Data Mining and Accouting Analytics -Data Preprocessing

Dr. Yi Long (Neal)

Most contents (text or images) of course slides are from the following textbook
Provost, Foster, and Tom Fawcett. Data Science for Business: What you need to know about data mining and data-analytic thinking. " O'Reilly Media, Inc.", 2013

Python Basics (Overall)

- Python Basics
  - Values
  - Operations on values
  - Assignments
  - Input/output operations
  - Control Flow (If, for, def function)
  - Data structure
  - Pandas/Numpy
- Data Understanding

Code example: sentiment_analysis.py

Types of Data

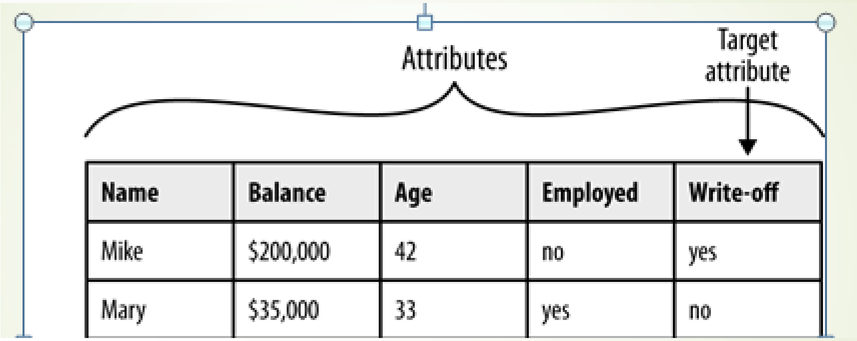
Labeled

Positive

Negative

广汽集团携手腾讯发展智能汽车

深交所发函质疑大连友谊资产重组


Attributes
Target attribute

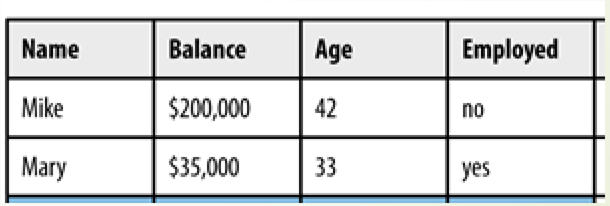
Name	Balance	Age	Employed	Write-off
Mike	$200,000	42	no	yes
Mary	$35,000	33	yes	no

Unstructured

Structured

广汽集团携手腾讯发展智能汽车

深交所发函质疑大连友谊资产重组

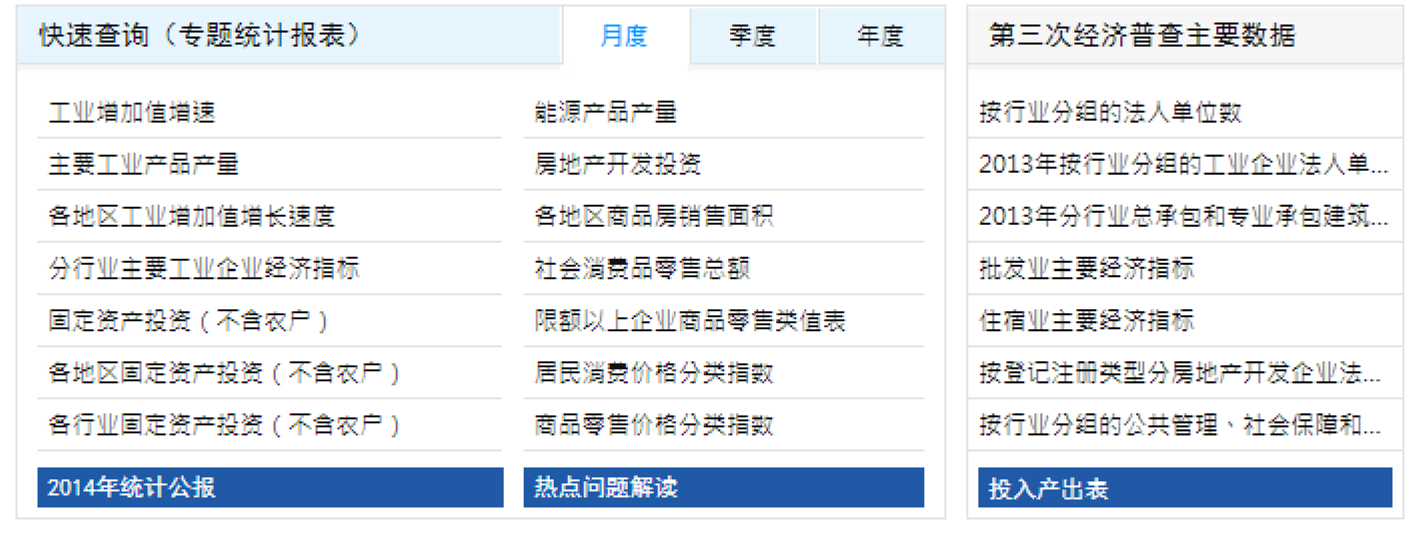


Name	Balance	Age	Employed
Mike	$200,000	42	no
Mary	$35,000	33	yes

Unlabeled

Data Collection

- There are tremendous public data set
  - ✓ AWS Public Datasets https://aws.amazon.com/public-datasets
  - ✓ UC Irvine Machine Learning Repositoy http://archive.ics.uci.edu/ml/index.php
  - ✓ 中国国家数据 http://data.stats.gov.cn/ ,
  - ✓ **Dataset List : https://github.com/awesomedata/awesome-public-datasets**
  - ✓ Business Database: Wind, CSMAR, WRDS …



快速查询（专题统计报表）

工业增加值增速
主要工业产品产量
各地区工业增加值增长速度
分行业主要工业企业经济指标
固定资产投资（不含农户）
各地区固定资产投资（不含农户）
各行业固定资产投资（不含农户）
2014年统计公报

能源产品产量
房地产开发投资
各地区商品房销售面积
社会消费品零售总额
限额以上企业商品零售类值表
居民消费价格分类指数
商品零售价格分类指数
热点问题解读

第三次经济普查主要数据

按行业分组的法人单位数
2013年按行业分组的工业企业法人单...
2013年分行业总承包和专业承包建筑...
批发业主要经济指标
住宿业主要经济指标
按登记注册类型分房地产开发企业法...
按行业分组的公共管理、社会保障和...
投入产出表

Data Quality – Completeness

- **Missing records** : selection bias is serious
  - ✓ Biased Sample: sampling population for children number
  - ✓ Issue credit card without those rejected users
- **Missing value:** values of a part of entries are missing
  - ✓ Missing at random: somehow better
  - ✓ Missing not at random : low –income participants are less likely to fill in the income
- **Handling missing value:**
  - ✓ Imputation with common/average/recent value
  - ✓ Drop records with missing value
  - ✓ Pandas treat missing value has NaN: isnull(), dropna(), fillna()

Data Quality – Unbalanced Data

- **Unbalanced data:** the data set might be biased extremely to one type of records
  - ✓ Fraud detection: about 2% of credit card accounts are defrauded per year.

Issues in Data Collection

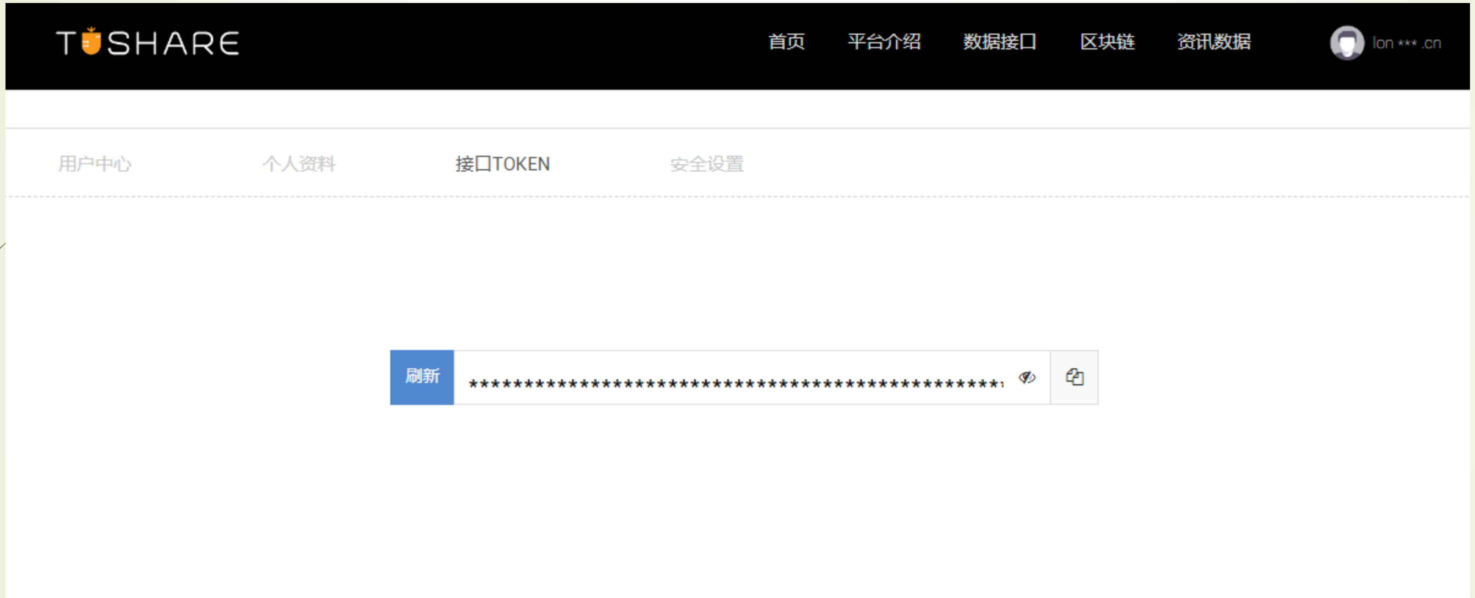
- ✓ Sampling bias
- ✓ Missing data
- ✓ Imbalanced data
- ✓ Privacy control
- ✓ Storage and manage
- ✓ Cross-check design
- ✓ ….

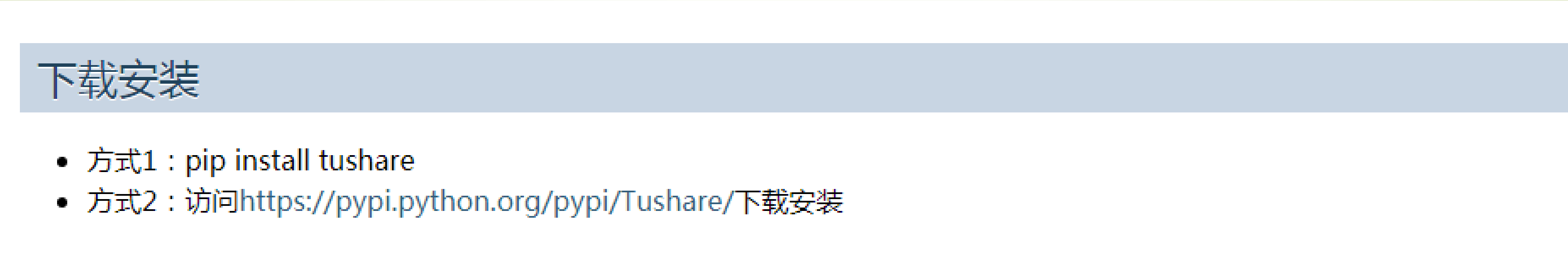
Types of Data – Structure (Example)

- **Structured data** refers to any data that resides in a fixed field within a record.
  - ✓ http://quotes.money.163.com/f10/zycwzb_600795.html#01c01
  - ✓ **CSV, database, Pandas dataframe (excel) (next lecture)**
- **Semi-structured data : Json, XML,HTML …**
  - ✓ http://api.money.126.net/data/feed/0000001,0600795,money.api?callback=_ntes_quote_callback5959502
  - ✓ https://www.xbrl-cn.org/xbrl/yingyong/
- **Unstructured data** does not have a pre-defined data model or is not organized in a pre-defined manner
  - ✓ http://quotes.money.163.com/f10/ggmx_600795_5188199.html

Tushare API

https://tushare.pro/



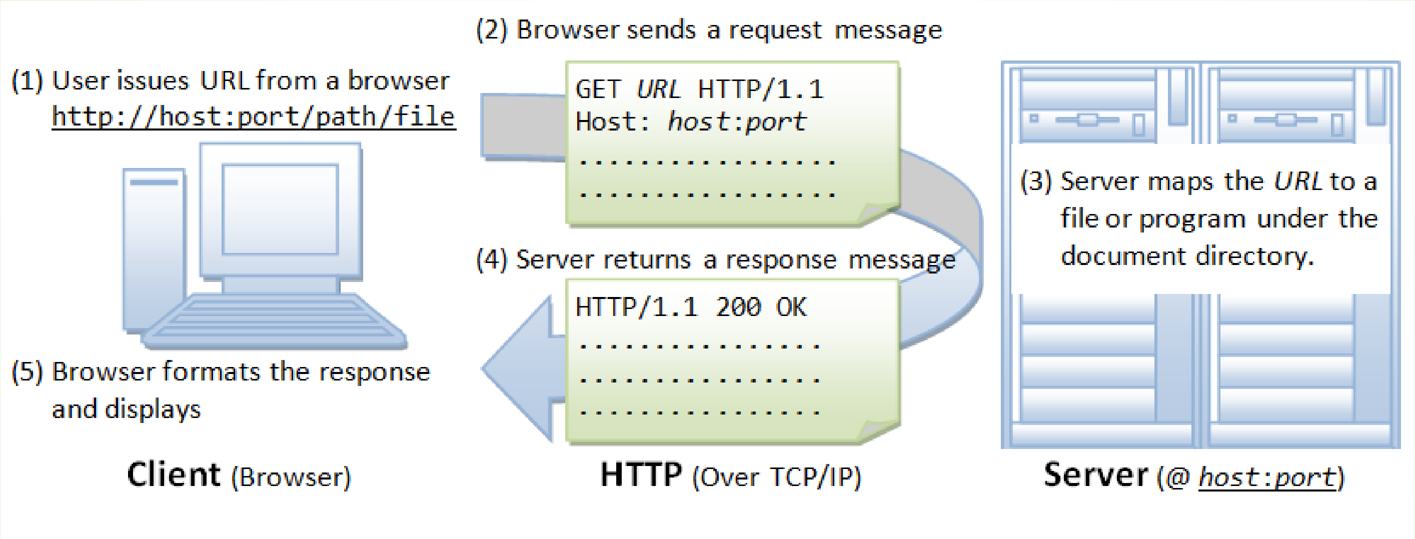


下载安装

- 方式1：pip install tushare
- 方式2：访问https://pypi.python.org/pypi/Tushare/下载安装

Web Scraping – Process

- Send well-prepared HTTP requests to the desired webpage
- Receive response from webpage server
- Check the response
- Parse the webpage into structured data if necessary
- Store the raw results/webpage


(2) Browser sends a request message
(1) User issues URL from a browser
http://host:port/path/file
GET URL HTTP/1.1
Host: host:port
(3) Server maps the URL to a file or program under the document directory.
(4) Server returns a response message
HTTP/1.1 200 OK
(5) Browser formats the response and displays
Client (Browser)
HTTP (Over TCP/IP)
Server (@ host:port)

HTTP Request

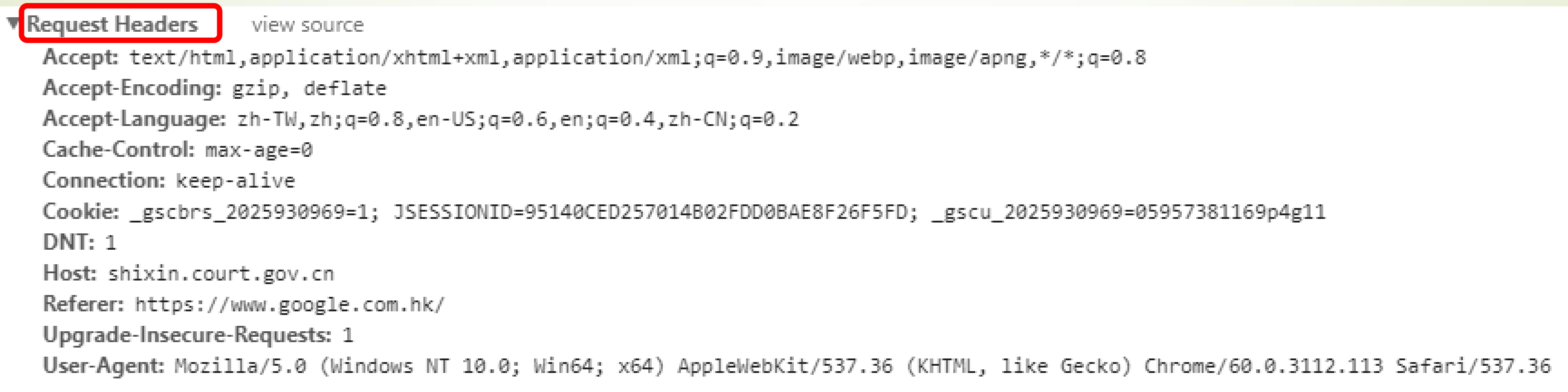
- Send HTTP requests to websites to download the page
  - ✓ **URLs with parameter:**

    http://httpbin.org/get?key1=value1&key2=value2&key2=value3

What do I want

  - ✓ **User-agent** to tell server what kind of client send this request
  - ✓ **Cookie** to verify the identify of senders (especially verify logged-in users)
  - ✓ IP address, Referer, Accept …

Who am I



```
Accept: text/html,application/xhtml+xml,application/xml;q=0.9,image/webp,image/apng,*/*;q=0.8
Accept-Encoding: gzip, deflate
Accept-Language: zh-TW,zh;q=0.8,en-US;q=0.6,en;q=0.4,zh-CN;q=0.2
Cache-Control: max-age=0
Connection: keep-alive
Cookie: _gscbrs_2025930969=1; JSESSIONID=95140CED257014B02FDD0BAE8F26F5FD; _gscu_2025930969=05957381169p4g11
DNT: 1
Host: shixin.court.gov.cn
Referer: https://www.google.com.hk/
Upgrade-Insecure-Requests: 1
User-Agent: Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/60.0.3112.113 Safari/537.36
```

Handling HTTP Request and Response

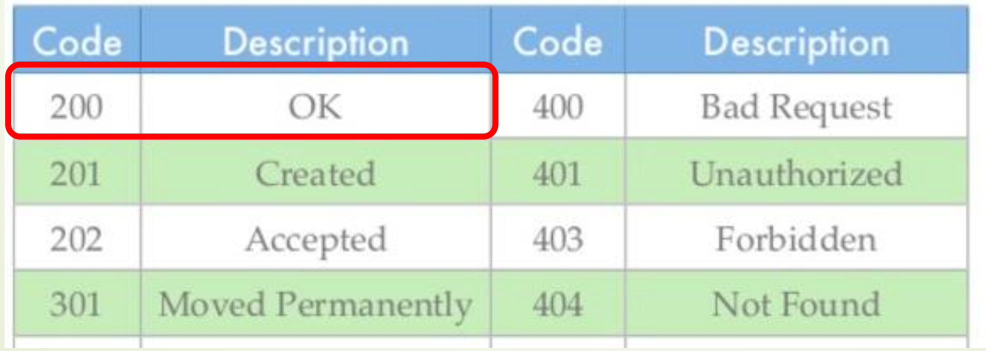
- Pacakge requests can help us prepare and send requests very easily
  - ✓ Import requests package
  - ✓ Send http request and get response in one command

```
response = requests.get('http://XXX.com”)
```

  - ✓ Send request

```
response = requests.get('http://XXX.com”, herders = header_dict)
```

  - ✓ Check the status of response by response.status_code
  - ✓ Read the content of response by response.text



Code	Description	Code	Description
200	OK	400	Bad Request
201	Created	401	Unauthorized
202	Accepted	403	Forbidden
301	Moved Permanently	404	Not Found

Advanced Usage

- Session object for making several requests to the same host
  - Allows you to persist certain parameters across requests. It also persists **cookies** across all requests made from the Session instance,
  - The underlying TCP connection will be reused, which can result in a significant performance increase

```
s = requests.Session()

s.get('https://httpbin.org/cookies/set/sessioncookie/123456789')
r = s.get('https://httpbin.org/cookies')

print(r.text)
# '{"cookies": {"sessioncookie": "123456789"}}'
```

HTTP + JSON

- JSON is a syntax for storing and exchanging data.
- JSON is text, written with JavaScript object notation.
- Python has a built-in package called json, which can be used to work with JSON data
  - If you have a JSON string, you can parse it by using the json.loads() method.
  - If you have a Python object, you can convert it into a JSON string by using the json.dumps() method
  - You can convert Python objects of the following types, into JSON strings:

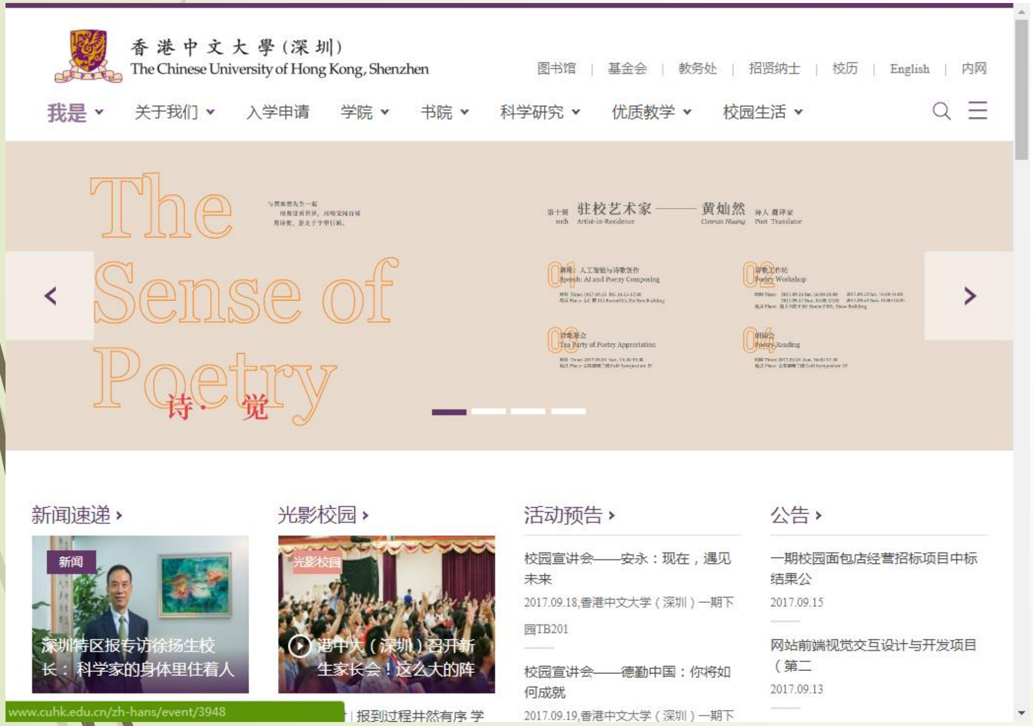
dict
list
Tuple

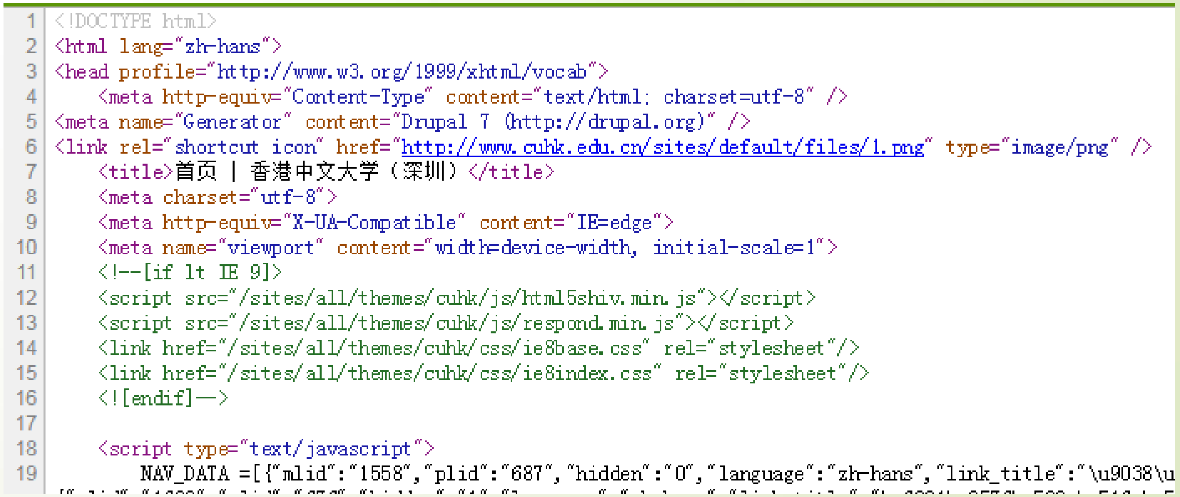
String
int
float

True
False
None

Huge Data Hides in Webpages

- Most webpages are written in HTML(similar XML)
  - ✓ HTML stands for **H**yper **T**ext **M**arkup **L**anguage


香港中文大學（深圳）
The Chinese University of Hong Kong, Shenzhen
图书馆 | 基金会 | 教务处 | 招贤纳士 | 校历 | English | 内网
我是 关于我们 入学申请 学院 书院 科学研究 优质教学 校园生活
The Sense of Poetry
诗·觉
驻校艺术家 —— 黄灿然
新闻速递
光影校园
活动预告
公告
深圳特区报专访徐扬生校长：科学家的身体里住着人
港中大（深圳）召开新生家长会！这么大的阵
校园宣讲会——安永：现在，遇见未来
校园宣讲会——德勤中国：你将如何成就
一期校园面包店经营招标项目中标结果公
网站前端视觉交互设计与开发项目（第二
www.cuhk.edu.cn/zh-hans/event/3948



```
<!DOCTYPE html>
<html lang="zh-hans">
<head profile="http://www.w3.org/1999/xhtml/vocab">
    <meta http-equiv="Content-Type" content="text/html; charset=utf-8" />
<meta name="Generator" content="Drupal 7 (http://drupal.org)" />
<link rel="shortcut icon" href="http://www.cuhk.edu.cn/sites/default/files/1.png" type="image/png" />
    <title>首页 | 香港中文大学（深圳）</title>
    <meta charset="utf-8">
    <meta http-equiv="X-UA-Compatible" content="IE=edge">
    <meta name="viewport" content="width=device-width, initial-scale=1">
    <!--[if lt IE 9]>
    <script src="/sites/all/themes/cuhk/js/html5shiv.min.js"></script>
    <script src="/sites/all/themes/cuhk/js/respond.min.js"></script>
    <link href="/sites/all/themes/cuhk/css/ie8base.css" rel="stylesheet"/>
    <link href="/sites/all/themes/cuhk/css/ie8index.css" rel="stylesheet"/>
    <![endif]-->

    <script type="text/javascript">
        NAV_DATA =[{"mlid":"1558","plid":"687","hidden":"0","language":"zh-hans","link_title":"\u9038\u
```

Introduction to HTML (1)

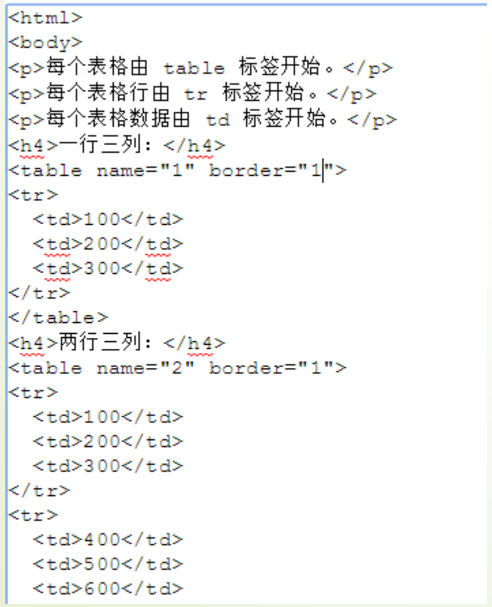
- HTML describes the structure/display of Web pages using markup
  - HTML elements are the building blocks of HTML pages
  - HTML elements are represented by **tags**
  - Each tag has a **tag name** and **other attributes** with values
- HTML tags are element names surrounded by angle brackets:

<tagname>content goes here...</tagname>

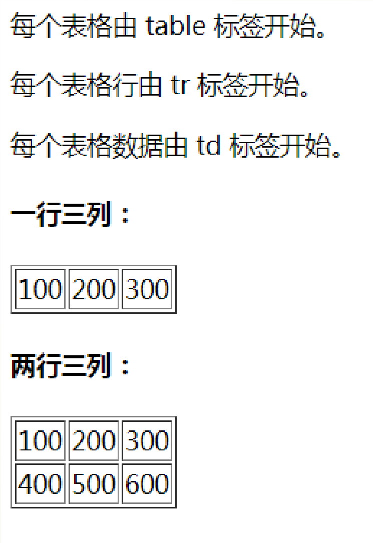
  - HTML tags normally **come in pairs** like <p> and </p>
  - The first tag in a pair is the **start/opening tag,** the second tag is the **end/closing tag**
  - The end tag is written like the start tag, but with a **forward slash** inserted before the tag name

Introduction to HTML (2)

- The browser can render the content of a page based on its HTML content
  - ✓ HTML tags are predefined with display settings: <table> <h1> < title >
  - ✓ https://www.w3schools.com/html/tryit.asp
  - ✓ https://htmlformatter.com/



```
<html>
<body>
<p>每个表格由 table 标签开始。</p>
<p>每个表格行由 tr 标签开始。</p>
<p>每个表格数据由 td 标签开始。</p>
<h4>一行三列：</h4>
<table name="1" border="1">
<tr>
  <td>100</td>
  <td>200</td>
  <td>300</td>
</tr>
</table>
<h4>两行三列：</h4>
<table name="2" border="1">
<tr>
  <td>100</td>
  <td>200</td>
  <td>300</td>
</tr>
<tr>
  <td>400</td>
  <td>500</td>
  <td>600</td>
```



每个表格由 table 标签开始。

每个表格行由 tr 标签开始。

每个表格数据由 td 标签开始。

一行三列：

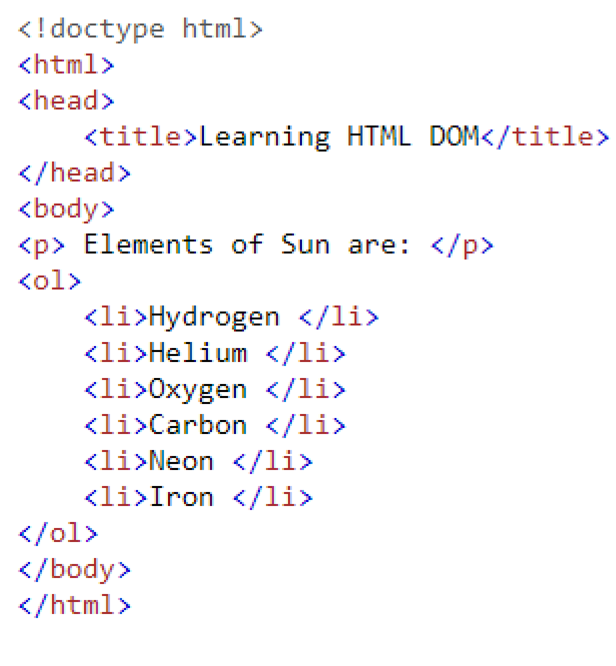
100	200	300

两行三列：

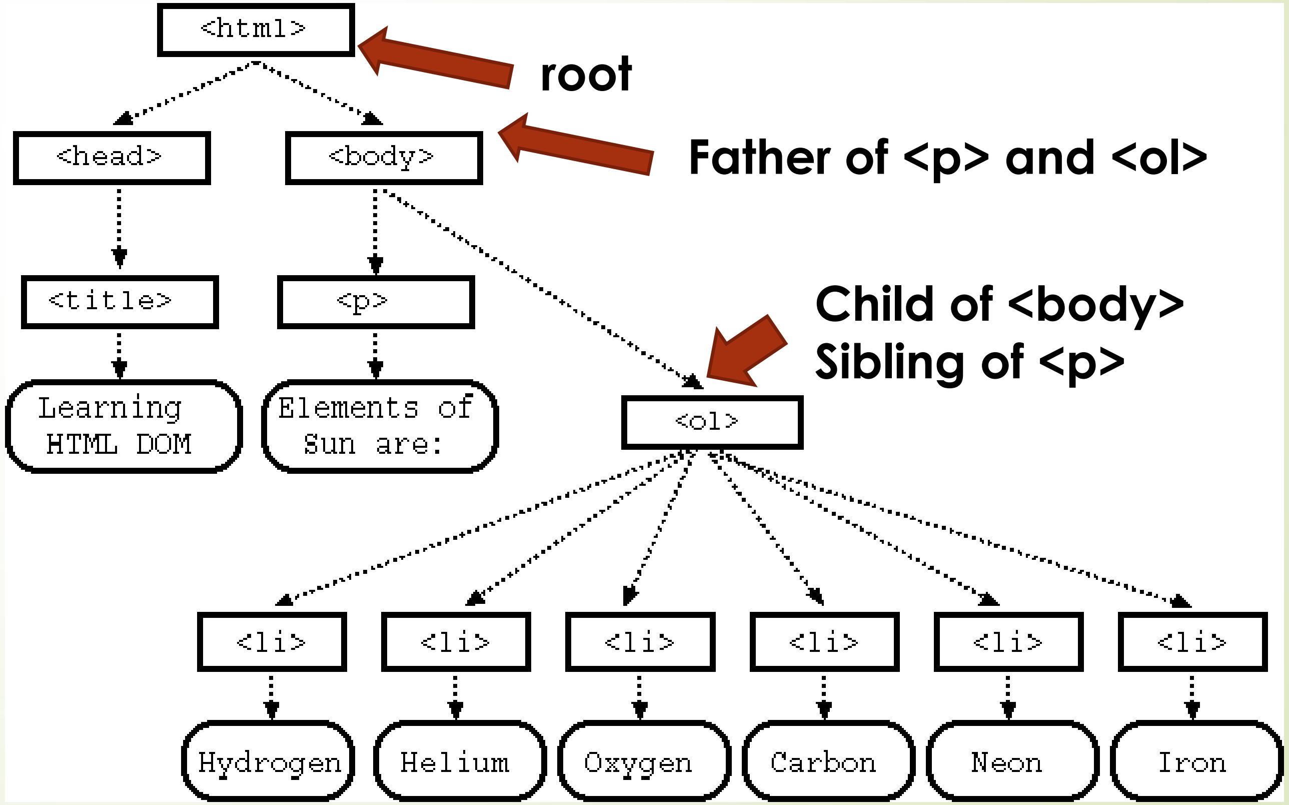
100	200	300
400	500	600

HTML DOM Tree

- HTML can be represented as a tree in Document Object Model （DOM）
  - ✓ Each tag can have multiple children tags



```
<!doctype html>
<html>
<head>
    <title>Learning HTML DOM</title>
</head>
<body>
<p> Elements of Sun are: </p>
<ol>
    <li>Hydrogen </li>
    <li>Helium </li>
    <li>Oxygen </li>
    <li>Carbon </li>
    <li>Neon </li>
    <li>Iron </li>
</ol>
</body>
</html>
```



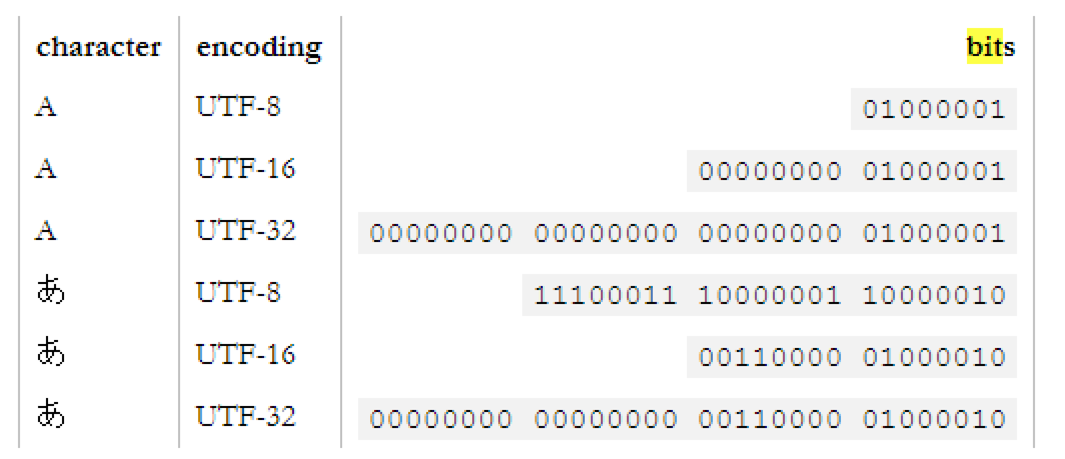
root

Father of <p> and <ol>

Child of <body>
Sibling of <p>

Character Encoding

- Character encoding is used to represent a repertoire of characters by some kind of encoding system.
  - Sometime referred as “character set", "character map", "codeset" and "code page“
  - ASCII can only handle 128 different characters , UTF-8 can hanldle 1,114,112 possible **characters**



character	encoding	bits
A	UTF-8	01000001
A	UTF-16	00000000 01000001
A	UTF-32	00000000 00000000 00000000 01000001
あ	UTF-8	11100011 10000001 10000010
あ	UTF-16	00110000 01000010
あ	UTF-32	00000000 00000000 00110000 01000010

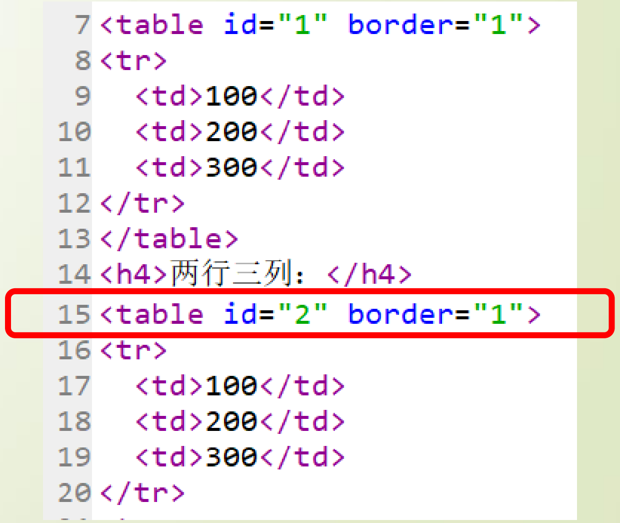
HTML Parser (Beautifulsoup)

- We can easily get the content in different tags and their structures
  - ✓ HTML Parser: beautifulsoup (easy)、lxml (fast)
  - ✓ Package **Beautifulsoup** can parse and build DOM for html
  - ✓ tag.find_all()/find() can search children tags under the given tag by name and attribute(/values of children tags:

father.find(child_name, attrs={‘key1’:’val1’,’key2’:’val2’})

  - ✓ tag.get_text() to get its text content
  - ✓ tag.name get its tag name
  - ✓ tag.children, tag.next_sibling …

table = root.find('table',attrs={'id':2})



```
<table id="1" border="1">
<tr>
  <td>100</td>
  <td>200</td>
  <td>300</td>
</tr>
</table>
<h4>两行三列：</h4>
<table id="2" border="1">
<tr>
  <td>100</td>
  <td>200</td>
  <td>300</td>
</tr>
```

Web Scraping – requests+bs4+csv

- Sending well-prepared HTTP requests to the desired webpage
- Receive response from webpage server
- check the response

requests

- Parsing the webpage into structured data if necessary

json
beautifulsoup

- Store the raw results/webpage

CSV, Excel,
Database

Resources for Web Crawler

- https://www.dataquest.io/blog/web-scraping-beautifulsoup/
- http://httpbin.org/user-agent
- https://requests.readthedocs.io/en/master/
- https://www.crummy.com/software/BeautifulSoup/bs4/doc/
- https://www.w3schools.com/python/python_json.asp
- Chrome

Outline

- Data Collection(web scraping)
- Data Storage
- Lab Quiz

Data Storage

- Structured data can be stored in various data structure in memory
  - ✓ Python list, tuple, set, dictionary …
  - ✓ Pandas dataframe, Numpy ndarray …
- How to store these data persistently and share
  - ✓ Python to python: pickle
  - ✓ Database: SQLite, MySQL, Oracle, MS SQL, Hbase, MangoDB …
  - ✓ Text files: txt, csv, tsv
  - ✓ Data interchange format: XML(/**html**), json,
  - ✓ Others: xls/xlsx (Excel), dta(STATA), ...

Read and Write Text Files

- Users can easily write/read content to/from files
  - ✓ Open file with proper status: f = open(file_path, ‘r/w/rb/wb’)
  - ✓ Write/reader content：content = f.read() / f.write(content)
  - ✓ Can read and write by lines as well
  - ✓ Close file, this is important to save changes/ release file: f.close()
  - ✓ Use the “encoding” parameter to deterring the character encoding
- Use **with** statement to close file automatically

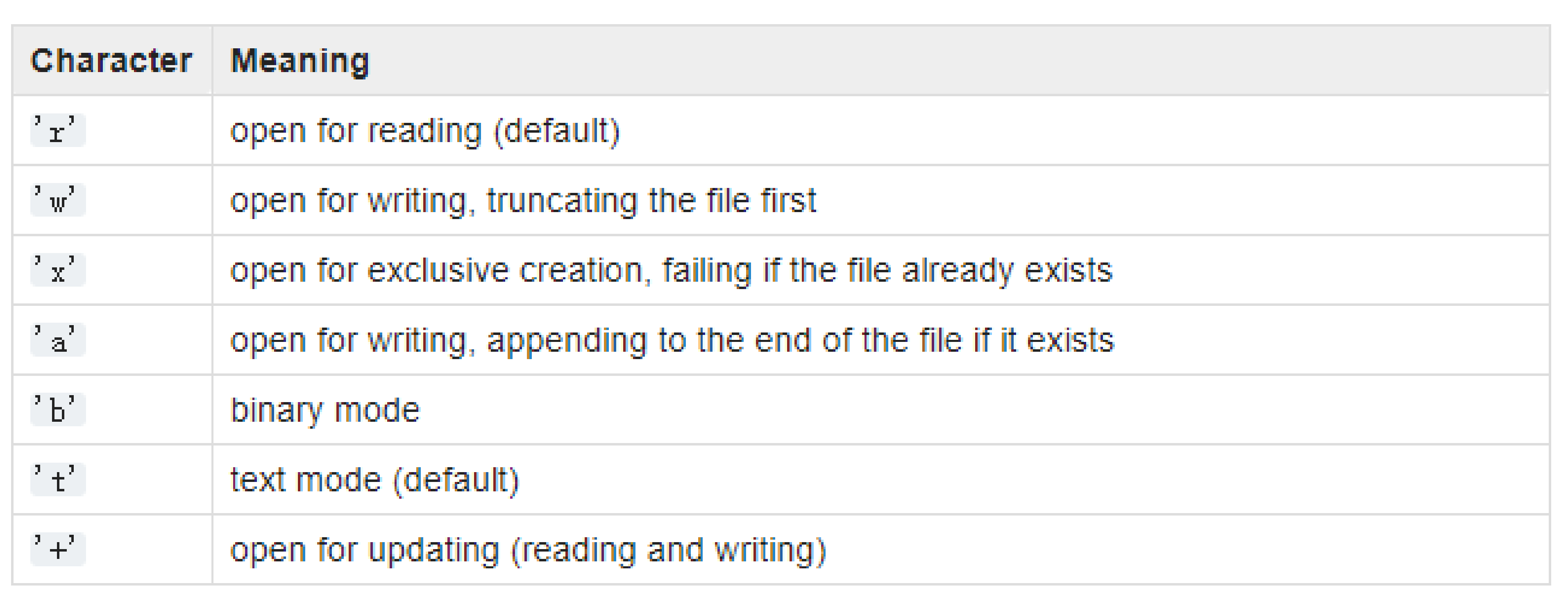
```
f=open(“work_file”,’w’)
f .write(haha)
#other operation
f.close()
```

→

```
with open(“work_file”,’w’) as f:
    f.write(haha)
    #other operation
```

Read and Write Other Files

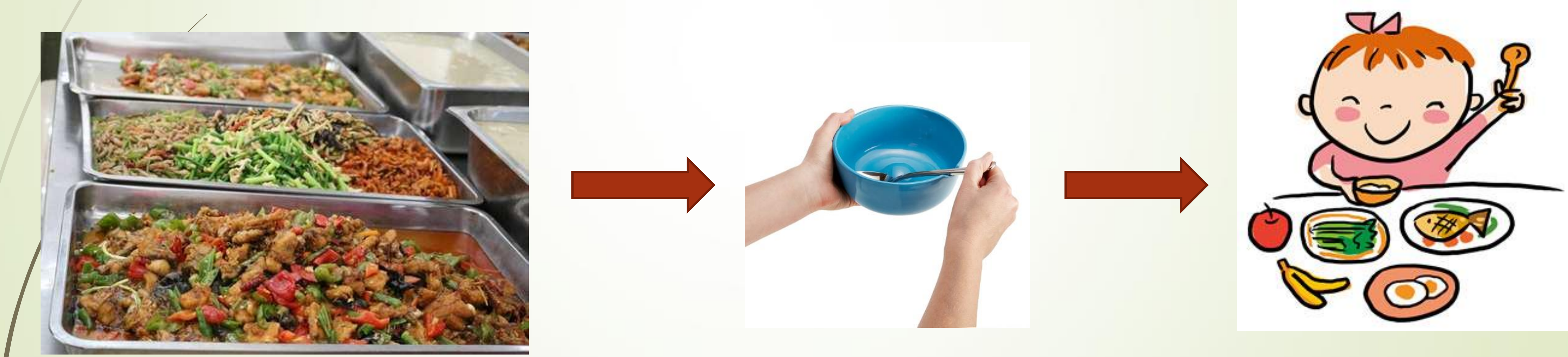
- It defaults to 'r' which means open for reading in text mode
- By default, 't' is included in the *mode* argument
- Can be used with combination, ‘rb’, ‘wb’, ‘w+b’…



Character	Meaning
’r’	open for reading (default)
’w’	open for writing, truncating the file first
’x’	open for exclusive creation, failing if the file already exists
’a’	open for writing, appending to the end of the file if it exists
’b’	binary mode
’t’	text mode (default)
’+’	open for updating (reading and writing)

Memory VS. Disk

- Memory is usually much smaller than disk
  - ✓ Processing data in small batches (row by row) is favorable



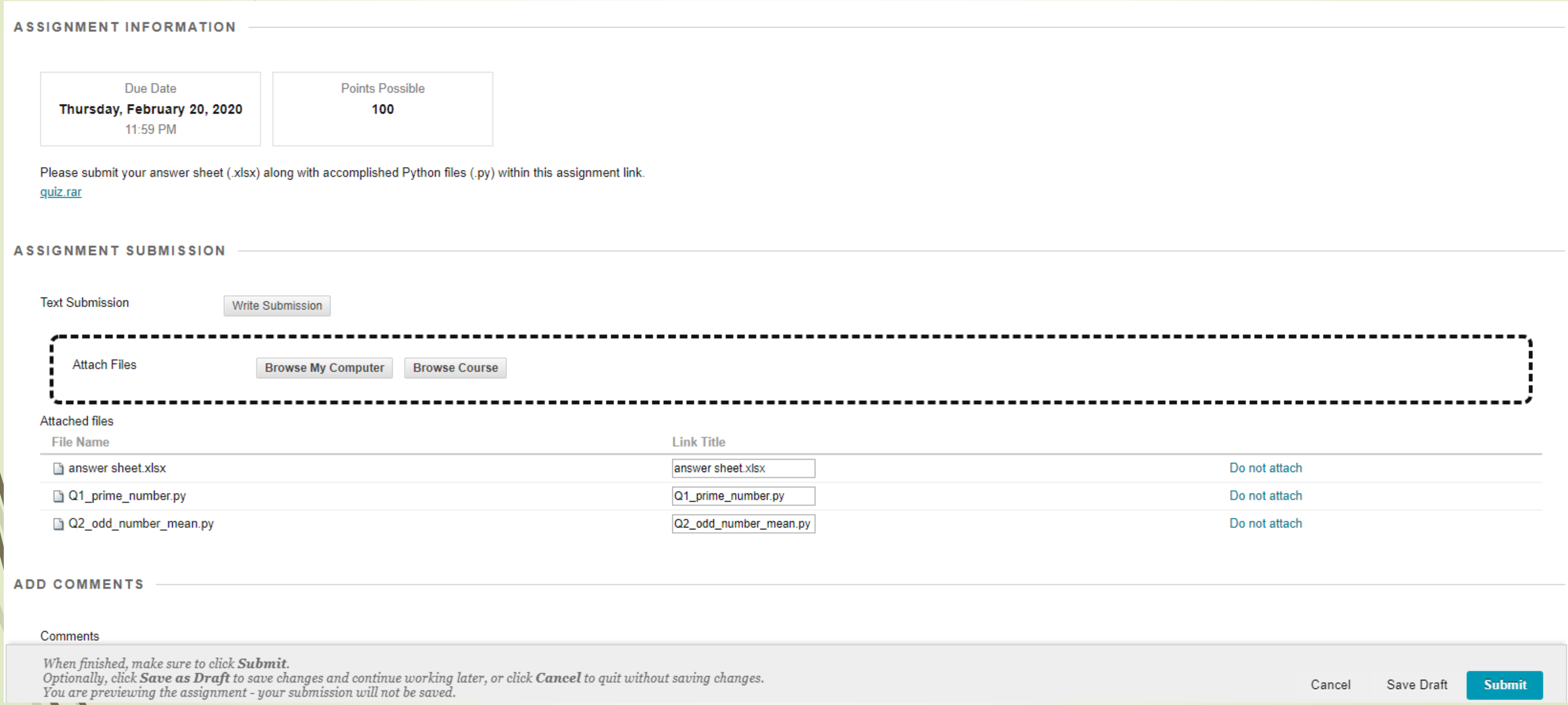
Outline

- Data Collection(web scraping)
- Data Storage (to be continued)
- Lab Quiz

Lab Quiz

- **Deadline**: 11:59 a.m., Jan. 21, 2020
- Two question accounting for 2% of overall score
- **Upload** the **answer worksheet** and the accomplished **Python files** to the **Blackboard**
- You may submit **unlimited times** but only the **LAST** submission will be considered
- Note： **MUST attach ALL** the required files in every submission/resubmission, otherwise other files will be missing.

Lab Quiz Submission



ASSIGNMENT INFORMATION

Due Date	Points Possible
Thursday, February 20, 2020 11:59 PM	**100**

Please submit your answer sheet (.xlsx) along with accomplished Python files (.py) within this assignment link.

quiz.rar

ASSIGNMENT SUBMISSION

Text Submission

Attach Files

Attached files

File Name	Link Title	
answer sheet.xlsx	answer sheet.xlsx	Do not attach
Q1_prime_number.py	Q1_prime_number.py	Do not attach
Q2_odd_number_mean.py	Q2_odd_number_mean.py	Do not attach

ADD COMMENTS

Comments

*When finished, make sure to click **Submit**.*
*Optionally, click **Save as Draft** to save changes and continue working later, or click **Cancel** to quit without saving changes.*
You are previewing the assignment - your submission will not be saved.