The Art of Informational Retrieval

Rhys Chouinard, MSc Physics
Tyler Dauphinee, MSc Math Physics

LINK TO THIS TALK*:

https://bit.ly/2M8G67g

* *This link may be dead at the time of reading check the website for the summer school for an updated link.

Introduction

- Rhys Chouinard
  - Experimental Physics
- Tyler Dauphinee
  - Mathematical Physics

Structure

- We designed this session to be a crash-course in practical methods of data gathering.

- This should be as informal as possible and more of a workshop then a lecture.

- We encourage everyone to get their hands dirty and give each use-case a try.

Environment

- Docker container based on ubuntu 18.04 with a jupyterlab interface.

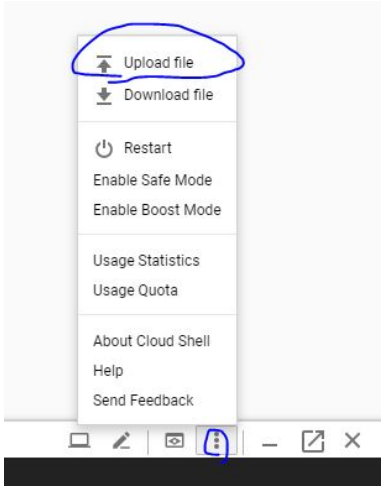
- Tested only on GCP in the cloud shell, so no guarantee of it working elsewhere.

Getting GCP Access

- Signup for a free google cloud account (https://cloud.google.com/)
- Create a new project (https://cloud.google.com/docs/overview/)
  - Don’t worry we will be utilizing only the cloud shell for this session so there will be no charges incurred.
- Open a cloud shell session (https://cloud.google.com/shell/docs/quickstart)

Downloading and Uploading the Code

- Download the repository*:
  - https://drive.google.com/open?id=1I_0Nu9QLmKFEXnSvDtghN2rCcHO1Aod-
  - If link does not work, check the hosting website for a link.
- Upload the zip to the cloud shell
  - Click the “three-dots” icon and select “Upload file”
  - Alternatively you can download the zip directly to the shell with wget if you prefer.
- Unzip the file in the cloud shell
  - Run the command `unzip <filename-here>` in the terminal



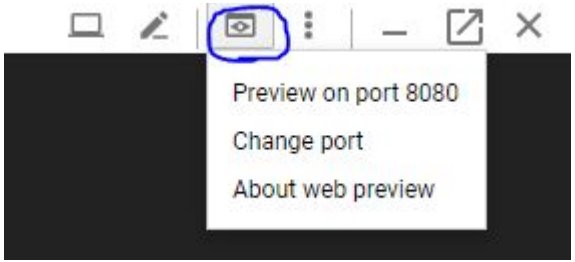
*This link may be dead at the time of reading check the website for the summer school for an updated link.

Running the container

- Navigate to the unzipped folder
  - It should contain three child directories “image”,”scripts”, and “work”
- Build the docker image
  - Locate the scripts folder and run the build script `bash build.sh`
- Start the container
  - In the same scripts folder run `bash run.sh`

Connecting to the Notebook

- Connect
  - In the top right of the cloud shell hit the “eye” icon and select preview on port 8080.
- Authenticate
  - The default token is ‘data-science’



API Overview

- What is an API?
- Common usage in data science?
- Hands-on intro
- Case Study: March Madness of Cartoons(?)

What is an API

- API stands for application programming interface
- It's a "common language" that many systems can decipher and use.
- Typically communicated via JSON objects (JavaScript Object Notation)
    - https://en.wikipedia.org/wiki/JSON


```
{
  "firstName": "John",
  "lastName": "Smith",
  "isAlive": true,
  "age": 27,
  "address": {
    "streetAddress": "21 2nd Street",
    "city": "New York",
    "state": "NY",
    "postalCode": "10021-3100"
  },
  "phoneNumbers": [
    {
      "type": "home",
      "number": "212 555-1234"
    },
    {
      "type": "office",
      "number": "646 555-4567"
    },
    {
      "type": "mobile",
      "number": "123 456-7890"
    }
  ],
  "children": [],
  "spouse": null
}
```

APIs in data science

- Used for serving models
- Used for data gathering
- Used for data “cleaning” (ex. OCR)

Hands-on Intro

- Python requests library and demo of a few open APIs.
  - https://2.python-requests.org/en/master/

Case study: Cartoon March Madness

- Local radio station put on a “march madness” for cartoons, where listeners voted on their favorite cartoons.
- Suppose you wanted to try and predict this bracket?
  - What data would you use?
  - How would you get that data?
  - How would you model this data?
  - How would you validate this model?

Data

- We sought to model using open data from IMDb
- Primary drivers could be time on tv (number of seasons), number of votes, overall rating and age.

How to get this data?

- We can grab it with an unofficial client library (imdbpy) in conjunction with the manually curated list of cartoons.
  - https://imdbpy.sourceforge.io/

What about the model

- With only one example we chose good ol’ fashioned human intuition for this task (harder to ML in this case, however you could take a bayesian approach).

- We chose a piecewise linear model to score each cartoon.

- We then use this score to “simulate” the bracket (higher score wins a faceoff).

Hands on demo

- Let’s run through the notebook and see what went down!

Scraping

- What is scraping
- Common usage in data science?
- Disclaimer

What is Scraping?

- Refers to the act of capturing raw HTML and programmatically extracting information from the resource.
- Google and other search engines regularly scrape websites to curate their search engine results.
- Many companies provide this service for a fee, intelligently crawling the requested website and gathering the data.

Disclaimer

- Web scraping is legal gray area in many jurisdictions.
- There have been cases of legal action taken against companies for actively scraping web pages.
- Most websites will state in their terms of service and on their robots.txt file whether or not they allow scraping and at what capacity.
- Always check the website in question before engaging in any scraping activity.

Common Usage in Data Science

- Data gathering from “unstructured” sources.
- Real-time aggregation of sources.

Hands-on Intro

- Python requests library and beautifulsoup to scrape a sandbox website: http://books.toscrape.com/

Case Study: NHL Playoffs

- Caught wind of hockey pool starting soon.
- Had two days until the deadline
- Wanted to see what we could do but we needed data...

Data

- Player data was available via an open source client library.
- Team data was available for download but only year by year and through a dialog, how do we automate this?
- The answer is some web scraping!

Model

- We defined a simple metric for choosing players: the expected number of points throughout the playoffs!
- If we assume that points per game and games played are independent (not great but not terrible) then...

  E[points] = E[(points per game)(games played)] = E[PPG]E[GP]

Model Continued

- Given this framework we sought to model each the PPG and GP separately:
    - PPG would be the average points over a season
    - GP would be the result of a monte carlo simulation of the bracket
- This monte carlo would take historical pair-wise performance and treat games as bernoulli trials.
- We can then estimate the probability that team i would beat team j.

Hands On Case Study

- Let’s run through the notebook and see what went down!

OCR

- What is OCR?
- What is the common usage?
- Hands on intro
- Case study

What is OCR?

- OCR (Optical Character Recognition) is a class of algorithms tasked with extracting text data from images.
- Used widely in financial services, government organizations (think canada post) and other organizations with a “paper-heavy” workflow.

Common Uses

- Information Retrieval
  - parsing and extracting data from medical records, packing slips, standard forms etc.
- Data Entry Minimization
  - Ex. take a picture of your ID instead of typing it in.

Hands-on Intro

- Tesseract OCR and OpenCV for a demo on real-world data.
- We’ll see the challenges of real world data:
  - Image Quality
  - Orientation
  - Skew
  - Structure

Case Study: Document Classification

- A pared down case study on a document classification pipeline using OCR
  - Synthetic clean data set.
  - Training set “pre-extracted”
- Problem:
  - Given images of book excerpts can you predict the author?

Hands On Case Study

- Let’s run through the notebook and see what went down!