# Introduction

Create a data set that contains data from websites/reddits/twitter of a certain topic.

# Spec

Choose a topic. For example, “sports”.

### Prework

Find 10~20 websites about this topic. Suggested ways to do this is

* Define some keywords of this topic. Then, use google search to find a set of websites.
* (Feel free to do anything fancier if you feel like doing it.)

### Part 1 (20 pts)

For each website, create a JSON object of the page that have

* images (urls or the images themselves)
* sentences containing a few keywords.
* # outgoing links
* website last-modified (from header)

The output can be a list of the JSON objects described above. For example,

|  |
| --- |
| [  {  “Images”: [“http://some\_image\_url/1”, “http://some\_image\_url/2”, “http://some\_image\_url/3”,]  “sentences”: [“The domestic dog is a member of the genus Canis.”, “Domestic dogs have been selectively bred for millennia for various behaviors.”]  “num\_outgoing\_links”: 5  “website\_last\_modified”: “10:20:40 pm Apr 3 2019”  },  {  // …..  }  ] |

### Part 2 (10 pts)

Part 2: Flatten the output of part 1 by sentences and write to a CSV. In this part, please define your own header. For example:

(sentence, image1, image2, image3, # outgoing links, last modified)

In this example, the output CSV file may look like.

|  |
| --- |
| sentence, image1, image2, image3, image4, # outgoing links, last modified  “The domestic dog is a member of the genus Canis.”, “http://some\_image\_url/1”, “http://some\_image\_url/2”, “http://some\_image\_url/3”, ?, 5, 10:20:40 pm Apr 3 2019  “Domestic dogs have been selectively bred for millennia for various behaviors.”, “http://some\_image\_url/1”, “http://some\_image\_url/2”, “http://some\_image\_url/3”, ?, 5, ? |

In this example, “?” means missing value.

The time format can be any format. I recommend to use epoch.

# Grading

You get full grades for submitting executable python code and data output satisfying the requirements above.

The assignment takes up 20% of the final grade.

# Submission

Please submit the following files through canvas

* Code
  + A notebook file or
  + A saved notebook with everything executed. (please use this option if you cannot run your code on Google colab.) or
  + A folder of python files with documentations on how to run it. (not recommended)
    - “data.json”, the output of the part 1.
    - “data.csv”, the output of the part 2.

Deadline: 10/4/2019

# Expectations

I expect this project can be done by someone knowing Python in 3 hours. If you are stuck, please ask on the canvas forum. I will do my best to help. I also encourage you to help your classmates.