**2021 Database System**

Project #3 Text mining with MongoDB

|  |  |
| --- | --- |
| 학과 | 컴퓨터공학과 |
| 학번 | 20181593 |
| 이름 | 계인혜 |

1. MongoDB 질의 및 결과
2. 전체 트윗의 개수를 구하시오.

|  |  |
| --- | --- |
| 질의문 | db.tweet.count() |
| 결과 |  |

1. retweet된 트윗의 개수를 구하시오.

|  |  |
| --- | --- |
| 질의문 | db.tweet.find({‘retweeted’ : true}).count() |
| 결과 | 텍스트이(가) 표시된 사진  자동 생성된 설명 |

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

|  |  |
| --- | --- |
| 질의문 | db.tweet.find({‘favorite\_count’ : {$gte : 2}}).count() |
| 결과 |  |

1. 전체 트윗을 날짜 별로 정렬해서 가장 빠른 날짜의 트윗을 출력하시오.

|  |  |
| --- | --- |
| 질의문 | db.tweet.find({}, {\_id:1,created\_at:1}).sort({‘created\_at’ : 1})[0] |
| 결과 |  |

1. 2015년 6월 1일부터 2015년 6월 30일까지 6월 한 달 동안 실린 기사의 개수를 구하시오.

|  |  |
| --- | --- |
| 질의문 | db.tweet.find({‘created\_at’ : {$gte : ISODate(“2015-06-01T00:00:00Z”)}, ‘created\_at’ : {$lt: ISODate(“2015-07-01T00:00:00Z”)}}).count() |
| 결과 |  |

1. RDB vs. NoSQL DB에 대한 비교
2. RDB(Relational DataBase)

관계형 데이터 모델에 기초를 둔 데이터베이스이다. 관계형 데이터 모델이란 데이타를 구성하는데 필요한 방법 중 하나로 모든 데이타를 2차원의 테이블 형태로 표현하는 것이다. RDB를 사용하게 되면 데이터의 독립성이 높고, 높은 수준의 데이터 조작언어를 사용하여 결합, 제약, 투영 등의 관계 조작에 의해 표현 능력을 높일 수 있다. 또한 이들의 관계 조작에 의해 자유롭게 구조 변경이 가능하다. RDB의 대표 사례로는 oracle, MySQL이 있다.

다음으로는 RDB의 특징에 대해 살펴보도록 하자. 먼저 RDB는 모든 데이터를 2차원 테이블로 표현한다. 이때 테이블은 row, column으로 이루어진 기본 데이터 저장 단위이다. 또한 RDB는 상호 관련성이 있는 테이블들의 집합으로, 만들거나 사용하기에 비교적 쉽지만 무엇보다도 확장이 용이하다는 장점을 갖는다. 마지막으로 RDB는 ERD(Entity Relationship Diagram)을 이용하여 데이터베이스를 만들고, 이때 데이터베이스는 하나 이상의 테이블로 구성된다.

1. NoSQL DB(Not Only SQL)

NoSQL은 관계형 DBMS가 가지고 있는 특성 뿐만 아니라 다른 특성들을 부가적으로 지원한다. NoSQL은 기존의 RDB보다 융통성있는 데이터 모델을 사용하여 데이터의 저장 및 검색에 특화되어있다. 이러한 성질을 이용하여 NoSQL은 단순 검색 및 추가 작업에 있어서 속도 및 처리 효율이 매우 뛰어난 특성을 갖는다. NoSQL의 대표 사례로는 MongoDB, Riak, hBase가 있다.

다음으로는 NoSQL의 특징에 대해 살펴보도록 하자. 먼저 NoSQL에는 관계형 모델이 사용되지 않으며 따라서 조인 기능이 없다. 또한 데이터 조회를 위해서 직접 프로그래밍을 해야하고, 여러 데이터베이스 서버를 묶어 하나의 데이터베이스를 구성하는 경우가 잦다. 그리고 스키마가 없기 때문에 유연하며, 수직 및 수평 확장이 가능하고, 데이터를 읽어오는 속도도 빠르다. 마지막으로 NoSQL은 데이터베이스의 서비스가 중단되지 않으며 자동 복구 기능을 갖는 것을 특징으로 한다.

1. 형태소 분석 및 불용어 처리

다음은 MorphAnalysis 메뉴를 선택한 결과를 캡처한 것으로 트윗의 object id를 입력받아 해당하는 트윗의 형태소를 출력하는 것을 볼 수 있다.

텍스트이(가) 표시된 사진

자동 생성된 설명

1. Word count 구하기

텍스트이(가) 표시된 사진

자동 생성된 설명

위 사진은 먼저 MorpAnalysis를 통해 특정 트윗에 어떤 형태소들이 저장되어 있는지 확인한다.(love, beautiful, perfect)

이후 wordcount 메뉴를 통해 같은 트윗에 대한 wordcount 값을 출력한다. ([1,1,1]) 이는 morph에 저장된 순서대로 각각 love, beautiful, perfect의 wordcount 값을 나타낸다.

텍스트이(가) 표시된 사진

자동 생성된 설명

위 사진 역시 먼저 MorpAnalysis를 통해 특정 트윗에 어떤 형태소들이 저장되어 있는지 확인한다.(baseball, team, boy, terrified, baseball, life)

이후 wordcount 메뉴를 통해 같은 트윗에 대한 wordcount 값을 출력한다. ([2,1,1,1,1,2,1]) 이는 morph에 저장된 순서대로 각각 baseball, team, boy, terrified, baseball, life의 wordcount 값을 나타낸다. 이때 두 번 등장하는 baseball에 해당하는 wordcount 값이 2로 올바르게 출력되는 것을 확인할 수 있다.