Challenge & Pitfalls of Using API for Data Collection

* Authentication
* Blacklisting (exceed a rate limit -> IP is broken for that request)

JSON

Captures “typed” data:

* String; Booleans; Numbers
* Lists(ordered objects); Dictionaries

import json

json.load(open(“example.json”,r))

json.dumps(dic)

Issues with JSON

* Bloat
* Hard to detect schema inconsistency {“author” “text”}; {“author” “content”}

Collecting Data from Reddit

Text

Description automatically generated

Python library: requests

Web scraping

Sometimes we cannot use API: (i) doesn’t exist (ii) doesn’t expose the info we want

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Sampling

**Bias** is when we disproportionately weight factors that impact our measurement.