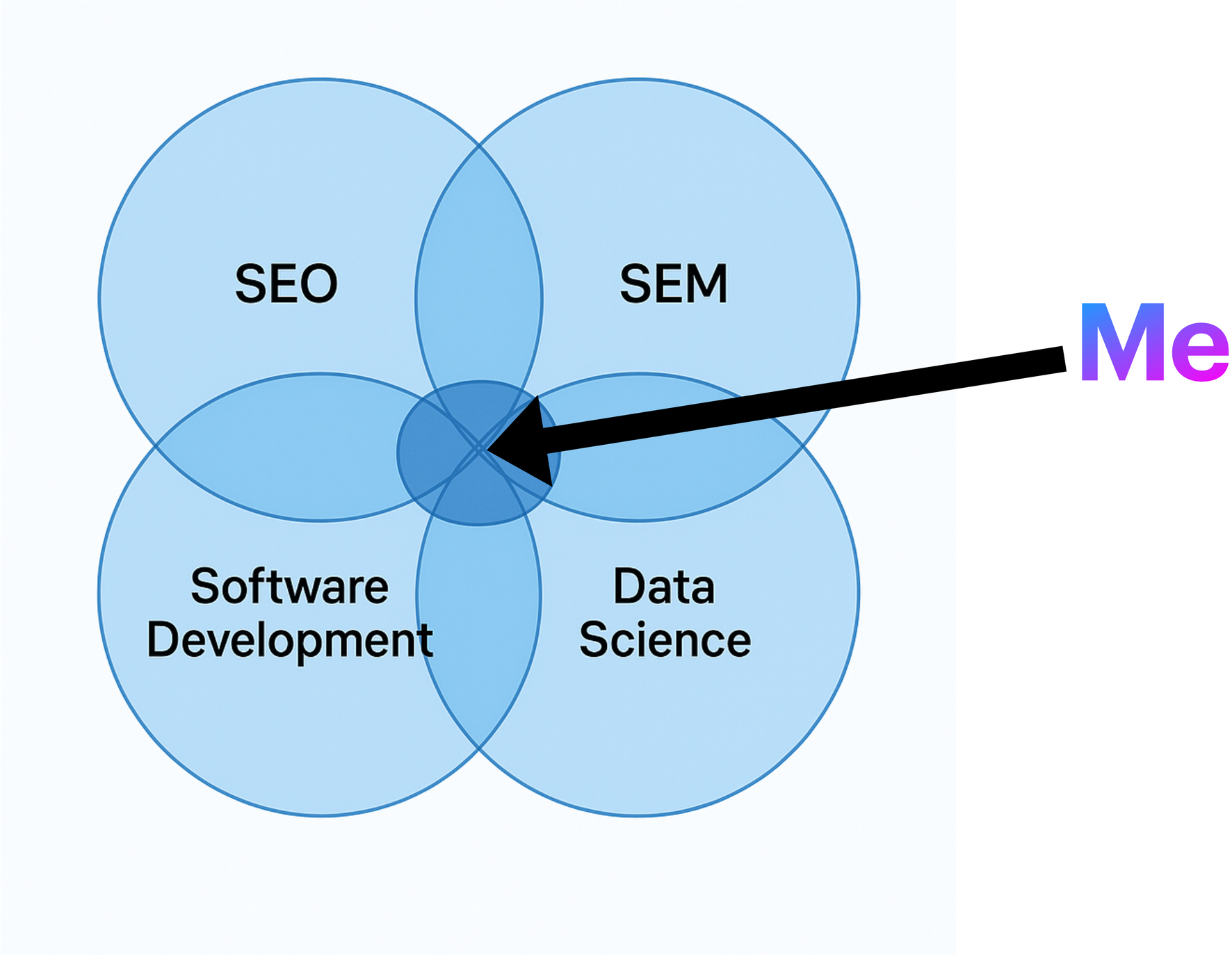
The Rise of the SEO Data Scientist

Elias Dabbas

SEOWeek

New York - April 28, 2025


SEO
SEM
Software Development
Data Science
Me

advertools

0.16.6 | 3.70M

Digital Marketing productivity and analysis tools.

Subscribe

Information | Downloads | Badge

Category

By Version

View Type

Line

Time Range

3 months

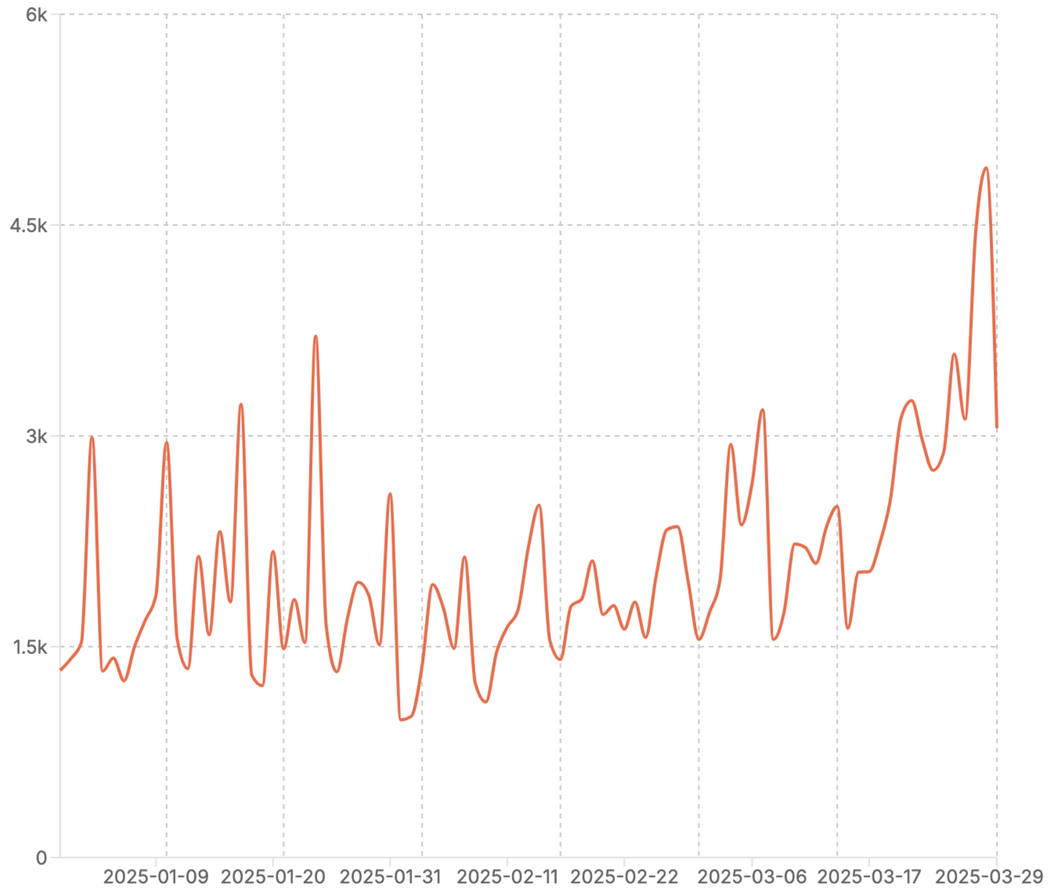
Time Granularity

Daily

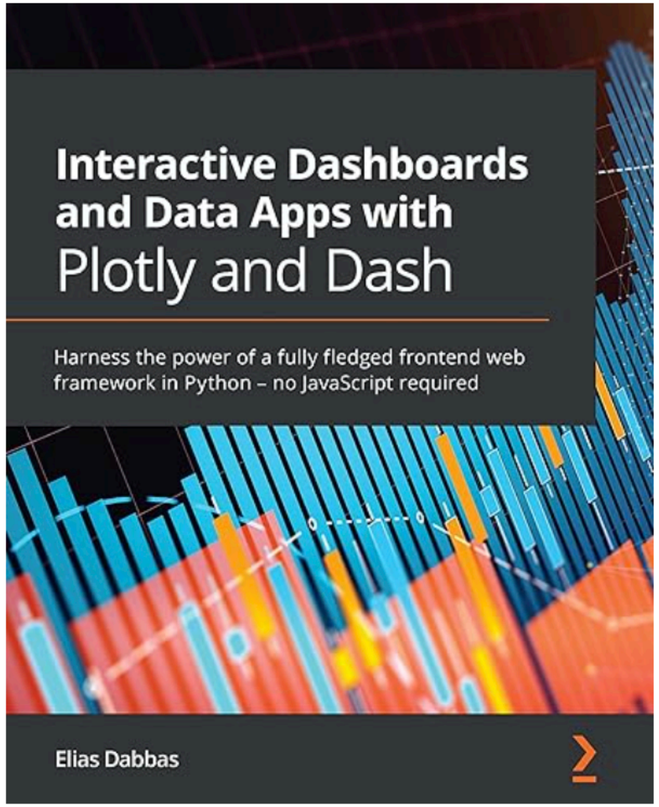
Include CI Downloads

Package Version

0.*



*3.7M installs (not people)



Read sample

Interactive Dashboards and Data Apps with Plotly and Dash: Harness the power of a fully fledged frontend web framework in Python – no JavaScript required 1st Edition, Kindle Edition

by Elias Dabbas (Author) | Format: Kindle Edition

4.4 ★★★★½ (64) See all formats and editions

Build web-based, mobile-friendly analytic apps and interactive dashboards with Python

Key Features

- Develop data apps and dashboards without any knowledge of JavaScript
- Map different types of data such as integers, floats, and dates to bar charts, scatter plots, and more
- Create controls and visual elements with multiple inputs and outputs and add functionality to the app as per your requirements

Book Description

Plotly's Dash framework is a life-saver for Python developers who want to develop complete data apps and interactive dashboards without JavaScript, but you'll need to have the right guide to make sure you're getting the most of it. With the help of this book, you'll be able to explore the functionalities of Dash for visualizing data in different ways.

Houston, we have a [keyword] problem!

"When some field is just getting started and you don't really understand it very well [*Data Science*], it's very easy to confuse the essence of what you're doing with the tools that you use [*Python*]."

Harold Abelson - MIT

SEO

- Keyword research
- Crawling
- Indexing
- Structured data
- Content strategy
- Other...
- **HTML**

SEO

- Keyword research
- Crawling
- Indexing
- Structured data
- Content strategy
- Other...
- **HTML**

Data Science

- Data manipulation
- Data visualization
- Machine learning
- Deep learning
- Math/stats
- Other...
- **Python**
- Also: “coding”, “automation”

You don't need
Python to do SEO!

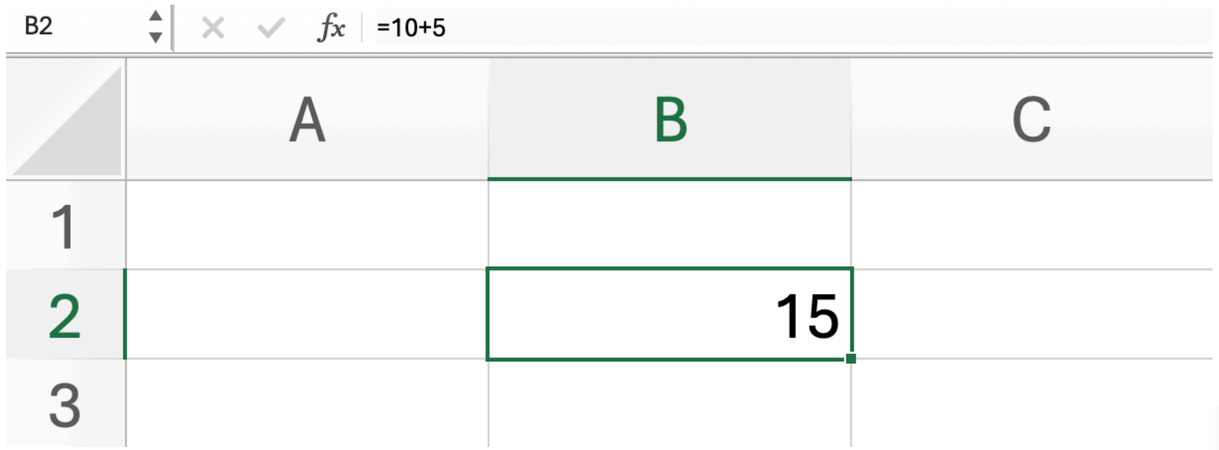
I don't.
But I'm enjoying it &
you're not.

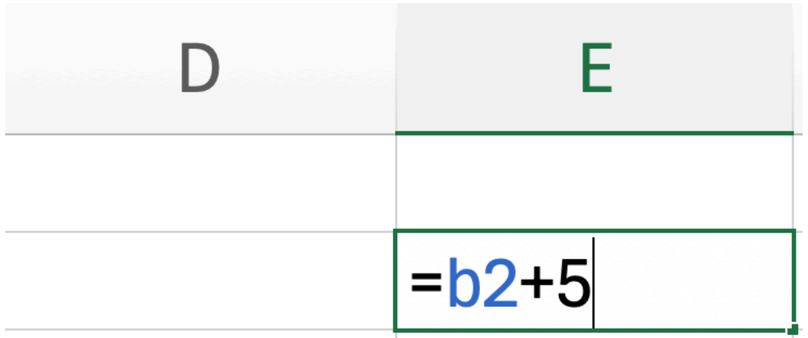
Data skills (data manipulation)

in the trenches

Variables

`name = value`





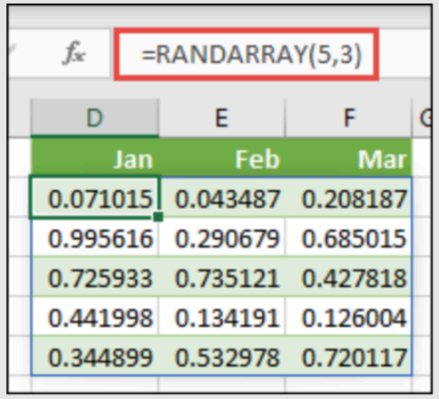
- `B2 = 10+5`
- `name = value` (potentially a complex formula)
- = is the “assignment operator”

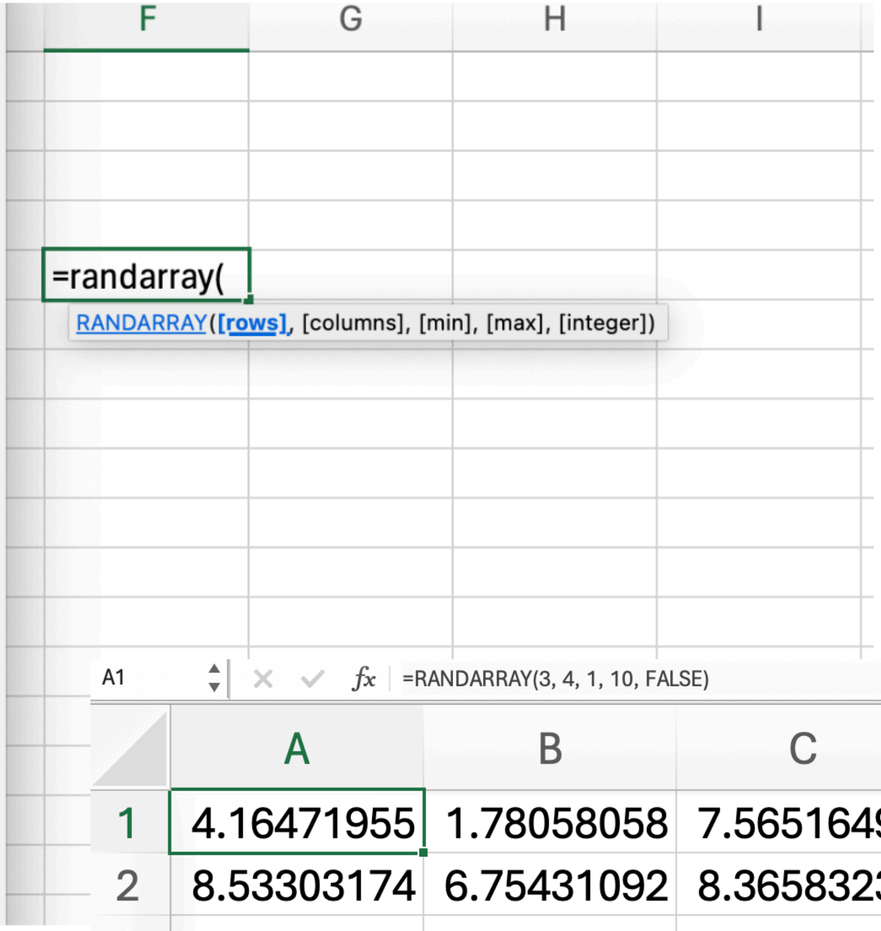
Functions

RANDARRAY function

The **RANDARRAY** function returns an array of random numbers. You can specify the number of rows and columns to fill, minimum and maximum values, and whether to return whole numbers or decimal values.

In the following examples, we created an array that's 5 rows tall by 3 columns wide. The first returns a random set of values between 0 and 1, which is RANDARRAY's default behavior. The next returns a series of random decimal values between 1 and 100. Finally, the third example returns a series of random whole numbers between 1 and 100.



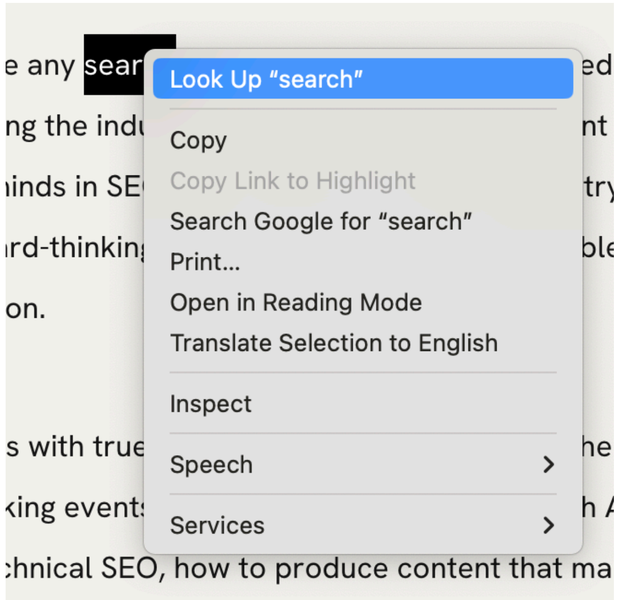


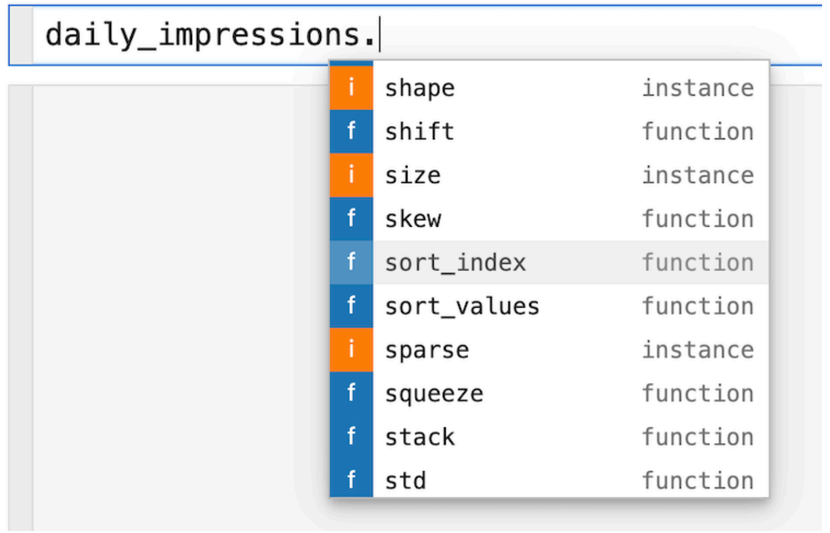

```
=FUNCTIONNAME(parameter_1, parameter_2, parameter_3, …)
```

Right-click (contextual functions)

Dot-notation

- Available actions depend on the type of object





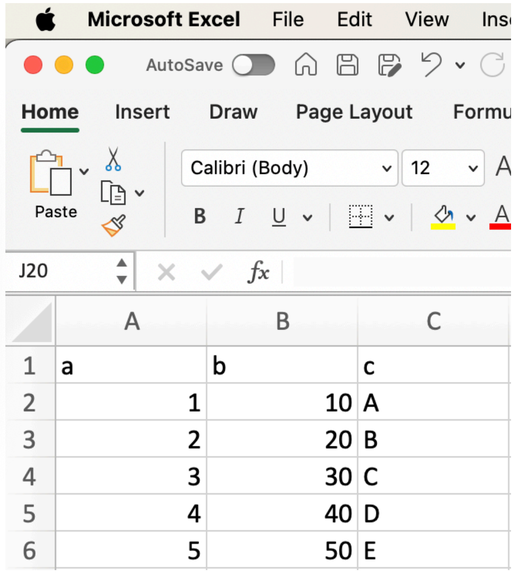
Jupyter Notebook

File format

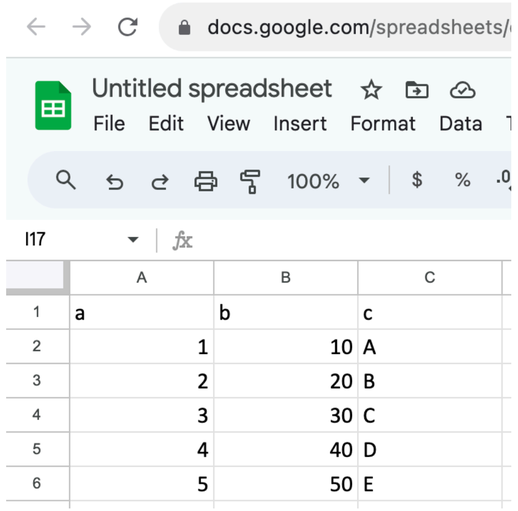
exampledf.csv

```
a,b,c
1,10,A
2,20,B
3,30,C
4,40,D
5,50,E
```

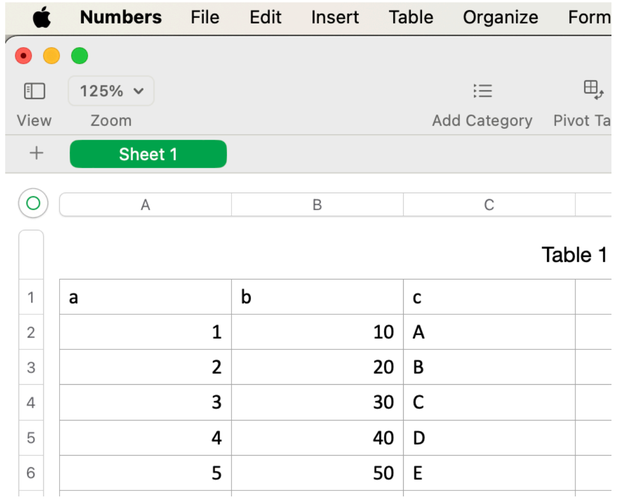
Supporting applications



	A	B	C
1	a	b	c
2	1	10	A
3	2	20	B
4	3	30	C
5	4	40	D
6	5	50	E

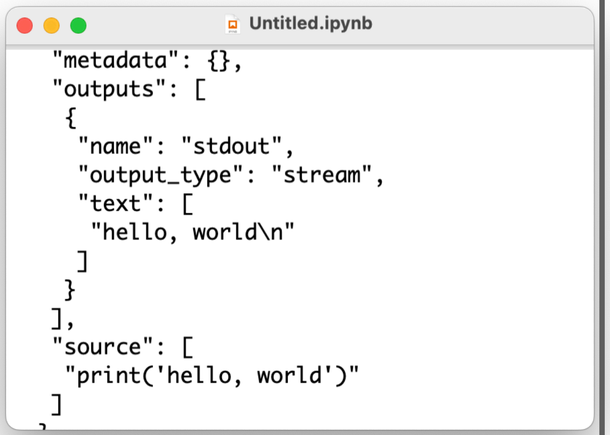


	A	B	C
1	a	b	c
2	1	10	A
3	2	20	B
4	3	30	C
5	4	40	D
6	5	50	E



	A	B	C
1	a	b	c
2	1	10	A
3	2	20	B
4	3	30	C
5	4	40	D
6	5	50	E

File format

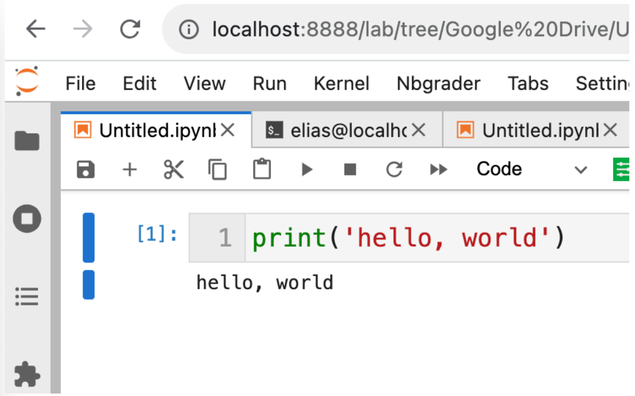


`.ipynb`

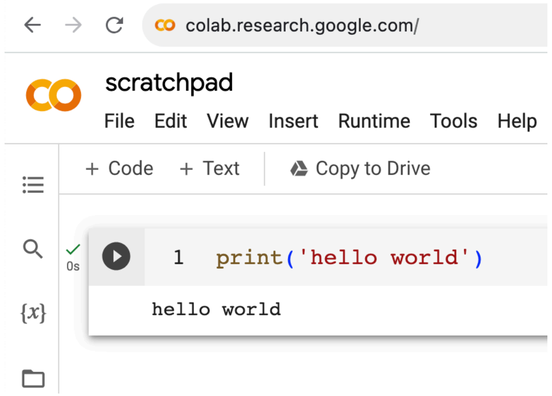
Interactive Python notebook

Jupyter Notebook

Supporting applications



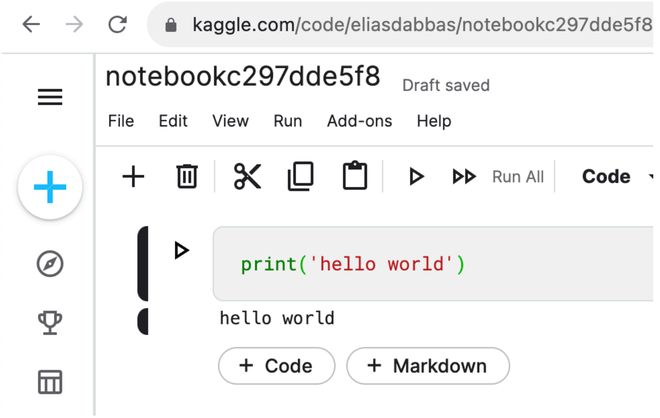
JupyterLab



Google Colab

https://colab.research.google.com/

Start here (online, no setup required)



Kaggle

Introducing `advertools`

Installation

In a code cell:

```
1 !python3 -m pip install advertools
```

Importing

Like starting an application

```
import advertools as adv

adv.
```

f	combinations	function
s	COMMA	statement
c	Counter	class
f	crawl	function
f	crawl_headers	function
f	crawl_images	function
f	crawllogs_to_df	function
m	crawlytics	module
i	CURRENCY	instance
<	CURRENCY_RAW	<unknown>

/robots.txt

Convert a robots.txt file to a DataFrame (table)

```
nyt_robots = adv.robotstxt_to_df("https://www.nytimes.com/robots.txt")
```

```
nyt_robots
```

	directive	content	robotstxt_last_modified	robotstxt_url	download_date
0	comment	New York Times content is made available for y...	2025-03-27 15:42:51	https://www.nytimes.com/robots.txt	2025-04-04 08:59:26.964814+00:00
1	comment	use subject to our Terms of Service here:	2025-03-27 15:42:51	https://www.nytimes.com/robots.txt	2025-04-04 08:59:26.964814+00:00
2	comment	https://help.nytimes.com/hc/en-us/articles/115...	2025-03-27 15:42:51	https://www.nytimes.com/robots.txt	2025-04-04 08:59:26.964814+00:00
3	comment	Use of any device, tool, or process designed t...	2025-03-27 15:42:51	https://www.nytimes.com/robots.txt	2025-04-04 08:59:26.964814+00:00
4	comment	using automated means is prohibited without pr...	2025-03-27 15:42:51	https://www.nytimes.com/robots.txt	2025-04-04 08:59:26.964814+00:00
...	...	...	...	...	...
217	Sitemap	https://www.nytimes.com/athletic/sitemap-verti...	2025-03-27 15:42:51	https://www.nytimes.com/robots.txt	2025-04-04 08:59:26.964814+00:00
218	Sitemap	https://www.nytimes.com/athletic/sitemap-teams...	2025-03-27 15:42:51	https://www.nytimes.com/robots.txt	2025-04-04 08:59:26.964814+00:00
219	Sitemap	https://www.nytimes.com/athletic/sitemap-citie...	2025-03-27 15:42:51	https://www.nytimes.com/robots.txt	2025-04-04 08:59:26.964814+00:00
220	Sitemap	https://www.nytimes.com/athletic/sitemap.xml	2025-03-27 15:42:51	https://www.nytimes.com/robots.txt	2025-04-04 08:59:26.964814+00:00
221	Sitemap	https://www.nytimes.com/games-assets/v2/assets...	2025-03-27 15:42:51	https://www.nytimes.com/robots.txt	2025-04-04 08:59:26.964814+00:00

222 rows × 5 columns

/robots.txt

Simple filter(s) to get the rows you want

```
nyt_robots[nyt_robots['content'].str.contains("Google")]
```

	directive	content	robotstxt_last_modified	robotstxt_url	download_date
14	User-agent	Googlebot	2025-03-27 15:42:51	https://www.nytimes.com/robots.txt	2025-04-04 08:59:26.964814+00:00
114	comment	Googlebot Specific Rules	2025-03-27 15:42:51	https://www.nytimes.com/robots.txt	2025-04-04 08:59:26.964814+00:00
115	User-agent	Googlebot	2025-03-27 15:42:51	https://www.nytimes.com/robots.txt	2025-04-04 08:59:26.964814+00:00
155	User-agent	Google-CloudVertexBot	2025-03-27 15:42:51	https://www.nytimes.com/robots.txt	2025-04-04 08:59:26.964814+00:00
158	User-agent	Google-Extended	2025-03-27 15:42:51	https://www.nytimes.com/robots.txt	2025-04-04 08:59:26.964814+00:00

/robots.txt

Extract XML sitemap URLs

```
1 nyt_sitemaps = nyt_robots[nyt_robots['directive'].eq("Sitemap")]['content'].tolist()
2 nyt_sitemaps
```

```
['https://www.nytimes.com/sitemaps/new/news.xml.gz',
 'https://www.nytimes.com/sitemaps/new/sitemap.xml.gz',
 'https://www.nytimes.com/sitemaps/new/collections.xml.gz',
 'https://www.nytimes.com/sitemaps/new/video.xml.gz',
 'https://www.nytimes.com/sitemaps/new/cooking.xml.gz',
 'https://www.nytimes.com/sitemaps/new/recipe-collects.xml.gz',
 'https://www.nytimes.com/sitemaps/new/regions.xml',
 'https://www.nytimes.com/sitemaps/new/best-sellers.xml',
 'https://www.nytimes.com/sitemaps/new/weather.xml.gz',
 'https://www.nytimes.com/sitemaps/new/espanol.xml.gz',
 'https://www.nytimes.com/sitemaps/new/espanol-collects.xml.gz',
 'https://www.nytimes.com/wirecutter/sitemapindex.xml',
 'https://www.nytimes.com/athletic/sitemap-live-blogs.xml',
 'https://www.nytimes.com/athletic/sitemap-authors.xml',
 'https://www.nytimes.com/athletic/sitemap-verticals.xml',
 'https://www.nytimes.com/athletic/sitemap-teams.xml',
 'https://www.nytimes.com/athletic/sitemap-cities.xml',
 'https://www.nytimes.com/athletic/sitemap.xml',
 'https://www.nytimes.com/games-assets/v2/assets/sitemap/games.xml']
```

/robots.txt

Extract User-agents

```
1 nyt_useragents = nyt_robots[nyt_robots['directive'].eq("User-agent")]['content'].tolist()
2 nyt_useragents
```

```
['*',
 'Googlebot',
 'Googlebot',
 'Amazonbot',
 'anthropic-ai',
 'Applebot-Extended',
 'AwarioRssBot',
 'AwarioSmartBot',
 'Bytespider',
 'CCBot',
 'ChatGPT-User',
 'ClaudeBot',
 'Claude-Web',
 'cohere-ai',
 'DataForSeoBot',
 'Diffbot',
 'DuckAssistBot',
 'FacebookBot',
 'FriendlyCrawler',
 'Google-CloudVertexBot',
 'Google-Extended',
 'GPTBot',
 'ImagesiftBot',
 'magpie-crawler',
 'Meta-ExternalAgent',
 'meta-externalagent',
 'Meta-ExternalFetcher',
 'meta-externalfetcher',
 'NewsNow',
 'news-please',
 'OAI-SearchBot',
 'omgili',
 'omgilibot',
 'peer39_crawler',
 'peer39_crawler/1.0',
 'PerplexityBot',
 'Quora-Bot',
 'Scrapy',
 'Timpibot',
 'TurnitinBot',
 'YouBot',
 'facebookexternalhit',
 'Twitterbot']
```

/robots.txt (bonus)

Supports multiple robots URLs

```
1 import advertools as adv
```

```
1 sneaker_robots = adv.robotstxt_to_df(
2     robotstxt_url=[
3         "https://www.nike.com/robots.txt",
4         "https://us.puma.com/robots.txt",
5         "https://www.reebok.com/robots.txt",
6         "https://www.champion.com/robots.txt"
7     ])
```

```
1 sneaker_robots[sneaker_robots['directive'].str.contains("Sitemap")]
```

	robotstxt_url	directive	content
41	https://www.champion.com/robots.txt	Sitemap	https://www.champion.com/sitemap.xml
97	https://www.champion.com/robots.txt	Sitemap	https://www.champion.com/sitemap.xml
139	https://www.champion.com/robots.txt	Sitemap	https://www.champion.com/sitemap.xml
185	https://www.reebok.com/robots.txt	Sitemap	https://www.reebok.com/sitemap.xml
241	https://www.reebok.com/robots.txt	Sitemap	https://www.reebok.com/sitemap.xml
283	https://www.reebok.com/robots.txt	Sitemap	https://www.reebok.com/sitemap.xml
306	https://us.puma.com/robots.txt	Sitemap	https://us.puma.com/assets/sitemaps/us/sitemap...
381	https://www.nike.com/robots.txt	Sitemap	https://www.nike.com/sitemap-us-help.xml
382	https://www.nike.com/robots.txt	Sitemap	https://www.nike.com/sitemap-v2-landingpage-in...
383	https://www.nike.com/robots.txt	Sitemap	https://www.nike.com/sitemap-v2-pdp-index.xml
384	https://www.nike.com/robots.txt	Sitemap	https://www.nike.com/sitemap-v2-snkrsweb-index...
385	https://www.nike.com/robots.txt	Sitemap	https://www.nike.com/sitemap-v2-gridwall-index...
386	https://www.nike.com/robots.txt	Sitemap	https://www.nike.com/sitemap-v2-article-index.xml
387	https://www.nike.com/robots.txt	Sitemap	https://www.nike.com/sitemap-locator-index.xml

XML Sitemaps

Convert a sitemap to a DataFrame

```
nyt_news = adv.sitemap_to_df("https://www.nytimes.com/sitemaps/new/news.xml.gz")
```

```
nyt_news
```

	loc	lastmod	publication_name	news_publication_date	news_title
0	https://www.nytimes.com/athletic/live-blogs/ma...	2025-04-12 09:10:05+00:00	The New York Times	2025-04-12T08:00:06Z	Masters 2025 live updates: Round 3 latest incl...
1	https://www.nytimes.com/athletic/6272876/2025/...	2025-04-12 09:03:33+00:00	The New York Times	2025-04-12T09:03:33Z	NASCAR Cup Series at Bristol odds, predictions...
2	https://www.nytimes.com/2025/04/12/arts/televi...	2025-04-12 09:03:26+00:00	The New York Times	2025-04-12T09:03:26Z	‘The Last of Us’: What to Remember Ahead of Se...
3	https://www.nytimes.com/2025/04/12/upshot/musk...	2025-04-12 09:03:23+00:00	The New York Times	2025-04-12T09:03:23Z	Musk’s Latest Fraud Finding Isn’t What It Seems
4	https://www.nytimes.com/2025/04/12/arts/music/...	2025-04-12 09:03:18+00:00	The New York Times	2025-04-12T09:03:18Z	‘Khovanshchina’ Is Finished in Time to Be Newl...
...	...	...	...	...	...
666	https://www.nytimes.com/es/2025/04/10/espanol/...	2025-04-10 18:52:33+00:00	New York Times en Español	2025-04-10T18:52:33Z	EE. UU. y China van camino a una ruptura ‘monu...
667	https://www.nytimes.com/es/2025/04/10/espanol/...	2025-04-10 18:06:10+00:00	New York Times en Español	2025-04-10T18:06:10Z	Una señal de que tu pez podría estar drogado: ...
668	https://www.nytimes.com/es/2025/04/10/espanol/...	2025-04-10 17:41:20+00:00	New York Times en Español	2025-04-10T17:41:20Z	5 ‘trucos’ con evidencia para vivir más y sin ...
669	https://www.nytimes.com/es/2025/04/10/espanol/...	2025-04-10 17:12:13+00:00	New York Times en Español	2025-04-10T17:12:13Z	De la calma al nerviosismo: lo que sabemos sob...
670	https://www.nytimes.com/es/2025/04/10/espanol/...	2025-04-10 15:45:32+00:00	New York Times en Español	2025-04-10T15:45:32Z	Trump firma órdenes para castigar a los que se...

671 rows × 5 columns

- Supports single sitemaps
- Sitemapindex files
- Zipped sitemaps
- Runs recursively and combines all sitemaps into one DataFrame
- Supports robots.txt URLs

%whos there?

Current “inventory” of variables

```
%whos
```

```
Variable         Type          Data/Info
-----------------------------------------
adv              module        <module 'advertools' from<...>/advertools/__init__.py'>
nyt_news         DataFrame                                <...>\n[671 rows x 16 columns]
nyt_robots       DataFrame         directive              <...>n\n[222 rows x 5 columns]
nyt_sitemaps     list          n=19
nyt_useragents   list          n=43
```

Robots tester

Bulk tester for all User-agent/URL combinations

```
nyt_robots_test = adv.robotstxt_test(
    robotstxt_url="https://www.nytimes.com/robots.txt",
    user_agents=nyt_useragents,
    urls=nyt_news['loc']
)
```

Robots tester

Bulk tester for all User-agent/URL combinations

```
nyt_robots_test
```

	robotstxt_url	user_agent	url_path	can_fetch
0	https://www.nytimes.com/robots.txt	*	https://www.nytimes.com/2025/04/10/arts/design...	True
1	https://www.nytimes.com/robots.txt	*	https://www.nytimes.com/2025/04/10/arts/design...	True
2	https://www.nytimes.com/robots.txt	*	https://www.nytimes.com/2025/04/10/arts/music/...	True
3	https://www.nytimes.com/robots.txt	*	https://www.nytimes.com/2025/04/10/arts/music/...	True
4	https://www.nytimes.com/robots.txt	*	https://www.nytimes.com/2025/04/10/arts/music/...	True
...	...	...	...	...
28848	https://www.nytimes.com/robots.txt	peer39_crawler/1.0	https://www.nytimes.com/interactive/2025/04/12...	False
28849	https://www.nytimes.com/robots.txt	peer39_crawler/1.0	https://www.nytimes.com/live/2025/04/10/nyregi...	False
28850	https://www.nytimes.com/robots.txt	peer39_crawler/1.0	https://www.nytimes.com/live/2025/04/10/us/tru...	False
28851	https://www.nytimes.com/robots.txt	peer39_crawler/1.0	https://www.nytimes.com/live/2025/04/11/busine...	False
28852	https://www.nytimes.com/robots.txt	peer39_crawler/1.0	https://www.nytimes.com/live/2025/04/11/us/tru...	False

28853 rows × 4 columns

Robots tester

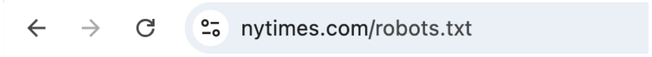
Bulk tester for all User-agent/URL combinations

```
1 nyt_robots_test['can_fetch'].value_counts()
```

```
can_fetch
False    27398
True      3605
Name: count, dtype: int64
```

```
1 nyt_robots_test['can_fetch'].value_counts(normalize=True)
```

```
can_fetch
False    0.883721
True     0.116279
Name: proportion, dtype: float64
```




```
# Disallow Rules

User-agent: Amazonbot
Disallow: /

User-agent: anthropic-ai
Disallow: /

User-agent: Applebot-Extended
Disallow: /

User-agent: AwarioRssBot
User-agent: AwarioSmartBot
Disallow: /

User-agent: Bytespider
Disallow: /

User-agent: CCBot
Disallow: /

User-agent: ChatGPT-User
Disallow: /

User-agent: ClaudeBot
Disallow: /

User-agent: Claude-Web
Disallow: /

User-agent: cohere-ai
Disallow: /

User-agent: DataForSeoBot
Disallow: /

User-agent: Diffbot
Disallow: /
```

URL Analysis

Convert a set of URLs to a DataFrame `adv.url_to_df()`

	loc
0	https://www.nytimes.com/2010/01/31/opinion/31sun2.html
1	https://www.nytimes.com/2010/01/25/sports/olympics/25whockey.html
2	https://www.nytimes.com/2010/01/19/nyregion/19faber.html
3	https://www.nytimes.com/2010/01/31/magazine/31ecopsych-t.html
4	https://www.nytimes.com/2010/01/27/theater/27arts-ATODOATMUCHA_BRF.html
5	https://www.nytimes.com/2010/01/22/opinion/22iht-edcohen.html
6	https://www.nytimes.com/2010/01/01/us/01lilly.html
7	https://www.nytimes.com/2010/01/03/opinion/03ohanlontext.html
8	https://www.nytimes.com/2010/01/25/sports/ncaabasketball/25syracuse.html
9	https://www.nytimes.com/2010/01/05/sports/football/05hearing.html

	scheme	netloc	path	query	fragment	dir_1	dir_2	dir_3	dir_4	dir_5	dir_6	last_dir
0	https	www.nytimes.com	/2010/01/31/opinion/31sun2.html	None	None	2010	01	31	opinion	31sun2.html	None	31sun2.html
1	https	www.nytimes.com	/2010/01/25/sports/olympics/25whockey.html	None	None	2010	01	25	sports	olympics	25whockey.html	25whockey.html
2	https	www.nytimes.com	/2010/01/19/nyregion/19faber.html	None	None	2010	01	19	nyregion	19faber.html	None	19faber.html
3	https	www.nytimes.com	/2010/01/31/magazine/31ecopsych-t.html	None	None	2010	01	31	magazine	31ecopsych-t.html	None	31ecopsych-t.html
4	https	www.nytimes.com	/2010/01/27/theater/27arts-ATODOATMUCHA_BRF.html	None	None	2010	01	27	theater	27arts-ATODOATMUCHA_BRF.html	None	27arts-ATODOATMUCHA_BRF.html
5	https	www.nytimes.com	/2010/01/22/opinion/22iht-edcohen.html	None	None	2010	01	22	opinion	22iht-edcohen.html	None	22iht-edcohen.html
6	https	www.nytimes.com	/2010/01/01/us/01lilly.html	None	None	2010	01	01	us	01lilly.html	None	01lilly.html
7	https	www.nytimes.com	/2010/01/03/opinion/03ohanlontext.html	None	None	2010	01	03	opinion	03ohanlontext.html	None	03ohanlontext.html
8	https	www.nytimes.com	/2010/01/25/sports/ncaabasketball/25syracuse.html	None	None	2010	01	25	sports	ncaabasketball	25syracuse.html	25syracuse.html
9	https	www.nytimes.com	/2010/01/05/sports/football/05hearing.html	None	None	2010	01	05	sports	football	05hearing.html	05hearing.html

URL Analysis

Convert a set of URLs to a DataFrame `adv.url_to_df()`

```
nyt_jan_2010 = adv.sitemap_to_df("https://www.nytimes.com/sitemaps/new/sitemap-2010-01.xml.gz")
nyt_jan_2020 = adv.sitemap_to_df("https://www.nytimes.com/sitemaps/new/sitemap-2020-01.xml.gz")
```

```
nyt_jan_2010_urldf = adv.url_to_df(nyt_jan_2010['loc'])
nyt_jan_2020_urldf = adv.url_to_df(nyt_jan_2020['loc'])
```

URL Analysis

adviz.value_counts()

```
1 import adviz
2 adviz.value_counts(nyt_jan_2010_urldf['dir_1'])
```

NYTimes.com XML Sitemap - Jan, 2010

rank	dir_1	count	cum. count	%	cum. %
1	2010	5,867	5,867	96.6%	96.6%
2	slideshow	156	6,023	2.6%	99.2%
3	interactive	48	6,071	0.8%	100.0%

```
1 adviz.value_counts(nyt_jan_2020_urldf['dir_1'])
```

NYTimes.com XML Sitemap - Jan, 2020

rank	dir_1	count	cum. count	%	cum. %
1	2020	4,259	4,259	91.7%	91.7%
2	interactive	139	4,398	3.0%	94.7%
3	slideshow	99	4,497	2.1%	96.8%
4	es	88	4,585	1.9%	98.7%
5	wirecutter	20	4,605	0.4%	99.2%
6	live	14	4,619	0.3%	99.5%
7	2019	8	4,627	0.2%	99.6%
8	article	5	4,632	0.1%	99.7%
9	athletic	3	4,635	0.1%	99.8%
10	audio	3	4,638	0.1%	99.9%
11	Others:	6	4,644	0.1%	100.0%

URL Analysis

`adviz.value_counts()`

NYTimes.com XML Sitemap - Jan, 2010

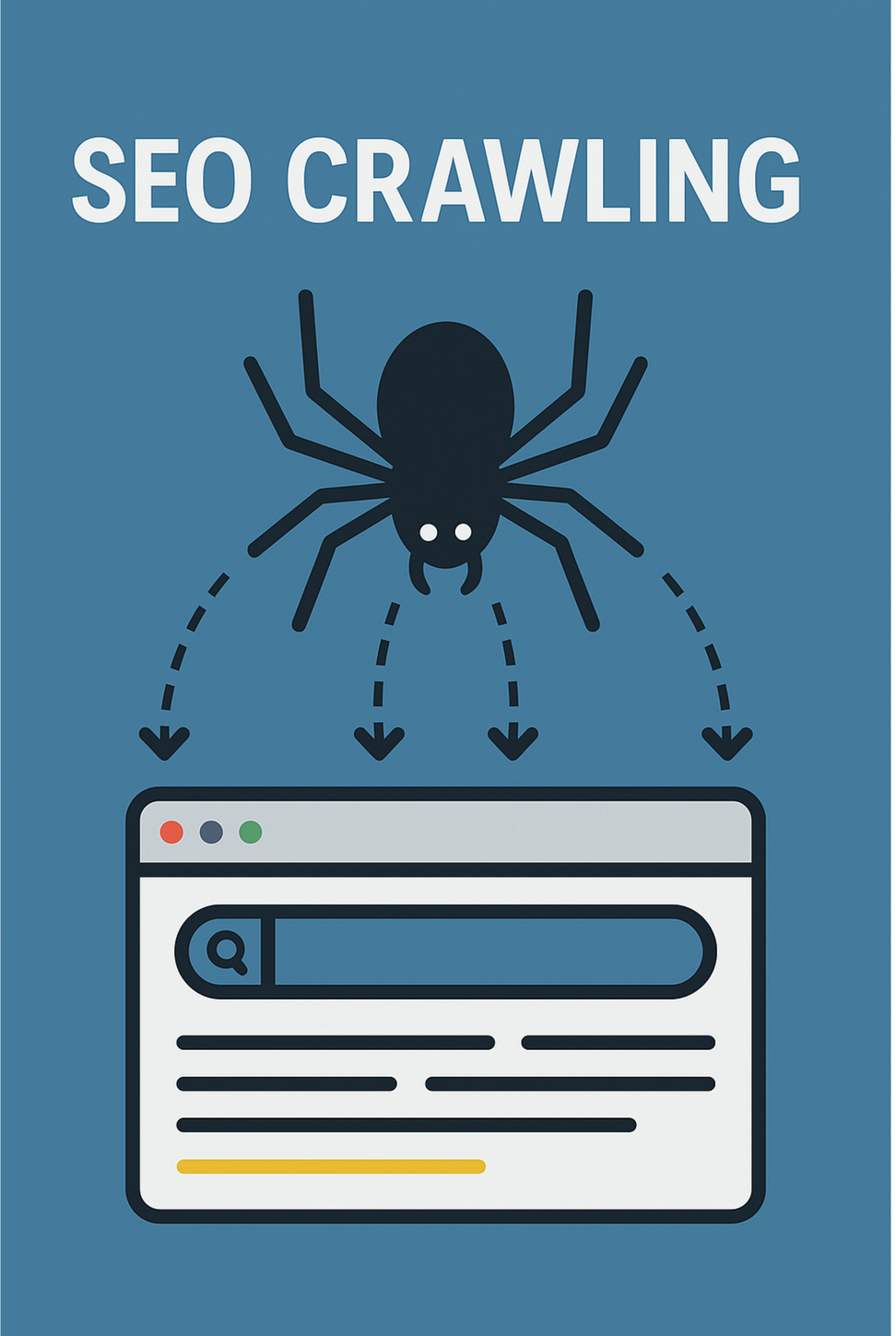
rank	dir_4	count	cum. count	%	cum. %
1	sports	903	903	14.9%	14.9%
2	world	668	1,571	11.0%	25.9%
3	business	637	2,208	10.5%	36.4%
4	nyregion	583	2,791	9.6%	46.0%
5	arts	529	3,320	8.7%	54.7%
6	us	528	3,848	8.7%	63.4%
7	opinion	502	4,350	8.3%	71.7%
8	fashion	172	4,522	2.8%	74.5%
9	books	171	4,693	2.8%	77.3%
10	technology	122	4,815	2.0%	79.3%
11	Others:	1,256	6,071	20.7%	100.0%

NYTimes.com XML Sitemap - Jan, 2020

rank	dir_4	count	cum. count	%	cum. %
1	us	692	692	14.9%	14.9%
2	opinion	471	1,163	10.1%	25.0%
3	world	470	1,633	10.1%	35.2%
4	arts	345	1,978	7.4%	42.6%
5	business	300	2,278	6.5%	49.1%
6	sports	237	2,515	5.1%	54.2%
7	nyregion	190	2,705	4.1%	58.2%
8	books	172	2,877	3.7%	62.0%
9	style	130	3,007	2.8%	64.8%
10	movies	128	3,135	2.8%	67.5%
11	Others:	1,509	4,644	32.5%	100.0%

Reproducibility

- Trust
- Collaboration and building on each other’s work
- Scalability



Crawling

Image: GPT-4o*

*But I wrote the prompt!

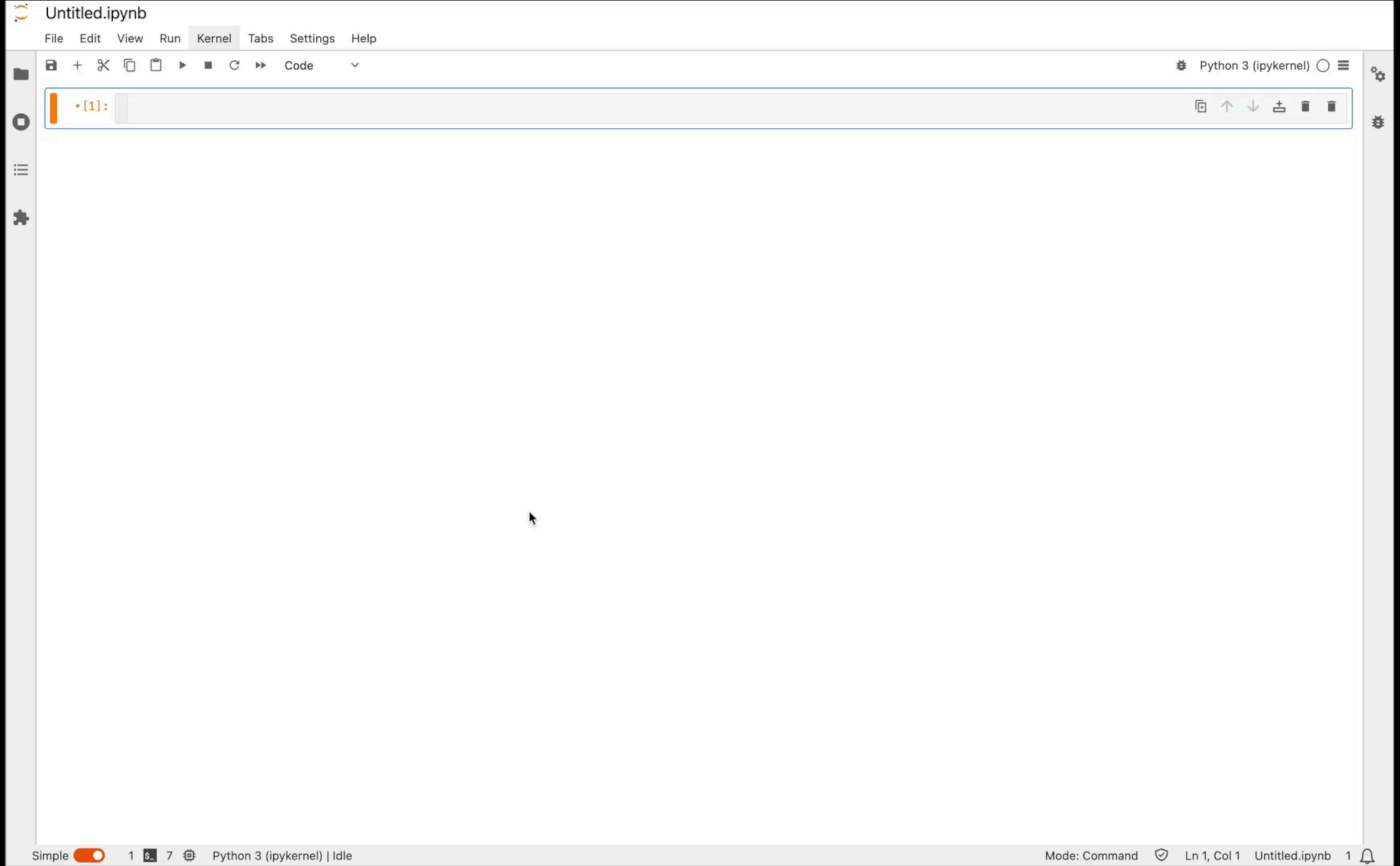
Crawling

- List mode (default):

```
adv.crawl(url_list="https://example.com", output_file="output_file.jsonl")
```

- Spider mode:

```
adv.crawl(url_list="https://example.com", output_file="output_file.jsonl", follow_links=True)
```


Untitled.ipynb
File Edit View Run Kernel Tabs Settings Help
Code
Python 3 (ipykernel)
[1]:
Simple 1 7 Python 3 (ipykernel) | Idle
Mode: Command Ln 1, Col 1 Untitled.ipynb 1

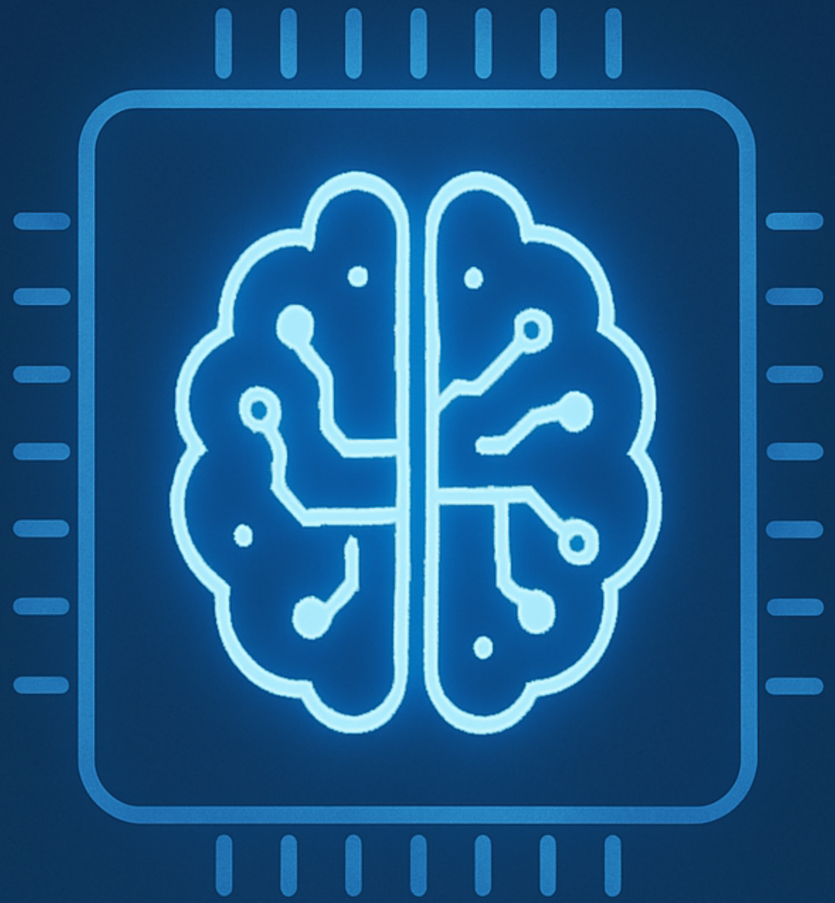
Crawling

- One function contains all the logic
- Single crawl file produced (JSON Lines `.jsonl` OR `.jl`)
- Structured data: JSON-LD, OpenGraph, Twitter
- All request/response headers
- Additional `crawlytics` module with specialized functions to analyze the file:
  - Links (internal and external)
  - Redirects
  - Handling large files
  - Analyze images
  - Compare crawls
  - Convert to parquet

```
adv.crawl(
    url_list=[
        "https://example.com/shopping/bags/",
        "https://example.com/shopping/shoes/"
    ],
    output_file="example_crawl_YYYY_MM_DD.jsonl",
    follow_links=True,
    include_url_regex="/product/",
    exclude_url_regex="comments|reviews",
    exclude_url_params=True,
    xpath_selectors={
        "price": "//div[@class='price']/text()",
        "in_stock": "//div[@class='instock']text()"
    },
    css_selectors={
        "key_1": "value_1",
        "key_2": "value_2",
    },
    custom_settings={
        "CLOSESPIDER_PAGECOUNT": 5000,
        "CONCURRENT_REQUESTS_PER_DOMAIN": 3,
        "LOG_FILE": "example_crawl_YYYY_MM_DD.log",
        "USER_AGENT": "YOUR_CUSTOM_USER_AGENT",
        "DOWNLOAD_DELAY": 2.5,
        "DEPTH_LIMIT": 3,
    }
)
```

- Everything in one place
- Auditable
- Shareable
- Scaleable

Prompt engineering



AI

Craftsmanship

Making a set of four chairs and a table

Engineering

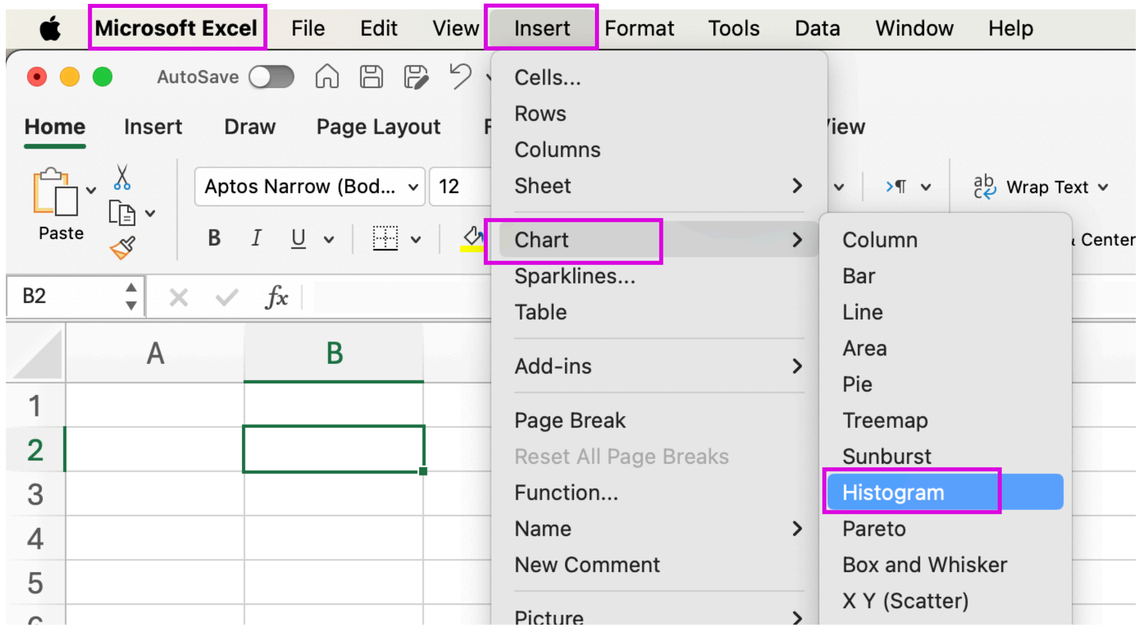
Manufacturing a thousand sets (4 chairs + 1 table)

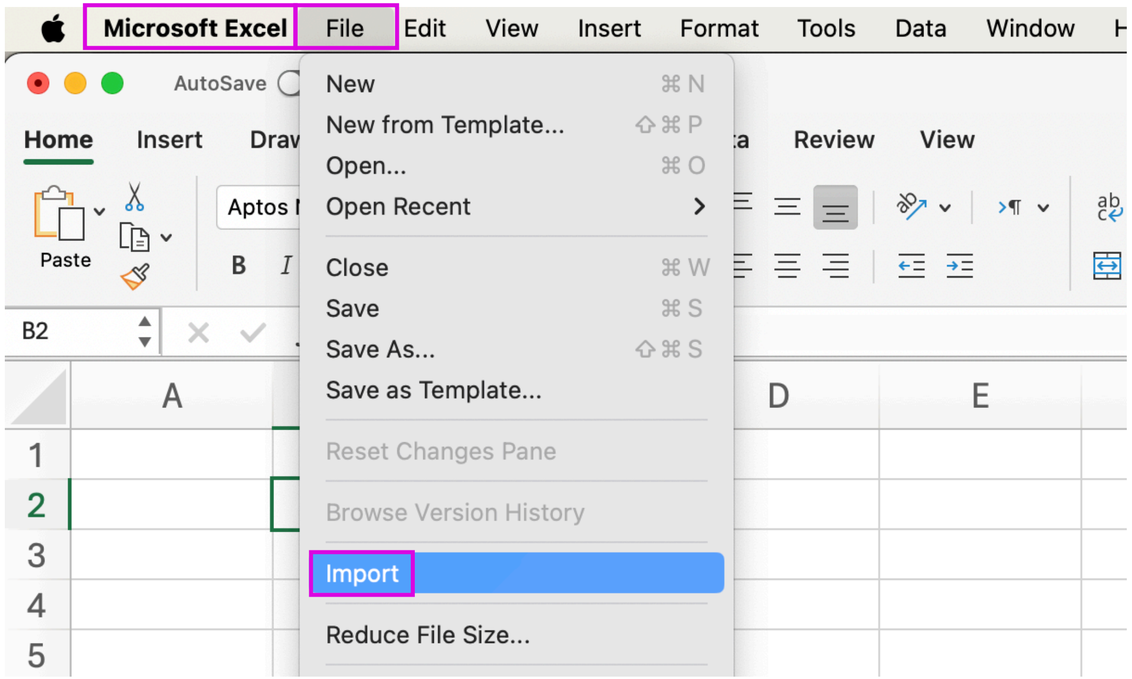
Bulk prompting

- Prompt template
- A mechanism to run that template with dynamic data values

Prompt templates

f“Please create an article about the mobile phone `{product_model}`, and make sure you mention its price: $`{price}`, color: `{product_color}`, and mention our great customer service team”





If Excel was a Python package

```
import excel
excel.insert.chart.histogram()
```

```
excel.file.import("file.csv")
```

```python
import openai

client = openai.OpenAI(api_key="YOUR-OPENAI-API-KEY")
```

```python
client.chat.completions.create(
    model="gpt-4o",
    messages=[
        {"role": "user",
         "content": "Please write an article."}]
)
```

```python
client.audio.transcriptions.create(
    model="whisper-1",
    file=open("audio_file.mp3", "rb")
)
```

```python
client.images.generate(
    model="dall-e-3",
    prompt="A delicious pizza",
    n=3,
    size="1024x1024")
```

Bulk prompt template

Evaluate content using Google's Helpful Content guidelines

```
prompt_intro = """
Please answer the following questions about this article.
Respond in JSON where questions are keys and answers are values.
Send the JSON string only.
Answers should be boolean only."""
```

```
questions = [
    'Does the content provide original information, reporting, research, or analysis?',
    'Does the content provide a substantial, complete, or comprehensive description of the topic?',
    'Does the content provide insightful analysis or interesting information that is beyond the obv
    'If the content draws on other sources, does it avoid simply copying or rewriting those sources
    'Does the main heading or page title provide a descriptive, helpful summary of the content?',
    'Does the main heading or page title avoid exaggerating or being shocking in nature?',
    "Is this the sort of page you'd want to bookmark, share with a friend, or recommend?",
    'Would you expect to see this content in or referenced by a printed magazine, encyclopedia, or
    'Does the content provide substantial value when compared to other pages in search results?',
    'Is the content free of spelling or stylistic issues?',
    'Is the content well-produced, appearing polished and thoughtfully created?',
    'Is the content carefully crafted with individual pages or sites receiving proper attention and
    "Does the content present information in a way that makes you want to trust it, such as clear s
    'If someone researched the site producing the content, would they come away with an impression
    'Is this content written or reviewed by an expert or enthusiast who demonstrably knows the topi
    'Is the content free of easily-verified factual errors?',
    'Do you have an existing or intended audience for your business or site that would find the con
    'Does your content clearly demonstrate first-hand expertise and a depth of knowledge (for examp
    "After reading your content, will someone leave feeling they've learned enough about a topic to
    "Will someone reading your content leave feeling like they've had a satisfying experience?"
]
```

Bulk prompt template

Evaluate content using Google's Helpful Content guidelines

```
title = "This is the title of the article"

body = """
Lorem ipsum dolor sit amet, consetetur sadipscing elitr,
sed diam nonumy eirmod tempor invidunt ut labore et dolore
magna aliquyam erat, sed diam voluptua. At vero eos et
accusam et justo duo dolores et ea rebum. Stet clita kasd
gubergren, no sea takimata sanctus est Lorem ipsum dolor
sit amet. Lorem ipsum dolor sit amet, consetetur
sadipscing elitr, sed diam nonumy eirmod tempor invidunt
ut labore et dolore magna aliquyam erat, sed diam voluptua.
At vero eos et accusam et justo duo dolores et ea rebum.
Stet clita kasd gubergren, no sea takimata sanctus est
Lorem ipsum dolor sit amet.
"""
```

Bulk prompt template

Evaluate content using Google's Helpful Content guidelines

```
                f"""

        {prompt_intro}

        Questions: {questions}

        -------------
        Article Title: {title}

        -------------
        Article Text: {body}

    """
```

Bulk prompt template

Evaluate content using Google's Helpful Content guidelines

```
response = client.chat.completions.create(
    model="gpt-4o",
    messages=[{
        "role": "user",
        "content": f"""

        {prompt_intro}

        Questions: {questions}

        ------------
        Article Title: {title}

        ------------
        Article Text: {body}

    """
    }],
    response_format={"type": "json_object"})
```

AI-powered bulk content audit

```
1 seoweek = pd.read_json("seoweek.jsonl", lines=True)
```

```
1 seoweek
```

	url	title	meta_desc	viewport	charset	h2	h3	h6	canonical
0	https://seoweek.org/	SEO Week | SEO's Next Chapter in NYC	Join SEO Week 2025 in New York City for a week...	width=device-width, initial-scale=1, shrink-to...	UTF-8	AGENDA@@PODCAST@@PLAN YOUR TRIP@@SPONSOR@@WHY ...	Michael King@@Rand Fishkin@@Jori Ford@@Lily Ra...	SEO / TECHNICAL / AI@@DIGITAL STRATEGY / SEO@@...	https://seoweek.org/
1	https://seoweek.org/crystal-carter/	Diving into Deepseek Generative Search Optimiz...	Crystal Carter shares her take on the current ...	width=device-width, initial-scale=1, shrink-to...	UTF-8	AGENDA@@PODCAST@@PLAN YOUR TRIP@@SPONSOR@@Abou...	\n\t\t\tRelated Posts\t\t\t@@Leave a Comment C...	NaN	https://seoweek.org/crystal-carter/
2	https://seoweek.org/cindy-krum/	Word to Your MUM Featuring Cindy Krum	Cindy Krum shares her take on the current and ...	width=device-width, initial-scale=1, shrink-to...	UTF-8	AGENDA@@PODCAST@@PLAN YOUR TRIP@@SPONSOR@@Abou...	\n\t\t\tRelated Posts\t\t\t@@Leave a Comment C...	NaN	https://seoweek.org/cindy-krum/
3	https://seoweek.org/jeff-coyle/	Authoritative Intelligence: Evolving IR, NLP, ...	Jeff Coyle shares his take on the current and ...	width=device-width, initial-scale=1, shrink-to...	UTF-8	AGENDA@@PODCAST@@PLAN YOUR TRIP@@SPONSOR@@Abou...	\n\t\t\tRelated Posts\t\t\t@@Leave a Comment C...	NaN	https://seoweek.org/jeff-coyle/
4	https://seoweek.org/bianca-anderson/	F$%@ Traffic: Prioritizing Conversions Over Va...	Bianca Anderson will challenge your old-school...	width=device-width, initial-scale=1, shrink-to...	UTF-8	AGENDA@@PODCAST@@PLAN YOUR TRIP@@SPONSOR@@Abou...	\n\t\t\tRelated Posts\t\t\t@@Leave a Comment C...	NaN	https://seoweek.org/bianca-anderson/

Bulk prompt template

for-loops

```
1 colors = ["Blue", "Green", "Red", "Yellow"]
```

```
1 for c in colors:
2     print(c)
```

```
Blue
Green
Red
Yellow
```

```
1 for c in colors:
2     print(c.lower())
```

```
blue
green
red
yellow
```

```
1 for c in colors:
2     print(f"Please create an article about the color {c}")
```

```
Please create an article about the color Blue
Please create an article about the color Green
Please create an article about the color Red
Please create an article about the color Yellow
```

Prompt template

```
response = client.chat.completions.create(
model="gpt-4o",
messages=[{
    "role": "user",
    "content": f"""
        {prompt_intro}

        Questions: {questions}

        ------------
        Article Title: {title}

        ------------
        Article Text: {body}

"""}],
response_format={"type": "json_object"})
```

Prompt template

```
all_responses = [] # create an empty list
for title, body in seoweek[['title', 'body_text']].values: # loop through the titles and body texts
    response = client.chat.completions.create(
    model="gpt-4o",
    messages=[{
        "role": "user",
        "content": f"""
            {prompt_intro}

            Questions: {questions}

            ------------
            Article Title: {title}

            ------------
            Article Text: {body}

    """}],
    response_format={"type": "json_object"})
    all_responses.append(response) # append each response to the all_responses list
```

Evaluations by page

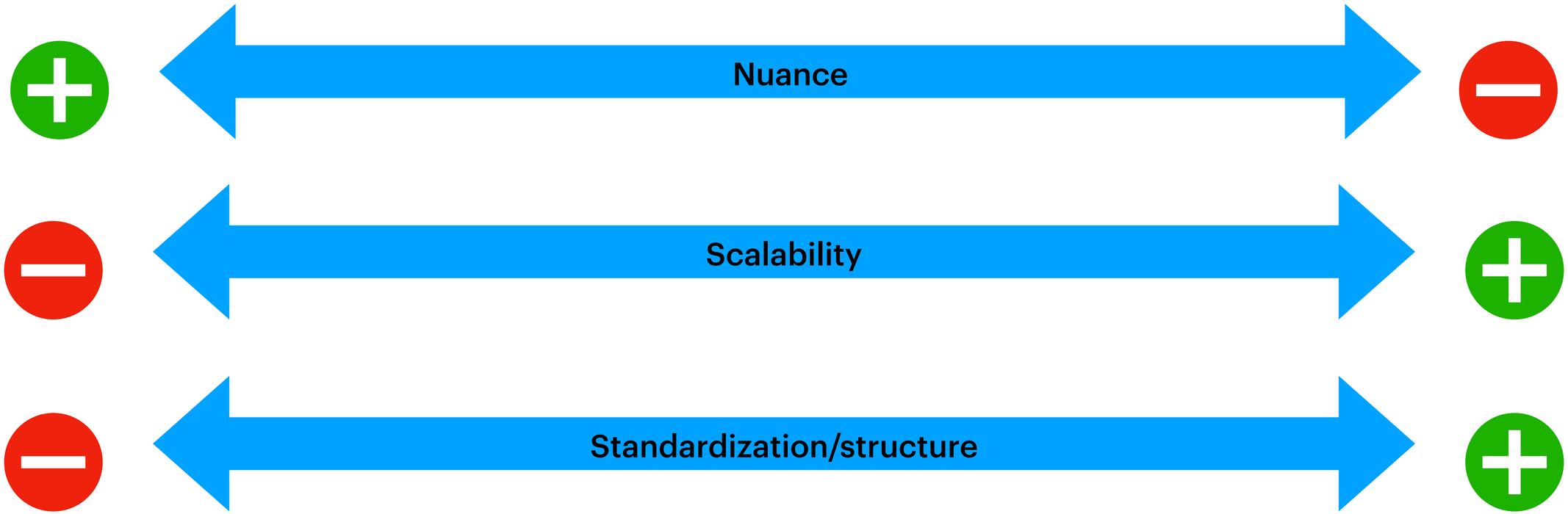
	url	title	question	answer
191	https://seoweek.org/tom-critchlow/	Executive Presence: How To Get Buy In and Budg...	Is the content carefully crafted with individu...	True
210	https://seoweek.org/ross-hudgens/	The Evolving Content Marketing Playbook Featur...	Is the content well-produced, appearing polish...	True
159	https://seoweek.org/cindy-krum/	Word to Your MUM Featuring Cindy Krum	Will someone reading your content leave feelin...	True
3	https://seoweek.org/devin-bramhall/	Stop Doing Marketing Featuring Devin Bramhall	If the content draws on other sources, does it...	True
300	https://seoweek.org/elias-dabbas/	The Rise of the SEO Data Scientist Featuring E...	Does the content provide original information,...	True
252	https://seoweek.org/dan-petrovic/	Beyond Rank Tracking: Analyzing Brand Percepti...	Does the content present information in a way ...	True
155	https://seoweek.org/cindy-krum/	Word to Your MUM Featuring Cindy Krum	Is the content free of easily-verified factual...	True
9	https://seoweek.org/devin-bramhall/	Stop Doing Marketing Featuring Devin Bramhall	Is the content free of spelling or stylistic i...	True
204	https://seoweek.org/ross-hudgens/	The Evolving Content Marketing Playbook Featur...	Does the main heading or page title provide a ...	True
103	https://seoweek.org/jori-ford/	Hybrid Engine Optimization: A Crawler Driven A...	If the content draws on other sources, does it...	True
78	https://seoweek.org/bianca-anderson/	F$%@ Traffic: Prioritizing Conversions Over Va...	After reading your content, will someone leave...	True
189	https://seoweek.org/tom-critchlow/	Executive Presence: How To Get Buy In and Budg...	Is the content free of spelling or stylistic i...	True
294	https://seoweek.org/talia-wolf/	Stop Chasing Conversions: Win More Customers w...	Is this content written or reviewed by an expe...	True
233	https://seoweek.org/ross-simmonds/	Search, AI & UGC: Navigating the Future of Goo...	If someone researched the site producing the c...	True
184	https://seoweek.org/tom-critchlow/	Executive Presence: How To Get Buy In and Budg...	Does the main heading or page title provide a ...	True

Aggregate evaluation by question

	question	answer
0	After reading your content, will someone leave feeling they've learned enough about a topic to help achieve their goal?	100%
1	Do you have an existing or intended audience for your business or site that would find the content useful if they came directly to you?	100%
2	Does the content present information in a way that makes you want to trust it, such as clear sourcing, evidence of the expertise involved, background about the author or the site that publishes it, such as through links to an author page or a site's About page?	100%
3	Does the content provide a substantial, complete, or comprehensive description of the topic?	100%
4	Does the content provide insightful analysis or interesting information that is beyond the obvious?	100%
5	Does the content provide original information, reporting, research, or analysis?	100%
6	Does the content provide substantial value when compared to other pages in search results?	100%
7	Does the main heading or page title avoid exaggerating or being shocking in nature?	94%
8	Does the main heading or page title provide a descriptive, helpful summary of the content?	100%
9	Does your content clearly demonstrate first-hand expertise and a depth of knowledge (for example, expertise that comes from having actually used a product or service, or visiting a place)?	100%
10	If someone researched the site producing the content, would they come away with an impression that it is well-trusted or widely-recognized as an authority on its topic?	100%
11	If the content draws on other sources, does it avoid simply copying or rewriting those sources, and instead provide substantial additional value and originality?	100%
12	Is the content carefully crafted with individual pages or sites receiving proper attention and care, rather than being mass-produced or widely outsourced?	100%
13	Is the content free of easily-verified factual errors?	100%
14	Is the content free of spelling or stylistic issues?	100%
15	Is the content well-produced, appearing polished and thoughtfully created?	100%
16	Is this content written or reviewed by an expert or enthusiast who demonstrably knows the topic well?	100%
17	Is this the sort of page you'd want to bookmark, share with a friend, or recommend?	100%
18	Will someone reading your content leave feeling like they've had a satisfying experience?	100%
19	Would you expect to see this content in or referenced by a printed magazine, encyclopedia, or book?	100%

Single long-form prompting
(Focus group)

API binary questions
(Binary survey)



Thank you!

@EliasDabbas

https://github.com/eliasdabbas/seoweek