1 序言 暖个场子

为什么

- 普及知识
- 做个整理
- 提高个人能力

讲什么

- 以数据工程为核心
- 个人学习经历
- 工具、方法、作品
- 大众向

怎么讲

- 数量多，内容集中
- 录屏
- 不定期更新

2 序言 数据工程和编程语言

数据工程

采集、存储、清洗、分析、可视化

编程语言

C++和 Java

python 大法

R

web：php、html、css、javascript

结合

- 采集：python
- 存储：python + 数据库
- 清洗：python
- 分析：python + R
- 可视化：R + web

3 序言 带好装备 Python 和 Sublime

程序员的两件装备

编辑器，Sublime，http://www.sublimetext.com/2

编程语言，Python，https://www.python.org/

Sublime 的安装和使用

下载，安装，Sublime Text 2

如何安装插件：

1. 按 Ctrl+`调出 console
2. 粘贴以下代码到底部命令行并回车：import urllib2,os;pf='Package Control.sublime-package';ipp=sublime.installed_packages_path(); os.makedirs(ipp) if not os.path.exists(ipp) else None;open(os.path.join(ipp,pf),'wb').write(urllib2.urlopen('http://sublime.wbond.net/'+pf.replace(' ','%20')).read())
3. 重启 Sublime Text 2
4. 如果在 Perferences->package settings 中看到 package control 这一项，则安装成功

安装插件：

1. 按下 Ctrl+Shift+P 调出命令面板
2. 输入 install 调出 Install Package 选项并回车，然后在列表中选中要安装的插件。

插件能干什么：提供某些功能或者对某些语言的支持，例如 jquery、latex 等

使用和操作

Python 的安装和使用

Mac、Linux、Windows，下载、安装

安装 pip：

Windows，http://www.tuicool.com/articles/eiM3Er3/

Mac，http://www.xuebuyuan.com/593678.html

更好的选择：

Anaconda，https://www.continuum.io/downloads

运行 python 代码的方法：

1. 在命令行中输入 python，交互式执行
2. 使用 ipython notebook 等交互式编程工具
3. 使用 Sublime 编辑和运行代码，或者编辑好后在命令行中运行

来一个 Hello World

4 Python 先学会基本语法

教程命名规则

编号 + 主题 + 名称

Python 2 基本语法

解释型（无需编译）、交互式、面向对象、跨平台、简单好用

中文编码：http://www.cnblogs.com/huxi/archive/2010/12/05/1897271.html

变量名：可以包括英文、数字以及下划线，但不能以数字开头，区分大小写

变量类型：弱类型语言、无需声明

- 数字 Number：整型和浮点型
- 字符串 String：字符串拼接、长度、切片
- 列表 List：添加元素、求长、切片、删除
- 元组 Tuple：readonly
- 字典 Dictionary：赋值、判断是否存在某个 key

注释：#，三引号

保留字符：and，not，class，def，等等等等

行和缩进

运算符：

- 算术运算符：+，-，* ，/，%
- 比较运算符：==，!=，>，<，>=，<=
- 赋值运算符：=，+=，-=，*=，/=，%=
- 逻辑运算符：and，or，not

条件：

- if...
- if...else...
- if...elif...else

循环：

- while
- for，for 遍历 list 和 dict

循环控制：

- break
- continue
- pass

时间：time.time()

文件：读写文件

异常

函数：def

补充学习资料

菜鸟教程：http://www.runoob.com/python/python-tutorial.html

廖雪峰python教程：http://www.liaoxuefeng.com/wiki/0014316089557264a6b348958f449949df42a6d3a2e542c000/

5 实战 西游记用字统计

目的

通过一个简单的项目

来巩固上次视频

所讲的 python 基础

数据

xyj.txt，《西游记》的文本，2.2MB

致敬吴承恩大师，4020 行（段）

目标

统计《西游记》中：

1. 共出现了多少个不同的汉字；
2. 每个汉字出现了多少次；
3. 出现得最频繁的汉字有哪些。

涉及内容：

1. 读文件；
2. 字典的使用；
3. 字典的排序；
4. 写文件

6 数据 解读数据结构和类型

数据的结构

举个栗子：地铁数据

静态数据：线路、站点（不一定有时间戳，更新慢）

动态数据：刷卡记录（必有时间戳，不断产生）

时间戳：从 1970 年 1 月 1 日 0 时 0 分 0 秒到现在所经历的秒数

行：记录、观测

列：字段、属性

二维数组、表

数据的类型

TXT：纯文本

CSV：逗号分隔值

JSON：键值对

SQL：数据库文件（后续教程再详细介绍）

7 爬虫 Http 请求和 Chrome

访问一个网页

http://kaoshi.edu.sina.com.cn/college/scorelist?tab=batch&wl=1&local=2&batch=&syear=2013

url：协议 + 域名 / IP + 端口 + 路由 + 参数

ping

通过 url 能得到什么

在浏览器中打开

墙裂推荐大家使用 Chrome 浏览器

渲染效果、调试功能都是没话说的

http://www.google.cn/intl/zh-CN/chrome/browser/desktop/index.html

开发者工具

显示网页源代码、检查

1. Elements：页面渲染之后的结构，任意调整、即时显示；
2. Console：打印调试；
3. Sources：使用到的文件；
4. Network：全部网络请求。

Http 请求

Http 是目前最通用的 web 传输协议

1. GET：参数包含在 url 中；
2. POST：参数包含在数据包中，url 中不可见。

http://shuju.wdzj.com/plat-info-59.html

Url 类型

1. html：返回 html 结构页面，通过浏览器渲染后呈现给用户；
2. API：Application Programming Interfaces，请求后完成某些功能，例如返回数据。

http://kaoshi.edu.sina.com.cn/?p=college&s=api2015&a=getAllCollege

8 爬虫 使用 urllib2 获取数据

Python 中的 Urllib2

https://docs.python.org/2/library/urllib2.html

我的 python 版本：2.7

发起 GET 请求

http://kaoshi.edu.sina.com.cn/college/scorelist?tab=batch&wl=1&local=2&batch=&syear=2013

```
request = urllib2.Request(url=url, headers=headers)
response = urllib2.urlopen(request, timeout=20)
result = response.read()
```

发起 POST 请求

http://shuju.wdzj.com/plat-info-59.html

```
data = urllib.urlencode({'type1': x, 'type2': 0, 'status': 0, 'wdzjPlatId': int(platId)})
request = urllib2.Request('http://shuju.wdzj.com/depth-data.html', headers)
opener = urllib2.build_opener(urllib2.HTTPCookieProcessor())
response = opener.open(request, data)
```

result = response.read()

处理返回结果

Html：BeautifulSoup，需要有一些 CSS 基础

API：JSON

https://www.crummy.com/software/BeautifulSoup/bs4/doc/

9 实战 爬取豆瓣电影数据

聊会天

三大目标：链家、豆瓣、点评

三月爬虫

矛与盾：伪装浏览器、IP 限制、登陆、验证码（CAPTCHA）

通用思路

一个汇总页

一堆详情页

找链接

从汇总页一步一步下钻到详情页

找字段

在详情页中需要哪些字段

动手

10 数据库 用 MAMP 和 WAMP 搭建 Web 环境

Web 环境

Web 服务器：Apache、Nginx，处理 Web 请求

数据库：MySQL，存储和管理数据

后端：PHP

Web 服务启动后，就可以在浏览器中访问根目录中的网站项目

MAMP：Mac，Apache，MySQL，PHP，https://www.mamp.info/en/

WAMP：Windows，Apache，MySQL，PHP，http://www.wampserver.com/en/

偏好设置

端口设置：Apache、MySQL，端口只是一个后缀，不同服务使用不同端口，彼此不冲突

根目录：访问 http://localhost:port/之后所得到的目录

Hello World

使用 Html

使用 PHP

11 数据库 MySQL 使用方法

基本概念

数据库和数据表

CURD 操作：Create、Update、Read、Delete

1 命令行

直接在终端或 cmd 中敲命令

2 Web 工具

phpmyadmin：简单、轻量、好用

新建数据库

新建数据表、定义字段（Int、Float、Varchar、Text）

3 本地软件

Navicat：功能更强大

数据的导入、导出

4 使用代码

mysql-python：读写更新数据

我的习惯

使用 phpmyadmin 新建数据库和数据表

使用 python 插入、读取、更新数据

使用 Navicat 导出数据库

使用 phpmyadmin 导入数据库

12 数据库 使用 Python 操作 MySQL

MySQLdb

安装：pip install mysql-python

加载包

```
import MySQLdb
import MySQLdb.cursors
```

建立连接

```
db = MySQLdb.connect(host='127.0.0.1', user='root', passwd='root', db='douban', port=8889,
charset='utf8', cursorclass = MySQLdb.cursors.DictCursor)
db.autocommit(True)
cursor = db.cursor()
```

执行操作

CURD

cursor.execute(sql)

关闭连接

---

cursor.close()

db.close()

SQL 教程：http://www.runoob.com/sql/sql-tutorial.html

13 ggplot2 在 R 中进行可视化

R 是一门统计分析语言，有很多包、功能强大

安装和下载

R： https://www.r-project.org/，核心

RStudio：https://www.rstudio.com/，更丰富的界面

R 基础

初探 RStudio：各种窗口、命令行、编写 R 脚本

安装包和加载包

变量类型：向量、矩阵、数组、数据框、因子、列表

从 CSV 中读取数据为数据框

ggplot2

安装和加载

一个简单的例子：mtcars

R 学习笔记

http://zhanghonglun.cn/blog/tag/r/

14 ggplot2 基本语法和基础图形

背景

有哪些图形：散点图、折线图、面积图、条形图、直方图、箱线图……

绘图元素：shape、color、fill……

还有呢：x 轴、y 轴、标题、图例……

如何选择：根据 x 轴、y 轴将要展示的变量（连续或离散），以及展示的需求

基本语法

ggplot(data) + geom_type()

元素映射：将某一变量（连续或离散）映射到 shape、color、fill 等元素上

条形图 bar

BOD、diamonds、cabbage_exp

y 为频数、y 为变量值、分组条形图

折线图 line、面积图 area

BOD、uspopage

散点图 point

heightweight、mtcars

描述数据分布

直方图 histogram、密度图 density、箱线图 boxplot

分面

facet_wrap()

R 数据可视化

http://zhanghonglun.cn/blog/tag/r/

15 实战 Diamonds 数据集探索

查看数据

diamonds

截取子集

set.seed(123)

diamonds <- diamonds[sample(nrow(diamonds), 1000),]

查看概要

summary()、str()

探索

价格和克拉的关系：geom_point()，映射颜色和形状

价格分布：geom_histogram()，映射填充、position="fill"/"dodge"

透明度分布：geom_bar()

价格概率分布：geom_density()，映射颜色、填充

不同切工下的价格分布：geom_boxplot()，映射填充

坐标变换：scale_y_log10()

加上坐标轴标签和标题：labs(x="", y="", title="")

16 NLP 走近自然语言处理

概念

Natural Language Processing/Understanding，自然语言处理/理解

日常对话、办公写作、上网浏览

希望机器能像人一样去理解，以人类自然语言为载体的文本所包含的信息，并完成一些特定任务

内容

中文分词、词性标注、命名实体识别、关系抽取、关键词提取、信息抽取、依存分析、词嵌入……

应用

篇章理解、文本摘要、情感分析、知识图谱、文本翻译、问答系统、聊天机器人……

17 NLP 使用 jieba 分词处理文本

jieba 中文分词

https://github.com/fxsjy/jieba

即使效果不是最好的，但是，完全开源、简单易用

安装

pip install jieba

中国特色社会主义是我们党领导的伟大事业，全面推进党的建设新的伟大工程，是这一伟大事业取得胜利的关键所在。党坚强有力，事业才能兴旺发达，国家才能繁荣稳定，人民才能幸福安康。党的十八大以来，我们党坚持党要管党、从严治党，凝心聚力、直击积弊、扶正祛邪，党的建设开创新局面，党风政风呈现新气象。习近平总书记围绕从严管党治党提出一系列新的重要思想，为全面推进党的建设新的伟大工程进一步指明了方向。

中文分词

基于规则、基于统计

jieba：基于前缀词典进行词图扫描，构成全部可能分词结果的有向无环图，动态规划查找最大概率路径

```
import jieba
seg_list = jieba.cut("我来到北京清华大学", cut_all=True) print("Full Mode: " + "/ ".join(seg_list)) # 全模式 seg_list = jieba.cut("我来到北京清华大学", cut_all=False) print("Default Mode: " + "/ ".join(seg_list)) # 精确模式 seg_list = jieba.cut("他来到了网易杭研大厦") # 默认是精确模式 print(", ".join(seg_list)) seg_list = jieba.cut_for_search("小明硕士毕业于中国科学院计算所，后在日本京都大学深造") # 搜索引擎模式 print(", ".join(seg_list))
```

关键词提取

```
import jieba.analyse
```

基于 TF-IDF：jieba.analyse.extract_tags(sentence, topK=20, withWeight=False, allowPOS=())
基于 TextRank：jieba.analyse.textrank(sentence, topK=20, withWeight=False, allowPOS=('ns', 'n', 'vn', 'v'))

词性标注

```
import jieba.posseg as pseg
words = pseg.cut("我爱北京天安门")
for word, flag in words:
    print('%s, %s' % (word, flag))
```

词性列表

---

1. 名词 (1 个一类，7 个二类，5 个三类)

n 名词

nr 人名

nr1 汉语姓氏

nr2 汉语名字

nrj 日语人名

nrf 音译人名

ns 地名

nsf 音译地名

nt 机构团体名

nz 其它专名

nl 名词性惯用语

ng 名词性语素

2. 时间词(1 个一类，1 个二类)

t 时间词

tg 时间词性语素

3. 处所词(1 个一类)

s 处所词 (家中、门外、境内、西方……)

4. 方位词(1 个一类)

f 方位词

5. 动词(1 个一类，9 个二类)

v 动词

vd 副动词

vn 名动词

vshi 动词“是”

vyou 动词“有”

vf 趋向动词

vx 形式动词

vi 不及物动词（内动词）

vl 动词性惯用语

vg 动词性语素

6. 形容词(1 个一类，4 个二类)

a 形容词

ad 副形词

an 名形词

ag 形容词性语素

al 形容词性惯用语

7. 区别词(1 个一类，2 个二类)

b 区别词 (主要、整个、所有……)

bl 区别词性惯用语

8. 状态词(1 个一类)

z 状态词

9. 代词(1 个一类，4 个二类，6 个三类)

r 代词

rr 人称代词

rz 指示代词

rzt 时间指示代词

rzs 处所指示代词

rzv 谓词性指示代词

ry 疑问代词

ryt 时间疑问代词

rys 处所疑问代词

ryv 谓词性疑问代词

rg 代词性语素

10. 数词(1 个一类，1 个二类)

m 数词

mq 数量词

11. 量词(1 个一类，2 个二类)

q 量词

qv 动量词

qt 时量词

12. 副词(1 个一类)

d 副词

13. 介词(1 个一类，2 个二类)

p 介词

pba 介词“把”

pbei 介词“被”

14. 连词(1 个一类，1 个二类)

c 连词

cc 并列连词

15. 助词(1 个一类，15 个二类)

u 助词

uzhe 着

ule 了 喽

uguo 过

ude1 的 底

ude2 地

ude3 得

usuo 所

udeng 等 等等 云云

uyy 一样 一般 似的 般

udh 的话

uls 来讲 来说 而言 说来

uzhi 之

ulian 连 （“连小学生都会”）

16. 叹词(1 个一类)

e 叹词

17. 语气词(1 个一类)

y 语气词(delete yg)

18. 拟声词(1 个一类)

o 拟声词

19. 前缀(1 个一类)

h 前缀

20. 后缀(1 个一类)

k 后缀

21. 字符串(1 个一类，2 个二类)

x 字符串

xx 非语素字

xu 网址 URL

22. 标点符号(1 个一类，16 个二类)

w 标点符号

wkz 左括号，全角：（ 〔 ［ ｛ 《 【 〖 〈 半角：([{ <

wky 右括号，全角：） 〕 ］ ｝ 》 】 〗 〉 半角：)] { >

wyz 左引号，全角：“ ‘ 『

wyy 右引号，全角：” ’ 』

wj 句号，全角：。

ww 问号，全角：？ 半角：?

wt 叹号，全角：！ 半角：!

wd 逗号，全角：， 半角：,

wf 分号，全角：； 半角： ;

wn 顿号，全角：、

wm 冒号，全角：： 半角： :

ws 省略号，全角：…… …

wp 破折号，全角：—— －－ ——－ 半角：--- ----

wb 百分号千分号，全角：％ ‰ 半角：%

wh 单位符号，全角：￥ $ £ °℃ 半角：$

---

18 NLP WordEmbedding 的概念和实现

背景

如何表示词语所包含的语义？

苹果？水果？Iphone？

苹果、梨子，这两个词相关吗？

语言的表示

符号主义：Bags-of-word，维度高、过于稀疏、缺乏语义、模型简单

分布式表示：Word Embedding，维度低、更为稠密、包含语义、训练复杂

Word Embedding

核心思想：语义相关的词语，具有相似的上下文环境，例如， 苹果和梨子

所做的事情：将每个词语训练成，词向量

实践

基于 gensim 包和中文维基语料

gensim：http://radimrehurek.com/gensim/models/word2vec.html

中文维基分词语料：链接 https://pan.baidu.com/s/1qXKIPp6 密码 kade

```
# 加载包
from gensim.models import Word2Vec
from gensim.models.word2vec import LineSentence
```

```
# 训练模型
sentences = LineSentence('wiki.zh.word.text')
model = Word2Vec(sentences, size=128, window=5, min_count=5, workers=4)

# 保存模型
model.save('word_embedding_128')

# 加载模型
model = Word2Vec.load("word_embedding_128")

# 使用模型
items = model.most_similar(u'中国')
model.similarity(u'男人', u'女人')
```

---

19 Web 基础 网页的骨骼 HTML

什么是 HTML

超文本标记语言：**H**yper **T**ext **M**arkup **L**anguage

这都不重要，重要的是：

HTML 是 Web 网页的基本组成部分

HTML 中定义的元素，决定了网页的内容和结构

Python：编程语言，编写程序

HTML：标记语言，像画画一样，画出网页的内容

基本结构

```
<!DOCTYPE html>
<html>
    <head>
```

```
    </head>
    <body>
    </body>
</html>
```

常用标签

单标签、双标签

```
<meta charset="UTF-8"/>
<title>我是一个标题</title>
```

块级标签、内联标签

```
<h1>我是一号标题</h1>，块级
<h6>我是六号标题</h6>，块级
<p>我是一个默默无闻的段落</p>，块级
<a href="http://zhanghonglun.cn" target="_blank">带你去一个好地方</a>，内联
<img src="logo.png" width="200" height="200"/>，内联
<br/>
<div>我是块级元素</div>
<span>我是内联元素</span>
```

表格：table、tr、th、td

列表：ul、ol、li

下拉：<select><option></option></select>

元素的属性

id、class、style

```
<a href="#id">跳转到某个 id 的元素</a>
```

HTML 注释

```
<!-- 这是一个注释 -->
```

表单

```
<form action="" method="post">
    用户名 <input type="text" placeholder="默认文本" name="username"/>
    密码 <input type="text" placeholder="密码" name="password"/>
    一大段文本 <textarea rows="10" cols="10" placeholder="想说的话" name="content"></textarea>
    <button type="submit">登陆</button>
</form>
```

---

input 的 type：button、checkbox、color、date、datetime、email、file、month、number、password、radio、range、submit、text、time

HTML 颜色

十六进制：#FFF

RGB：rgb(255, 255, 255)，rgba(255, 255, 255, 1)

颜色名称：red，green，blue

DOM

文档对象模型：Document Object Model

HTML5

新标签：canvas、svg、audio、video、embed

svg：http://www.runoob.com/svg/svg-tutorial.html

canvas：http://zhanghonglun.cn/blog/canvas 初探：基本语法

新的语义元素：header、nav、section、article、aside、figcaption、figure、footer

新功能：元素拖放、地理定位、video、audio、更丰富的 input type、Web 存储（localStorage 和 sessionStorage）

HTML 补充学习

http://www.runoob.com/html/html-tutorial.html

20 Web 基础 网页的血肉 CSS

什么是 CSS

层叠样式表：**C**ascading **S**tyle **S**heets

这都不重要，重要的是：

CSS 决定了如何显示 HTML 元素

基本结构

选择器 + 样式（key: value）

```
p {
    color: red;
    font-size: 20px;
}
```

使用 CSS

1. 引入外部.css 文件
2. 在 html 中定义 css
3. 在元素中使用内联 css

常用选择器

- 元素名
- id
- class
- 后代选择器
- 子元素选择器
- 相邻兄弟选择器、普通相邻兄弟选择器

- 伪类

常用样式

背景：background-color、background-image、background-repeat、background-size、background-attachment、background-position

大小：width、height

大小单位：px、%、em

文本：color、text-align、text-decoration、text-indent、line-height、font-size、font-family

留白：margin、padding

边框：border、border-radius、box-shadow

显示：display

定位：static、fixed、relative、absolute、float

CSS 注释

/* 这是一个注释 */

CSS3

新属性：渐变、transform（translate、rotate、scale、skew、matrix）、transition、animation（keyframes）

新功能：加载想要的字体

实例

美化一个 button、添加 hover 动画效果

CSS 补充学习

http://www.runoob.com/css/css-intro.html

21 Web 基础 网页的关节 JS

什么是 JS

HTML 中的脚本编程语言：JavaScript，但和 Java 毛关系没有

这都不重要，重要的是：

JS 决定了如何动态改变 HTML 元素

使用 JS

1. 在 html 中使用 js
2. 引入外部.js 文件

内容

- document.write()
- 变量 var：数值、字符、数组、字典/对象
- document.getElementById()
- onclick="myFunction()"
- innerHTML
- console.log()
- 运算符、条件、循环
- 注释
- 函数
- 作用域
- 事件

JS 补充学习

http://www.runoob.com/js/js-tutorial.html

22 Web 进阶 比 JS 更方便的 JQuery

简介

- JQuery 是一个 JS 库

- 极大地简化了 JS 编程
- JQuery 很容易学习

引入

- 下载下来并引入：http://jquery.com/download/
- 直接引用 CDN： http://cdn.bootcss.com/jquery/2.1.4/jquery.min.js

语法

$(document).ready(function(){});

$('选择器').action();

选择器可以是：元素名、id、class、子元素选择器、后代元素选择器、（相邻）兄弟选择器、属性选择器、this

action 可以是：click、dbclick、mouseenter/leave/over/out、hover、keypress/up/down、change、focus、blur，效果和动画，DOM 操作

效果：hide、show、toggle，fadeIn、fadeOut、fadeToggle、slideUp、slideDown、slideToggle

动画：animate

回调（callback）：完成某一函数之后再执行的操作

JQuery 链（Chaining）：连续写多个 action

DOM 操作

获取和设置内容：text()、html()、val()

获取属性：attr()

添加元素：append()、prepend()、before()、after()

删除元素：remove()、empty()

获取和设置属性：css()

遍历和关系：each()、parent()、children()、find()、siblings()

AJAX

异步 JavaScript 和 XML（Asynchronous JavaScript and XML）

JQuery 补充学习

http://www.runoob.com/jquery/jquery-tutorial.html

还有很多前端框架

Angular.js、Vue.js、React.js

23 实战 和 DT 财经合作的中秋节月饼项目

成品展示

http://zhanghonglun.cn/dt_moon_cake/

为手机端设计，PC 端访问请将浏览器调整至合适大小

项目链接

https://github.com/Honlan/dt_moon_cake

两个 html、三个 json、一些其他文件

首页：输入基本信息

月饼页：返回对应的月饼介绍

所涉及内容：html、css、js、jquery

代码时间

Let's Go!

24 Web 进阶 基于 ThinkPHP 的简易个人博客

什么是 ThinkPHP

一款基于 PHP 的后端框架

PHP 基础：http://www.runoob.com/php/php-tutorial.html

ThinkPHP 官网（中国人开发的 PHP 框架）：http://www.thinkphp.cn/

其他流行的 PHP 框架：

- CI：http://codeigniter.org.cn/
- Yii：http://www.yiiframework.com/，http://www.yiichina.com/
- Laravel：https://laravel.com/，http://www.golaravel.com/（推荐）

页面

首页、文章列表页、文章详情页

步骤

1. 下载 ThinkPHP
2. MVC
3. 数据库和 config.php
4. 函数和渲染
5. U 函数
6. 表单实现
7. 处理表单并跳转
8. 读取数据和渲染

'DB_TYPE' => 'mysql',

'DB_HOST' => 'localhost',

'DB_NAME' => 'dbname',

'DB_USER' => 'root',

'DB_PWD' => 'root',

'DB_PORT' => 8889,

'LAYOUT_ON' => true,

'LAYOUT_NAME' => 'layout',

接下来

如果你对 PHP 框架感兴趣并且希望进一步了解

去学习 Laravel 吧～

25 Web 进阶 基于 Flask 的简易个人博客

什么是 Flask

一款基于 Python 的轻量级后端框架

Flask 官网：http://flask.pocoo.org/docs/0.10/

Flask 安装：http://docs.jinkan.org/docs/flask/installation.html#virtualenv

其他流行的 Python 框架：

- Django：https://www.djangoproject.com/

页面

首页、文章列表页、文章详情页

步骤

1. 准备项目：static/、templates/、venv/、config.py、run.py
2. 数据库和 config.py
3. layout.html，{% block body %} {% endblock %} {% extends 'layout.html' %}
4. 函数和渲染
5. url_for()
6. 表单实现
7. 处理表单并跳转
8. 读取数据和渲染

26 动态可视化 国内开源良心之作 ECharts

ECharts 是什么

基于 Canvas 的一款 js 图形可视化工具

http://echarts.baidu.com/index.html

国内开源良心之作，Github 上 15000+stars

从 ECharts2 更新到 ECharts3，更加简单、功能更强

加载 ECharts

下载完整版 js 或使用 CDN，http://www.bootcdn.cn/echarts/

第一个 ECharts 图形

- html 中的 div
- js 中的 init
- 设置 option
- 显示图形

使用其他主题

http://echarts.baidu.com/download-theme.html

更多探索

这是一个充满想象的世界

http://echarts.baidu.com/examples.html

27 实战 再谈豆瓣电影数据分析项目

成果展示

项目链接：http://zhanghonglun.cn/data-visualization/

Github 地址：https://github.com/Honlan/data-visualize-chain

项目内容

采集、清洗、存储、分析、可视化

再谈 BeautifulSoup

```
html = response.read()
html = BeautifulSoup(html)
html.select()
html.find_all("tag", attrs={"key": "value"})
```

历史遗留问题

1. 当时能力不足，很多代码写得不够好
2. 用的 ECharts2，略显麻烦
3. 希望你们能做得更好

28 动态可视化 数据可视化之魅 D3

什么是 D3

Data Driven Documents，数据驱动文档，https://d3js.org/

最流行的 js 可视化库之一，Github 上 58000+stars

D3 核心思想

1. 为 DOM 元素（多为 SVG）绑定数据
2. 利用数据确定 DOM 元素的外观和位置等属性
3. 当数据发生变化时，相应地更新 DOM 元素

一个简单的例子

和 D3 邂逅的第一眼

更加深入的理解

数据元素的添加、更新、删除

http://bl.ocks.org/mbostock/3808234

更多探索

这是一个充满想象的世界

https://github.com/d3/d3/wiki/Gallery

和 ECharts 有哪些区别？

29 实战 星战系列电影知识图谱可视化

成果展示

项目链接：http://zhanghonglun.cn/starwars/

Github 地址：https://github.com/Honlan/starwar-visualization

项目内容

PPT 分享

进一步了解

全过程详细讲解：

http://study.163.com/course/courseMain.htm?courseId=1003528010

接下来

1. 选择你感兴趣的数据
2. 打开你的脑洞
3. 创造出独具特色的 D3 可视化

30 动态可视化 艺术家的可视化工具 Processing

什么是 Processing

Processing 是一门用来生成图片、动画和交互软件的编程语言

非常简单，不只是程序猿，设计狮、艺术僧也在使用！

下载和安装

https://processing.org/download/

Processing 基础

软件界面：工具栏、文本编辑器、控制台

常用函数：

- setup()和 draw()
- size()
- frameRate()
- point()，line()，rect()，ellipse()
- background()，fill()，stroke()
- smooth()

变量（int、float、String）

运算符、判断、循环

一些常量：mouseX，mouseY，pmouseX，pmouseY，mousePressed，mouseButton

多媒体：图片、字体

函数、对象（class，构造函数、成员变量、成员函数）、数组

学习更多更详细

http://zhanghonglun.cn/blog/tag/processing/

31 实战 上海地铁的一天动态可视化

成果展示

https://gold.xitu.io/post/583a43eaac502e006ea02d64

项目代码

链接: https://pan.baidu.com/s/1hrHIGr6 密码: ny7j

代码时间

开始动手

接下来

1. 选择你感兴趣的数据
2. 打开你的脑洞
3. 创造出独具特色的 Processing 可视化

32 机器学习 明白一些基本概念

什么是机器学习

研究如何通过计算的手段，利用经验来改善系统自身的性能

通俗来讲，让代码学着干活

- 特征：自变量
- 标签：因变量

学习的种类

- 有监督学习：提供标签，分类、回归
- 无监督学习：无标签，聚类
- 增强学习 :也称强化学习 ,马尔科夫决策过程(Markov Decision Processes ,MDP)
- 主动学习：边学习边标注
- 迁移学习：从一个域（Domain）迁移（Transfer）到另一个域
- 集成学习：Ensemble，三个臭皮匠赛个诸葛亮，Boosting 和 Bagging

两大痛点

- 维度灾难：数据量和特征数
- 过拟合：模型泛化能力

学习的流程

- 预处理：数据重塑、缺失值处理（补全、统计为缺失特征）
- 特征工程：特征没做好，参数调到老。在已有的特征上生成新的特征，数值、类别
- 特征选择、降维：基于 MIC、Pearson 相关系数、正则化方法、模型，PCA、tSNE
- 训练模型、调参：单模型，多模型融合，集成
- 评估模型 :正确率 (Acurracy)、准确值 (Pecision)、召回值 (Recall)、F 值、AUC

代码实现

你需要的都在这里：http://scikit-learn.org/

33 机器学习 常用经典模型及其实现

常用经典模型

- 线性回归：有监督回归，y=WX+b，X 为 m 维向量，y、b 为 n 维向量，W 为 n*m 维矩阵
- Logistic 回归：有监督回归，y=logit(WX+b)
- 贝叶斯：有监督分类，最可能的分类是概率最大的分类
- k 近邻：有监督分类，kNN，距离的定义
- 决策树：有监督分类，树形判断分支，非线形边界，+集成=随机森林
- 支持向量机：有监督分类，将原空间变换到另一空间，在新空间里寻找 margin 最大的分界面（hyperplane）
- k-means：无监督聚类，初始化中心，不断迭代，EM 算法
- 神经网络：有监督和无监督都有，详情参见下一章，深度学习

实现之前的准备

安装 scikit-learn：http://scikit-learn.org/

sklearn、numpy

>>> from sklearn import svm >>> X = [[0, 0], [1, 1]] >>> y = [0, 1] >>> clf = svm.SVC() >>> clf.fit(X, y) SVC(C=1.0, cache_size=200, class_weight=None, coef0=0.0, decision_function_shape=None, degree=3, gamma='auto', kernel='rbf', max_iter=-1, probability=False, random_state=None, shrinking=True, tol=0.001, verbose=False)

>>> clf.predict([[2., 2.]]) array([1])

一个简单的例子：http://scikit-learn.org/stable/auto_examples/svm/plot_iris.html#sphx-glr-auto-examples-svm-plot-iris-py

34 机器学习 调参比赛大杀器 XGBoost

为什么要调参

步骤，大家都熟；模型，大家都懂；方法，大家都会

所以，调参很重要！一套不同的参数，最后的结果可能千差万别

什么是 XGBoost

eXtreme Gradient Boosting，Gradient Boosting 算法的一种升级版

安装：https://xgboost.readthedocs.io/en/latest/build.html

以 mac os 为例，找一个地方，例如桌面上

1. 编译：
    a. git clone --recursive https://github.com/dmlc/xgboost
    b. cd xgboost
    c. cp make/minimum.mk ./config.mk
    d. make -j4
2. 系统级别安装：
    . cd python-package
    a. sudo python setup.py install
3. 删除安装文件

XGBoost 的参数

- General Parameters：
    - booster：所使用的模型，gbtree 或 gblinear
    - silent：1 则不打印提示信息，0 则打印，默认为 0
    - nthread：所使用的线程数量，默认为最大可用数量
- Booster Parameters（gbtree）:
    - eta：学习率，默认初始化为 0.3，经多轮迭代后一般衰减到 0.01 至 0.2
    - min_child_weight：每个子节点所需的最小权重和，默认为 1
    - max_depth：树的最大深度，默认为 6，一般为 3 至 10
    - max_leaf_nodes：叶结点最大数量，默认为 2^6
    - gamma：拆分节点时所需的最小损失衰减，默认为 0
    - max_delta_step：默认为 0
    - subsample：每棵树采样的样本数量比例，默认为 1，一般取 0.5 至 1
    - colsample_bytree：每棵树采样的特征数量比例，默认为 1，一般取 0.5 至 1
    - colsample_bylevel：默认为 1

    - lambda：L2 正则化项，默认为 1
    - alpha：L1 正则化项，默认为 1
    - scale_pos_weight：加快收敛速度，默认为 1
- Learning Task Parameters：
    - objective：目标函数，默认为 reg:linear，还可取 binary:logistic、multi:softmax、multi:softprob
    - eval_metric：误差函数，回归默认为 rmse，分类默认为 error，其他可取值包括 rmse、mae、logloss、merror、mlogloss、auc
    - seed：随机数种子，默认为 0

来一个 XGBoost 调参实例：https://github.com/Honlan/fullstack-data-engineer/tree/master/data/Parameter_Tuning_XGBoost_with_Example

补充阅读：https://www.analyticsvidhya.com/blog/2016/03/complete-guide-parameter-tuning-XGBoost-with-codes-python/

35 实战 微额借款用户人品预测

做什么

竞赛主页：微额借款用户人品预测大赛

通过数据挖掘来分析小额微贷申请借款用户的信用状况，以分析其是否逾期

数据在这里，链接: https://pan.baidu.com/s/1b2WZnS 密码: crka

数据来源于 CashBUS 现金巴士赞助的微额借款用户人品预测大赛，经 CashBUS 授权使用

冠军团队：不得直视本王

参考资料：https://github.com/wepe/DataCastle-Solution

数据概况

- train_x.csv：训练数据特征，共 1138 维特征（1045 为数值，93 为类别），15000 行
- train_y.csv：训练数据标签，1 为正常，0 为有问题，二分类
- test_x.csv：测试数据特征，共 1138 维特征，5000 行，待分类
- train_unlabeled.csv：无标签训练数据，共 1138 维特征，50000 行
- features_type.csv：1138 维特征的类型说明，数值 or 类别

缺失值处理

统计 train_x、test_x、train_unlabeled 中，1138 维特征的缺失情况并绘图

train_x

test_x

train_unlabeled

缺失值数量区间化，去掉缺失值数量大于 194 的行（可能引入噪声，造成过拟合）

特征工程

1. 排序特征：基于 7W 原始数据，对数值特征排序，得到 1045 维排序特征
2. 离散特征：将排序特征区间化（等值区间化、等量区间化），这里采用等量区间化为 1-10，得到 1045 维离散特征
3. 计数特征：统计每一行中，离散特征 1-10 的个数，得到 10 维计数特征
4. 类别特征编码：将 93 维类别特征用 one-hot 编码
5. 交叉特征：特征之间两两融合，x+y、x-y、x*y、x^2+y^2 等，由于时间复杂度较高，暂时跳过

特征选择

基于 XGBoost，在训练模型时，对特征重要性进行排序，以进行特征选择

模型设计

1 单模型

XGBoost、SVM 等，0.717

2 Bagging of XGBoost

36 个 XGBoost 模型：

- 特征多样：保留 topN1 个原始特征、topN2 个排序特征、topN3 个离散特征、10 个计数特征，N1、N2、N3 分别在 300-500、300-500、64-100 的范围内随机选择
- 模型多样：XGBoost 的各项参数在经调优的最佳值附近小范围抖动
- 融合！0.725

3 多模型融合

XGboost 的 Py、R、Java 版本，BoX，SVM，加权融合，0.7279

4 迭代半监督

用最好的模型预测无标签数据，并保留融合后能提升性能的数据

5 暴力半监督

- 每次从无标签数据中无放回选择 10 条，共有 2^10=1024 种可能的标签，保留融合后性能最好的一组标签，从而获得 5000 组即 5W 条标注数据
- 取 5000 组中的 top500 共 5000 条，每次选择 20-50 条，保留融合后能提升性能的选择，得到最终模型，0.7341

更加详细的内容

全过程演示+手敲代码，课程筹备中，敬请期待！

36 深度学习 揭开 DL 的神秘面纱

什么是深度学习

深度学习=深度神经网络+机器学习

人工智能 > 机器学习 > 表示学习 > 深度学习

神经元模型

输入信号、加权求和、加偏置、激活函数、输出

全连接层

输入信号、输入层、隐层（多个神经元）、输出层（多个输出，每个对应一个分类）、目标函数（交叉熵）

待求的参数：连接矩阵 W、偏置 b

训练方法：随机梯度下降，BP 算法（后向传播）

Python 中深度学习实现：Keras

官网：https://keras.io/

安装：pip install Keras

优点：高度集成和封装，上手快、使用方便

内容：Model、Layer、Objective、Metric、Optimizer、Activation、Initialization、Regularizer

全连接层：Dense

37 深度学习 用于处理图像的 CNN

什么是 CNN

Covolutional Neural Network，卷积神经网络

卷积是指将一些数线性加权，卷起来

一维卷积：

- 三个数 a1、a2、a3
- 权值 w1、w2、w3
- 卷起来，w1*a1+w2*a2+w3*a3
- 卷积窗口大小为 3

二维卷积：

- 九个数 a11、a12、a13、a21、a22、a23、a31、a32、a33
- 权值 w11、w12、w13、w21、w22、w23、w31、w32、w33
- 卷起来，
  w11*a11+w12*a12+w13*a13+w21*a21+w22*a22+w23*a23+w31*a31+w32*a32+w33*a33
- 卷积窗口大小为 3*3

所以，卷积的本质，是进行滑动的融合（一维沿着一个方向滑动，二维沿着两个方向滑动）

CNN 的核心

- 局部连接：仅卷积的部分连接起来，而不像全连接层那样，下一层的每个神经元都和上一层的每个神经元相连
- 权值共享：每一个卷积层（filter）所用的权值是相同的

看懂以下的例子，你就懂 CNN 了：
卷积用以融合和抽象，子采样用以提取

CNN 通用套路

1. 原始数据：二维
2. 卷积、子采样、卷积、子采样……

3. 接上全连接层
4. 接上分类层，输出

Keras 中的实现

一维卷积：Convolution1D

二维卷积：Convolution2D

池化层：MaxPooling1D、MaxPooling2D、AveragePooling1D、AveragePooling2D

38 深度学习 用于处理序列的 RNN

什么是 RNN

Recurrent Neural Network，循环神经网络

还有一个东西叫 RecNN，Recursive Neural Network，递归神经网络，感兴趣的童鞋自行搜索

输入长什么样？

- 一个序列（例如一句话）
- 每个元素都是一个向量（例如每个词的 WordEmbedding）

循环是指，网络只有一层（全连接层），且其隐态 h(t)取决于：

- 当前时刻的输入 x(t)
- 上一时刻的隐态 h(t-1)
- h(t)=sigmoid(Wx(t)+Uh(t-1))

LSTM

Long Short-Term Memory，长短时记忆

解决梯度爆炸和梯度消失问题，学习长程依赖

模型更复杂、参数更多、学习能力更强

RNN 通用套路

- 整理输入数据，每句文本处理成固定长度，做 Embedding
- 将一句话的每个词逐一输入到 RNN 中，得到每一步的输出
- 最后一步的输出可以视为整句话的一个融合
- 接上分类器，输出
- 可以和 CNN 结合，应用：看图说话、VQA 等

Keras 中的实现

SimpleRNN、GRU、LSTM

39 实战 多种手写数字识别模型

手写数字数据集

MNIST：http://yann.lecun.com/exdb/mnist/

训练集 6W，测试集 1W，特征 28*28 的黑白像素点，标签 0-9

Keras Examples

https://github.com/fchollet/keras，examples 文件夹下

读取数据

Keras 里面的 datasets 中已经准备好了 mnist 数据集，直接使用即可

实践

讲三个代码：

- 全连接层
- CNN
- IRNN

如果感兴趣的话，把 Keras 文档（https://keras.io/）完整读一遍，把提供的 examples 逐个试一遍

40 PPT 我把故事讲给你听

为什么要做 PPT

做了好的工作还不够，更重要的是分享给别人看

试想一下：

- 你做了很多工作，现在参加决赛答辩
- 下面坐了一排评委，以及满满一大片的围观群众
- 你只有 10 分钟，如何尽可能完整地展示你的工作，抓住评委和观众的眼球，并得到肯定？

讲一个引人入胜的故事

首先应当有内容：

- 完整的故事框架，问题背景、问题痛点、应用场景、需求分析、理论核心、技术实现、商业价值、可行性分析
- 一个贯穿始终的核心，避免堆积工作量

如何做一个美观的 PPT

- 字体：避免一些看着很不舒服很 low 的字体
- 颜色：字体、背景、图形，颜色搭配

- 元素：页面上不要出现太多文字，适量地搭配一些 icon
- 布局：布局错落有致，元素间注意对齐

我的进步

从 SODA 到 拍拍贷魔镜杯、天池公益云图 和 上海 BOT

多么痛的领悟

术业有专攻，大全栈还是太难，找个设计师合作吧！

对比一下设计妹子美化之后的 SODA，相形见绌、简直汗颜