# Web Scraping Tutorial

Allison J.B. Chaney

DUKE FUQUA SCHOOL OF BUSINESS

# What will we cover?

At the end of this tutorial you should:

- Understand the difference between using an API and scraping a raw webpage
- Have sense of how to use APIs to obtain data
- Have sense of how to scrape webpages to obtain data

We will **NOT** cover:

- What to do with data once you have it (e.g., how to clean text data in preparation for analysis)

Getting data from the web is like running an experiment to get data: you need a **clear vision**.

# API vs. raw web data

API:

- stands for “application programming interface”
- a set of functions you can call that returns data
- you usually have to have an account / agree to terms

raw web data:

- HTML / CSS / Javascript / JSON
- the public code behind what you see when you load a website

# API

developer key
api terms
getting started
developer forums
contact us

## INTRODUCTION

The Goodreads API allows developers access to Goodreads data in order to help websites or applications that deal with books be more personalized, social, and engaging. The API can be used in many ways, including:

- **Goodreads Connect**: Let members connect to their Goodreads accounts, and you'll have full access to the books in their shelves, their ratings, their reviews, and their friends – the social reading graph. Use this to personalize an ecommerce store, power recommendations, show a widget of a member's favorite books, build a mobile or desktop client app, and more. Learn more about Goodreads Connect.

## GETTING STARTED

- Most API methods will require you to register for a developer key.
- All developers using the API must adhere to the Goodreads API Terms & Conditions.
- Read examples about how to use the API.
- For questions and discussion, visit the Developer Group or contact us.

## API METHODS

auth.user — Get id of user who authorized OAuth.

author.books — Paginate an author's books.

author.show — Get info about an author by id.

author_following.create — Follow an author.

author_following.destroy — Unfollow an author.

author_following.show — Show author following information.

book.isbn_to_id — Get Goodreads book IDs given ISBNs.

book.id_to_work_id — Get Goodreads work IDs given Goodreads book IDs.

book.review_counts — Get review statistics given a list of ISBNs.

book.show — Get the reviews for a book given a Goodreads book id.

book.show_by_isbn — Get the reviews for a book given an ISBN.

book.title — Get the reviews for a book given a title string.

Search all documentation...

Basics

Accounts and users

Tweets

Direct Messages

Media

Trends

Geo

Ads

Metrics

Publisher tools

Twitter for Websites

Labs

Developer utilities

# API reference index

## Basics

### Authentication

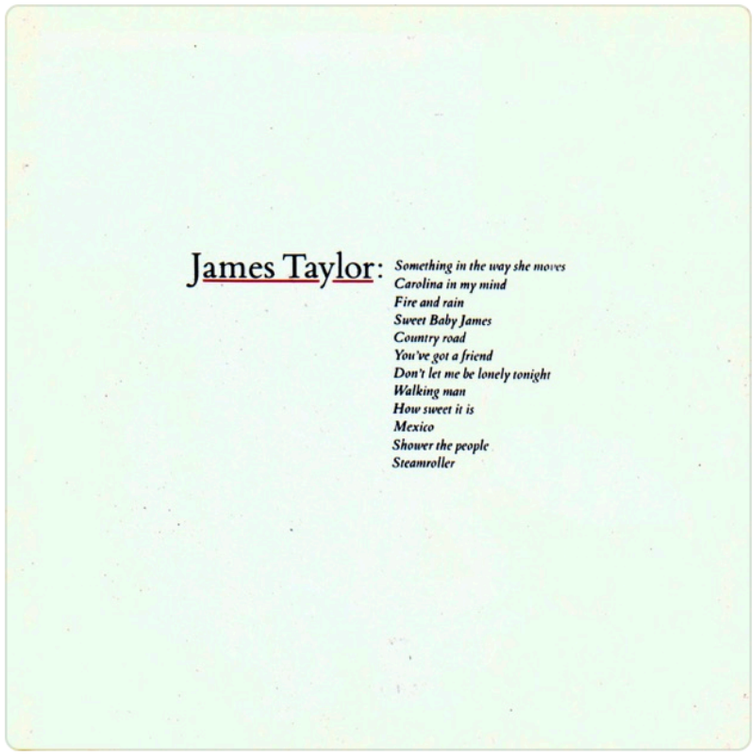
- GET oauth/authenticate
- GET oauth/authorize
- POST oauth/access_token
- POST oauth/invalidate_token
- POST oauth/request_token
- POST oauth2/invalidate_token
- POST oauth2/token

## Accounts and users

### Create and manage lists

- GET lists/list
- GET lists/members
- GET lists/members/show
- GET lists/memberships
- GET lists/ownerships
- GET lists/show
- GET lists/statuses

# iTunes Preview




12 Songs, 41 Minutes

Preview

## Greatest Hits, Vol. 1

James Taylor

Singer/Songwriter · 1976

★★★★½ 4.5, 504 Ratings

$9.99

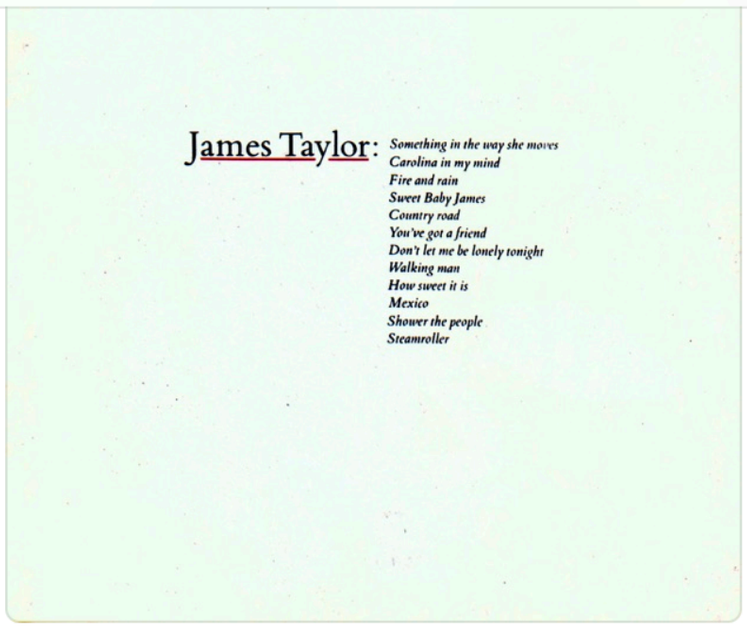
View in iTunes ↗

**EDITORS' NOTES**

*James Taylor: Greatest Hits, Vol. 1* opens gently with a re-recording of "Something In the Way She Moves" from his self-titled debut album (this take without the psychedelic harpsichord intro). "Carolina In My Mind" also gets the soft rock r more

| | | | |
|---|---|---|---|
| ▶ | Something In the Way She Moves | -0:29 | $1.29 |
| 2 | Carolina In My Mind | 3:58 | $1.29 |
| ★ 3 | Fire and Rain | 3:20 | $1.29 |
| 4 | Sweet Baby James | 2:51 | $1.29 |
| 5 | Country Road | 3:25 | $1.29 |
| 6 | You've Got a Friend | 4:28 | $1.29 |
| 7 | Don't Let Me Be Lonely Tonight | 2:35 | $1.29 |
| 8 | Walking Man | 3:32 | $1.29 |
| 9 | How Sweet It Is (To Be Loved By You) | 3:35 | $1.29 |
| 10 | Mexico | 2:58 | $1.29 |

iTunes Preview




12 Songs, 41 Minutes  Preview

James Taylor
Singer/Songwriter · 1976
★★★★½ 4.5, 504 Ratings
$9.99

View in iTunes ↗

| | | | |
|---|---|---|---|
| ▶ | Something In the Way She Moves | -0:29 | $1.29 |
| 2 | Carolina In My Mind | 3:58 | $1.29 |
| ★ 3 | Fire and Rain | 3:20 | $1.29 |

Console Sources Network Performance Memory Application Security Audits AdBlock

```
▶<tr id="ember321" class="table__row is-now-playing we-selectable-item is-available we-selectable-item--allows-interaction ember-view">…
</tr>
▼<tr data-metrics-click="{"actionType":"preview","targetType":"button","targetId":"320083852"}" id="ember328" class="table__row we-selectable-item is-available we-selectable-item--allows-interaction ember-view">
  ▶<td class="table__row__track">…</td>
  ▶<td class="table__row__name we-selectable-item__link-text">…</td>
  ▶<td class="table__row__share">…</td>
  ▼<td class="table__row__duration we-selectable-item__endcopy" data-duration-width="-0:00">
     <time data-test-we-duration datetime="PT3M58S" aria-label="TIME" class="table__row__duration-counter">3:58</time>
     ::after
   </td>
   <td class="table__row__price we-selectable-item__endcopy small-hide large-show-tablecell">$1.29</td> == $0
 </tr>
▶<tr data-metrics-click="{"actionType":"preview","targetType":"button","targetId":"320083858"}" id="ember335" class="table__row popularity-star we-selectable-item is-available we-selectable-item--allows-interaction ember-view">…</tr>
▶<tr data-metrics-click="{"actionType":"preview","targetType":"button","targetId":"320083866"}" id="ember342" class="table__row we-selectable-item is-available we-selectable-item--allows-interaction ember-view">…</tr>
▶<tr data-metrics-click="{"actionType":"preview","targetType":"button","targetId":"320083891"}" id="ember349" class="table__row we-selectable-item is-available we-selectable-item--allows-interaction ember-view">…</tr>
▶<tr data-metrics-click="{"actionType":"preview","targetType":"button","targetId":"320083897"}" id="ember356" class="table__row we-selectable-item is-available we-selectable-item--allows-interaction ember-view">…</tr>
▶<tr data-metrics-click="{"actionType":"preview","targetType":"button","targetId":"320083932"}" id="ember363" class="table__row we-selectable-item is-available we-selectable-item--allows-interaction ember-view">…</tr>
▶<tr data-metrics-click="{"actionType":"preview","targetType":"button","targetId":"320083935"}" id="ember370" class="table__row we-selectable-item is-available we-selectable-item--allows-interaction ember-view">…</tr>
▶<tr data-metrics-click="{"actionType":"preview","targetType":"button","targetId":"320084024"}" id="ember377" class="table__row we-selectable-item is-available we-selectable-item--allows-interaction ember-view">…</tr>
▶<tr data-metrics-click="{"actionType":"preview","targetType":"button","targetId":"320084033"}" id="ember384" class="table__row we-
```

```
295 #ember297 div div div #ember318 table tbody #ember328 td.table__row__price.we-selectable-item__endcopy.small-hide.large-show-tablecell
```

Styles Computed

Filter

```
element.style {
}
@media only screen
.table__row__duratio
.table__row__price
  padding-left: 28p
}
@media only screen
.table__row__duratio
.table__row__price
  padding-left: 22p
}
.table__row__duratio
.table__row__price
  padding-left: 10p
}
@media only screen
.table__row__duratio
nth-last-child(-n+1
.table__row__price:
  text-align: right
}
.table__row td {
  height: 51px;
```

# Let’s get started with APIs



Search all documentation...



# Twitter libraries

These libraries, while not necessarily all built or tested by Twitter, should support the standard Twitter API.

## Libraries built by Twitter

### Java

- twitter-kit-android — Twitter Kit for Android is a multi-module gradle project containing several Twitter SDKs including TweetComposer, TwitterCore, and TweetUi.
- hbc — A Java HTTP client for consuming Twitter’s Streaming API

### Objective-C & Swift

- Twitter Kit for iOS — Twitter Kit for iOS includes API wrappers, Tweet display, Log in with Twitter, and a Tweet composer view. Can be included in Swift apps.

## Libraries built for the Twitter Platform

Inclusion in the list of libraries below is not an endorsement or recommendation of those organizations by Twitter. In addition, such inclusion is not intended to imply, directly or indirectly, that those organizations endorse or have any affiliation with Twitter.

### Multi-platform

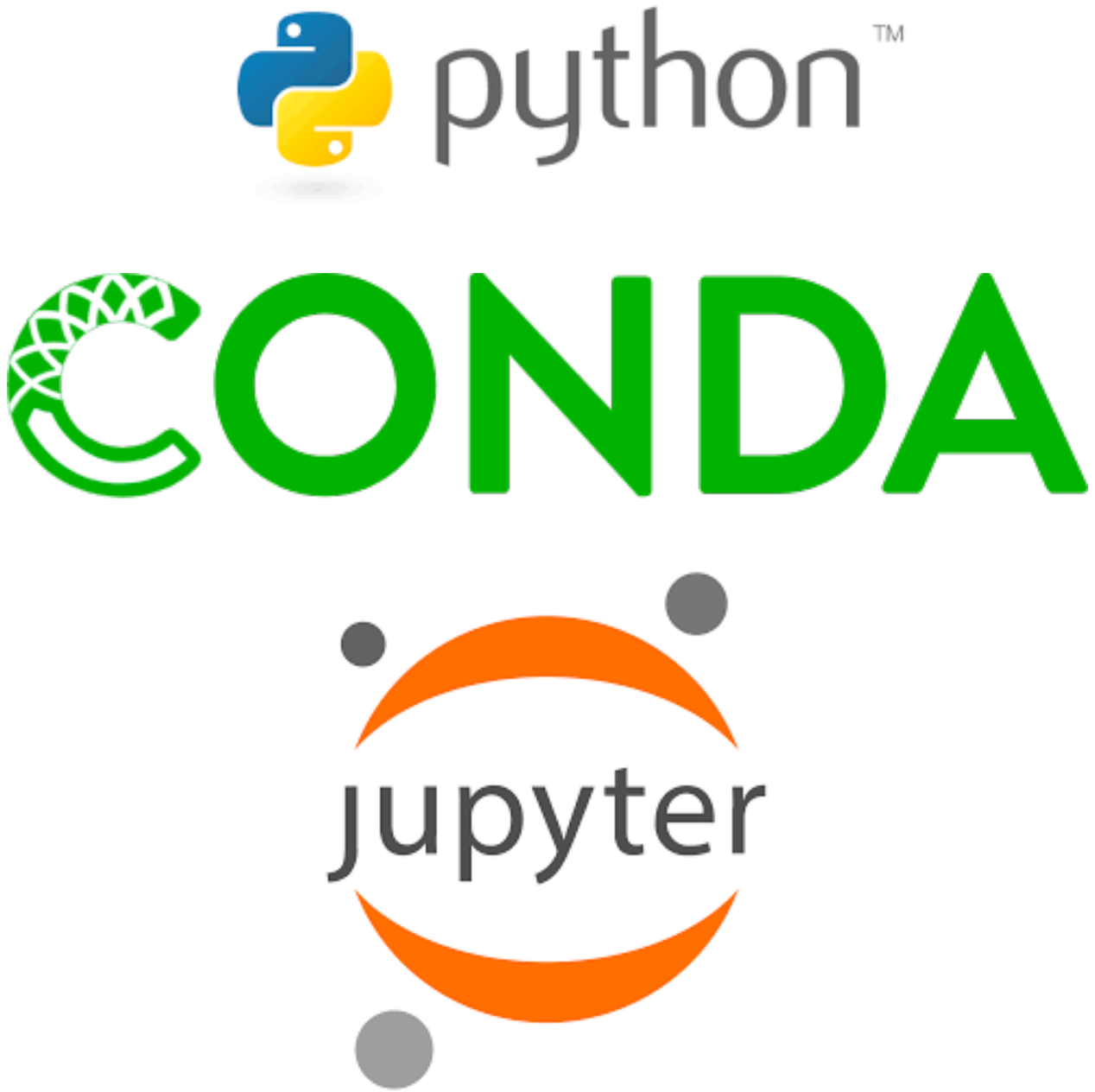
- Temboo — *by @temboo* — Framework for working with Twitter via many platforms including iOS, Android, Java, PHP, Python, Ruby, and Node.js

### ASP

- asptwitter *by @timacheson* — “the simplest possible way to implement Twitter within a classic ASP website”.

## Python

- python-twitter *maintained by @bear* — this library provides a pure Python interface for the Twitter API (documentation)
- tweepy *maintained by @applepie & more* — a Python wrapper for the Twitter API (documentation) (examples)
- TweetPony *by @Mezgrman* — A Python library aimed at simplicity and flexibility.
- Python Twitter Tools *by @sixohsix* — An extensive Python library for interfacing to the Twitter REST and streaming APIs (v1.0 and v1.1). Also features a command line Twitter client. Supports Python 2.6, 2.7, and 3.3+. (documentation)
- twitter-gobject *by @tchx84* — Allows you to access Twitter's 1.1 REST API via a set of GObject based objects for easy integration with your GLib2 based code. (examples)
- TwitterSearch *by @crw_koepp* — Python-based interface to the 1.1 Search API.
- twython *by @ryanmcgrath* — Actively maintained, pure Python wrapper for the Twitter API. Supports both normal and streaming Twitter APIs. Supports all v1.1 endpoints, including dynamic functions so users can make use of endpoints not yet in the library. (docs)
- TwitterAPI *by @boxnumber03* — A REST and Streaming API wrapper that supports python 2.x and python 3.x, TwitterAPI also includes iterators for both API's that are useful for processing streaming results as well as paged results.
- Birdy *by @sect2k* — "a super awesome Twitter API client for Python"




https://www.python.org/about/gettingstarted/

https://www.datacamp.com/courses/introduction-to-data-science-in-python

https://docs.conda.io/projects/conda/en/latest/index.html

https://www.codecademy.com/learn/paths/analyze-data-with-python

https://jupyterlab.readthedocs.io/en/stable/

Developer

developer.twitter.com

Developer Use cases Products Docs More Labs Apply

Developers

# Tap into what's happening.

**Publish and analyze Tweets, optimize ads, and create unique customer experiences.**

Apply — Twitter Developers

developer.twitter.com/en/application/use-case

Developer Use cases Products Docs More Labs Apply Apps

# Get access to the Twitter API

## #welcome

We're excited you want to use Twitter APIs and data!

As a developer platform, our first responsibility is to our users: to provide a place that supports the health of conversation on Twitter.

**This application process helps us to:**

1. Prevent abuse of the Twitter platform.
2. Better understand and serve our developer community.

**Thank you for your time and thoughtful responses.**
Applications are final once submitted and can't be edited.

## What is your primary reason for using Twitter developer tools?

We'll help you on your path to getting the most out of Twitter APIs and data.

**Professional**
...for commercial uses

- Building B2B products
- Building consumer products
- Build customized solutions in-house

**Hobbyist**
...for a personal project

- Making a bot
- Building tools for Twitter users
- Exploring the API

**Academi...**
...f...ion or research

- Doing academic research
- doing research to advance huma... ...tanding of a topic through Tw...
- Teaching
- Student

**Other**
I don't fit any of those

- Embedding Tweets on a website
- Doing something else

Next

Search all documentation...

Basics

Getting started

Things every developer should know

Frequently asked questions

Twitter developer apps

Developer portal

Authentication

Rate limits

Rate limiting

Response codes

Cursoring

Security

Twitter IDs (snowflake)

Counting characters

t.co links

Accounts and users

Tweets

# Rate limits

## Standard API rate limits per window

Standard

### POST endpoints

The standard API rate limits described in this table refer to POST endpoints. These rate limits apply to the standard API endpoints only, does not apply to premium APIs.

| Endpoint | Resource family | POST limit window | POST per user limit | POST per app limit |
|---|---|---|---|---|
| POST statuses/update | create content | 3 hours* | 300* | 300* |
| POST statuses/retweet/:id | create content | 3 hours* | 300* | 300* |
| POST favorites/create | favorites | 24 hours | 1000 | 1000 |
| POST friendships/create | friendships | 24 hours | 400 | 1000 |
| POST direct_messages/events/new | direct messages | 24 hours | 1000 | 15000 |

*Please note* - The 300 per 3 hours is with the POST statuses/update and POST statuses/retweet/:id endpoints is a combined limit. You can only post 300 Tweets or Retweets during a 3 hour period.

# GET endpoints

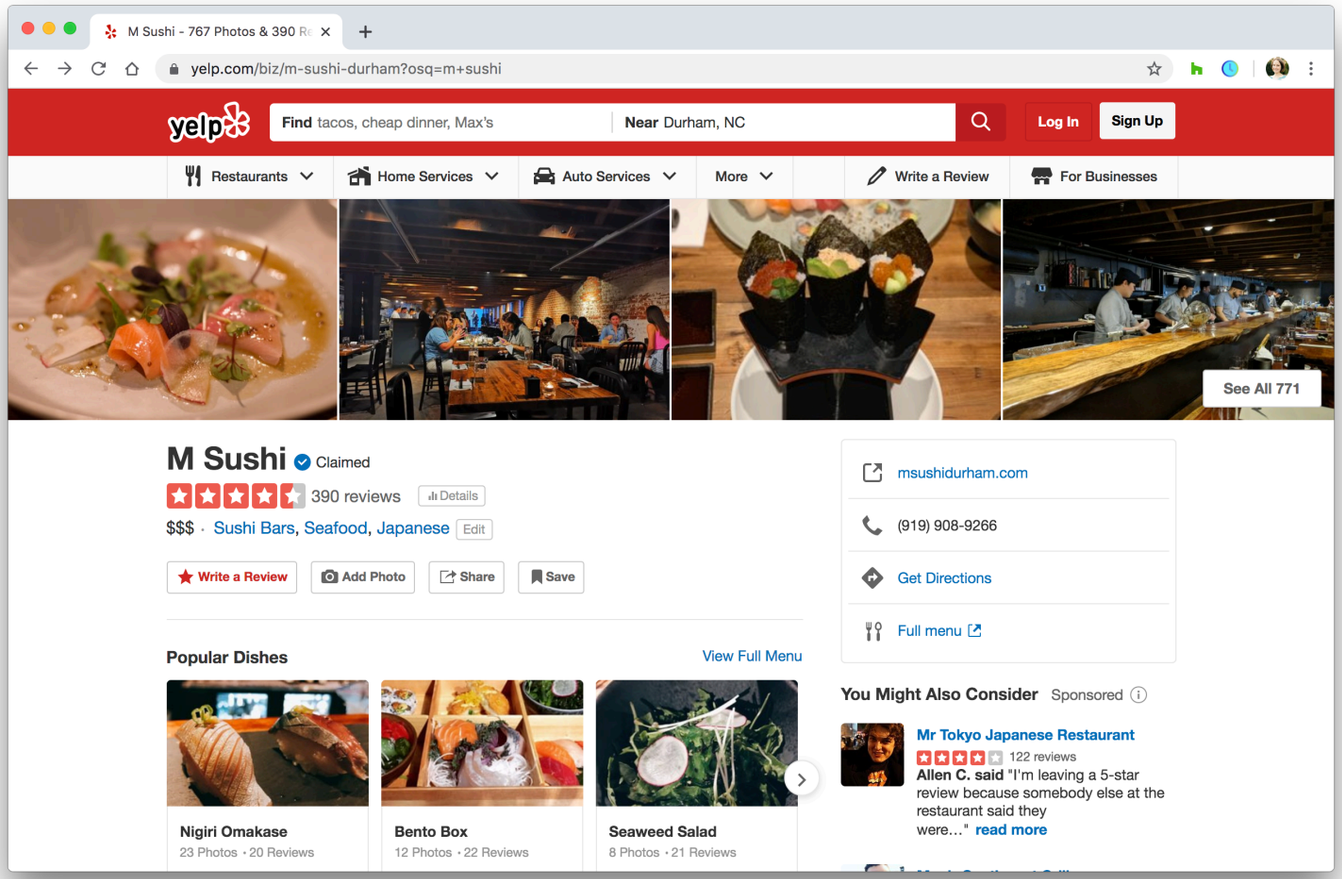
The standard API rate limits described in this table refer to GET (read) endpoints. Note that endpoints not listed in the chart default to 15 requests per allotted user. All request windows are 15 minutes in length. These rate limits apply to the standard API endpoints only, does not apply to premium APIs.

| Endpoint | Resource family | Requests / window (user auth) | Requests / window (app auth) |
|---|---|---|---|
| GET account/verify_credentials | application | 75 | 0 |
| GET application/rate_limit_status | application | 180 | 180 |
| GET favorites/list | favorites | 75 | 75 |
| GET followers/ids | followers | 15 | 15 |
| GET followers/list | followers | 15 | 15 |
| GET friends/ids | friends | 15 | 15 |
| GET friends/list | friends | 15 | 15 |
| GET friendships/show | friendships | 180 | 15 |
| GET geo/id/:place_id | geo | 75 | 0 |
| GET help/configuration | help | 15 | 15 |
| GET help/languages | help | 15 | 15 |
| GET help/privacy | help | 15 | 15 |
| GET help/tos | help | 15 | 15 |
| GET lists/list | lists | 15 | 15 |
| GET lists/members | lists | 900 | 75 |

# Demo

# Scraping webpages

# Anatomy of a webpage




HTML: main content

CSS: add styling

Images (e.g., JPG and PNG)

Javascript: interactivity

JSON: data for javascript

# Anatomy of a webpage

```
<!DOCTYPE html>
<html>
<body>

<h1>M Sushi</h1>
<p>A description.</p>
<img src="my_image.png" />

</body>
</html>
```

HTML: main content

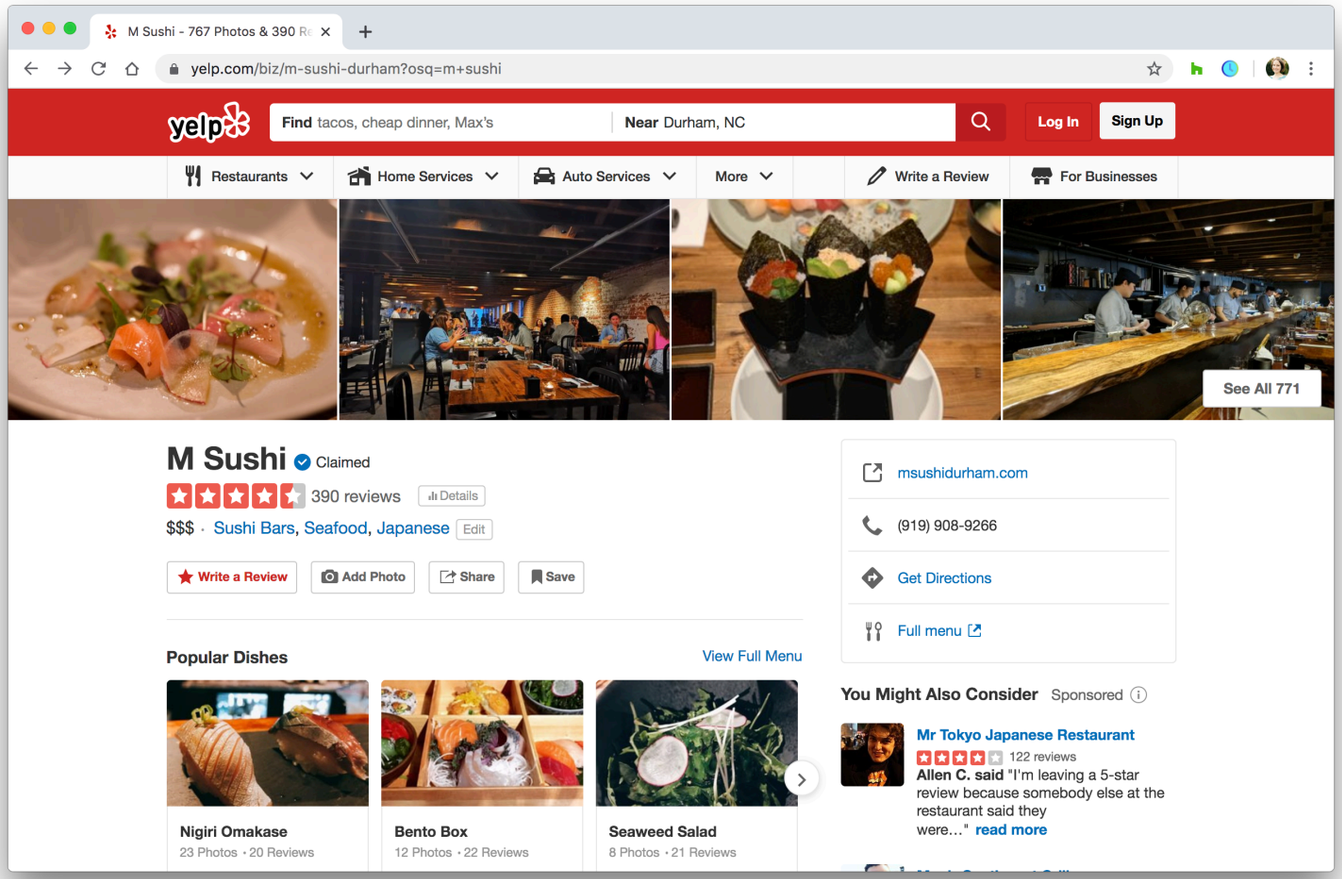
CSS: add styling

Images (e.g., JPG and PNG)

Javascript: interactivity

JSON: data for javascript

# Anatomy of a webpage




HTML: main content

CSS: add styling

Images (e.g., JPG and PNG)

Javascript: interactivity

JSON: data for javascript