Technology Fundamentals for Business Analytics

Jason Kuruzovich

Issues

- Ipython Notebooks
  - Wakaria.io
  - Rstudio
- Vagrant
  - Windows 8.1 – Disable Hyper-V
  - Ensure Hardware Virtualization is enabled in your machine's BIOS

Agenda

- Python – Review and Extension
- Overview of the Web, ultimate big data source
- Accessing Data
  - File Access/Bulk Download
  - API
  - Web Scraping
- Lab 3
  - Twitter
  - Web Mining

Python

- Work through Chapter 1-8 of Basic Python Tutorial

https://docs.python.org/2/tutorial/

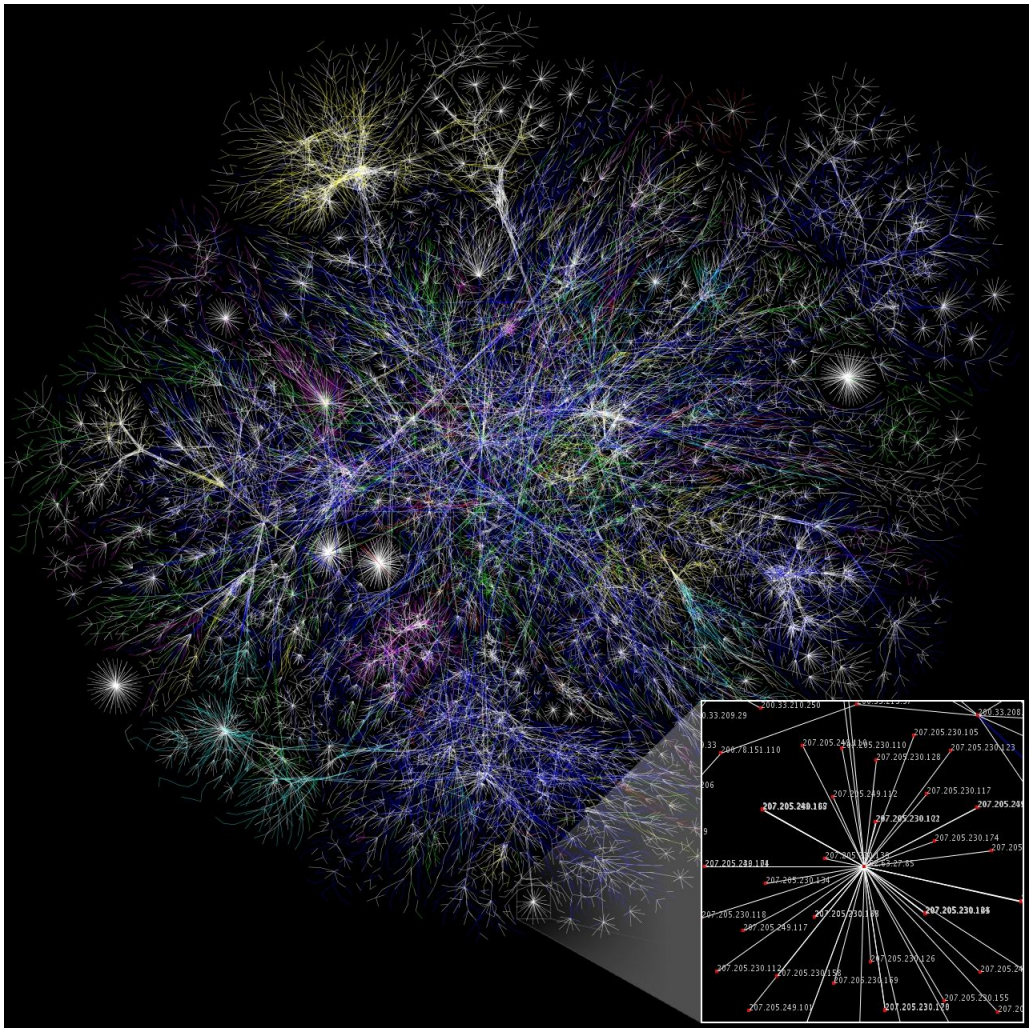
Indexed Files on Google

- 100,000,000 GB
- 920,000 Servers
- 30,000,000,000,000 Pages

(estimates vary depending on source)

http://www.statisticbrain.com/total-number-of-pages-indexed-by-google/

Web as Information Source



Google processes about 24 petabytes (1PB=1024 TB) of information a day

Approximate size of web 1024 petabytes (1 exobyte)

To think of it easier; if you filled a room that was 8' X 10' X 8' (ceiling) you could fit about 450 or so hard drives in there. Assuming you used even 2 TB hard drives you would still need over 1000 of those rooms filed to " download the internet".

http://wiki.answers.com/Q/How_large_is_the_Internet

But the “Deep Web” May Include Many More Pages

https://www.wilsoncenter.org/sites/default/files/stip_dark_web.pdf

What Can We Do With All this Data?

Methods for Getting Data

1. Bulk Data Download/Cloud Access
   - Easiest way to get large datasets, but won't work for dynamic
2. API
   - Make specific calls to data resources
3. Screen Scraping
   - Parse data from html

Bulk Data Download

- Specialty Sites
  - Baseball Databank (CSV file or .sql database archive)
- Government (Data.gov)
  - Education Data link
  - Economic Data link
- Data Marketplaces
  - Infochimps link
  - Amazon link (Mapping drives so don't have to do upload/download)

Why Would Companies Make Data Available Via API?

Data/API Strategy

- Twitter
  - Make data available, enabling an ecosystem of developers to surround the platform, making it more valuable
- Facebook
  - Make data available, enabling an ecosystem of developers to surround the platform, making it more valuable

API Economy

- API
- **“Application programming interfaces (APIs):** Programming hooks, specifications, or guidelines published by firms that tell other programs how to get a service to perform a task such as send or receive data”
- “Empowering developers to build against your platform doesn’t just create value for partners; the API provider wins as well by expanding the ecosystem, increasing retention, and driving up the value of the platform.”

http://venturebeat.com/2013/08/31/api-economy/#z1H2PHajvumblUla.99

Facebook as a Platform

- In May 2007, at a conference called F8, Mark Zuckerberg announced that he was opening up the screen real estate on Facebook to other application developers
- Facebook published a set of application programming interfaces (APIs) that specified how programs could be written to run within and interact with Facebook
  - Any programmer could write an application that would run inside a user's profile

Facebook as a Platform

- Developers can charge for their wares, offer them for free, and even run ads
- Facebook let developers keep what they made
- A key distinction: MySpace initially restricted developer revenue on the few products designed to run on their site, at times even blocking some applications
  - The choice was clear, and developers flocked to Facebook
- To promote the new apps, Facebook runs an applications area on the site where users can browse offerings

Facebook as a Platform

- Each application potentially added more value and features to the site without Facebook lifting a finger
- Some applications were accused of spamming friends with invites to install them
- There were security concerns and apps that violated the intellectual property of other firms

RESTful APIs

- Many initial visions of how systems would work together using web services (SOAP/XML)
- Transition to RESTful APIs that utilize HTTP
  - Client-Server
  - Stateless
  - Cache
  - Uniform Interface
  - Layered System

Source Roy Thomas Felding
http://www.ics.uci.edu/~fielding/pubs/dissertation/rest_arch_style.htm

What do we mean by an API utilizing HTTP?

HTTP

- The Hypertext Transfer Protocol (HTTP) is an application protocol for distributed, collaborative, hypermedia information systems. HTTP is the foundation of data communication for the World Wide Web.
- Hypertext is structured text that uses logical links (hyperlinks) between nodes containing text. HTTP is the protocol to exchange or transfer hypertext.

WIKIPEDIA ENTRY)

HTTP

- Use of HTTP simplifies data resource management because http can be used from whatever language the individual is using

Example

https://explore.data.gov/views/3vvm-4qnc/rows.json?accessType=DOWNLOAD

While Databases…

- HTTP API
  - Port is typically available
  - Driver Common
- Databases
  - Different databases, different ports may not be available (MySQL 3306)
  - Different types of drivers

API Authentication

- No Authentication
- Username/Password
- API Key Authentication (Data.gov)
- Oauth (Twitter/Facebook)
  - https://davidlyness.com/blog/wp-content/uploads/2013/04/oauth-authentication.png

Web Scraping

Data Scraping

- Web Pages often have a similar structure, this can be leveraged to “scrape” data in such a way that it can be used (in matrix)
- Tools
  - Connotatehttp://pages.connotate.com/replace-outdated-web-scraping.html
- Packages
  - Beautiful Soup

Data Scraping

PROBLEM: Any changes in html structure can break a web scraper

Sample Procedure

- Index
- Download HTML file
- Parse data and store in relational database
- Analyze data

Lab 3