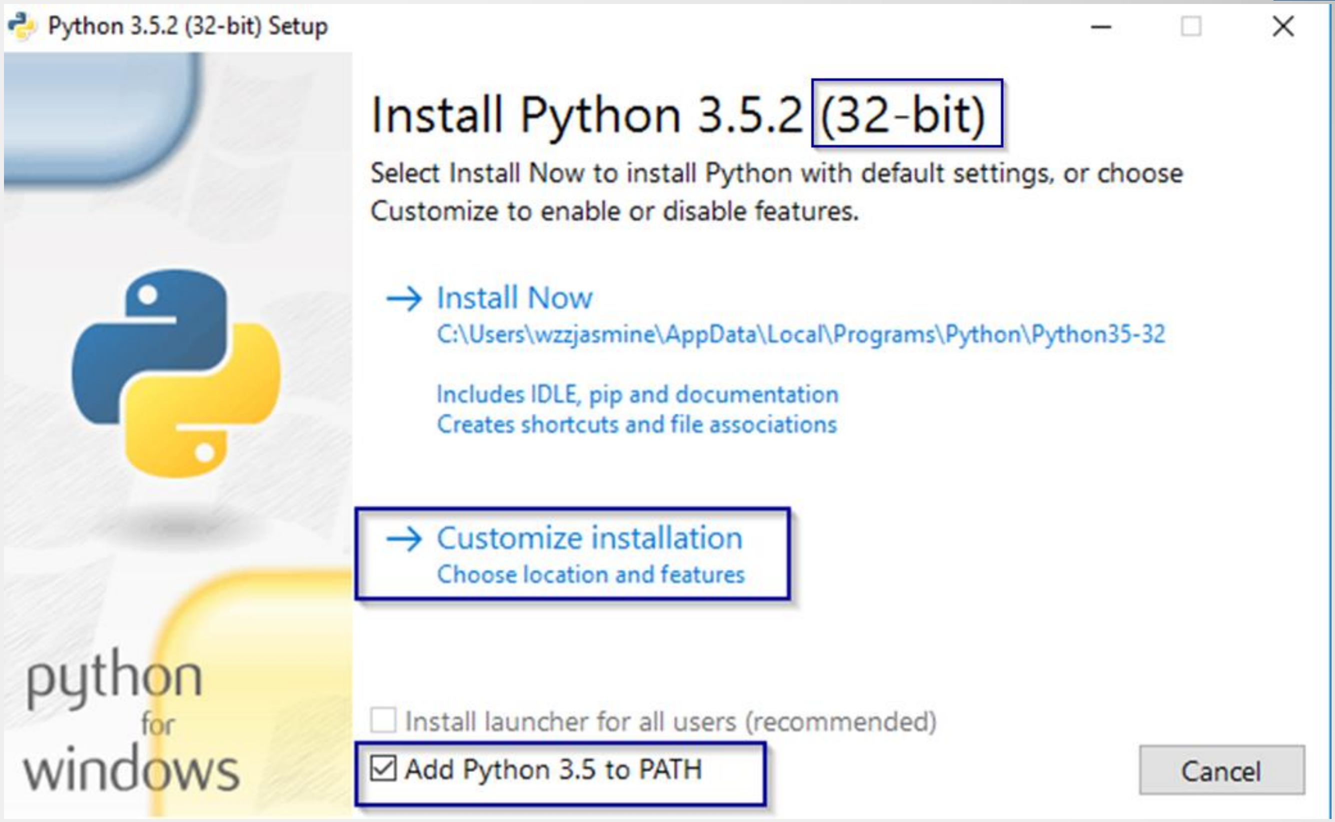
Sentiment Analysis

Wang Zhenzhen

Working Environment

- Download and install Python3.5.2



Python 3.5.2 (32-bit) Setup

Install Python 3.5.2 (32-bit)

Select Install Now to install Python with default settings, or choose Customize to enable or disable features.

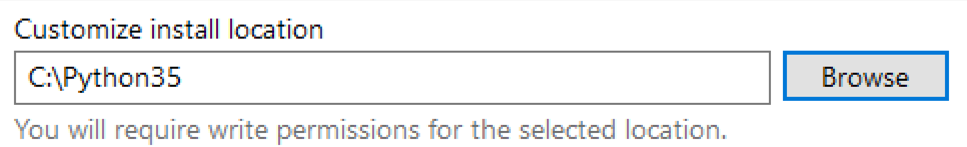
→ Install Now
C:\Users\wzzjasmine\AppData\Local\Programs\Python\Python35-32

Includes IDLE, pip and documentation
Creates shortcuts and file associations

→ Customize installation
Choose location and features

□ Install launcher for all users (recommended)

☑ Add Python 3.5 to PATH



Customize install location

C:\Python35

You will require write permissions for the selected location.

Package to install

- Textblob
  - pip install textblob
  - python -m textblob.download_corpora
- Snownlp
  - pip install snownlp
- Jieba
  - pip install jieba
- Beautifulsoup
  - pip install beautifulsoup4
- PyNotebook
  - pip install jupyter

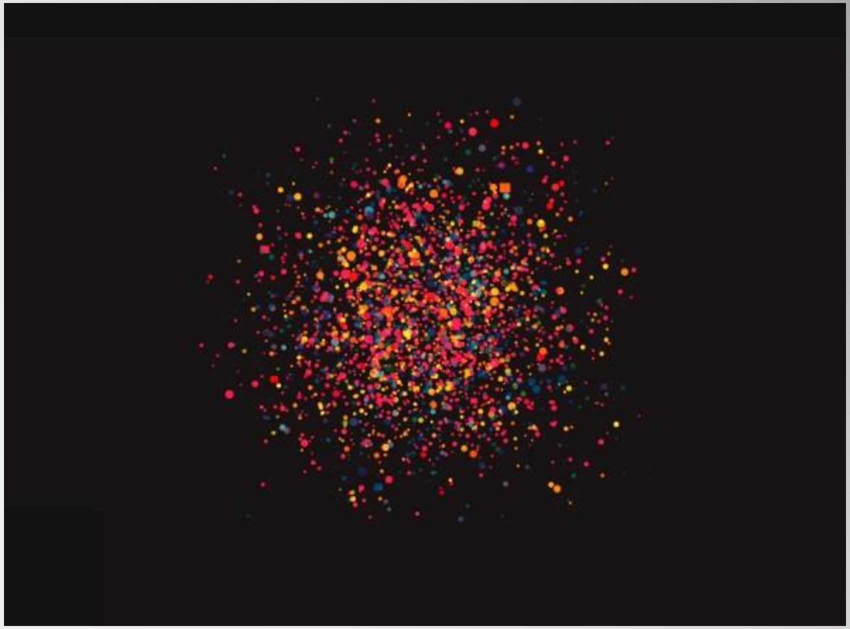
Introduction

- “Sentimental analysis, also called opinion mining, is the field of study that analyzes people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes.” (Liu, 2012)
  - – opinion mining
  - – opinion extraction
  - – sentiment mining
  - – subjectivity analysis
  - – affect analysis
  - – emotion analysis
  - – review mining
  - – ...

Why Sentimental Analysis

- We want to know others' opinion: surveys, opinion polls, and focus groups
- Explosive growth of social media: reviews, forum discussions, blogs, micro-blogs, Twitter

- We feel fine(video)



Problem Definition

Posted by: John Smith Date: September 10, 2011

(1) I bought a Canon G12 camera six months ago. (2) I simply love it. (3) The picture quality is amazing. (4) The battery life is also long. (5) However, my wife thinks it is too heavy for her.

Problem Definition

Posted by: John Smith Date: September 10, 2011

(1) I bought a Canon G12 camera six months ago. (2) I simply love it. (3) The picture quality is amazing. (4) The battery life is also long. (5) However, my wife thinks it is too heavy for her.

- Definition (Opinion): An opinion is a quintuple, (g, a, s, h, t), where g is the opinion (or sentiment) target, a is attribute or component of the target, s is the sentiment about the target, h is the opinion holder and t is the time when the opinion wasexpressed.

Problem Definition

- Objective of sentiment analysis: Given an opinion document d, discover all opinion quintuples (g, a, s, h, t) in d

- Task 1: Extract all entities/aspects/holders in D, and categorize or group synonymous expressions into entity clusters (or categories)
- Task 2: Extract the times when opinions are given and standardize different time formats
- Task 3: Determine whether an opinion on an aspect is positive, negative or neutral, or assign a numeric sentiment rating to the aspect.
- Task 4: Produce all opinion quintuples (g, a, s, h, t) expressed in document d based on the results of the above tasks

Example (1)

Posted by: bigJohn Date: Sept. 15, 2011

(1) I bought a Samsung camera and my friends brought a Canon camera yesterday. (2) In the past week, we both used the cameras a lot. (3) The photos from my Samy are not that great, and the battery life is short too. (4) My friend was very happy with his camera and loves its picture quality. (5) I want a camera that can take good photos. (6) I am going to return it tomorrow.

Example (2)

- Four opinion quintuples
  - (Samsung, picture_quality, negative, bigJohn, Sept-15-2011)
  - (Samsung, battery_life, negative, bigJohn, Sept-15-2011)
  - (Canon, GENERAL, positive, bigJohn's_friend, Sept-15-2011)
  - (Canon, picture_quality, positive, bigJohn's_friend, Sept-15-2011)

Problem Definition

- Objective of sentiment analysis: Given an opinion document d, discover all opinion quintuples (g, a, s, h, t) in d

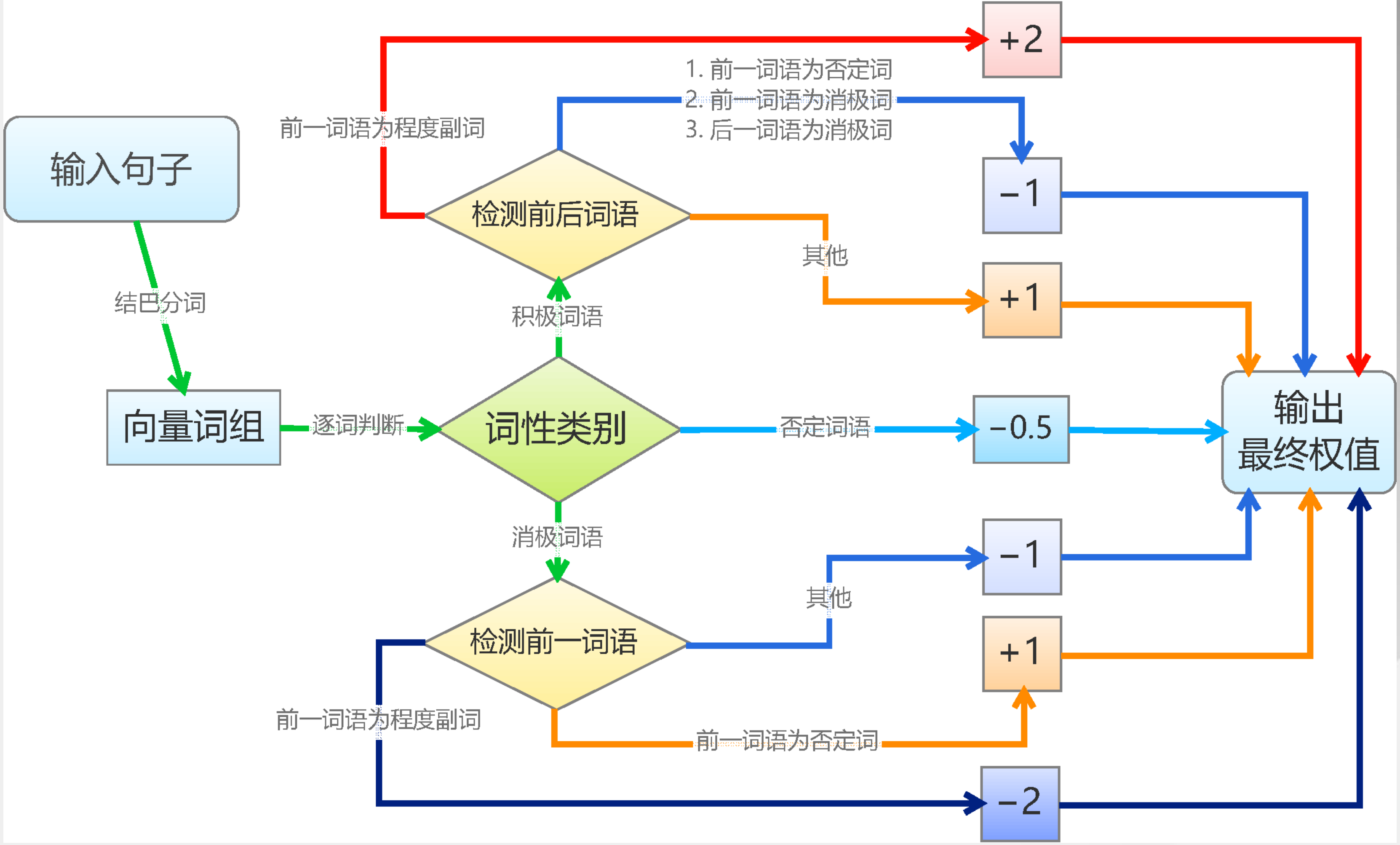
- Task 1: Extract all entities/aspects/holders in d, and categorize or group synonymous expressions into entity clusters (or categories)
- Task 2: Extract the times when opinions are given and standardize different time formats
- Task 3: Determine whether an opinion on an aspect is positive, negative or neutral, or assign a numeric sentiment rating to the aspect.
- Task 4: Produce all opinion quintuples (g, a, s, h, t) expressed in document d based on the results of the above tasks

MINE SENTIMENT

Unsupervised Learning

- Dictionary based approach
  - Sentiment words dictionary
  - Amplifier words dictionary
  - Shifter words dictionary

Dictionary Based Sentiment Analysis



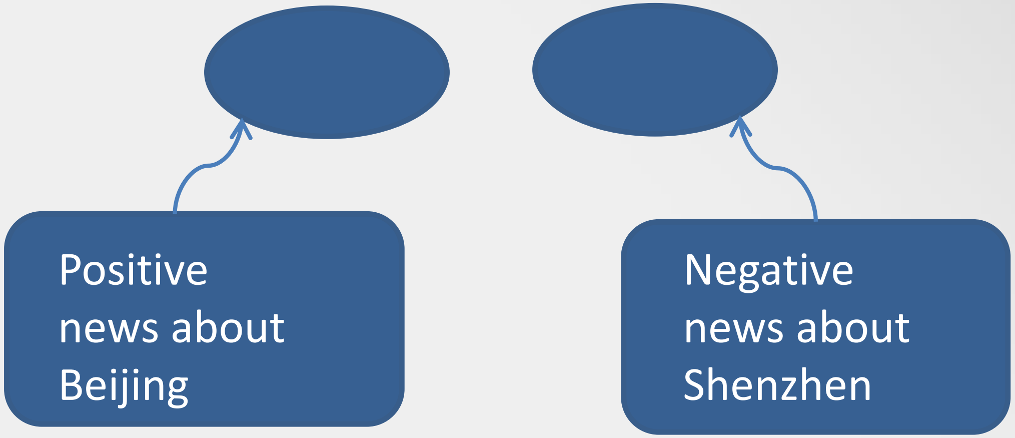
(Source: http://spaces.ac.cn/archives/3360/)

Mine Sentiment Dictionary

- Feature selection
- Word association

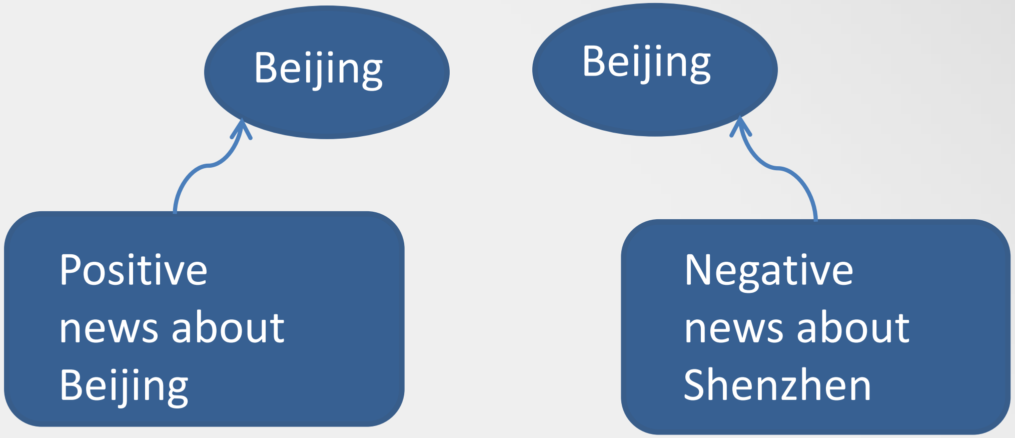
Feature Selection: Starter Edition

- Finding terms that best represent texts in each category.
- Starter edition: pick up the most popular term in each category.
- Problem: What if the most popular terms are the same across categories?



Feature Selection: Starter Edition

- Finding terms that best represent texts in each category.
- Starter edition: pick up the most popular term in each category.
- Problem: What if the most popular terms are the same across categories?



Feature Selection: Advanced Edition

- Picking up the most discriminant terms in each category.

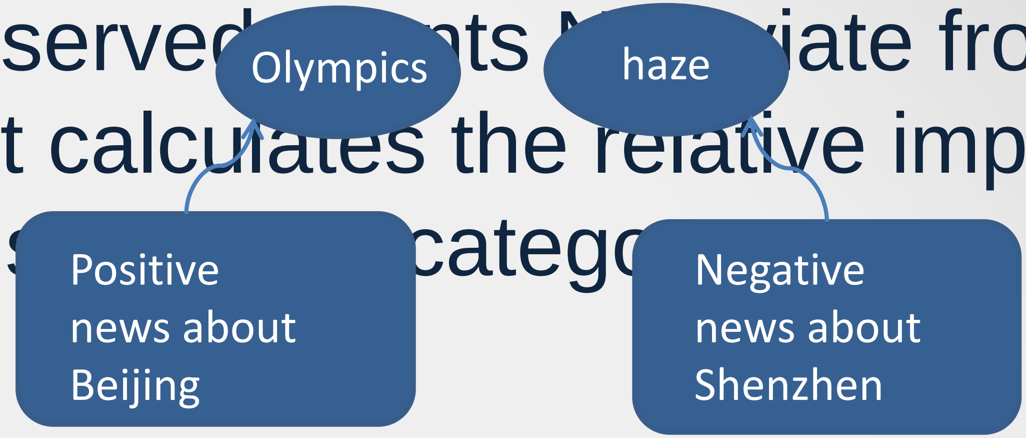
- χ^2 calculates whether the occurrence of the term and occurrence of the category are independent.

Feature Selection: Using χ^2 to Realize the Advanced Edition

- $\chi^2(t,c) = \sum_{e_t \epsilon \{0,1\}} \sum_{e_c \epsilon \{0,1\}} \frac{(N_{e_t e_c} - E_{e_t e_c})^2}{E_{e_t e_c}}$
  - Et=1 (the document contains term t)
  - Et=0 (the document does not contain term t)
  - Ec=1 (the document is in class c)
  - Et=0 (the document is not in class c)
- χ^2measures how much expected counts E and observed counts N deviate from each other. It calculates the relative importance of terms for each category.

Feature Selection: Using χ^2 to Realize the Advanced Edition

- $\chi^2(t,c) = \sum_{e_t \epsilon \{0,1\}} \sum_{e_c \epsilon \{0,1\}} \frac{(N_{e_t e_c} - E_{e_t e_c})^2}{E_{e_t e_c}}$
  - Et=1 (the document contains term t)
  - Et=0 (the document does not contain term t)
  - Ec=1 (the document is in class c)
  - Et=0 (the document is not in class c)
- χ^2measures how much expected counts E and observed nts iate from each other. It calcu es the relative importance of terms catego



Olympics

haze

Positive news about Beijing

Negative news about Shenzhen

Exercise

- $\chi^2(t,c) = \sum_{e_t \epsilon \{0,1\}} \sum_{e_c \epsilon \{0,1\}} \frac{(N_{e_t e_c} - E_{e_t e_c})^2}{E_{e_t e_c}}$
  - Et=1 (the document contains term t)
  - Et=0 (the document does not contain term t)
  - Ec=1 (the document is in class c)
  - Et=0 (the document is not in class c)
- A corpus of 801,948 news articles
- 27,701 articles are classified as “positive”. Of them, 49 articles contain the word “Olympics”.
- Of the rest 774,247 articles, 141 contain the word “Olympics”.

Exercise

	$e_{positive}$=1	$e_{positive}$=0
$e_{olympics}$=1	N_{11}=49 E_{11}=6.6	N_{10}=141 E_{10}=183.4
$e_{olympics}$=0	N_{01}=27652 E_{01}=27694.4	N_{00}=774106 E_{00}=774063.6

$$\chi^2(t,c) = \sum_{e_t \epsilon \{0,1\}} \sum_{e_c \epsilon \{0,1\}} \frac{(N_{e_t e_c} - E_{e_t e_c})^2}{E_{e_t e_c}} = 284$$

Word Association

- Select benchmark words
- Calculate PMI of the target word and benchmark words
- Estimate sentiment orientation of the target word based on above calculation

Word Association

- PMI: pointwise mutual information

$$PMI(term_1, term_2) = \log_2 \left(\frac{\Pr(term_1 \wedge term_2)}{\Pr(term_1)\Pr(term_2)} \right)$$

  - Pr(term1^term2) is the actual co-occurrence probability of term1 and term2
  - Pr(term1)Pr(term2) is the co-occurrence probability of the two terms if they are statistically independent.

- PMI measures the degree of statistical dependence between two terms.

Word Association

- Estimate sentiment orientation

$$SO(phrase) = \log_2\left(\frac{hits(\text{phrase } NEAR \text{ "excellent"})hits(\text{"poor"})}{hits(\text{phrase } NEAR \text{ "poor"})hits(\text{"excellent"})}\right)$$

Exercise

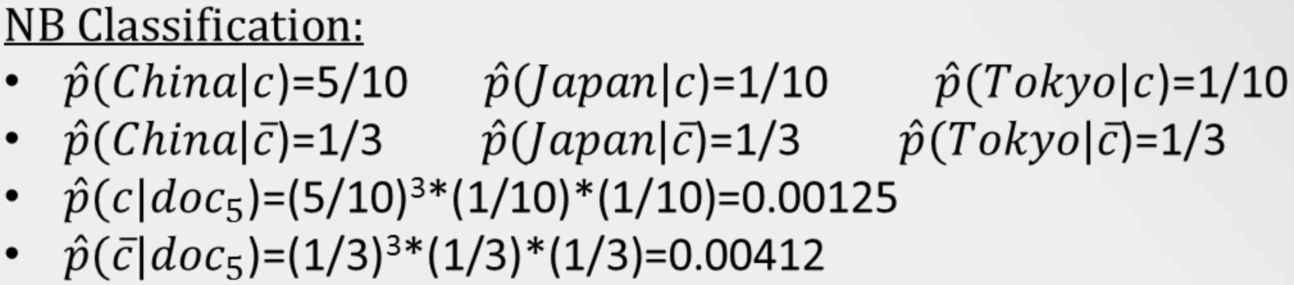
- Use Google/Baidu to mine sentiment orientation of ‘Olympics’

Supervised Learning

- Naive Bayes (NB) Classification
  - Retrieving a training set (usually manually), with texts already assigned to a known list of categories
  - Based on the training set, determining the contribution of each term for each category
  - Based on the terms in texts, assigning new texts into categories.

NB Classification: Example

Doc Type	Doc ID	Terms in doc	in *c = China*?
Training Set	1	China Beijing China Tokyo	yes
	2	China China Shanghai	yes
	3	China Macao Japan	yes
	4	Tokyo Japan China	no
Testing Set	5	China China China Tokyo Japan	?



NB Classification:

- $\hat{p}(China|c)$=5/10 $\hat{p}(Japan|c)$=1/10 $\hat{p}(Tokyo|c)$=1/10
- $\hat{p}(China|\bar{c})$=1/3 $\hat{p}(Japan|\bar{c})$=1/3 $\hat{p}(Tokyo|\bar{c})$=1/3
- $\hat{p}(c|doc_5)$=(5/10)^3*(1/10)*(1/10)=0.00125
- $\hat{p}(\bar{c}|doc_5)$=(1/3)^3*(1/3)*(1/3)=0.00412

(Revised from Manning, Raghavan, & Schütze, 2008)

Use NB Classification to do Sentiment Analysis (1)

Doc Type	Doc ID	Terms in doc	Classification
Training Set	1	I love this car	Positive
	2	This view is amazing	Positive
	3	I feel great this morning	Positive
	4	I am so excited about the concert	Positive
	5	He is my best friend	Positive
	6	I do not like this car	Negative
	7	This view is horrible	Negative
	8	I feel tired this morning	Negative
	9	I am not looking forward to the concert	Negative
	10	He is my enemy	Negative

(Revised Wang Chengjun’s blog)

Use NB Classification to do Sentiment Analysis (2)

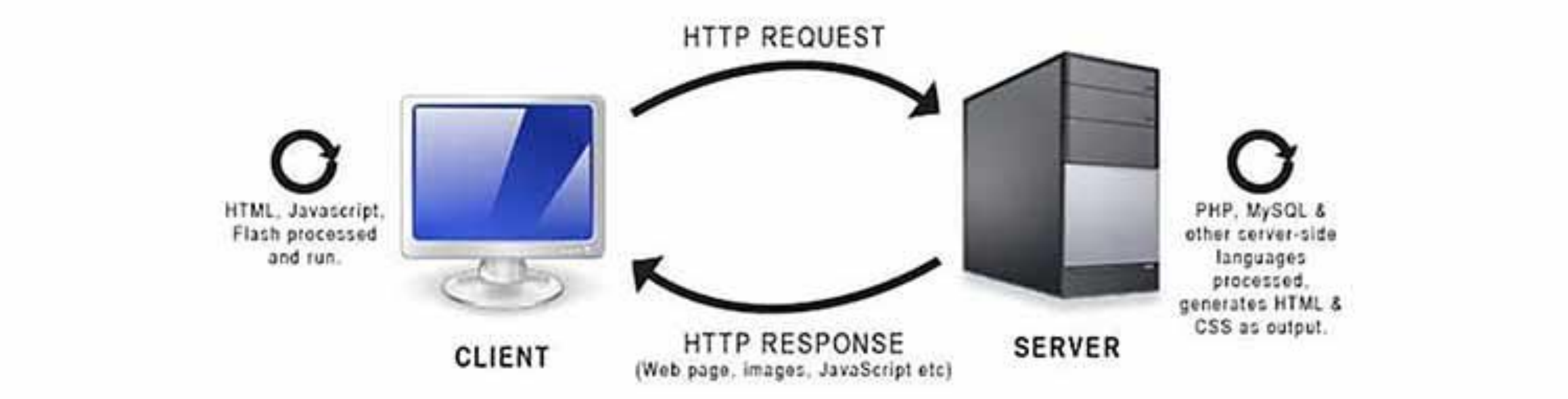
Doc Type	Doc ID	Terms in doc	NB Classification	Human coding
Testing Set	1	feel happy this morning	N	P
	2	Oh I love my friend	P	P
	3	not like that man	N	N
	4	house not great	N	N
	5	your song annoying	N	N

Accuracy: 80%

(Revised Wang Chengjun's blog)

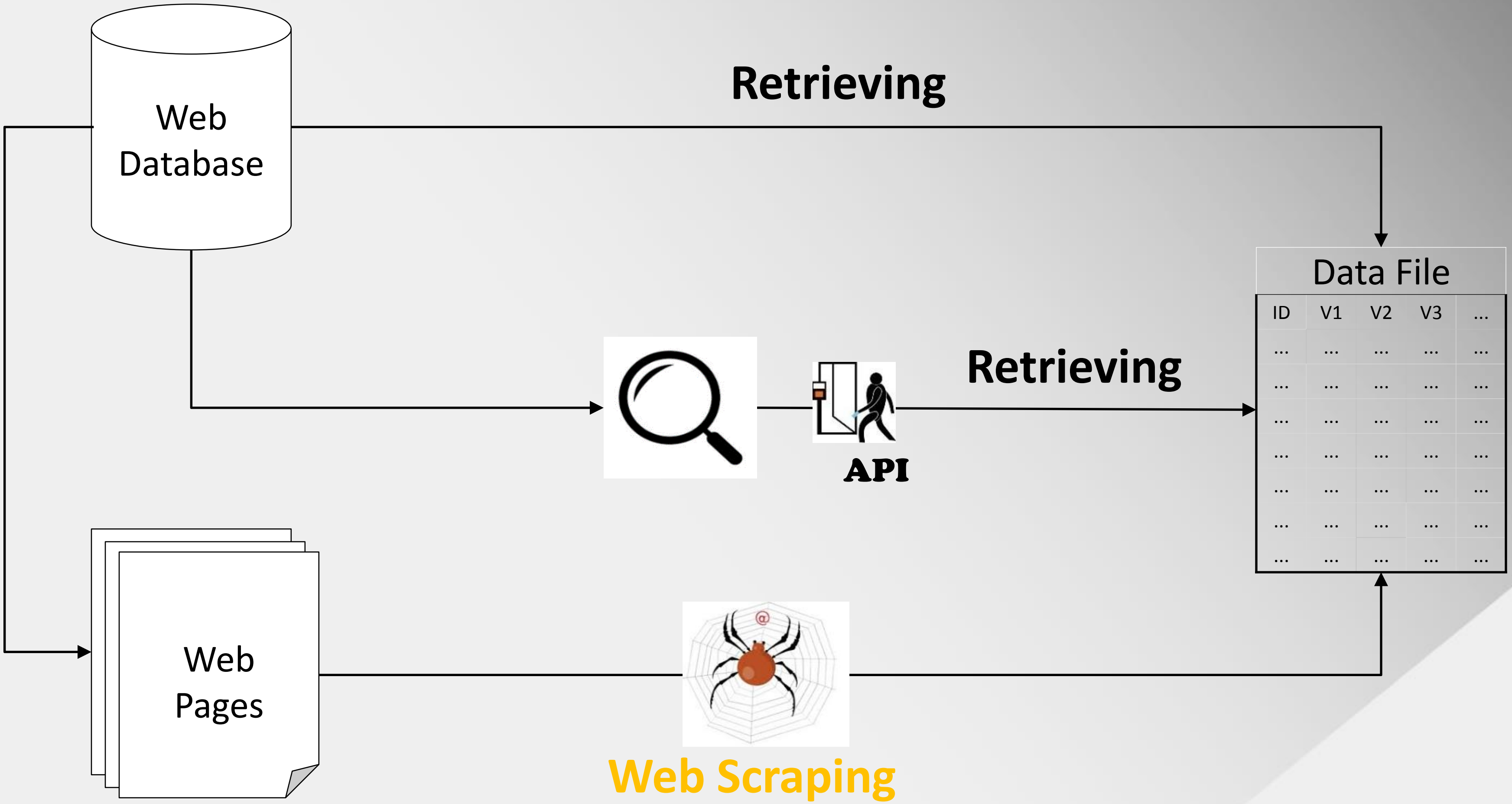
RETRIEVE SENTIMENT TEXT

Access the Internet: The Human Way


HTTP REQUEST
HTML, Javascript, Flash processed and run.
PHP, MySQL & other server-side languages processed, generates HTML & CSS as output.
CLIENT
HTTP RESPONSE
(Web page, images, JavaScript etc)
SERVER

(The client-server model)

Access Internet: Alternatives



Web Database

Retrieving

Retrieving

API

Web Pages

Web Scraping

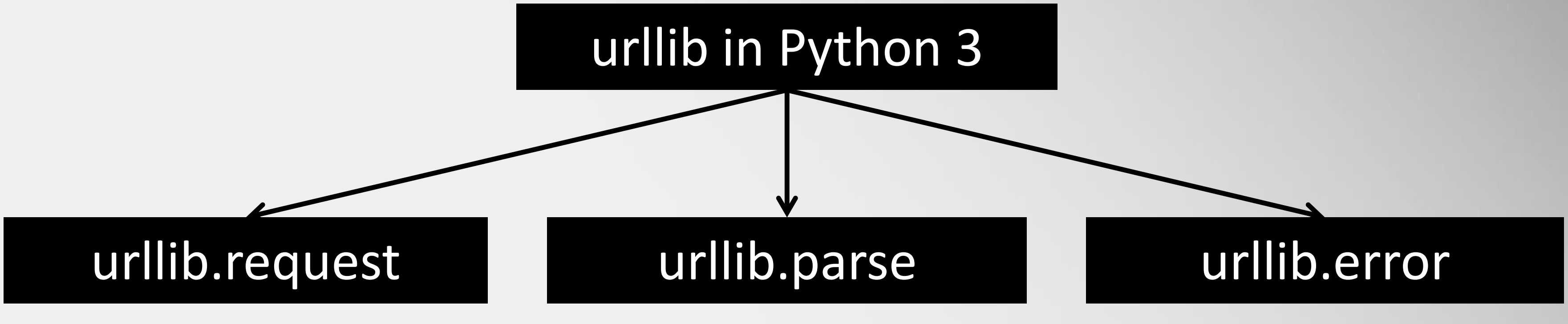
Data File

ID	V1	V2	V3	…
…	…	…	…	…
…	…	…	…	…
…	…	…	…	…
…	…	…	…	…
…	…	…	…	…
…	…	…	…	…
…	…	…	…	…

What does Web Scraper do?

- Retrieving HTML data from a domain name
- Parsing that data for target information
- Storing the target information
- Optionally, moving to another page to repeat the process

Your First Web Scrapper



urllib in Python 3

urllib.request

urllib.parse

urllib.error

http://pythonscraping.com/pages/page1.html

Introduction to HTML (1)

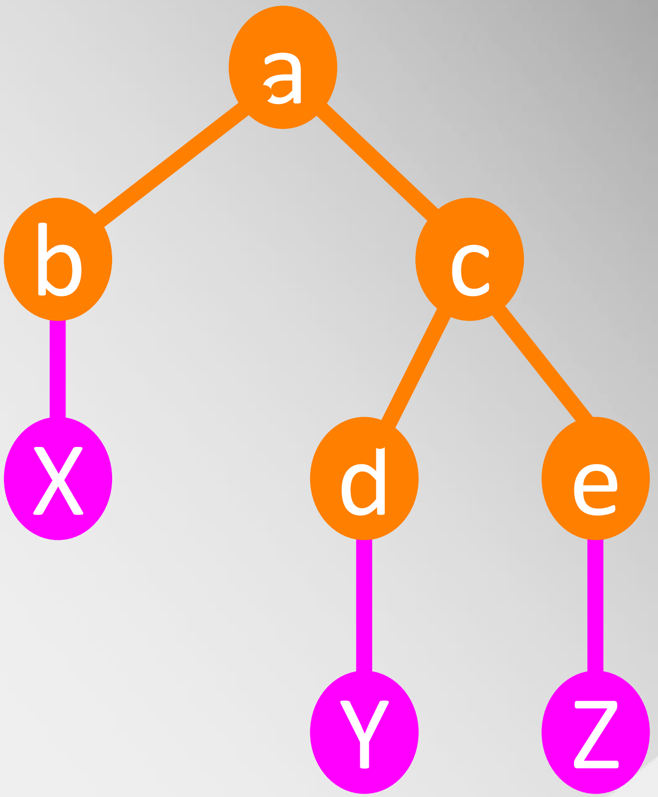
- HTML is a markup language for describing web documents (web pages).
  - HTML stands for Hyper Text Markup Language
  - A markup language is a set of markup tags
  - HTML documents are described by HTML tags
  - Each HTML tag describes different document content

(http://www.w3schools.com/html/html_intro.asp)

Introduction to HTML (2)

- HTML as Tree

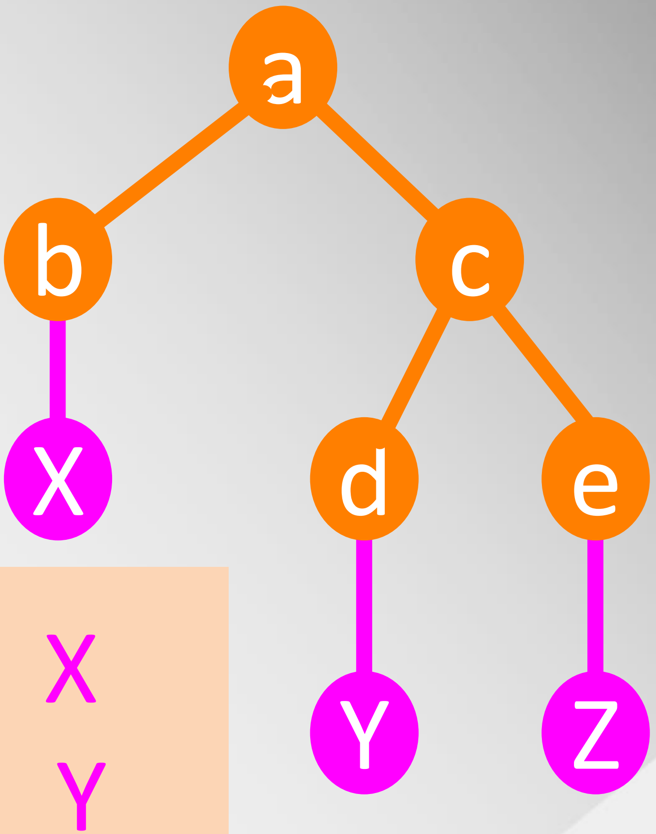
```
<a>
 <b>X</b>
  <c>
    <d>Y</d>
    <e>Z</e>
  </c>
</a>
```



Introduction to HTML (3)

- HTML as Path

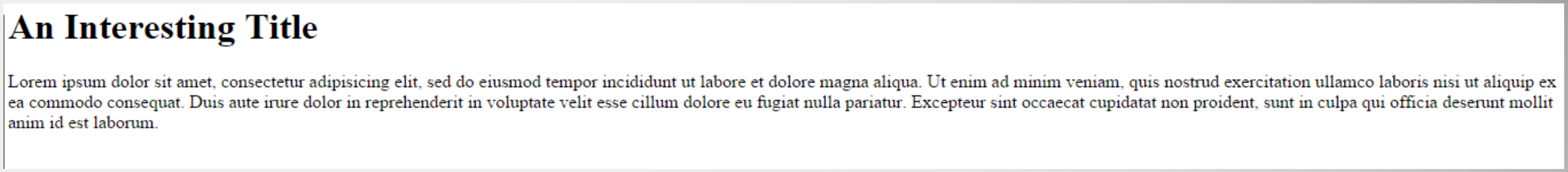
```
<a>
  <b>X</b>
   <c>
       <d>Y</d>
       <e>Z</e>
   </c>
</a>
```



/a/b	X
/a/c/d	Y
/a/c/e	Z

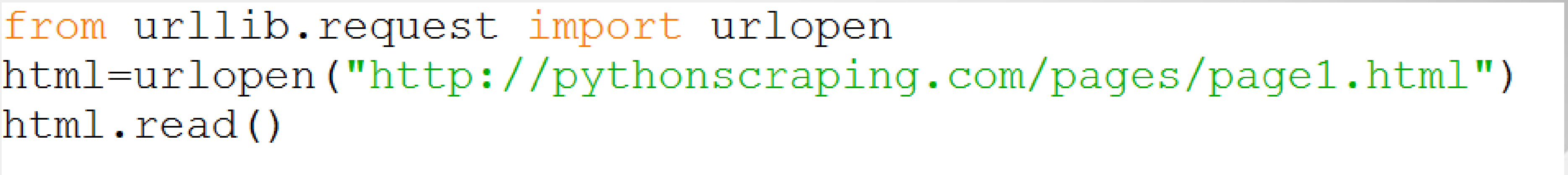
Your First Web Scrapper (1)

http://pythonscraping.com/pages/page1.html



An Interesting Title

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.



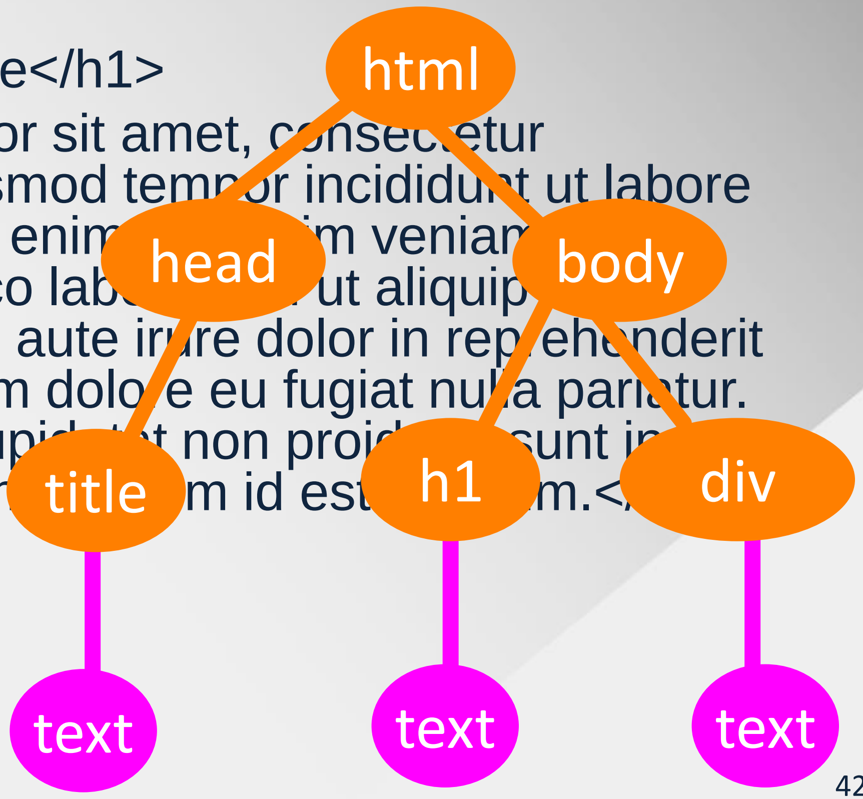
```
from urllib.request import urlopen
html=urlopen("http://pythonscraping.com/pages/page1.html")
html.read()
```

Your First Web Scrapper (2)

```
<html>
        <head>
                <title>A Useful Page</title>
        </head>
        <body>
                <h1>An Interesting Title</h1>
                <div>Lorem ipsum dolor sit amet, consectetur
            adipisicing elit, sed do eiusmod tempor incididunt ut labore
            et dolore magna aliqua. Ut enim ad minim veniam, quis
            nostrud exercitation ullamco laboris nisi ut aliquip ex ea
            commodo consequat. Duis aute irure dolor in reprehenderit
            in voluptate velit esse cillum dolore eu fugiat nulla pariatur.
            Excepteur sint occaecat cupidatat non proident, sunt in
            culpa qui officia deserunt mollit anim id est laborum.</div>
        </body>
</html>
```

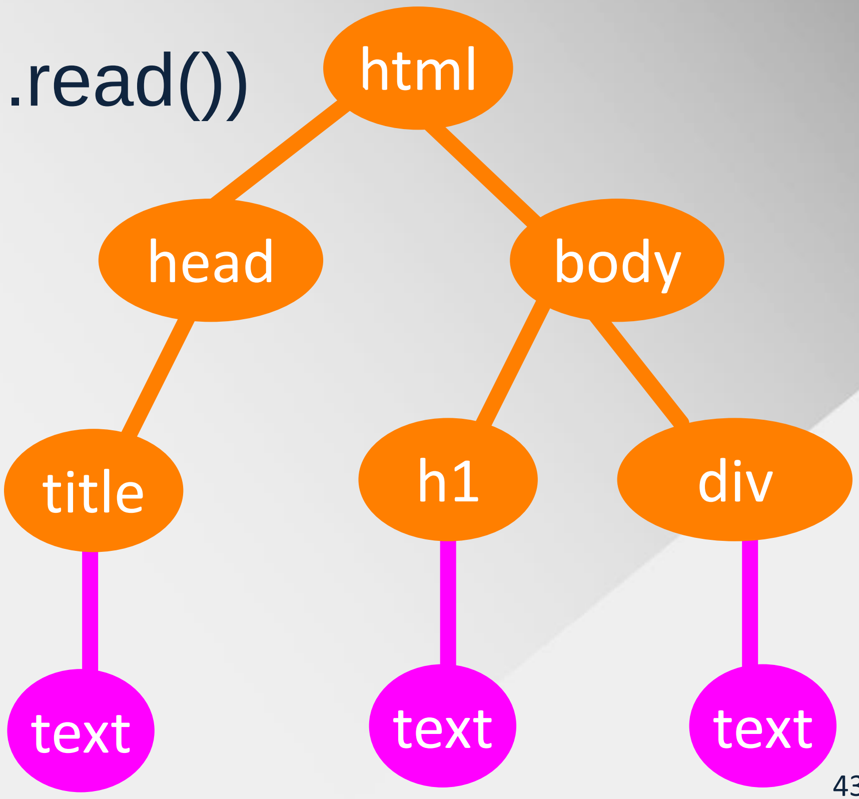
Your First Web Scrapper (2)

```
<html>
        <head>
                <title>A Useful Page</title>
        </head>
        <body>
                <h1>An Interesting Title</h1>
                <div>Lorem ipsum dolor sit amet, consectetur
    adipisicing elit, sed do eiusmod tempor incididunt ut labore
    et dolore magna aliqua. Ut enim ad minim veniam, quis
    nostrud exercitation ullamco laboris nisi ut aliquip ex ea
    commodo consequat. Duis aute irure dolor in reprehenderit
    in voluptate velit esse cillum dolore eu fugiat nulla pariatur.
    Excepteur sint occaecat cupidatat non proident, sunt in
    culpa qui officia deserunt mollit anim id est laborum.</div>
        </body>
</html>
```



Your First Web Scrapper (3)

- from bs4 import BeautifulSoup
- html=urlopen("http://pythonscraping.com/pages/page1.html")
- bsobj=BeautifulSoup(html.read())
- print (bsobj.html.body.h1)
- print (bsobj.body.h1)
- print (bsobj.html.h1)



Save Webpages (1)

Write to a file

```
fileName="D://example.txt"
p=open(fileName,"w") #open for writing, truncating the file first
print("hello",file=p)
print("world",file=p)
p.close()

p=open(fileName,"a") #open for writing, appending to the end of the file if it exists
print("hello world again",file=p)
p.close()
```

Save Webpages (2)

Crawl a webpage and write it to a file

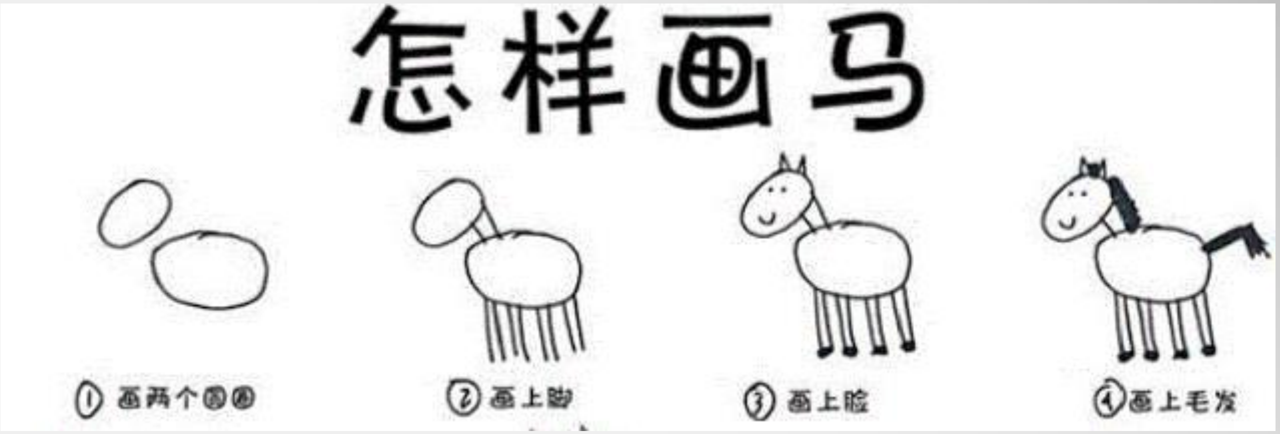
```
html=urlopen("http://www.pythonscraping.com/pages/page1.html")
fileName="D://page1.txt"
content=html.read().decode("utf-8")
p=open(fileName,"w")
print(content,file=p)
p.close()
```

Coding rule

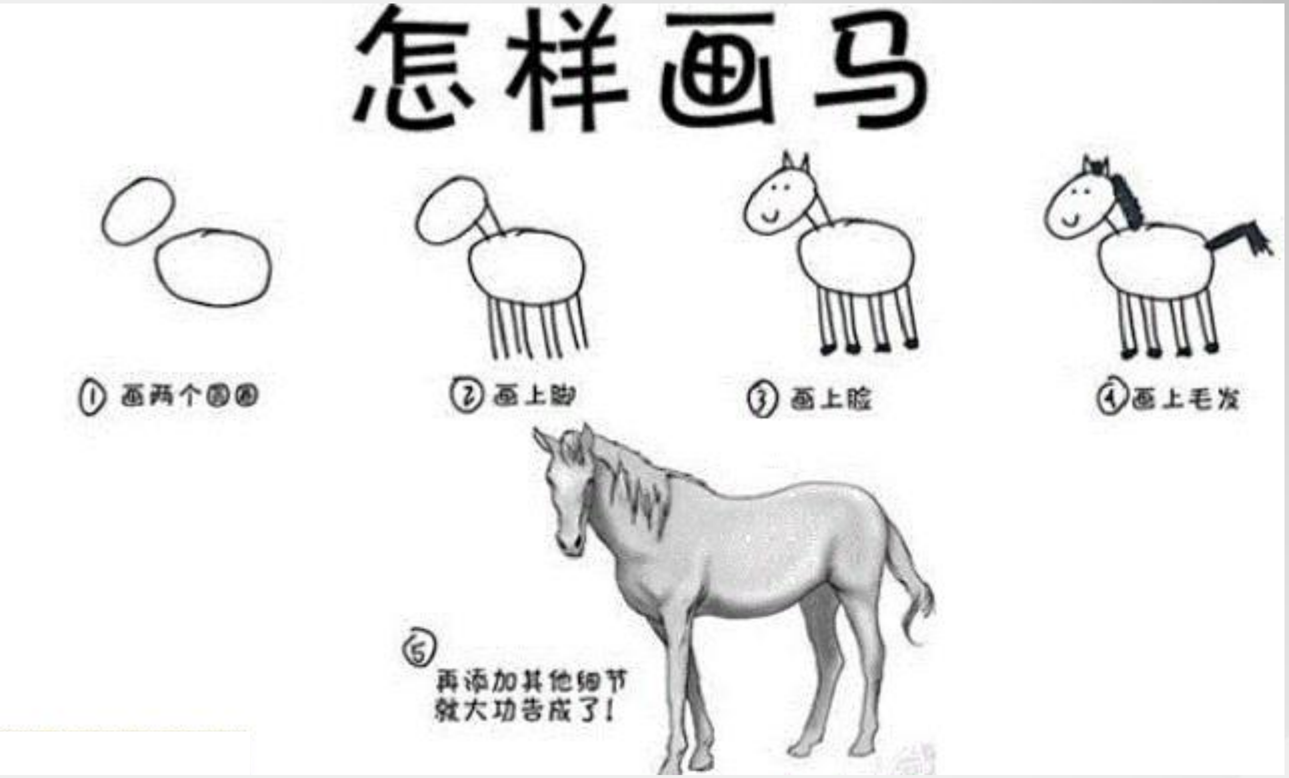
Details about encode/decode

LEARN BY DOING

Drawing Horse



Drawing Horse



Comparative Opinion

- Coke tastes better than Pepsi.
- Coke tastes the best.

Implicit Opinion

- I bought the mattress a week ago, and a valley has formed.
- The battery life of Nokia phones is longer than Samsung phones.

Standing Point

- The housing price has gone down, which is bad for the economy.

Sarcastic Sentences

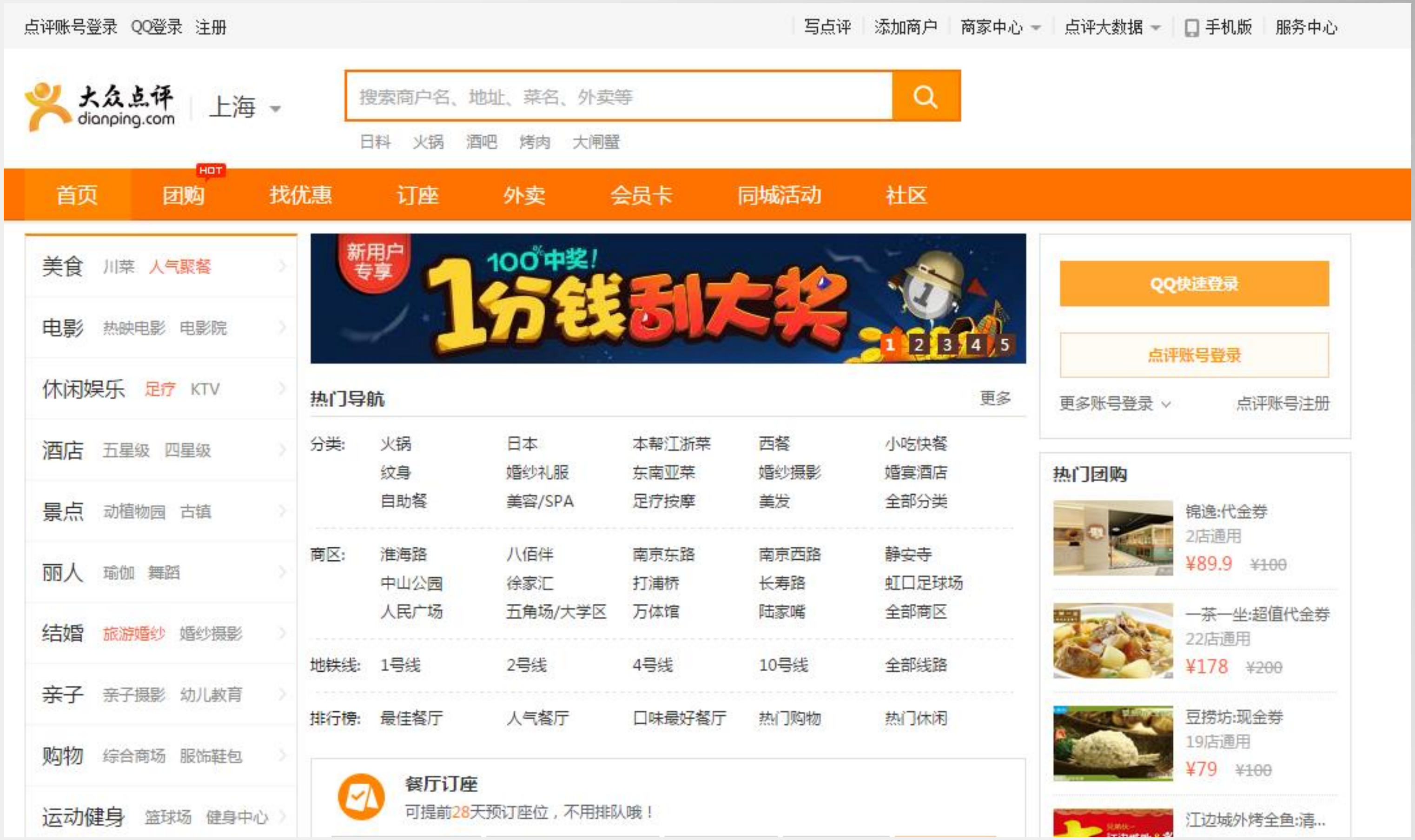
- What a great car! It stopped working in two days.

References(1)

- Liu, B. (2012).Sentiment analysis and opinion mining.

A REAL EXAMPLE: SCRAPING DIANPING

Crawl Dianping



点评账号登录 QQ登录 注册

写点评 | 添加商户 | 商家中心 | 点评大数据 | 手机版 | 服务中心

大众点评 dianping.com 上海

搜索商户名、地址、菜名、外卖等

日料 火锅 酒吧 烤肉 大闸蟹

首页 团购 HOT 找优惠 订座 外卖 会员卡 同城活动 社区

美食 川菜 人气聚餐

电影 热映电影 电影院

休闲娱乐 足疗 KTV

酒店 五星级 四星级

景点 动植物园 古镇

丽人 瑜伽 舞蹈

结婚 旅游婚纱 婚纱摄影

亲子 亲子摄影 幼儿教育

购物 综合商场 服饰鞋包

运动健身 篮球场 健身中心

热门导航 更多

分类:	火锅	日本	本帮江浙菜	西餐	小吃快餐
	纹身	婚纱礼服	东南亚菜	婚纱摄影	婚宴酒店
	自助餐	美容/SPA	足疗按摩	美发	全部分类
商区:	淮海路	八佰伴	南京东路	南京西路	静安寺
	中山公园	徐家汇	打浦桥	长寿路	虹口足球场
	人民广场	五角场/大学区	万体馆	陆家嘴	全部商区
地铁线:	1号线	2号线	4号线	10号线	全部线路
排行榜:	最佳餐厅	人气餐厅	口味最好餐厅	热门购物	热门休闲

餐厅订座

可提前28天预订座位，不用排队哦！

QQ快速登录

点评账号登录

更多账号登录 点评账号注册

热门团购

锦逸:代金券
2店通用
¥89.9 ~~¥100~~

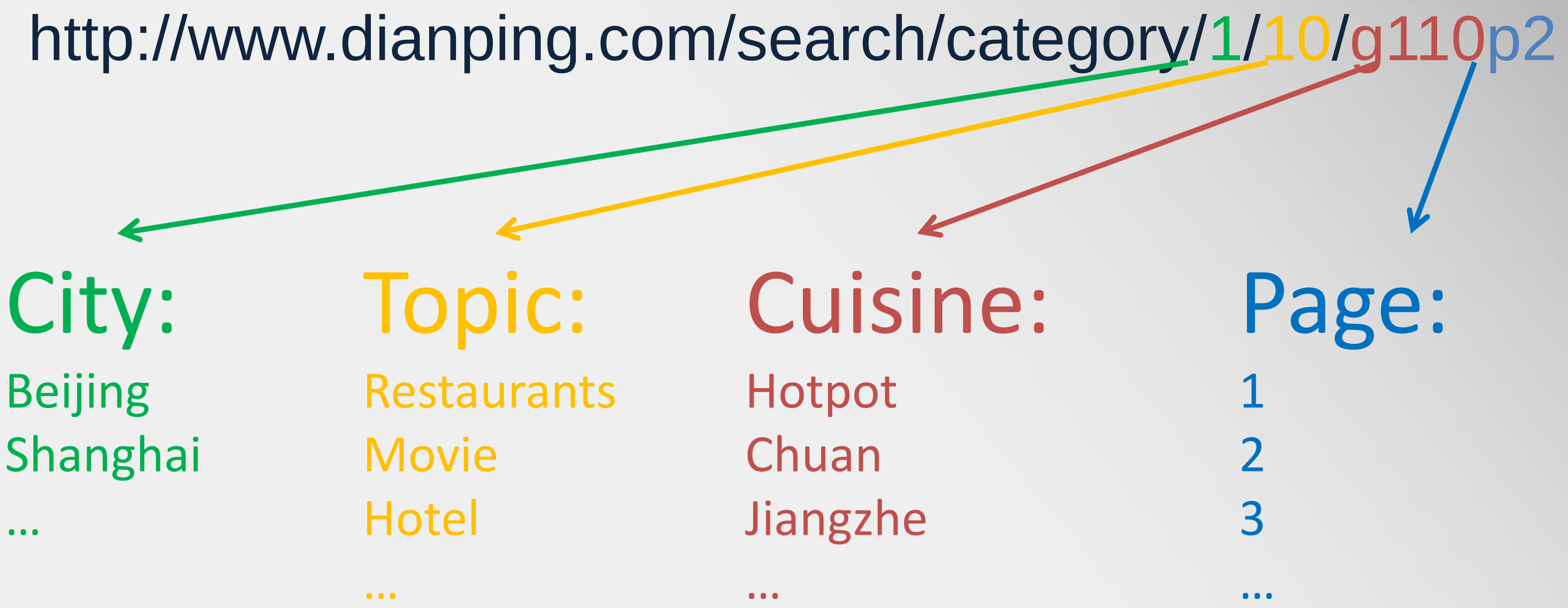
一茶一坐:超值代金券
22店通用
¥178 ~~¥200~~

豆捞坊:现金券
19店通用
¥79 ~~¥100~~

江边城外烤全鱼:清...

Crawl Dianping

上海>>火锅>>page2



http://www.dianping.com/search/category/1/10/g110p2

City:

Beijing

Shanghai

…

Topic:

Restaurants

Movie

Hotel

…

Cuisine:

Hotpot

Chuan

Jiangzhe

…

Page:

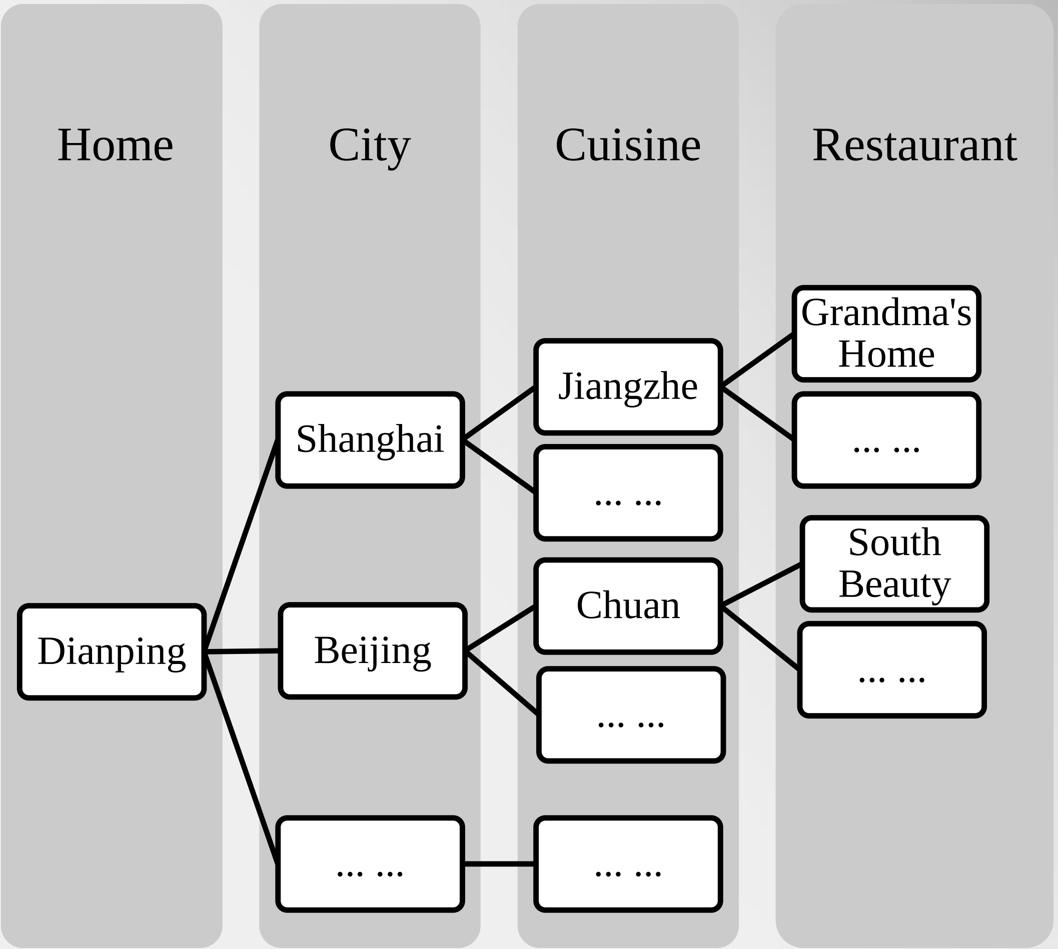
1

2

3

…

Crawl Dianping



Crawl Dianping

- Preparation:
  - – Generate a list of cities (or just one city)
  - – Generate a list of topics (or just one topic)
  - – Generate a list of cuisines (or just one cuisine)
- Crawl:
  - – Combine city, topic, cuisine, and generate the url of index page
  - – Extract shops url on the index page
  - – crawl shop webpage one by one

Sampling Problem

- 北京>>美食>>素菜
  - http://www.dianping.com/search/category/2/10/g109
  - 171 shops, 12 pages, 100% displayed
- 北京>>美食>>江浙菜
  - http://www.dianping.com/search/category/2/10/g101
  - ??? shops, 50 pages, ?% displayed
- 北京>>美食>>粤菜
  - http://www.dianping.com/search/category/2/10/g103
  - ??? shops, 50 pages, ?% displayed
- 北京>>美食>>川菜
  - http://www.dianping.com/search/category/2/10/g102
  - ??? shops, 50 pages, ?% displayed
- 北京>>美食>>北京菜
  - http://www.dianping.com/search/category/2/10/g311
  - ??? shops, 50 pages, ?% displayed

ID-based System

- Shop URL: http://www.dianping.com/shop/XXX

a unique numeric value—shop ID

automatically generated by the web database

Draw a Random Sample (1)

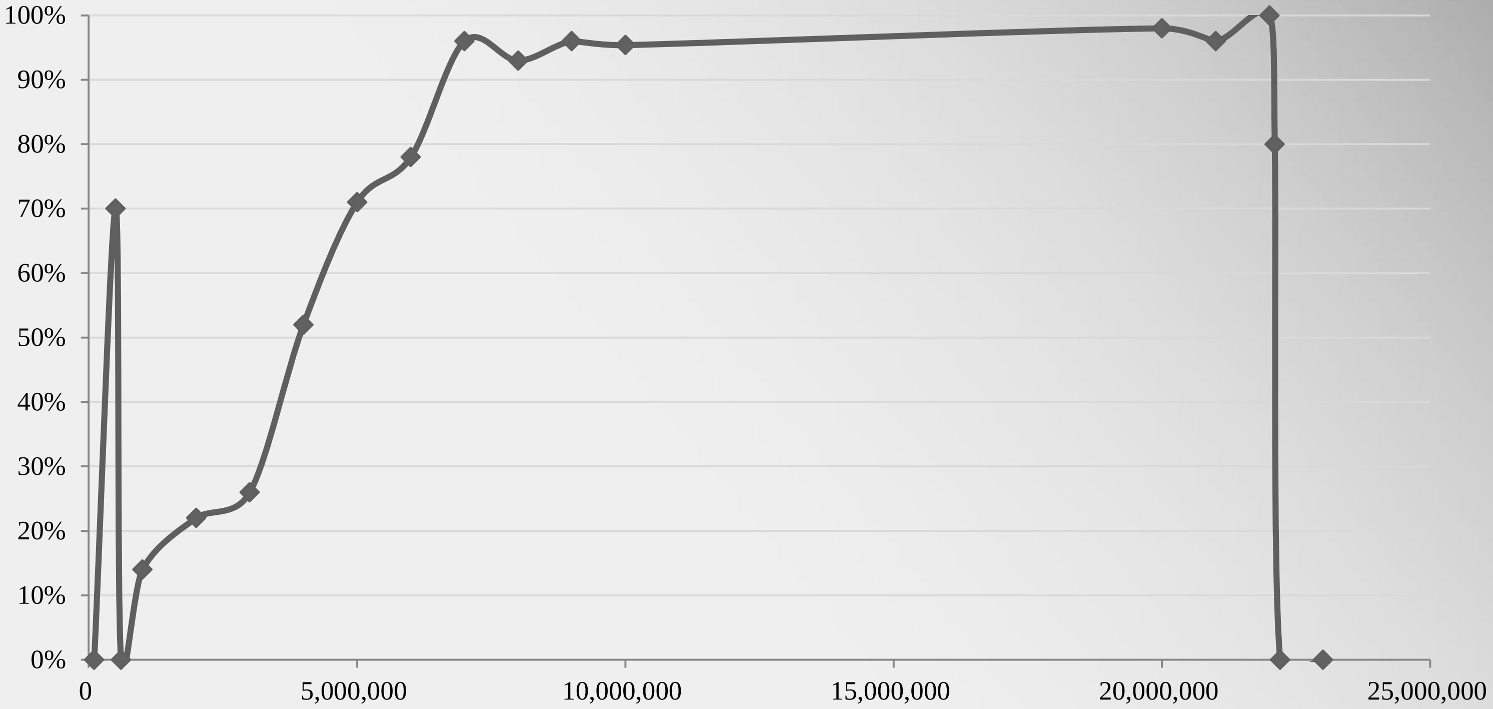
- 1, Detect ID ranges: the likely minimal and maximal number of digits of IDs

No. of digits	Population	Sample size	Sampling fraction	Hit
1	10	10	100%	NA
2	100	100	100%	NA
3	1,000	1,000	100%	NA
4	10,000	10,000	100%	NA
5	100,000	10,000	10%	NA
6	1,000,000	10,000	1%	776
7	10,000,000	10,000	0.1%	6,074
8	100,000,000	10,000	0.01%	1,273
9	1,000,000,000	10,000	0.001%	NA
10	10,000,000,000	10,000	0.0001%	NA

Draw a Random Sample (2)

- 2, Draw a small sample to calculate occupancy rate of the study population.
  - – Occupancy Rate (OR) measures the proportion of potential user IDs being valid bloggers.
  - – OR will be instrumental for projecting the population, designing sampling strategies for future studies, weighting resultant samples, and other practical purposes.

Draw a Random Sample (3)



Draw a Random Sample (4)

- 3, Draw a substantive sample based on the occupancy rate.

References(2)

Ryan, M. (2015) Web scraping with Python

Zhu, J. J. H., Mo, Q., Wang, F., & Lu, H. (2011). A random digit search (RDS) method for sampling of blogs and other user-generated content. Social Science Computer Review, 29 (3), 327-339.

THANKS!
QUESTIONS?