论使用python实现推荐引擎的优越性

Zalora South East Asia 架构师 翁伟

关于我

- 汕头人
- 常驻新加坡
- .net、python、GO
  - 对python感情非常深厚！
- 新浪微博 @Wuvist 吃货慎关注！

Zalora South East Asia

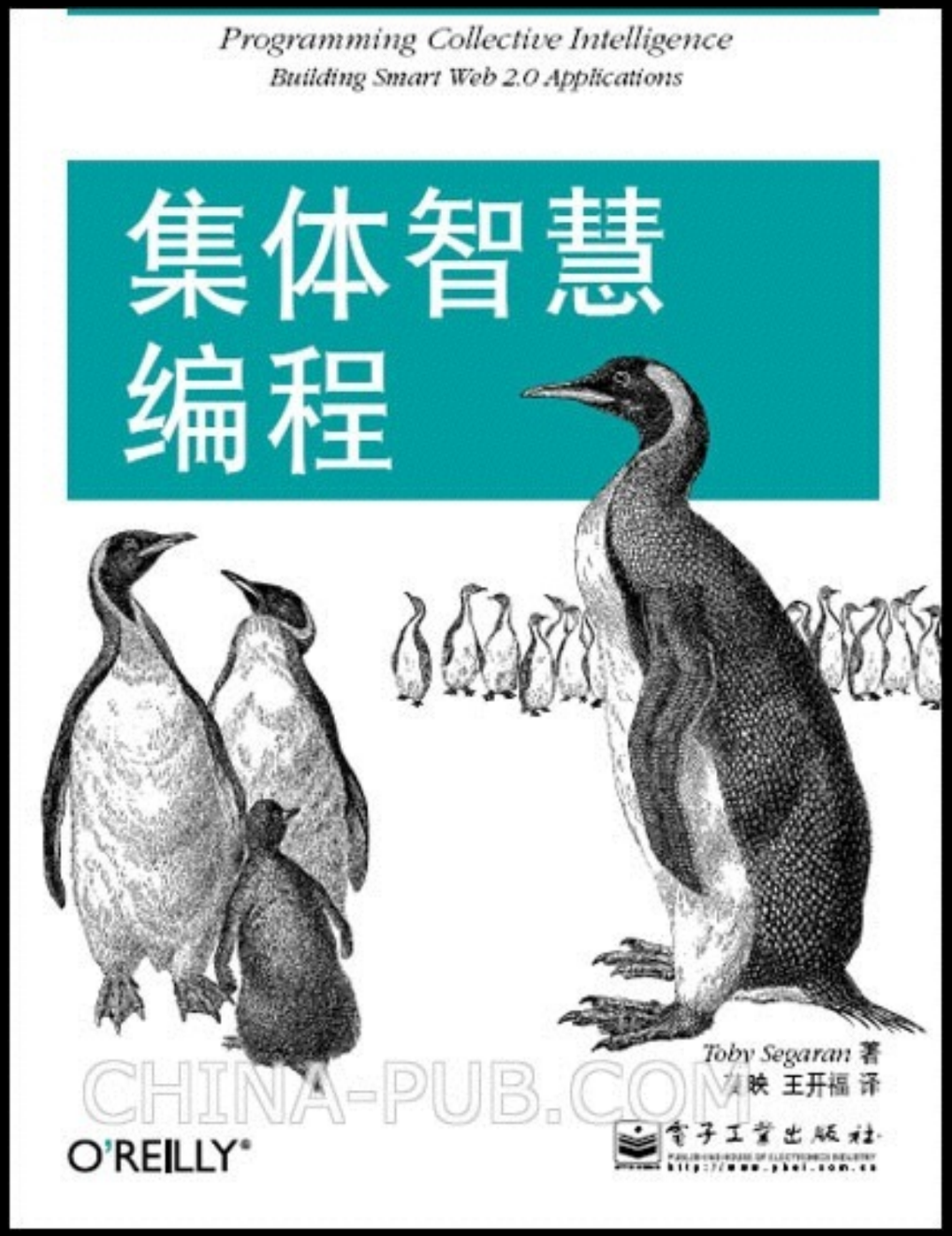
- 2012年初成立
- 东南亚时尚电商
  - 新加坡（总部）马来西亚 印尼 香港 越南 泰国 菲律宾 文莱
- 约1000名员工（IT部门约40人）

推荐引擎

- 今年上半年以试验性（20%时间）项目开发
- **最终**使用了python

学习的优越性

- 协同过滤、推荐、聚类、甚至搜索都有介绍
- 以python实现为例



资源的优越性

- phpserialize
- tornado
- supervisord
- ....

调试的优越性

- IPython
  - http://ipython.org
- 交互式的编码、调试

IPy IPython Dashboard | IPy spectrogram

127.0.0.1:8888/a5222740-848b-4ac1-b212-d732c9f8f78b

IP[y]: Notebook spectrogram Last saved: Mar 07 11:14 PM

File Edit View Insert Cell Kernel Help

Markdown

Simple spectral analysis

An illustration of the Discrete Fourier Transform

$$X_k = \sum_{n=0}^{N-1} x_n e^{-\frac{2\pi i}{N} kn} \qquad k = 0, \ldots, N-1$$

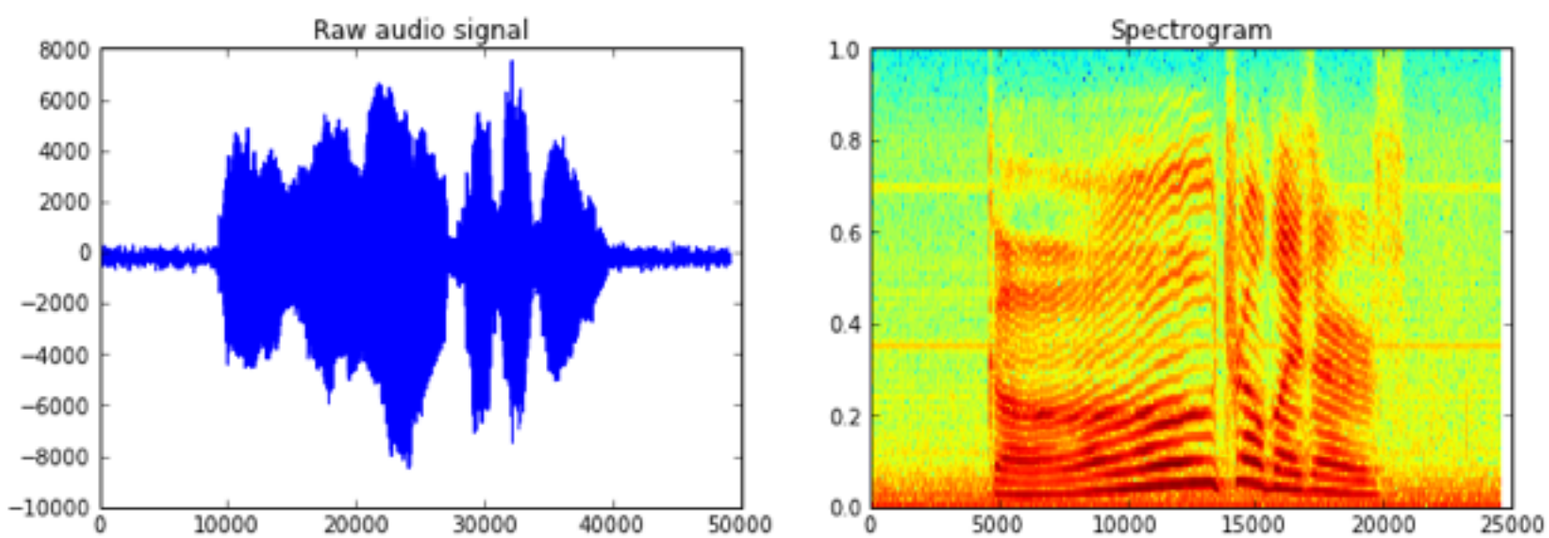
using windowing, to reveal the frequency content of a sound signal.

We begin by loading a datafile using SciPy's audio file support:

In [1]:
```
from scipy.io import wavfile
rate, x = wavfile.read('test_mono.wav')
```

And we can easily view its spectral structure using matplotlib's builtin `specgram` routine:

In [2]:
```
fig, (ax1, ax2) = plt.subplots(1, 2, figsize=(12, 4))
ax1.plot(x); ax1.set_title('Raw audio signal')
ax2.specgram(x); ax2.set_title('Spectrogram');
```



性能的优越性

- pandas
  - http://pandas.pydata.org
  - 高级的数据操作抽象
  - 成组的数据计算极其快
  - 与IPython完美结合

最终使用了python

一开始用的是groovy...

来自Groovy

- 推荐计算极其慢
  - 分析20万订单需要5个小时

我是会Python的

- 花了一晚上重新把核心用PY重写
- 果然很快
  - 两分钟搞定

用PY推倒重来的理由

- 我不熟Groovy？
- Python比Groovy快
  - 为什么Groovy慢？

```
class FastEval {
    static scripts = [:]

    def synchronized static run(data, String scriptText) {
        def script = getScript(scriptText)
        def binding = script.getBinding()
        def util = data["util"]
        def keyPrefix = data["keyPrefix"]
        def tag = data["tag"]

        Closure addScore = { key, target, score ->
            return util.addScore(keyPrefix + key, target, score, tag)
        }
        addbinding(binding, data)
        binding.setVariable("addScore", addScore)

        return script.run()
    }
```

- Groovy/JVM版本比PY快50%
- （当然瓶颈不在此）

pandas

- py版本快3倍 （1小时 => 20分钟）
- （当然，也有架构上的优化）

IPython + pandas demo

兼容的优越性

- 对windows友好 （请批判的理解）
  - hg / IPython / pandas等等

部署的优越性

- vitualenv并不足够
  - 不同应用依赖不同版本的python肿么破?
- pyenv
  - https://github.com/yyuu/pyenv

zsh — 105×23

```
~$ pyenv versions
* 2.7.3 (set by /Users/yyuu/.pyenv/version)
  3.2.3
  jython-2.5.3
  pypy-1.9
~$ python --version
Python 2.7.3
~$ pyenv global pypy-1.9
~$ python --version
Python 2.7.2 (341e1e3821ff, Jun 07 2012, 15:42:54)
[PyPy 1.9.0 with GCC 4.2.1]
~$ cd /Volumes/geishatokyo/lowrise
/Volumes/geishatokyo/lowrise(master)$ pyenv version
3.2.2 (set by /Volumes/geishatokyo/.pyenv-version)
/Volumes/geishatokyo/lowrise(master)$ python --version
Python 3.2.3
/Volumes/geishatokyo/lowrise(master)$
```

@laiyonghao

#python#这两年有点尴尬，搞 web 干不过#php#，搞socket干不过#go#，搞游戏逻辑干不过#lua#，搞群众基础干不过#java#，搞windows应用干不过#csharp#，要肿么破？！

11月19日 21:21 来自新浪微博 | 举报 (3) | 转发(112) | 收藏 | 评论(39)

搞推荐引擎很有优越性！

- 单机可容纳所有数据
- 适合中小公司快速实现
- 利用20%时间开发推荐引擎
  - http://www.infoq.com/cn/presentations/twenty-percent-time-to-develop-recommendation-engine

谢谢～

wuvist@gmail.com