Data Science
Codecademy
Data Acquisition

Introduction: Data Acquisition
Exploring and defining the methods of obtaining data

The goal of nearly all data science endeavours is to answer a question with data.

Data acquisition (also called **data mining**) is the process of gathering data.

Some things need to be considered when acquiring data:

- What data is needed to achieve the business goal?
- How much data is needed to produce valuable insight and modeling?
- Where and how can this data be found?
- What legal and privacy parameters should be considered?

There are several methods we can utilize to acquire data. Some methods are:

- Public and Private data.
- Web scraping
- APIs
- Manual Data Acquisition

Public and Private Data

Public data:

Some popular public data sets are:

- Github
- Kaggle
- KDnuggets
- UCI Machine Learning Repository
- US Government's Open Data

- Five Thirty Eight
- Amazon Web Services

Each repository or dataset has its own terms of use.

Private data:

Private datasets are the ones available for purchase or licensing.

Web Scraping

Web scraping is extracting or copying data directly from a website. This data can then be stored and used just like any other dataset. We can scrape data two ways: manually and programmatically.

Manually just means literally copying and pasting it.

Programmatically means writing a bot or crawler that will systematically scan the page for data that fits the parameters we specify.

Some useful libraries of python for web scraping are:

- BeautifulSoup
- Selenium
- Scrapy

APIs

APIs or Application Programming Interfaces, are built around the HTTP Request/Response Cycle. A client sends a request to a website's server for data through an API call, then the server then searches its database and responds back to the client either with the data, or and error stating the request can not be fulfilled.

Manual Data Acquisition

When the required data is not available online we need to harvest it ourselves. We can make google forms and share it with others or use googles paid service of google surveys.

Ethics

With all the data acquisition methods, there are ethical considerations that are important to make. Issues of privacy and data ownership are top concerns among ethicists and politicians.

Questions like:

- Who owns the data uploaded to a website by users?
- When and how should users of services be notified that data about them is being acquired?
- What kinds of data should be restricted from being acquired about users?
- How can users protect their privacy and know when it has been breached?

As a data practitioner, it is essential to keep in mind the short term implications of data collection as well as the long-term effects of data analysis. By considering the ethical implications we can make sound data driven decisions.

https://towardsdatascience.com/how-to-build-your-own-dataset-for-data-science-projects-7f4ad0429de4