**데이터베이트 프로젝트 #3 보고서**

**20191657 최세은**

**1. MongoDB 질의문과 결과**

**1) 전체 트윗의 개수를 구하시오.**

db.tweet.count()



**2) retweet된 트윗의 개수를 구하시오.**

db.tweet.count({retweeted:true})



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

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



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

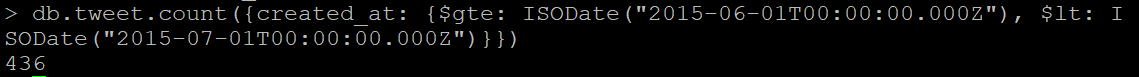
**력할 것.)**

db.tweet.find({}, {"\_id":true, "created\_at":true}).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.)**

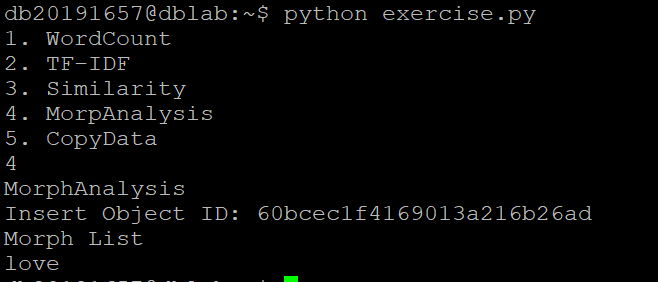
db.tweet.count({created\_at: {$gte: ISODate("2015-06-01T00:00:00.000Z"), $lt: ISODate("2015-07-01T00:00:00.000Z")}})



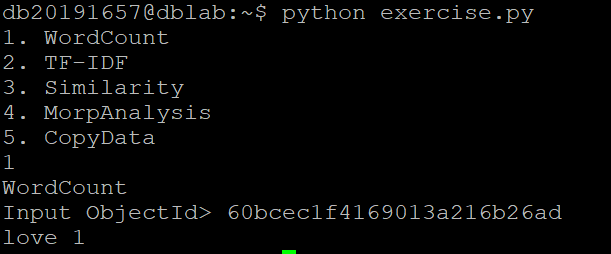
(결과: 436)

**2. 프로그램 결과**

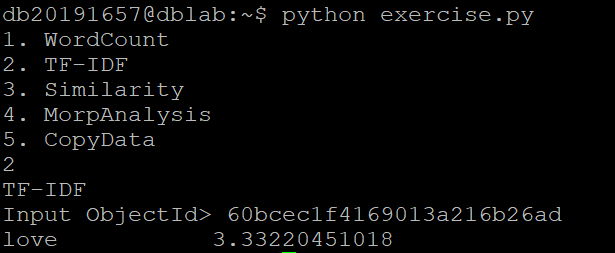
**1) 형태소 분석 및 불용어 처리**



**2) Word Count 구하기**



**3) TF-IDF 수행**



**3. RDB vs. NoSQL DB에 대한 비교: 텍스트 마이닝을 수행할 때에는 RDB보다 NoSQL DB를 사용하는 것이 좋은가? 이에 대한 답을 하고 이유를 기술 하시오(진행된 프로젝트와 관련 지어 기술).**

그렇다.

이 프로젝트에서는 NoSQL DB인 mongoDB로 텍스트 마이닝 기법 중 하나인 TF-IDF 가중치를 이용하여 제공된 트위터 데이터들을 분석하고 개별 트윗마다 핵심어를 추출함으로써 트윗들 사이의 유사도를 구할 수 있는 프로그램을 작성했다.

프로그램에서 TF-IDF 가중치를 계산하기 위해 형태소 등 다양한 형식의 데이터가 document에 저장되었다. NoSQL DB를 이용했으므로 다양한 형식의 데이터를 document에 쉽게 추가하고 변경할 수 있었다. 또한 프로젝트의 예제에선 약 1000개의 document가 있던 만큼 텍스트 마이닝에서 큰 데이터를 다루는데 NoSQL DB가 더 용이했다. 그리고 RDB보다 더 직관적이고 파이썬을 사용했기 때문에 프로그램을 짜기 쉬웠다.

물론 RDB는 더 정교하고 작업의 완전성을 보장하지만, 이 프로젝트의 데이터는 종류와 값이 계속 변하기 때문에 오히려 그 점이 유연하게 데이터를 다루는데 어려움을 주므로 NoSQL DB를 사용하는 것이 더 좋다.