This document describes the basic steps for having the NYT Archive dataset that we downloaded locally installed in a MongoDB server in your machine. We also give here an example of how to query a database in MongoDB.
However, we remark that query sintax in MongoDB is a bit tricky, specially if you don't have experience with any other query language as SQL. We can help you with any particular information extraction task that you want to perform. Please write to mbeiro@fi.uba.ar for anything you need regarding this.

1. Downloading the dataset

Our dataset of the NYT Archive API comprises **14.099.370 articles** from 1882 to 2017 and is contained in one single compressed file of **2.2 GB** (gigabytes). It can be downloaded from http://cnet.fi.uba.ar/digging/hasdf12313443asfdasdads14asaas13mklkj43924/NewYorkTimes.rar. When you uncompress this file in your operating system you will get a file named caratulas.bson (16 GB) and one small file caratulas.metadata.json. These files have the format used by the MongoDB database system. This implies that you will have to install MongoDB in your operating system if you want to perform local queries to the data.

2. Installing MongoDB and loading the dataset

MongoDB is a free database management system. Its latest version can be downloaded from https://www.mongodb.com/download-center. If you use an Ubuntu Linux system, you could also install MongoDB with the following command:

```
sudo apt-get install mongodb-server
```

Once MongoDB is installed, we are going to load our dataset into the MongoDB database. This is done with the following command:

```
mongorestore -d nyt --drop filename.bson
```

This might take a couple of minutes. If this command succeeds, then you have MongoDB correctly installed and the NYT dataset correctly loaded into it.

3. Accessing the dataset

We will perform an initial simple query into the dataset, in which we will ask for articles talking about 'cherries' in 1970. We will use the MongoDB client interface which is accesed with the command `mongo`. (However, in the next section we will propose a more fashionable way of interacting with MongoDB through Python).
MongoDB has a special syntax for writing queries[1, 2, 3]. In this case, our query in the `mongo` client should look like this:

```
usuario@localhost:~$ mongo nyt

> var results = db.caratulas.aggregate([
                    { $project : {
                       year: { $year: "$pub_date" },
                       lead_paragraph: "$lead_paragraph",
                       headline: "$headline",
                       articleId: "$_id"}
```

```
    },
{ $match : {
    year: 1970,
    lead_paragraph: { '$regex' : ".*cherri.*"} }
}])
```

```
> result.next()
```

4. Working from Python

Using the MongoDB API for Python is the way that we prefer. In order to do this you will have to install Python in your operating system. We suggest you use Anaconda, which is a great Python distribution containing all the necessary packages and with a very simple instalation process. You can install it from: https://www.anaconda.com/download/ (choose the 3.6 version).
Then, from the command line you should run the following command to install the MongoDB Python API, called `pymongo`:

```
pip install pymongo
```

Now the installation is complete. Python is 'launched' by executing the following command, which will open your Web browser in a new window:

```
jupyter notebook
```

From there, you can create a *"New Notebook"* and insert the following code:

```
from pymongo import MongoClient
conn = MongoClient()

import pprint

col_archive_nyt = conn['nyt']['caratulas']

result = col_archive_nyt.aggregate( [
     { "$project":
         { "year": { "$year": "$pub_date" },
           "lead_paragraph": "$lead_paragraph",
           "headline": "$headline",
           "pub_date": "$pub_date",
           "articleId": "$_id"} },
    {"$match": { "year": 1970, "lead_paragraph": { '$regex' : ".*cherri.*"} }
    }
])

for r in result:
    pprint.pprint(r)
```

When you run it, you will see the result of the query in your screen!

Remind the typical structure of an article in the NYT Archive API:

```
{'_id': '4fd220358eb7c8105d7a3946',
 'abstract': "Argentine soccer player Diego Maradona's choice of Havana for his recovery from heart
              problems and cocaine addiction is viewed in Buenos Aires as something of propaganda
              victory for Fidel Castro's government and its claims of superior medical services;
              Maradona's troubles have prompted new anti—drug campaign by Argentine government (M)',
 'blog': [],
 'byline': {'original': 'By CLIFFORD KRAUSS',
            'person': [{'firstname': 'Clifford',
                        'lastname': 'KRAUSS',
                        'organization': '',
                        'rank': 1,
                        'role': 'reported'}]},
 'document_type': 'article',
 'headline': {'main': 'Ex—Argentine Soccer Star Makes News Again, Now in Cuba'},
 'keywords': [{'name': 'persons', 'value': 'MARADONA, DIEGO'},
              {'name': 'persons', 'value': 'CASTRO, FIDEL'},
              {'name': 'glocations', 'value': 'CUBA'},
              {'name': 'glocations', 'value': 'ARGENTINA'},
              {'name': 'subject', 'value': 'SOCCER'},
              {'name': 'subject', 'value': 'MEDICINE AND HEALTH'},
              {'name': 'subject', 'value': 'COCAINE AND CRACK COCAINE'},
              {'name': 'subject', 'value': 'DRUG ABUSE AND TRAFFIC'}],
 'lead_paragraph': 'Diego Maradona, once the sensation of the soccer world, was on the cover of
                    one of Argentina's most prominent political magazines, Noticias, last week. He
                    looked vastly overweight, his hair was dyed orange, and there was a huge tattoo
                    of Che Guevara on his right arm. Beside him was the headline, ''Therapy Cuba Style.''
                    In recent days, Mr. Maradona, now 39, has made some of the biggest headlines of his
                    long soap—opera career. His choice of Havana for his recovery from heart problems
                    and cocaine addiction is viewed as something of a propaganda victory for Fidel Castro's
                    government and its claims of superior medical services. His newest troubles have
                    prompted a new anti—drug publicity campaign by the government here.',
 'multimedia': [],
 'news_desk': 'Foreign Desk',
 'print_page': '11',
 'pub_date': datetime.datetime(2000, 2, 10, 0, 0),
 'section_name': 'World; Health',
 'slideshow_credits': None,
 'snippet': 'Diego Maradona, once the sensation of the soccer world, was on '
            'the cover of one of Argentinas most prominent political '
            'magazines, Noticias, last week. He looked vastly overweight, his '
            'hair was dyed orange, and there was a huge tattoo of Che Gueva...',
 'source': 'The New York Times',
 'subsection_name': None,
 'type_of_material': 'News',
```

'web_url': 'https://www.nytimes.com/2000/02/10/world/ex-argentine-soccer-star-makes-news-again-now-in-cuba.html',
'word_count': 595}

Just as an example, we can do things as:

- Counting the number of times that words appear (possibly by year).
- Counting how keywords interact with each other (possibly by year).
- Extracting articles which contain a certain keyword or word.

References

[1] N. O'Higgins, *MongoDB and Python*. O'Reilly, 2011.

[2] M. D. K Chodorow, *MongoDB, The Definitive Guide*. O'Reilly, 2010.

[3] *MongoDB Web reference*, 2017. [Online]. Available: https://docs.mongodb.com/manual/aggregation/