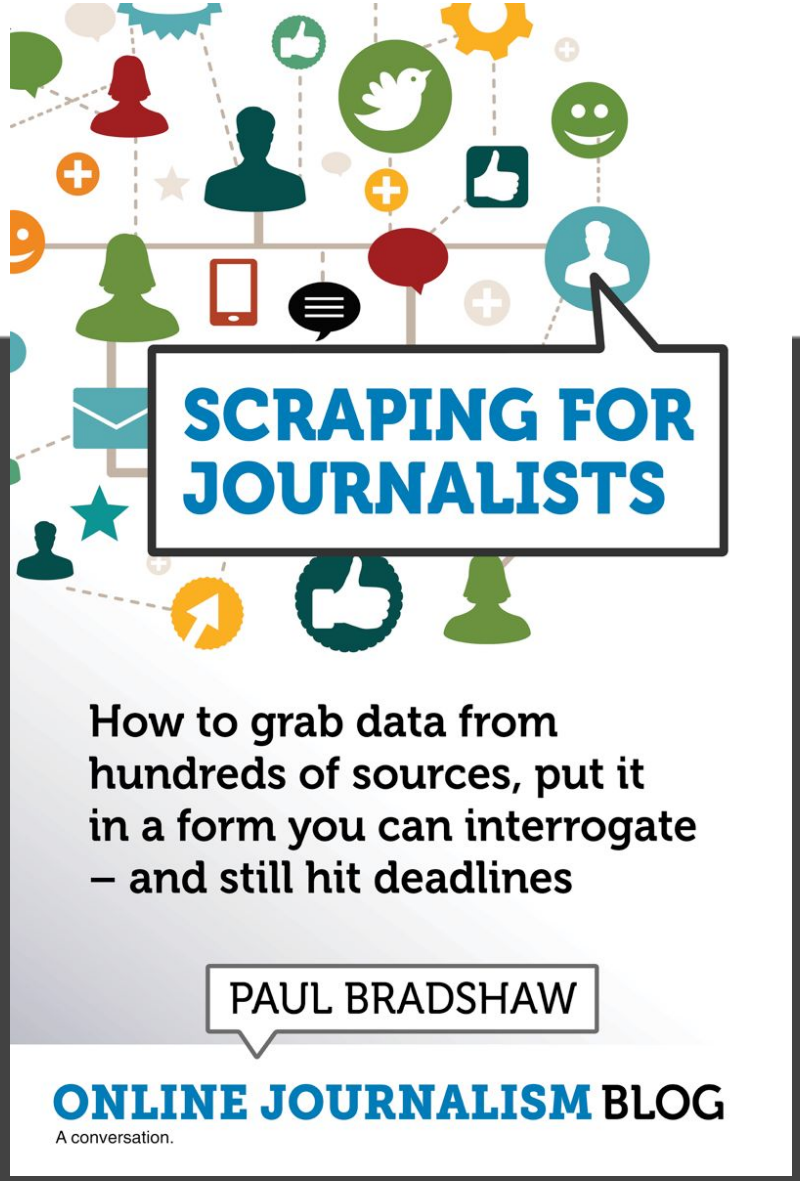
Scraping: intro to lists



Paul Bradshaw

Leanpub.com/scrapingforjournalists

What we'll cover

- How **lists** are used in scraping
- How to **generate** a list of URLs to scrape
- The difference between 1-stage and 2-stage scraping

Lists in scraping

- To scrape webpages or documents you need a list of URLs!
- A list of numbers can be used to **generate** some URLs (e.g. page numbers)
- Or a list of words (e.g. place names, categories)

For example:

[“https://www.bbc.co.uk/news/uk-scotland-56072396”,
“https://www.bbc.co.uk/news/health-56083905”,
“https://www.bbc.co.uk/news/uk-56082027”,
“https://www.bbc.co.uk/news/technology-56084575”]

Generating from words:

[“avon”, “dorset”, “essex”]

...can be used to generate:

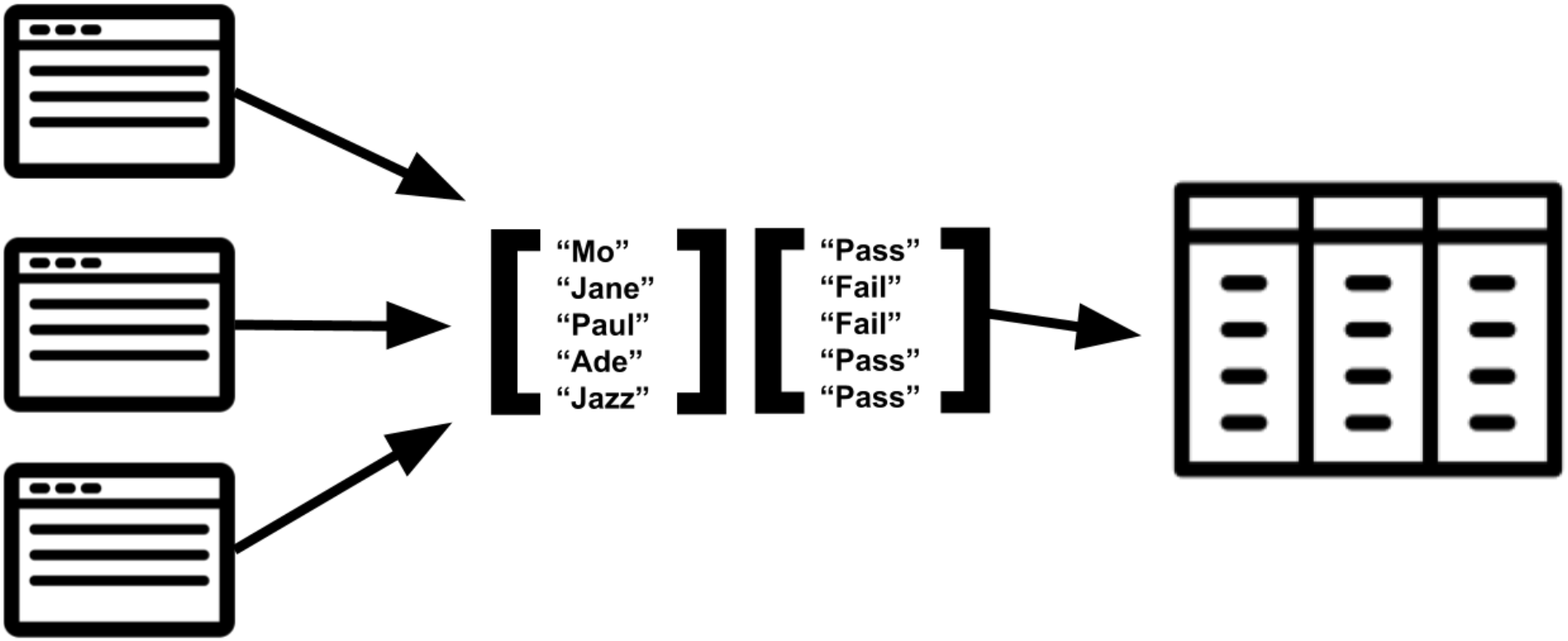
[“http://www.uk-go-karting.com/tracks/avon/”,
“http://www.uk-go-karting.com/tracks/dorset/”,
“http://www.uk-go-karting.com/tracks/essex/”]

Or from numbers:

[1, 2, 3]

...can be used to generate:

[“https://www.reed.co.uk/jobs/jobs-in-birmingham?pageno=1”, “https://www.reed.co.uk/jobs/jobs-in-birmingham?pageno=2”, “https://www.reed.co.uk/jobs/jobs-in-birmingham?pageno=3”]



Loop through each page -> scrape HTML -> extract specific info from HTML -> add row to table -> download

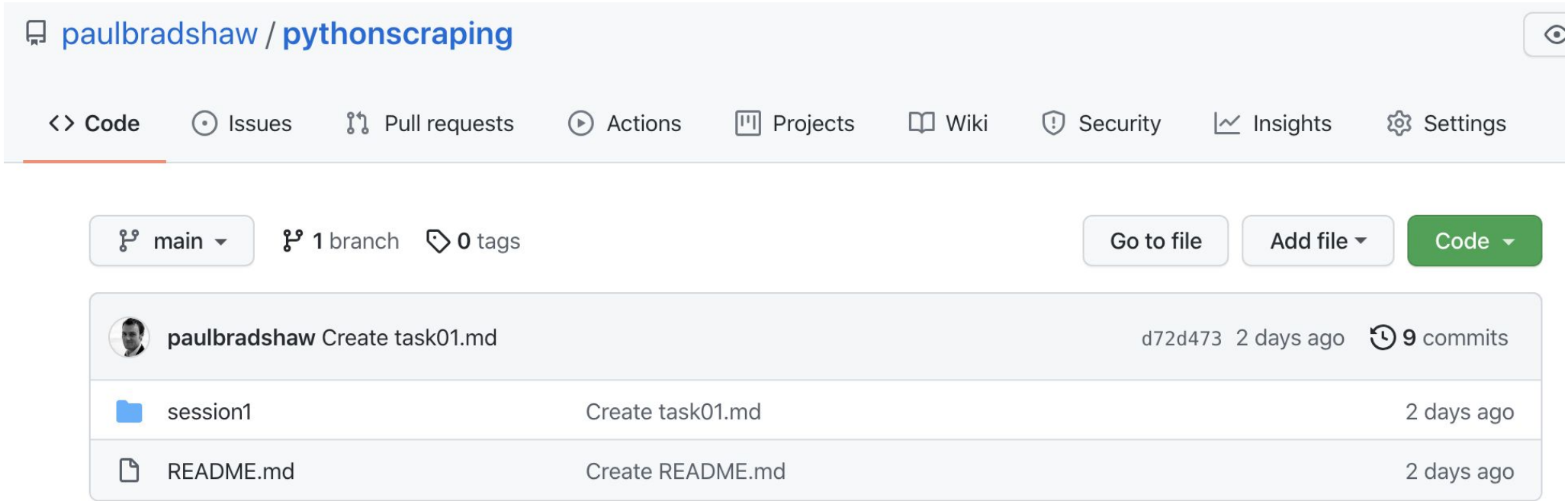
Where does the list come from?

- **1-stage scraping**: you have (or can generate) a list of URLs to scrape
- **2-stage scraping**: you need to scrape the list of URLs (stage 1) then scrape the information at each URL (stage 2)

Time to code.

Let’s look at Google Colab...

1. Go to: **colab.research.google.com**
2. A window will open showing any notebooks
3. Click the **NEW NOTEBOOK** option in the bottom right corner.



Python and scraping

This repo contains notebooks and other resources that outline how to get started with scraping in Python using Google Drive's Colab notebooks.

- Notebook 1: First steps in Colab and Python
- Session 1 tasks

https://github.com/paulbradshaw/pythonscraping/blob/main/session1/pythonFirstStepsColab.ipynb

Key points

- **Lists** are used in scraping to store URLs to scrape
- We **loop** through a list in Python to use each item
- **Google Colab** is one place to write Python in Google Drive