API

adam okulicz-kozaryn

adam.okulicz.kozaryn@gmail.com

this version: Thursday 14th April, 2022 09:29

outline

again, data management always comes first

- for traditional dat man, (still?) Stata is the best
- but for untypical and internet data, Py probably best:
  - getting data (APIs) and webscrapping
  - text processing
  - and other stuff (images, ets)

data data everywhere: internet!

- you've heard about:
  - information age, machine age, web 2.0, social media, etc
- data are everywhere:
  - blogs
  - wikis
  - government websites, etc
- soc sci traditionally just use the data that somebody puts in a spreadsheet for download
- but why not use ANY data that is online?
  - Python excels at it, and it's easy

API=Application Programming Interface

- basically an interface to get internet data
- most major websites have API
- you can connect to API and get data
- websites run on top of databases, eg google, twitter
- API is a way to access that database
- pretty much any company and organization has a database, is online, and has API

API v web scraping

- web scrping is getting data from internet by hand
  - flexible–can do any website
  - but has to write the code to extract information
  - time consuming, better use API
  - we'll just do API
- if you want data from website without API
  - email them first! they probably have API! say research only, otherwhise need to pay
  - and if that doesnt work, let me know and we'll try to scrap it