**Malini Mittal**

**W205**

**Assignment 3**

**Task 2 - NoSQL (MongoDB):**

**Data Cleaning:**

Each tweet that gets read in will be cleaned to store only the required fields:

for tweet in tweets:

d = {}

d['created\_at'] = parser.parse(tweet['created\_at']).isoformat()

d['userhashtags'] = [h[u'text'] for h in tweet['entities']['hashtags']]

d['userscreenname'] = tweet['user']['screen\_name'].encode('utf-8')

d['username'] = tweet['user']['name'].encode('utf-8')

# … any other operation … #

**Data Organization:**

Create a database called ‘test\_database’. Then create a collection in this database called ‘test\_tweet\_collection’.

**Storage:**

All the cleaned up tweets will be stored in this collection as separate documents.

**Retrieval:**

# get the database and the collection

client = MongoClient('mongodb://localhost:27017/**'**)  
db = client.test\_database

tweets = db.test\_tweet\_collection

1. Who tweeted the most during the conference?

def userTweetCount(tweets, outdb):

#count the number of tweets per user, and sort

pipe = [

{ '$group': {'\_id': '$userscreenname', 'total' : {'$sum' : 1}}},

{ '$sort': {'total': -1}},

{ '$out': outdb}

]

tweets.aggregate(pipeline=pipe)

userTweetCount(tweets, "users")

users = db.users

print "users:"

for user in users.find().limit(5):

print user

1. What were the top ten hashtags used?

def hashtagCount(tweets, outdb):

#count the number of times each hashtag occurs, and sort

pipe = [

{ '$project': {'\_id': 0, 'hashtags': '$userhashtags.text'}},

{ '$unwind' : '$hashtags'},

{ '$group': {'\_id': '$hashtags', 'total': {'$sum': 1}}},

{ '$sort': {'total': -1}},

{ '$out': outdb}

]

tweets.aggregate(pipeline=pipe)

hashtagCount(tweets, "hashtags")

hashtags = db.hashtags

for hashtag in hashtags.find().limit(10):

print hashtag

1. How many tweets were produced each hour?

def getNumTweets(tweets, timestart, timeend):

return tweets.find({"created\_at" : { "$gte" : timestart, "$lt" :

timeend}}).count()

date1 = datetime.datetime.strptime("2015-02-14T08:00:00", "%Y-%m-%dT%H:%M:%S")

date2 = date1 + datetime.timedelta(days=1)

for day in (date1, date2):

for hour in range(0,7):

starttime = day + datetime.timedelta(hours=hour)

endtime = day + datetime.timedelta(hours=hour+1)

numtweets = getNumTweets(tweets, starttime.isoformat(), endtime.isoformat())

print ("Number of tweets on {0} between {1} and {2} =

{3}").format(starttime.date(), starttime.time(), endtime.time(), numtweets)