4/5/21 → Basics of network protocols and python libraries available to extract data from web

1. Hyper text transfer Protocol (HTTP)

→ Its application protocol for distributed and hyper media information system to exchange or transfer hypertext

→ Its structured texts and uses logical hyperlinks btwn nodes containing text.

→ Foundation of data communication in world wide web

HTTP is the media through which we can retrieve we based data

↓

P - Protocol is rules that determines → who sends request
→ what action to be taken
→ what response will be given.

→ To send request and receive response, HTTP uses GET & POST methods

2. The world's simplest web browser

A socket is much like a file, except that a single socket provides a two-way connection between two programs

Socket → Bi-directional data path to remote system

→ If you write something to a socket, it is sent to application at the other end of the socket

→ If you read from the socket, you are given the data which other application has sent.

→ When you create a socket you have to specify its address family and then you can only use that address type with the socket

AF_INET ⟶ Its address family (IP) that is used to designate the type of address that your socket can communicate with.

SOCK_STREAM ⟶ (TCP) is a constant indicating the type of socket. It works as a file stream and most reliable over the network

PORT ⟶ Its a logical end-point
Port-80 is the most commonly used in TCP

protocol suite

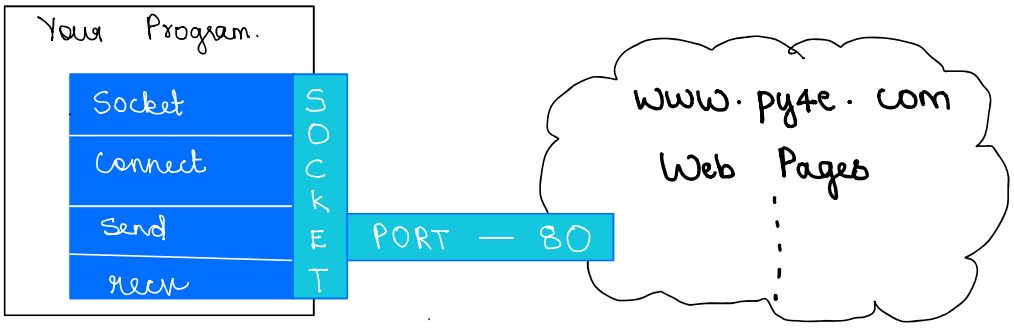
Command to retrive data must use CRLF and end in \r\n\r\n

CRLF → carriage return line feed
\r\n\r\n → line break in protocol specification

encode() or b at end ⟶ method applied on string and returns the byte-representation of the string

decode() ⟶ method return a string decoded from given bytes.

Socket connection between user program and web page



3) Retriving an image over http

→ We extract images in a chunck of 5120 bytes at a time.
→ Store the data in a string.
→ Trim of the edges
→ then store image file on a disk

4) Retriving web pages with Urllib

→ Here webpage is treated like a file
→ Urllib handles all HTTP protocols and header details

→ Once webpage is opened using urllib.urlopen() we can treat it like a file and read through it using for-loop.

→ the headers are still sent but urllib code consumes the headers and only returns the data to us

5) Reading Binary files using Urllib

→ the program reads all the data at once and stores it in variable img in main memory of computer

→ It works if size of file is less that size of RAM in

computer.

→ To avoid memory overflow we retrive data in blocks/buffers and write each block to the disk before retriving next block.

6) Parsing HTML and scraping the web

Web Scraping → program that pretends to be we browser and retrives pages then examines data in those pages looking for pattern.

a) Parsing HTML using regular expression

pattern of html (Hyper text Markup language)

<h1> & </h1> → beginning & end of header tags

 & → beginning & end of paragraph tags

<a> & → beginning & end of anchor tags

href is the attribute for anchor tags which takes the value as link for another page.

6/5/21 b) Parsing html using beautiful Soup

→ it converts incoming documents to unicode and outgoing documents to UTF-8
→ install beautiful soup package to use.

href → hypertext reference

BeautifulSoup to extract href attribute from anchor tag

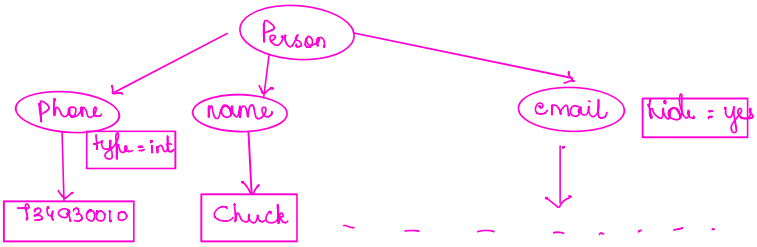
SSL → secure socket layer

6) Using Web services

a) XML → extensible markup language, common format used for exchange of data across the web.

→ suitable for exchanging document style data
→ similar to HTML but more structured than HTML.
→ its a tree structure
→ children tags are derived from parent nodes.

ex: Tree Structure of XML



i) Parsing XML

→ python library used is xml.etree.ElementTree

→ XML code is given as string to built in method fromstring() of ElementTree Class.

↓

convert XML code into tree structure of XML codes.

→ find() method searches XML and retrieves a node that matches the specified tag.

→ get() retrieves the value associated with specified attribute of that tag.

→ XML documents are hierarchical and contain multiple nodes. to process them we loop through these nodes.

→ findall() method returns a python list of subtrees that represent the user structure in XML tree.

b) JSON ⇒ Javascript Object Notation

→ its used when programs want to exchange dictonaries, lists or other internal information with each other.

→ in JSON we have simple key-value pairs

→ XML's tags are replaced by outer curly braces in JSON.

→ it has less capability than XML

→ its advantage is that it maps directly to some combination of dictonaries and list.

i) Parsing JSON

→ Module JSON is used to parse data in JSON pages.

→ method loads in JSON converts the string into a list of dictionaries.

*) API → Application Programming Interface

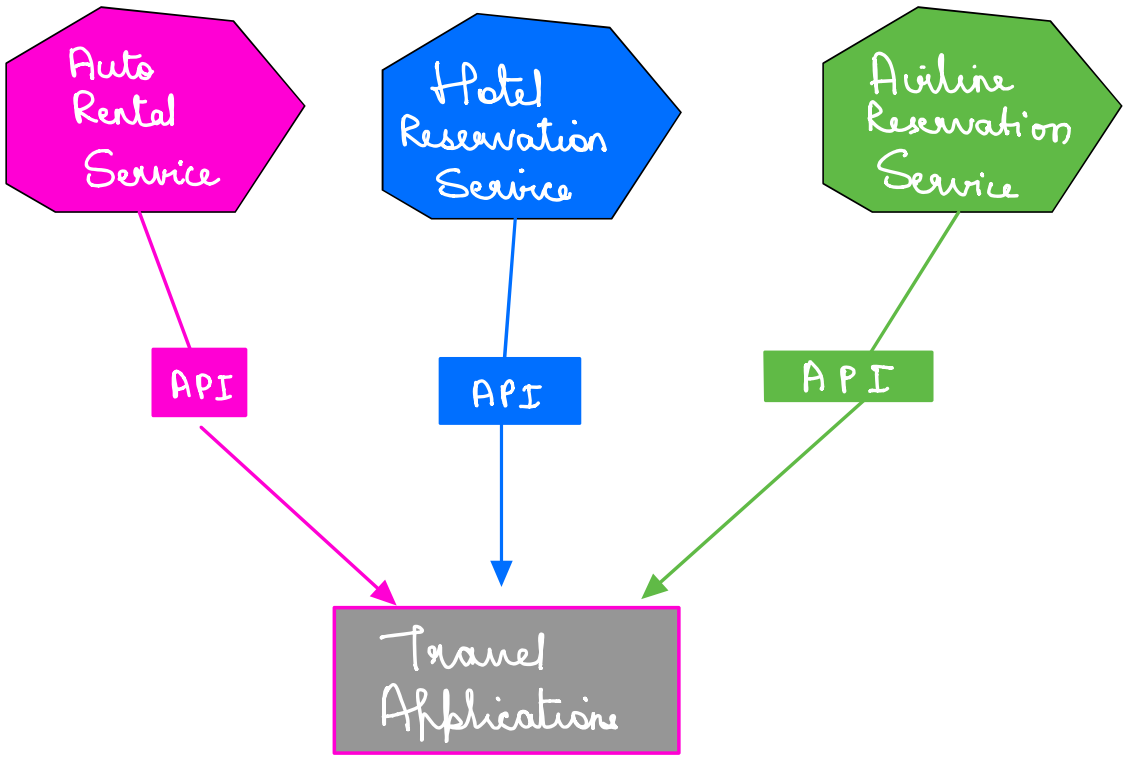
→ It defines documents that contracts between applications.

SOA → Service oriented Architecture

↳ approach of programming where our program includes services provided by other programs.

→ A non-SOA approach is when our application is a single stand-alone application which contains all of the code necessary to implement the application.

ex of SOA using travel application



advantages → Maintain only one copy of the data (security is necessary)

→ Owners of the data can set the rules about

use of their data.

8) Google Geocoding Web services

Program → (retrieve) Search String → (encoded + Google API link) = URL to fetch data from internet

→ data retrieved is now parsed to JSON as JSON object format.

→ If data is not retrieved we display

"Failure to retrieve"

→ If positively retrieved we will dump the data in JSON object

→ Use indexing on JSON to retrieve location address, longitude, latitude.

9) Security and API usage

→ Private API is need API key from vendos to use it.

→ Most common protocol used by Internet in signing requests is OAuth.

OAuth ?? ....

8/5/21 Using Databases and SQL

Database → A structured set of data stored in permanent storage (ex: hard disk)

→ They are like dictionaries and have key-value pairs

→ Database management softwares are designed to insert and retrieve data fast. However big the dataset is.

→ softwares builds indexes as data is added to index to build quicker access to a particular entry.

ex: Oracle, MySQL, Microsoft SQL, SQLite, PostgreSQL

→ SQLite is built into python.
→ It's a C-library that provides light weight disk based database
→ It doesn't require a seperate server process

→ It allows accessing the database using non-standard variant of SQL query language.

→ It's designed to be embedded into other applications to provide database support within the applications.

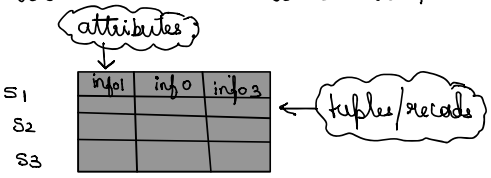
↳ It's used for data manipulation problems in Informatics such as twitter spidering applications.

1) Database concepts

→ Database is like a spreadsheet

→ Primary datastructures in a database are tables, rows and columns.

tables - relation
rows - tuple / record (m)
columns - attribute (n)



every cell (i, j) → j^{th} attribute for i^{th} tuple

→ table columns indicate the type of information to be stored.
→ table row indicates the value / record of each column for a particular entity.

(RDBMS) ⟶ Relational Database Management System

↳ It's a software that can maintain relationships between multiple tables in a single database.

2) SQL Summary (Structured Query Language)

→ SQL is the language for storing, manipulation and retriving data in databases irrespective of RDBMS software used.

→ Usage of SQL commands may vary from one RDBMS to other with little syntactical difference

SQL Commands

Command	Meaning
CREATE DATABASE	creates a new database
ALTER DATABASE	modifies a database
CREATE TABLE	creates a table

ALTER TABLE	modifies a table
DELETE TABLE	alters a table
INSERT INTO	inserts new data into database
SELECT	extracts data from database
UPDATE	updates data in a database
DELETE	deletes data from database

→ every RDBMS has its own way of storing the data in the tables
→ Each RDBMS uses its own set of data types for attribute values to be used.
→ SQL commands are case sensitive

Data types of SQLite

Data types	Description
NULL	The value is a null value
INTEGER	The value is a signed integer stored in 1, 2, 3, 4, 6, 8 bytes depending on the magnitude of the value
REAL	The value is a floating-point stored as 8-byte floating point number
TEXT	the value is a text-string, stored using database encoding (UTF-8, UTF-16BE, UTF-16LE)
BLOB	The value is BLOB (binary large object) data stored exactly as it was input

UTF-8 → Unicode transformation format 8 bit

UTF-16BE → Unicode transformation format 16-bit Big Endian

UTF-16LE → Unicode transformation format 16-bit little Endian

3) Database browser for SQLite

→ Simple operations on database will be done using database manager and more complex operations is done using Python

database manager : database :: text editor : text files

→ software is Database Browser for SQLite which is freely available

http://sqlitebrowser.org

4) Creating a database Table

→ Specify the names of colums and type of data that will be stored in it.

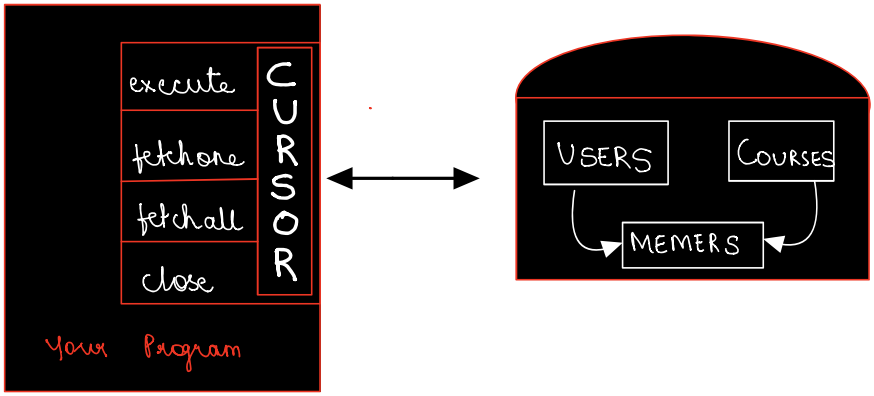
→ Once type is know, it will choose effective way To store and look up to the tube of data.

ex: Table - Tracks with two columns title and plays of type TEXT & INTEGER respectively

Connect() ⇒ makes connection to the database (creates if doesn't exist) in the current directory

cursor() ⇒ file handle that can perform operations on data stored in the database.

execute() ⇒ used the commands on the contents of database



commit() → method is used to store data permanently in database.

SELECT → command is used to retrieve the data from the database using for loop.

:memory: argument to connect() will be on memory (RAM) of the computer and not on harddisk.

5) Kinds of keys

→ We build data models by putting our data into multiple linked tables and linking the rows of these table using some keys.

→ three types of keys used in database models are:

a) Logical key → this used by the 'real world'
→ It defines the relationship between primary and foreign keys.

↳ Most of the times a unique constraint is added to logical keys

b) Primary key
- → It's the number assigned automatically by the database.
- → has no meaning outside the program
- → Used to link rows from different tables together.

Usually searching for a row using is the fastest way and consumes less storage & is sorted very quickly

c) foreign key → it points to primary key of an associated row in a different table.

6) Basic data Modelling

→ RDBMS has the power of linking multiple tables

data Modelling → It's the act of deciding how to break up your application data into multiple tables and establishing the relationship between these tables.

data model → Design document that shows the table and their relationships.

data modelling is based on the concept of data normalization that has certain set of rules.

ex: We never put same string of data in a database more than once.

If we need it more than once, we create a primary key & reference it when needed.

⇓

→ Because string needs more data than Integer
→ data retrival by 'comparing' is simpler for integers

*) Using JOIN to retrieve data

→ SQL uses JOIN to reconnect these tables
→ In JOIN we can specify the field that is used to reconnect rows between these tables.