



2013 Taiwan

PYCON

Big Data Analysis in Python

Jimmy Lai

r97922028 [at] ntu.edu.tw

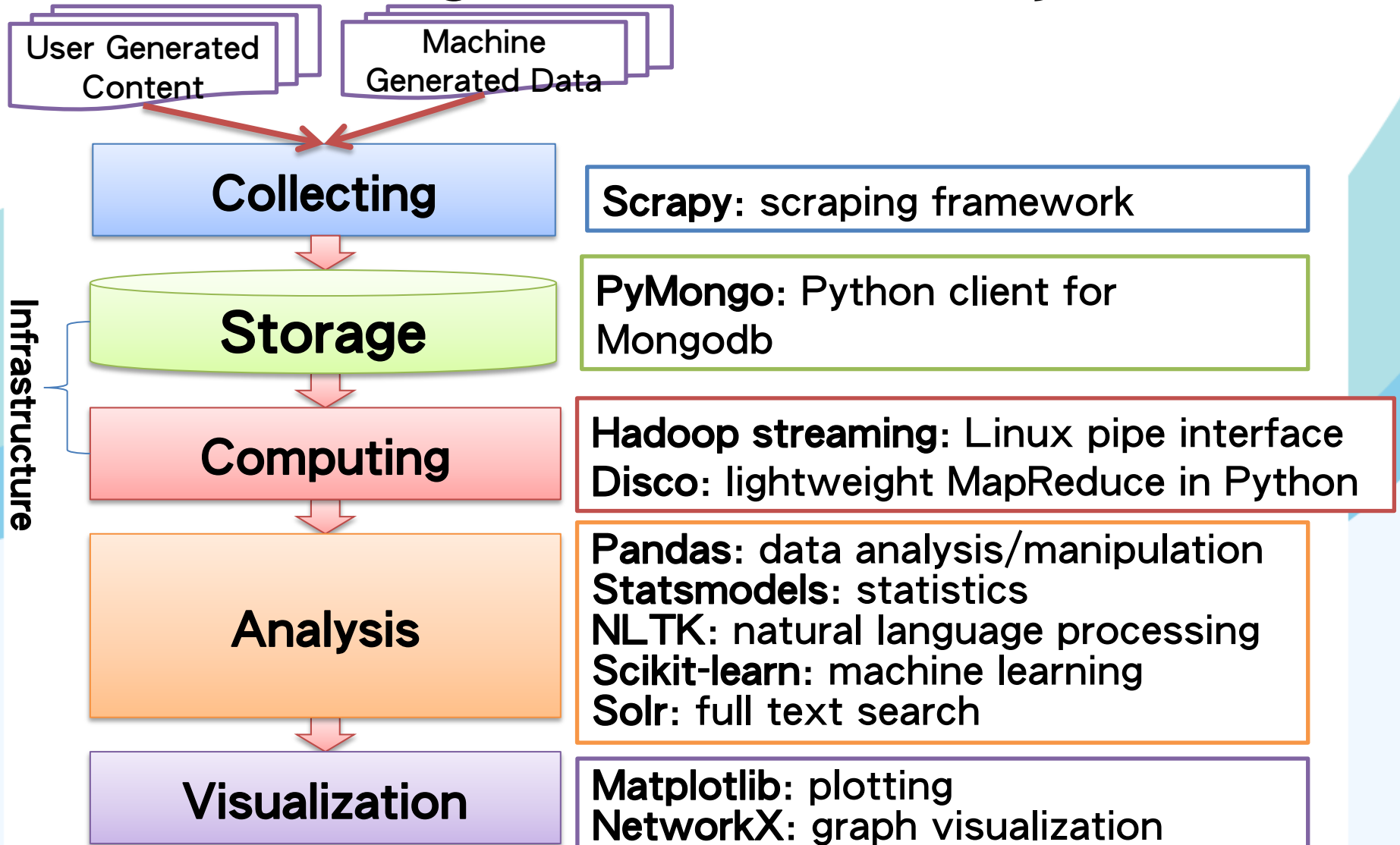
<http://tw.linkedin.com/pub/jimmy-lai/27/4a/536>

2013/05/26

Outlines

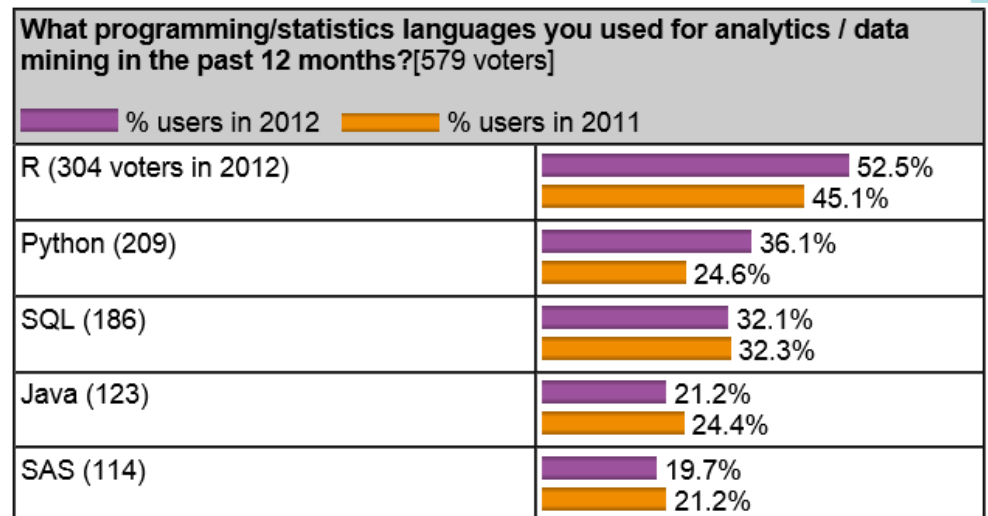
1. Overview
2. Big Data Analysis Example
3. Scrappy
4. MongoDB
5. Solr
6. Scikit-learn

When Big Data meet Python



Why Python?

- Good code readability for fast development.
- Scripting language: the less code, the more productivity.
- Application:
 - Web
 - GUI
 - OS
 - Science
- Fast growing among open source communities.
 - Commits statistics from ohloh.net



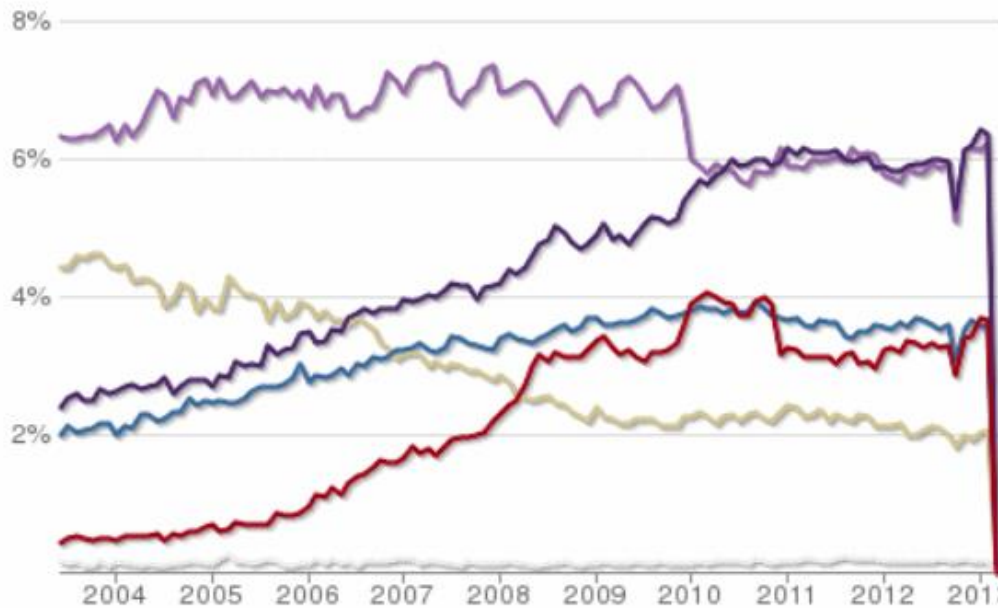
<http://www.kdnuggets.com/2012/08/poll-analytics-data-mining-programming-languages.html>



Why Python?

- The support of open source projects

Monthly Projects (Percent of Total)

The lines show the count of projects with at least one line of code changed in a month. [More](#)



	Java	<input type="button" value="v"/>
	Perl	<input type="button" value="v"/>
	PHP	<input type="button" value="v"/>
	Python	<input type="button" value="v"/>
	R	<input type="button" value="v"/>
	Ruby	<input type="button" value="v"/>
	[None]	<input type="button" value="v"/>
<input type="button" value="Update"/>		

<http://www.ohloh.net/languages/compare?measure=projects&percent=true&l0=java&l1=perl&l2=php&l3=python&l4=r&l5=ruby&l6=-1&l7=-1&commit=Update>

Big Data Analysis Example

www.douban.com/group/111410/?ref=sidebar

影 音乐 同城 小组 阅读 豆瓣FM 更多

豆瓣小组

我的小组



每月养成一个好习惯

创建于2008-05-26 组长: 传奇

活着就是为了庆祝生命

人生在世，最大的敌人不一定是外来的，而可能是我们自己！——你是不是难以把握机会，因为犹疑、拖延的毛病？你是不是容易满足现状，因为没有更高的理想？你是不是不敢面对未来，因为缺乏信心？你是不是未能突破，因为不想去突破？你是不是无法发挥潜能，因为不能超越？



每月养成一个好习惯

活着就是为了庆祝生命

人生在世，最大的敌人不一定是外来的，而可能是我们自己！——你是不是难以把握机会，因为犹疑、拖延的毛病？你是不是容易满足现状，因为没有更高的理想？你是不是不敢面对未来，因为缺乏信心？你是不是未能突破，因为不想去突破？你是不是...

220019 人聚集在这个小组，
你是否愿意成为其中的一员？

加入小组

珍惜时间！要变得完美，强大，自制，忍

作者	回应
川蔓	104415
始努力永远不晚！	3142
一路盛开	6777
左小姐~	4488
[已注销]	5890
be dreams	8785
D『2011.3.3~...	8075
.....	2523
蚕蚕	3903
传奇	3937
我是逆...	1924
shine	145

Big Data Analysis Example

www.douban.com/group/111410/?ref=sidebar

影 音乐 同城 小组 阅读 豆瓣FM 更多

豆瓣小组

我的小组

发现小组

发现



每月养成一个好习惯

我是这个小组

创建于2008-05-26 组长: 传奇

活着就是为了庆祝生命

.....

人生在世，最大的敌人不一定是外来的，而可能是我们自己！——
你是不是难以把握机会，因为犹疑、拖延的毛病？
你是不是容易满足现状，因为没有更高的理想？
你是不是不敢面对未来，因为缺乏信心？
你是不是未能突破，因为不想去突破？
你是不是无法发挥潜能，因为不能超越自己？

最近话题 / 最热话题

话题	作者	回应
↑ ZH说：叫每天早睡早起坚持贴 更好~~	川蔓	104415
↑ 爱自己，对自己的人生负责！从现在开始努力永远不晚！	折翼的蝴蝶	3142
↑ 【宣言饮酒】请看我头置簪花 一路走来 一路盛开	苏暖袖。	6777
↑ 那些寂寞的时光，用来构建强大的内心	左小姐~	4488
↑ 每天专心种番茄	[已注销]	5890
↑ 【夏日大作战】Don't let your dreams be dreams	夙 酱 *	8785
↑ 『五一快乐^_^』遇见更美好的自己:-D『2011.3.3~...	晓夕	8075
↑ 【当你醒来，请回答】我今天的目标是	conge	2523
↑ 【欢迎加入】提高学习效率PK自由赛	.蚕蚕	3903
↑ 日常养生之482：春季养颜吃什么？	传奇	3937
就一年，姑娘，要劲的时候到了。我是安妮，我是逆...	ada瑞银	1924
珍惜时间！要变得完美，强大，自制，忍	shine	145

Scrapy web scraping framework

<http://scrapy.org/>

- pip install scrapy
- Parse field by XPath
- Spider
 - Request
 - Response
 - Parse_function
 - Pipeline
- For Big Data:
 - Parallel crawling

Traverse web pages

www.douban.com/group/111410/discussion?start=50

【给自己的内心建一所小房子】	IM	63
so shut up man,I am the rule.	贱人喵。	14
淘宝指南，有了她在也不浪费时间买到假货了	丫爺、至尊K	
【夏日大作战。大一编辑学。】为了更接近太阳！	慕斯	273
做一个勤劳的园丁，静待满园花开	默默	22
在2013年，遇见努力的自己~~~~	左小魅	142
♥ 24 BETTER ME {打卡}	尼小采	2
微信练口语	学术淫娃美少年	45

<前页 1 2 3 4 5 6 7 8 9 ... 1465 1466 后页>

Scrapy web scraping framework

<http://scrapy.org/>

【考研打卡】天行健，君子以自强不息



来自: 采蘑菇的小姑娘(耐心是一种优秀的品质。) 2013-03-03 23:05:08

每星期一句

只要还有明天，今天就永远是起跑线。

目标：上海财经大学。

日语跨经济。偶要专业第一考进去！

分数目标：410（只有上400才没问题！）

政治：70

日语：75

数学：135

专业课：130

关于总结一些方法的链接戳这里：

<http://www.douban.com/group/topic/37047473/?start=700#457234511>

Parse article

分享到 推荐 13人

37人 喜欢

Big Data Anysis in Py



Sunny Smile (人生路那么多，选择一条，专心。) 2013-03-04

大四党，同考研失败，不过还好，路在脚下往前看。



采蘑菇的小姑娘(耐心是一种优秀的品质。) 2013-03-04 10:58

【书摘】但并非所有的思维转换都是积极的。例如我们前面提到的魅力，这种转换反而让我们偏离了通向成功与幸福的轨道。



鱼儿要跃龙门(奋斗吧青年，但我不知如何奋斗) 2013-03-04 1

我也大四了，找工作，毕业论文什么的，很烦很烦.....



采蘑菇的小姑娘(耐心是一种优秀的品质。) 2013-03-04 11:01

【书摘】一棵邪恶的大树，砍它枝叶千斧，不如砍它根基一斧。行是根基，抓住根本才能让生活产生实质性的进展

Parse coments



采蘑菇的小姑娘(耐心是一种优秀的品质。) 2013-03-04 11:03

【书摘】【以原则为中心的思维定式】公平、诚信、正直、服务、

```
class GroupSpider(BaseSpider):
```

```
    name = "group"
```

```
    allowed_domains = ["douban.com"]
```

```
    start_urls = ["http://www.douban.com/group/111410/discussion?start=%d" % i  
for i in range(10000, 20000, 25)]
```

Traverse web pages

```
def __init__(self):
```

```
    self.coll = get_coll()
```

```
def parse(self, response):
```

```
    hxs = HtmlXPathSelector(response)
```

```
    for tr in hxs.select('//tr'):
```

```
        title, url, author, author_url = None, None, None, None
```

```
        result = tr.select('td[@class="title"]')
```

```
        if len(result) == 1:
```

Extract field by XPath

```
            url = result.select('a/@href').extract()[0]
```

```
            if self.coll.find_one({'url': url}) is not None:
```

```
                self.log('duplicated url: %s' % (url))
```

```
            else:
```

```
                yield Request(url, callback=self.parse_article)
```





Document based NoSQL database

<http://www.mongodb.org/>

- Python client: `pip install pymongo`
- Json style document
- **For Big Data:**
 - Mongodb cluster with shard and replica
- **article:**
 - author
 - author_url
 - title
 - content
 - comments
 - url

```
import pymongo
```

```
conn = pymongo.Connection()
coll = conn.database.collection
coll.ensure_index('url', unique=True)
article = {'author': u'小明', 'author_url':
'http://host/path1', 'comments': [],
          'title': u'天天向上', 'content': '', 'url':
'http://host/path2'}
coll.insert(article)
results = [i for i in coll.find({'title': u'天天
向上'})]
```



Full-text search engine

<http://lucene.apache.org/solr/>

- Customize full-text indexing and searching by xml config
- RESTful API for select/update index
- Spatial Search
- For Big Data:
 - SolrCloud:
distributed indexes

The screenshot shows the Apache Solr Admin UI. On the left is a sidebar with the Solr logo and a list of navigation links: Dashboard, Logging, Core Admin, Java Properties, Thread Dump, a dropdown menu showing 'collection1', Overview (highlighted), Ping, Query, Schema, Config, Replication, Analysis, Schema Browser, and Plugins / Stats. The main content area is divided into two sections. The top section, titled 'Statistics', shows the last update was 25 minutes ago, and lists metrics: Modified, Num Docs: 18742, Max Doc: 18742, Deleted: 0, Docs, Version: 7, Segment: 3, Count, Optimized: (checked), and Current: (checked). The bottom section, titled 'Replication (Master)', contains a table with columns 'Version' and 'Gen'. The table lists two entries: 'Master (Searching)' and 'Master (Replicable)', both with version '1369407058081' and generation '4'. Below the table is a 'Healthcheck' section with a message: 'Ping request handler is not configured with a healthcheck file.' At the bottom of the page, there are links for 'Documentation' and 'Issue Tracker'.

	Version	Gen
Master (Searching)	1369407058081	4
Master (Replicable)	1369407058081	4

```

<field name="url" type="text_general" indexed="true" stored="true"/>
<field name="title" type="text_general" indexed="true" stored="true"/>
<field name="content" type="text_general" indexed="true" stored="true"/>
<field name="comments" type="text_general" indexed="true" stored="true"
multiValued="true"/>
<field name="author" type="text_general" indexed="true" stored="true"/>
<field name="author_url" type="text_general" indexed="true" stored="true"/>

```

Define fields and types

```

<copyField source="title" dest="text"/>
<copyField source="author" dest="text"/>
<copyField source="content" dest="text"/>
<copyField source="comments" dest="text"/>

```

Define index process

```

<fieldType name="text_smart_chinese" class="solr.TextField"
positionIncrementGap="100">
  <tokenizer class="solr.SmartChineseSentenceTokenizerFactory"/>
  <filter class="solr.SmartChineseWordTokenFilterFactory"/>
  <filter class="solr.StopFilterFactory" ignoreCase="true" words="stopwords.txt"
enablePositionIncrements="true" />
  <filter class="solr.LowerCaseFilterFactory"/>
  <filter class="solr.PorterStemFilterFactory"/>
</fieldType>

```





```
{
+ responseHeader: {...},
- response: {
  numFound: 7834,
  start: 0,
  docs: [
    - {
      author: "追筑",
      url: "http://www.douban.com/group/topic/28792422/",
      title: " 【好好生活】读书笔记 ",
      content: "每天坚持读书，坚持思考。",
      author_url: "http://www.douban.com/people/1393829/",
      _version_: 1435927368971583500
    },
    - {
      author: "Wise海綿",
      url: "http://www.douban.com/group/topic/26795128/",
      title: " 要早睡早起，坚持贴 ",
      content: "早睡早起，每日读书",
      author_url: "http://www.douban.com/people/NinaWang_/",
      _version_: 1435927371653841000
    },
    - {
      author: "糖糖",
      url: "http://www.douban.com/group/topic/27373910/",
      title: " 【送书】地址给我，我把书免费邮给你 （关于外公的） ",
      content: "读书会第二期：《外公与我》试读 点链接 ",
      author_url: "http://www.douban.com/people/57500408/",
      _version_: 1435927372410912800
    },
  ],
}
```


Popular Book Ranking List

By regex search in Mongoddb	By full-text search in Solr
少有人走的路	那些年我们一起追的女孩
百年孤独	了不起的盖茨比
遇见未知的自己	我不要你死于一事无成
平凡的世界	考拉小巫的英语学习日记
送你一颗子弹	高效能人士的7个习惯
小王子	哪来的天才
拖延心理学	被嫌弃的松子的一生
活着	与众不同的心理学
苏菲的世界	奇特的一生
红楼梦	普罗旺斯的一年
目送	接纳不完美的自己
如何阅读一本书	我就是想停下来，看看这个世界
围城	这些都是你给我的爱



Machine Learning

<http://scikit-learn.org/>

- Machine learn algorithms
 - Supervised learning
 - Unsupervised learning
- Processing
 - Model selection
 - Pipeline
- Application: Article Recommendation
 - Let computer figures out your interests, and then recommend articles for you.
 - Learning by OneClassSVM

Article Recommendation

Pickup some favorite articles



Extract features from article and let model learn from the features



Model predict on unseen articles and recommend for you

```
articles = [coll.find_one({'url': url})['content'] for url in urls]
tfidf = TfidfVectorizer(tokenizer=my_tokenizer,
                        ngram_range=(1, 3))
svm = OneClassSVM(kernel="linear", nu=0.3)
train_vectors = tfidf.fit_transform(articles)
svm.fit(train_vectors)
svm.predict(test_vectors)
```

- **Result: recommend 390 from 8000 articles, and the recommended articles are good to me.**

Reference (1/2)

- Book:
 - Wes McKinney, “Python for Data Analysis” , O’Reilly, 2012
 - Toby Segaran, “Programming Collective Intelligence: Building Smart Web 2.0 Applications” , O’Reilly, 2008
 - Philipp K. Janert, “Data Analysis with Open Source Tools” , O’Reilly, 2010
- Coursera:
 - Web Intelligence and Big Data, [url](#)
 - Introduction to Data Science, [url](#)
 - Data Analysis, [url](#)

Reference (2/2)

- Conference:
 - PyData 2013, <http://pydata.org/abstracts/>
 - PyData 2012, http://marakana.com/s/post/1090/2012_pydata_workshop
- My former shares:
 - When Big Data Meet Python, COSCUP 2012, [slides](#)
 - NLTK: natural language toolkit overview and application, PyCon.tw 2012, [slides](#)