ISA 414 – Managing Big Data

Lecture 9 – Data Collection

APIs (Part I)

Dr. Arthur Carvalho

arthur.carvalho@miamioh.edu

MIAMI UNIVERSITY

Lecture Objectives

- Quick review of Assignment 1
- Learn about APIs and how they are used during data collection
- Understand the data exchange format called JSON

Lecture Instructions

- Download the notebook "*Lecture 9.ipynb*" available on Canvas
  - Open the above file with VS Code

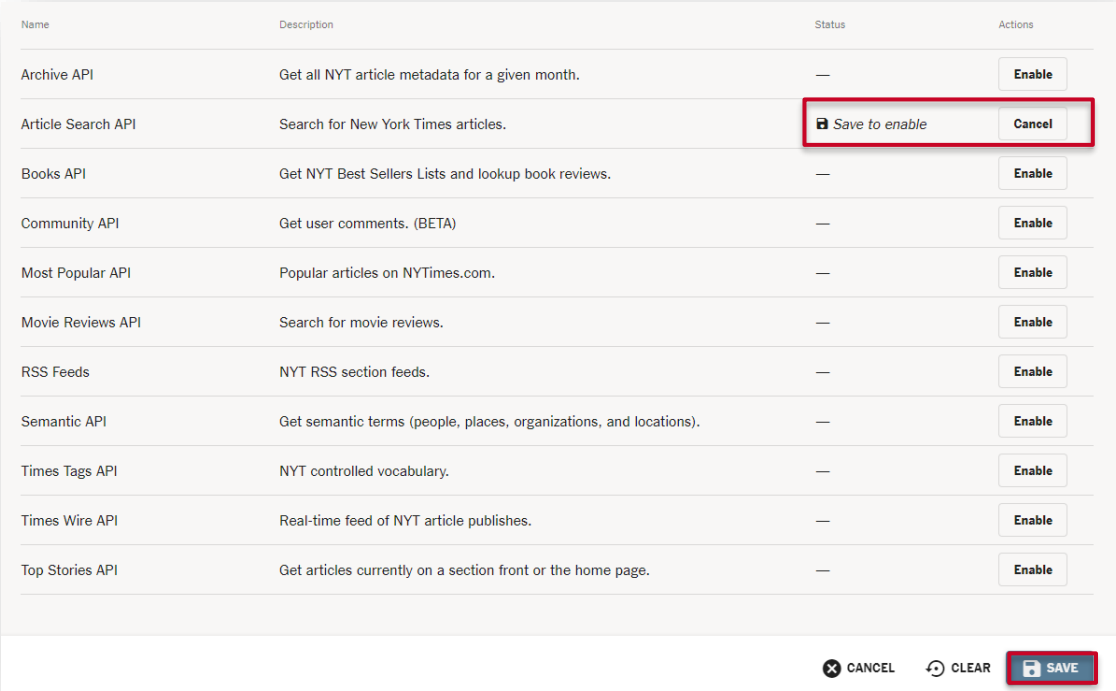
- Download the file NYTimes.json (optional)

Lecture Instructions

- Go to https://developer.nytimes.com/
  - Click on "*Get Started*" -> "*Create account*"
  - Fill in the text fields
  - Check your email and verify your identity
  - Click on "*Sign in,*" and use your email and password to sign in

Lecture Instructions

- Continuing …
  - Click on your email address (top right) -> "*Apps*"
  - Click on "*+ NEW APP*"
    - Pick an app name (e.g., ISA414)
    - Enable the "*Article Search API*" option
  - Click on *Save*

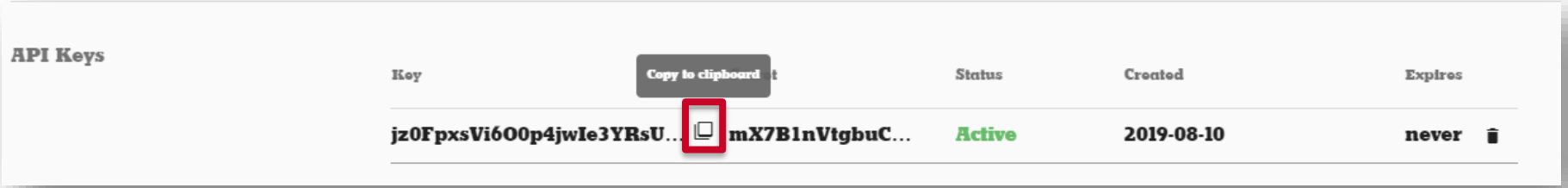


Name	Description	Status	Actions
Archive API	Get all NYT article metadata for a given month.	—	Enable
Article Search API	Search for New York Times articles.	*Save to enable*	Cancel
Books API	Get NYT Best Sellers Lists and lookup book reviews.	—	Enable
Community API	Get user comments. (BETA)	—	Enable
Most Popular API	Popular articles on NYTimes.com.	—	Enable
Movie Reviews API	Search for movie reviews.	—	Enable
RSS Feeds	NYT RSS section feeds.	—	Enable
Semantic API	Get semantic terms (people, places, organizations, and locations).	—	Enable
Times Tags API	NYT controlled vocabulary.	—	Enable
Times Wire API	Real-time feed of NYT article publishes.	—	Enable
Top Stories API	Get articles currently on a section front or the home page.	—	Enable

CANCEL CLEAR SAVE

Lecture Instructions

- Copy your API Key and save it somewhere



MIAMI UNIVERSITY

Data Collection

- Can one scrap all the data on the web? No!!!
  - Legal issues
    - Copyright


Contact Us

Farmer School of Business

The Farmer School of Business offers one of the world's best undergraduate business educations with small class sizes and faculty who are focused on teaching. We are consistently ranked among the nation's top undergraduate business programs for our commitment to our students' professional and personal growth and success. We place emphasis on experiential learning, international study and leadership development to prepare our students to become leaders in the workplace and the community.

Farmer School of Business
800 E. High St.
Oxford, OH 45056
deanofbusiness@miamioh.edu
513.529.3631

Directions and Parking

©2019 Miami University. All rights reserved.

  - Website owners can:
    - Block an IP address that is requesting too many webpages in a short period of time
    - Use captchas
    - Make small/random changes to HTML code
  - Controlled access to data is often done via API
    - Today's lecture

Data Collection

- Data are assets
  - Not all relevant data are freely available on the web
  - Many companies make money by directly or indirectly selling data, *e.g.*, Twitter, Google, Facebook
    - How can we access these companies' data?
      - Which protocol should we follow?
    - What is the format of the provided data?
      - Unlikely to be CSV
    - We will address the above questions in this lecture

API

- API: Application Programming Interface
  - Set of protocols that allow two+ software to communicate with each other
    - *E.g.*, to exchange data
  - Example:
    - When you write a script that formally requests data from the NY Times, that script must follow a certain protocol
      - Set of rules

API

- There are different types of APIs
  - Operating Systems (OS) APIs
    - Software/apps can access functionalities of the underlying OS
    - *E.g.*, an app can request Android to access a phone's camera
      - The request must follow a certain pre-defined protocol
      - The OS can approve or deny the request
  - Web APIs
    - Sometimes referred to as *web services*
    - Requests are often made via a protocol called REST
    - Responses are typically expressed in JSON or XML

API

- When using or developing web APIs, one must consider both:
  1. The protocol required when requesting data
  2. The format of the resulting responses

- We will learn how to use web APIs to request data, but not how to program/create APIs
  - Beyond the scope of this course

API

- Protocols
  - There are different ways of specifying an API protocol
  - Our focus in this course: **REST** or **RESTful** systems
    - Currently, the most popular method to create web APIs
    - Every resource (*e.g.*, data) is represented uniquely
      - For example, *via* URLs
    - Standard HTTP methods are very often used for retrieving, replacing, uploading, or deleting data
      - For example, <u>one can request data similarly to requesting a web page via URLs</u>

API

- **HTTP-based REST** operations

Uniform Resource Locator (URL)	GET	PUT	POST	DELETE
Collection of files, such as http://api.example.com/resources/	**List** the URLs	**Replace** the entire collection of files with another collection	**Create** a new entry in the collection of files	**Delete** the entire collection
Element, such as http://api.example.com/resources/item10	**Retrieve** an element of the collection	**Replace** an element of the collection	**Create** a new resource to an existent element	**Delete** an element of the collection

API

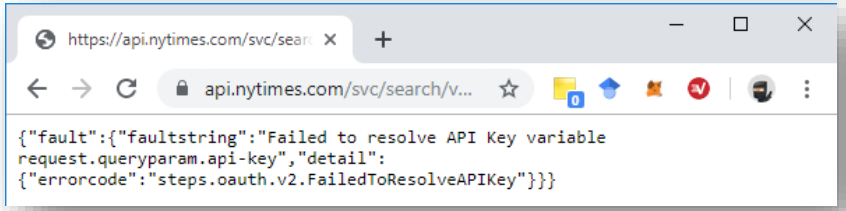
- We will focus primarily on request (GET) operations
  - We shall work with POST operations in the second half of the course
- Let's make some GET requests
  - Background story
    - Suppose you work for Microsoft
    - Your job: monitor social media and news outlets
      - Goal: quickly react to bad/good news
    - Let's find out what the NY Times is writing about Microsoft

API

- Let's make some GET requests
  - First, via a web browser to build intuition, then via Python
  - Base URL of NY Times API:

https://api.nytimes.com/svc/search/v2/articlesearch.json

  - After requesting data via the above URL, we get the following response
    - Some sort of an error



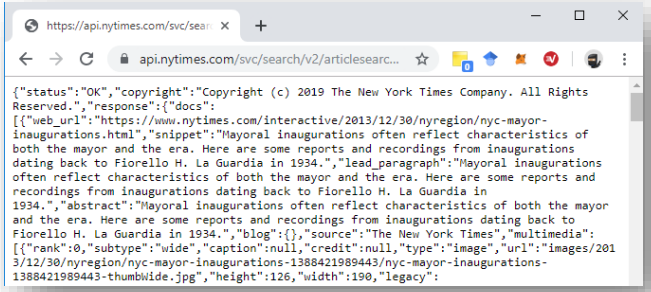
MIAMI UNIVERSITY

API

- The previous message says that the API Key is missing
  - That is, we are not authorized to request data
- API authorization is often tackled via login/pass and/or a token
  - Data access control
  - NY Times requires a token you obtained in the beginning of the lecture
    - Not always free
  - Let's try another URL: https://api.nytimes.com/svc/search/v2/articlesearch.json?api-key=add –your-token-here
    - The question mark '?' in the URL means that what comes next is a series of key-value pairs that will be used by the API to provide the right answer to the request

API

- It seems we get something
  - Data in a weird format (more on that soon)



- Let’s get data (articles) about Microsoft
  - Key = *q*, value = 'Microsoft'

https://api.nytimes.com/svc/search/v2/articlesearch.json?api-key=add –your-token-here&q='Microsoft'

MIAMI UNIVERSITY

API

- Once again, we get some data in a weird format, but this time about Microsoft


```
{"status":"OK","copyright":"Copyright (c) 2018 The New York Times Company. All Rights Reserved.","response":{"docs":[{"web_url":"https://topics.nytimes.com/top/news/business/companies/microsoft_corporation/index.html","snippet":"News about Microsoft Corporatio
including commentary and archival articles published in The New York Times.","blog":{},"multimedia":[],"headline":{"main":"Microsoft Corporation","kicker":null,"content_kicker":null,"print_headline":null,"name":null,"seo":null,"sub":null},"keywords":
[],"document_type":"topic","type_of_material":"timestopic","_id":"527ab68138f0d866066631f57","word_count":15,"score":3.6317077},{"web_url":"https://topics.nytimes.com/topic/company/microsoft-corporation","snippet":"News about Microsoft Corporation, including
commentary and archival articles published in The New York Times.","blog":{},"multimedia":[],"headline":{"main":"Microsoft Corporation","kicker":null,"content_kicker":null,"print_headline":null,"name":null,"seo":null,"sub":null},"keywords":
[],"document_type":"topic","type_of_material":"timestopic","_id":"56e0233c38f0d836dc036a59","word_count":15,"score":3.6236665},{"web_url":"https://topics.nytimes.com/top/news/business/companies/yahoo_inc/yahoo-microsoft-deal/index.html","snippet":"News about
Yahoo-Microsoft Deal , including commentary and archival articles published in The New York Times.","blog":{},"multimedia":[],"headline":{"main":"Yahoo-Microsoft
Deal","kicker":null,"content_kicker":null,"print_headline":null,"name":null,"seo":null,"sub":null},"keywords":[],"document_type":"topic","type_of_material":"timestopic","_id":"527ab88238f0d8660663221c","word_count":16,"score":2.8483667},
{"web_url":"https://www.nytimes.com/2018/02/15/technology/personaltech/microsoft-office-chromebook.html","snippet":"If you have a device running the Chrome OS but prefer Microsoft Office to Google Docs, you don't have to change your ways much.","blog":
{},"source":"The New York Times","multimedia":[{"rank":0,"subtype":"xlarge","caption":null,"credit":null,"type":"image","url":"images/2018/02/16/technology/personaltech/16techtipwebART/16techtipwebART-articleLarge.jpg","height":497,"width":600,"legacy":
{"xlargewidth":600,"xlarge":"images/2018/02/16/technology/personaltech/16techtipwebART/16techtipwebART-articleLarge.jpg","xlargeheight":497},"subType":"xlarge","crop_name":"articleLarge"},
{"rank":0,"subtype":"wide","caption":null,"credit":null,"type":"image","url":"images/2018/02/16/technology/personaltech/16techtipwebART/16techtipwebART-thumbWide.jpg","height":126,"width":190,"legacy":
{"wide":"images/2018/02/16/technology/personaltech/16techtipwebART/16techtipwebART-thumbWide.jpg","widewidth":190,"wideheight":126},"subType":"wide","crop_name":"thumbWide"},
```


- Note the use of '&' in the previous URL
  - Implies that another key-value term is coming
- How does one know the key-value pairs to use in a request?
  - *E.g., api-key* and *q*
  - Answer: read the API documentation: https://developer.nytimes.com/docs/articlesearch-product/1/overview

API

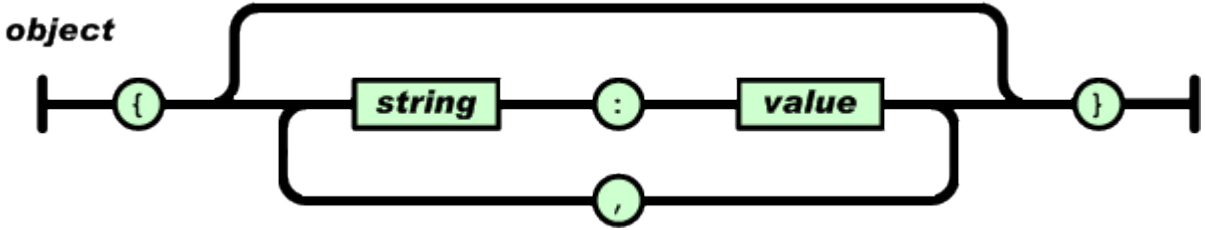
- To summarize
  - REST protocol to request data
    - GET, PUT, POST, DELETE operations
      - GET operations is similar to requesting a webpage, *i.e.,* one needs a URL
    - Common URL format: www.baseurl.com/somepage**?**key1=value1**&**key2=value2...
      - Example: https://api.nytimes.com/svc/search/v2/articlesearch.json**?**api-key=jz0FpxsVi6O0p4jwIe3YRsUcQHPFdDhx**&**q='Microsoft'

JSON

- Let's understand the retrieved data now
  - It is in a format called JSON (JavaScript Object Notation)
    - Data-interchange language
      - Widely used to communicate and store unstructured data
    - Textual (like CSV)
      - Format is completely language independent
    - Two key structures
      - A collection of key/value pairs, also known as *objects*
      - An ordered list of values, also known as *vector/array*

JSON

- Object
  - An unordered set of key-value pairs
  - Begins with { and ends with }
  - Each key is followed by : and the key-value pairs are separated by ,



JSON

- Object
  - Example: Professor (3 key-value pairs)

```
{"name":"Arthur",
 "dept":"ISA",
"position":“Paliwal Innov. Chair",
"position 2":"Assistant Professor"}
```

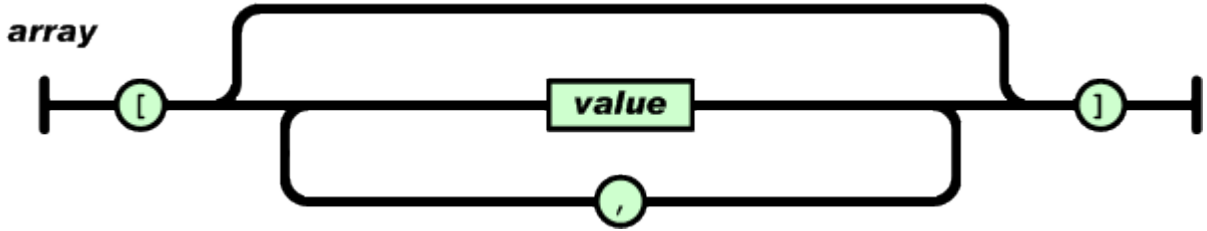
```
{ "name":"Skip",
  "dept":"ISA",
  "position":“Dept. Chair“ }
```

name	dept	position	position 2
Arthur	ISA	Assistant Professor	Paliwal Innov. Chair
Skip	ISA	Dept. Chair	

The above data could also be stored as a table. So why using JSON? Answer: flexibility

JSON

- Array/vector
  - Ordered collection of values
  - Begins with [and ends with] (*i.e.*, square brackets)
  - Values are separated by comma



JSON

- Array/vector
  - Example:

```
{ "TWD characters" : ["Rick", "Michonne", "Negan"] }
```

JSON

- We can always combine objects and vectors

```
{
 "menu": {
      "id": "night",
      "menuitem": [
                {"Item": "Rice", "Price": 3},
                {"Item": "Beef", "Price": 10},
                {"Item": "Chicken", "Price": 9}
      ]
 }
}
```

- Main object: menu
- Value associated with menu: an object containing 2 key-value pairs
  - Key-value pair #1: "id":"night"
  - Key-value pair #2: "menuitem":*array containing 3 objects*
    - Item: Rice, Price: 3
    - Item: Beef, Price: 10
    - Item: Chicken, Price:9

JSON

- Let's look at the JSON file you got from the NY Times
  - Open the file *NYTimes.json* (File -> Open File …) with VS Code

```
{
"response": {
    "meta": {
        "hits": [29997],
        "time": [9],
        "offset": [0]
     }
     "docs":[…],
  }
 "status": ["OK“],
  "copyright": ["Copyright (c) …"]
}
```

Line 1: the '{' indicates that the whole response is an object

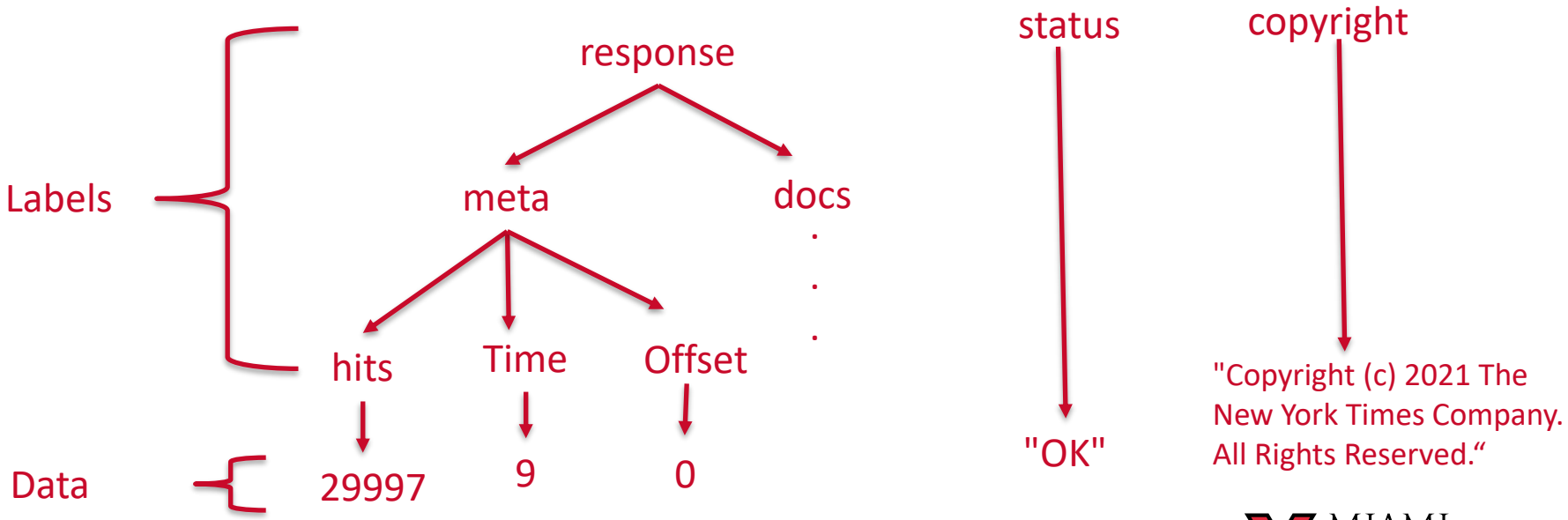
Black lines: the key *response* is associated with an object. That object in turn has two key-value pairs: *meta* and *docs*

Green line: 2nd key-value pair

Blue line: 3rd key-value pair

JSON

- The best way of interpreting JSON files is by means of a hierarchical tree



MIAMI UNIVERSITY

JSON

- Let's process JSON files with Python
  - Install the required modules: pip install requests
  - Request the JSON file

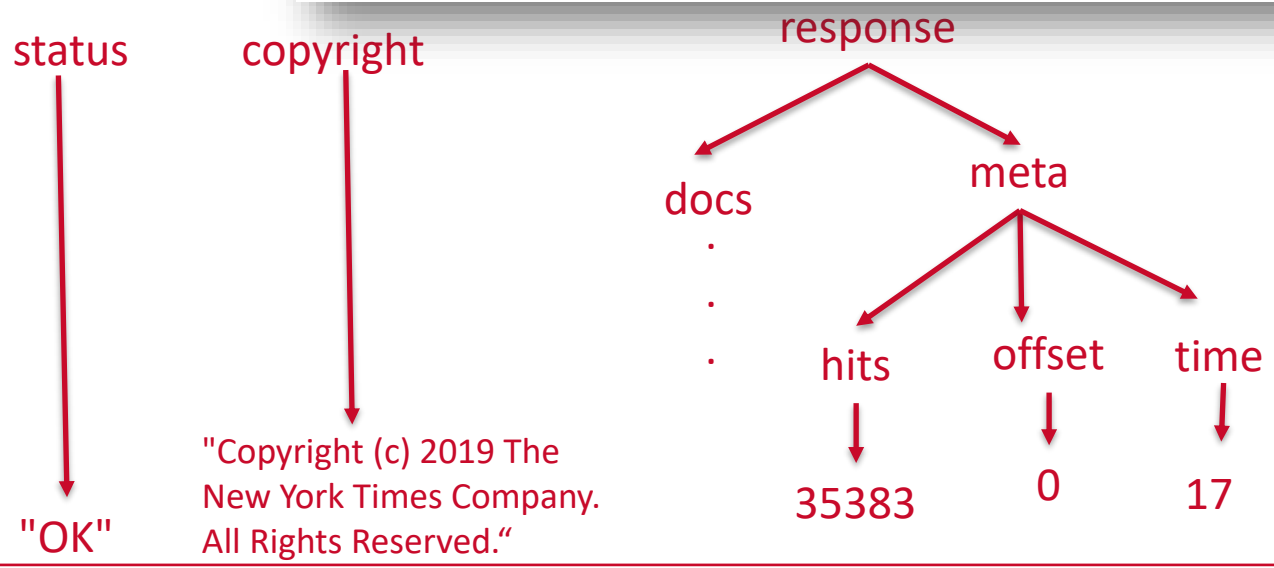
```
import requests

response = requests.get("https://api.nytimes.com/svc/search/v2/articlesearch.json?api-key=ADD YOUR TOKE HERE&q='Microsoft'")

json_data = response.json()
```

  - Python has a data structure called dictionary that handles the JSON format
    - We used that format before when creating data frames

JSON

- Let's navigate through the JSON file
  - With the help of VS Code, we can see that json_data is a **dictionary**

Name	▲ Type	Size	Value
json_data	dict	3	{'status': 'OK', 'copyright': 'Copyright (c) 2021 T



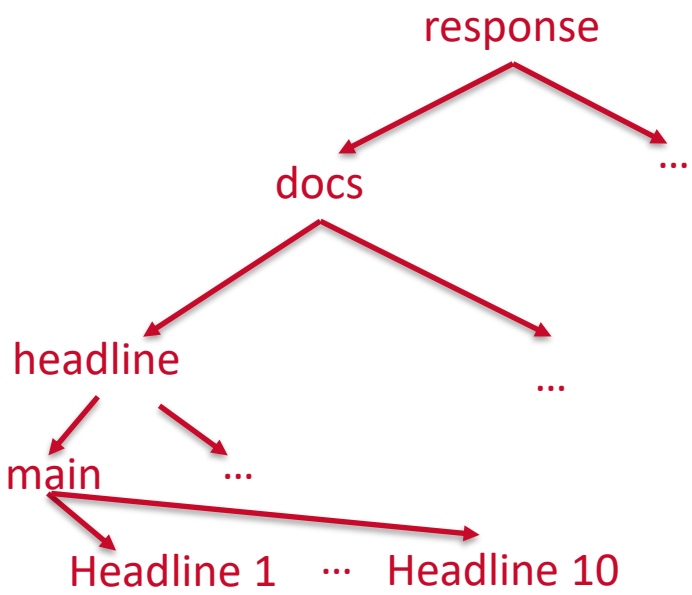
JSON

- The data we are looking for are inside/below *response* and *docs*
  - Let’s see what is inside response

```
response_data = json_data['response’]
response_data.keys()
```

JSON

- Obtaining the headlines

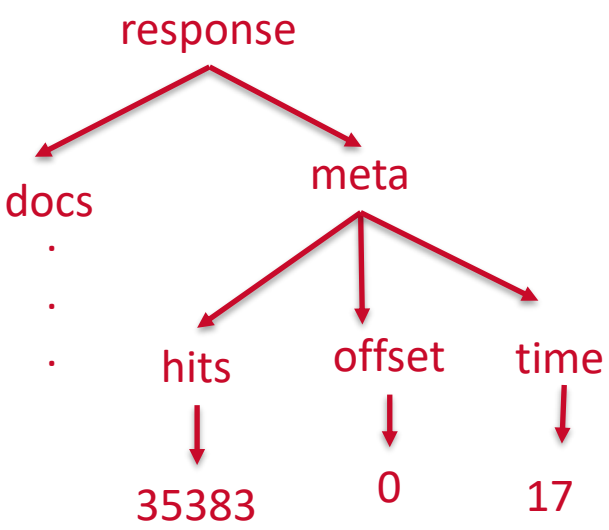



```
docs = response_data['docs']

headlines = []
for i in range(len(docs)):
    headlines.append(docs[i]['headline']['main'])
```

JSON

- Let's obtain the number of hits (articles about Microsoft)

```
meta = response_data['meta']
meta['hits']
```



API

- Final note
  - APIs make data collection very easy
  - Think about how cumbersome it would be to collect Microsoft-related articles from the NY Times via web scrapping (assuming this is allowed)
    - Download all articles
    - Filter the articles related to Microsoft
  - Can one get the content of a NYT article?
    - No (at the time of writing)

API

- Final note
  - There are a number of APIs out there
    - **YouTube**: https://developers.google.com/youtube/
    - **Google Maps**: https://developers.google.com/maps/
    - **Flicker**: https://www.flickr.com/services/api/
    - **Facebook**: https://developers.facebook.com/docs/graph-api
    - **Amazon Product Advertising**: https://affiliate-program.amazon.com/gp/advertising/api/detail/main.html
    - **Fitbit**: https://dev.fitbit.com/
    - **Twitter**: https://developer.twitter.com/en/docs.html
    - **KDnuggets 52 useful APIs**: http://www.kdnuggets.com/2017/02/machine-learning-data-science-apis-updated.html

Summary

- We learned about the concept of APIs
  - Request: REST protocol
  - Response: JSON file
- Next lecture
  - Introduction to APIs – Part II
    - Response: XML