate("2015-06-30")}}).count()



**4.1.2 RDB vs. NoSQL DB에 대한 비교**

**텍스트 마이닝을 수행할 때에는 RDB보다 NoSQL DB를 사용하는 것이 좋은가? 이에 대한 답을 하고 이유를 기술하시오(진행된 프로젝트와 관련 지어 기술).**

RDB란 Relational Database의 약자로 ‘관계형 데이터베이스’를 말한다. RDB는 모든 data를 2차원배열의 형태로 저장한다. 즉, data 끼리의 관계가 data 객체 간의 관계에서 보여진다. RDB는 배우기 쉽고, 효율적인 데이터베이스이다. 데이터의 독립성이 매우 강하며 DML(Data Manipulation Language)을 사용하여 데이터의 표현이 용이하다는 장점을 가진다. 마지막으로 data 끼리의 관계를 직접 조작하여 구조를 자유롭게 변경할 수 있다는 장점을 가지고 있다.

하지만 데이터의 양과 처리량이 계속해서 증가한다면 스키마 문제, 스케일업 한계 등의 문제점이 생긴다. 스키마 문제란 빅데이터를 RDM의 schema에 맞추어 변경해서 넣을 때 매우 긴 시간이 발생하는 문제점을 말하며, 스케일업의 한계란 RDBMS가 스케일아웃을 염두에 두고 설계되지 않았기 때문에 relation model과 트랜잭션의 연산, 일관성, 속성 등을 유지하며, 스케일아웃에서 RDBMS를 조작하는 것이 어렵다는 것을 뜻한다.

NoSQL DB는 Not Only SQL 데이터베이스를 말한다. SQL 계열 쿼리 언어를 사용할 수 있으며 RDB와는 다르게 2차원 배열의 형태로 데이터를 저장하지 않는다.

NoSQL DB는 RDB와 비교하였을 때, 데이터의 저장에서의 flexibility가 용이한 일관성 모델을 이용하여 데이터를 저장하고 검색을 위한 기능을 제공한다. 단순 검색 및 추가 작업에 최적화 되어 있으며 처리 속도나 효율이 매우 뛰어나다는 장점을 가진다. 마지막으로 분산형 구조를 통해 여러 대의 서버에 분산하여 저장하고 상호 복제하여 데이터 유실이나 서비스 중지에 대비할 수 있다는 장점이 있다.

RDB와 NoSQL DB 두 데이터베이스의 특징을 비교하였을 때, 이번 project3과 같은 텍스트마이닝을 수행할 때는 RDB보다 NoSQL DB를 사용하는 것이 적절하다고 생각한다.

텍스트 마이닝에서는 대용량 데이터를 처리해야한다. 이번 프로젝트에서도 많은 양의 tweet을 처리해야 했으며, 각 tweet 당 여러 개의 단어를 처리해야 하기 때문에 데이터의 용량이 매우 크다. 따라서 테이블 schema가 유동적이고 RDB보다 더 많은 데이터를 저장 가능하며, 데이터의 검색과 처리가 빠르게 될 수 있는 NoSQL DB의 사용이 더 효율적이라고 생각한다.

**4.1.3 exercise.py 실행 후 출력화면**

**1) 형태소 분석 및 불용어 처리**

텍스트이(가) 표시된 사진

자동 생성된 설명

**2) Word Count 구하기**

텍스트이(가) 표시된 사진

자동 생성된 설명

**3) TF-IDF 수행**

텍스트이(가) 표시된 사진

자동 생성된 설명

**4) 문서 유사도 구하기**

텍스트이(가) 표시된 사진

자동 생성된 설명