A Summary of Homework

耿新鹏

知识图谱学习小组一期

2016.7.10

自我介绍

字是不是不够大？ ;-)

耿新鹏

Github:xpgeng

坐标：大连

学生党

本科：数学

研究僧:运筹学与控制论

目录

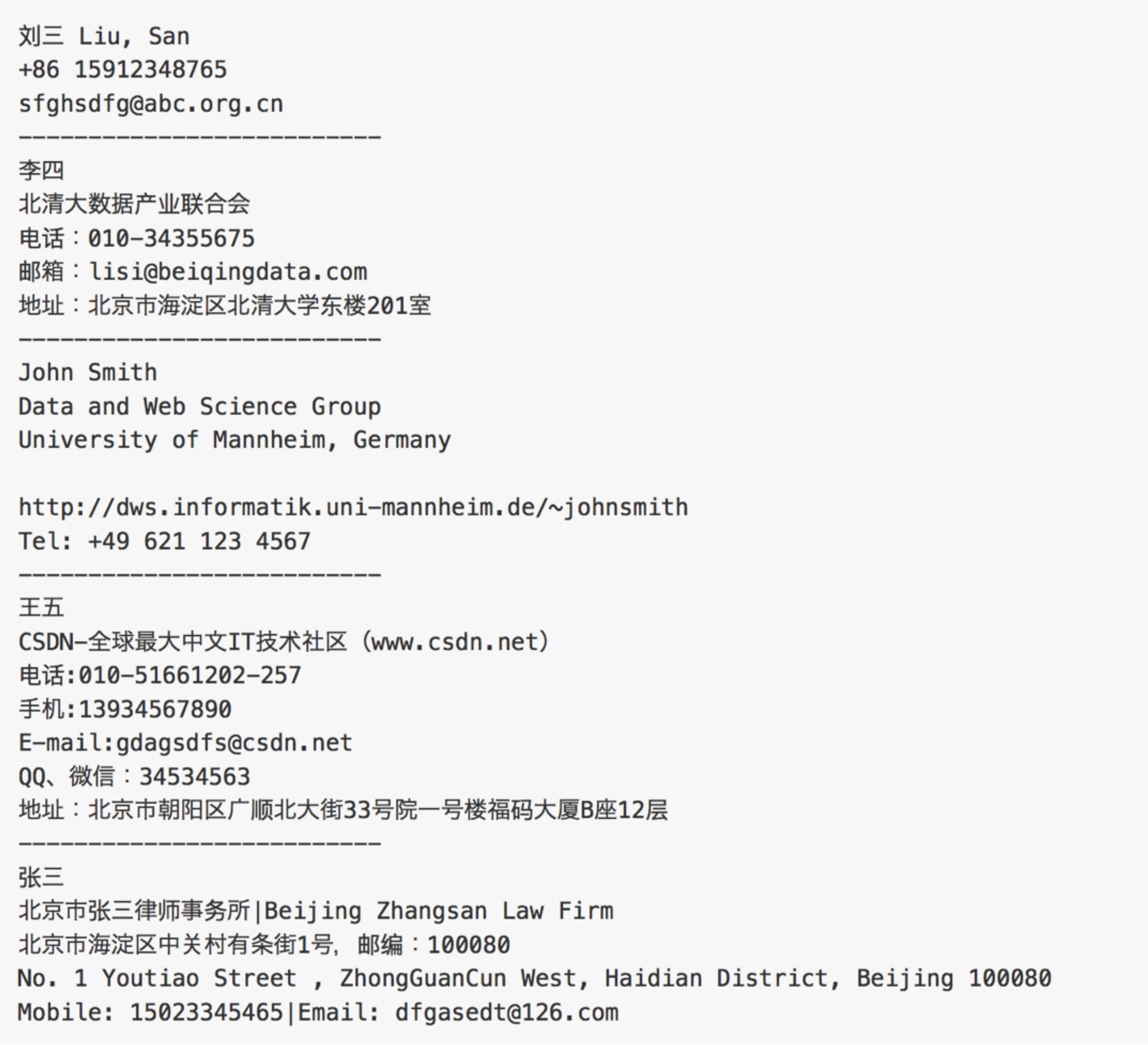
- 提取签名档
- 设计电子邮件的结构化表示
- 用PostgreSQL存储自己的电子邮件JSON

提取签名档

- Homework: 综合分词工具和正则表达式提取邮件签名档
    - 从5个签名档中提取关键字段
    - W3C邮件集

5个真实签名档

- 从中提取关键字段
  - 姓名
  - 单位
  - 电话号码
  - 电子邮件



```
刘三 Liu, San
+86 15912348765
sfghsdfg@abc.org.cn
--------------------------
李四
北清大数据产业联合会
电话：010-34355675
邮箱：lisi@beiqingdata.com
地址：北京市海淀区北清大学东楼201室
--------------------------
John Smith
Data and Web Science Group
University of Mannheim, Germany

http://dws.informatik.uni-mannheim.de/~johnsmith
Tel: +49 621 123 4567
--------------------------
王五
CSDN-全球最大中文IT技术社区（www.csdn.net)
电话:010-51661202-257
手机:13934567890
E-mail:gdagsdfs@csdn.net
QQ、微信：34534563
地址：北京市朝阳区广顺北大街33号院一号楼福码大厦B座12层
--------------------------
张三
北京市张三律师事务所|Beijing Zhangsan Law Firm
北京市海淀区中关村有条街1号，邮编：100080
No. 1 Youtiao Street , ZhongGuanCun West, Haidian District, Beijing 100080
Mobile: 15023345465|Email: dfgasedt@126.com
```

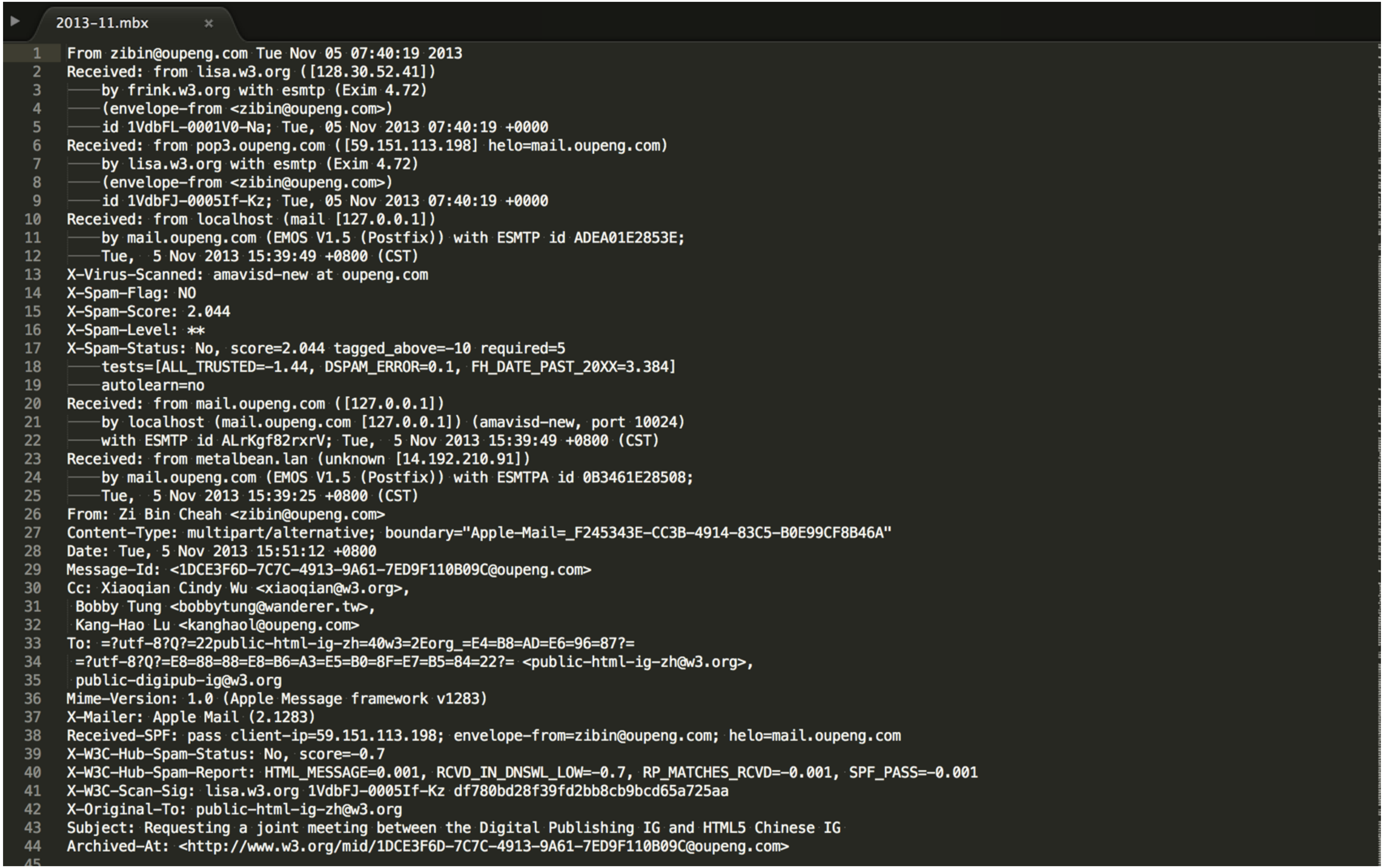
坑...

- 中英文签名档混杂
  - 使用函数: isalpha( )
- 中文单位名提取
  - jieba + user_dict

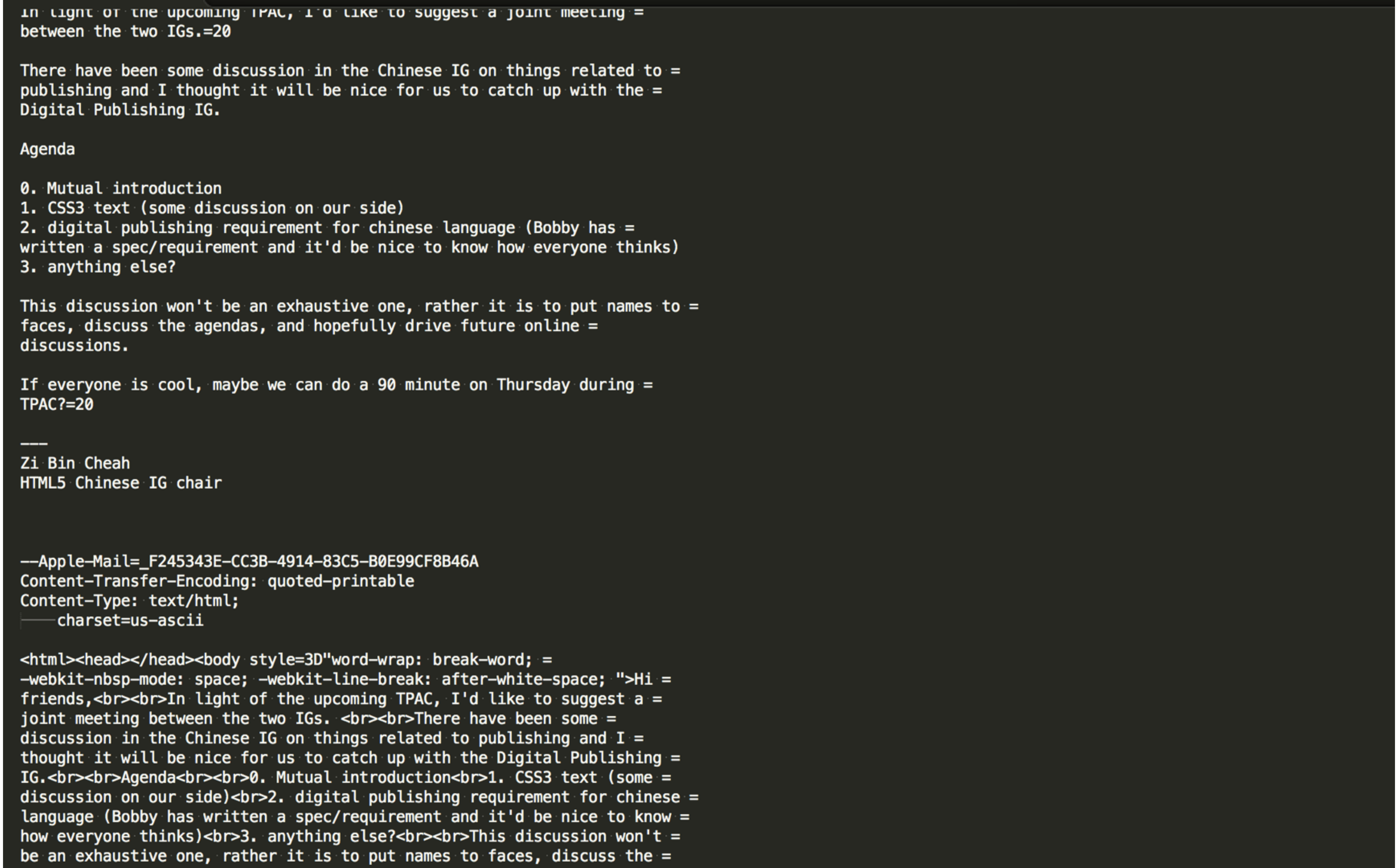
思路

- 判断中英文签名档, 提取姓名, 单位
    - 中文使用Jieba
    - 英文使用NLTK
- 使用Regex提取电话, 邮箱, 邮编等信息

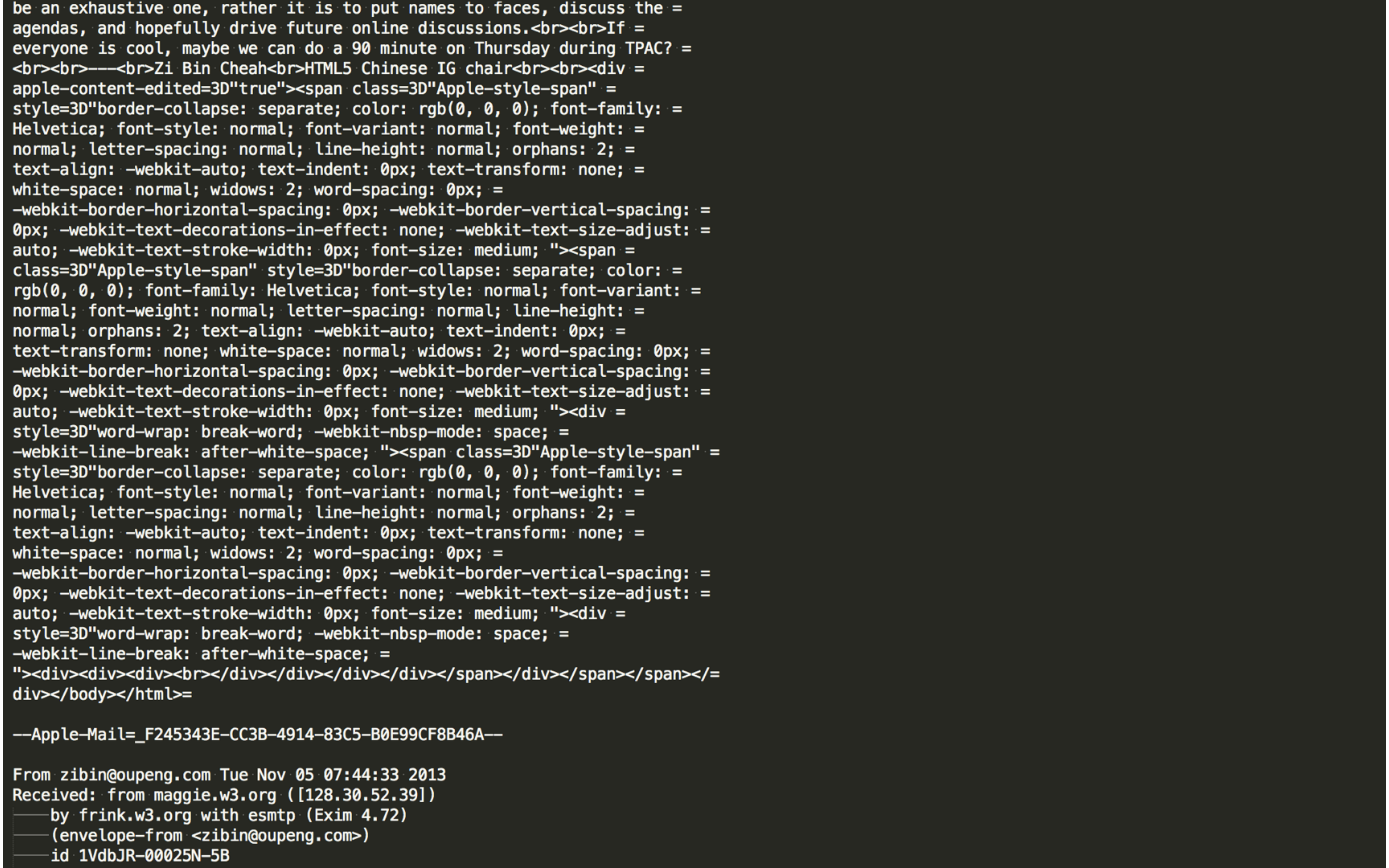
W3C邮件集



```
From zibin@oupeng.com Tue Nov 05 07:40:19 2013
Received: from lisa.w3.org ([128.30.52.41])
    by frink.w3.org with esmtp (Exim 4.72)
    (envelope-from <zibin@oupeng.com>)
    id 1VdbFL-0001V0-Na; Tue, 05 Nov 2013 07:40:19 +0000
Received: from pop3.oupeng.com ([59.151.113.198] helo=mail.oupeng.com)
    by lisa.w3.org with esmtp (Exim 4.72)
    (envelope-from <zibin@oupeng.com>)
    id 1VdbFJ-0005If-Kz; Tue, 05 Nov 2013 07:40:19 +0000
Received: from localhost (mail [127.0.0.1])
    by mail.oupeng.com (EMOS V1.5 (Postfix)) with ESMTP id ADEA01E2853E;
    Tue,  5 Nov 2013 15:39:49 +0800 (CST)
X-Virus-Scanned: amavisd-new at oupeng.com
X-Spam-Flag: NO
X-Spam-Score: 2.044
X-Spam-Level: **
X-Spam-Status: No, score=2.044 tagged_above=-10 required=5
    tests=[ALL_TRUSTED=-1.44, DSPAM_ERROR=0.1, FH_DATE_PAST_20XX=3.384]
    autolearn=no
Received: from mail.oupeng.com ([127.0.0.1])
    by localhost (mail.oupeng.com [127.0.0.1]) (amavisd-new, port 10024)
    with ESMTP id ALrKgf82rxrV; Tue,  5 Nov 2013 15:39:49 +0800 (CST)
Received: from metalbean.lan (unknown [14.192.210.91])
    by mail.oupeng.com (EMOS V1.5 (Postfix)) with ESMTPA id 0B3461E28508;
    Tue,  5 Nov 2013 15:39:25 +0800 (CST)
From: Zi Bin Cheah <zibin@oupeng.com>
Content-Type: multipart/alternative; boundary="Apple-Mail=_F245343E-CC3B-4914-83C5-B0E99CF8B46A"
Date: Tue, 5 Nov 2013 15:51:12 +0800
Message-Id: <1DCE3F6D-7C7C-4913-9A61-7ED9F110B09C@oupeng.com>
Cc: Xiaoqian Cindy Wu <xiaoqian@w3.org>,
 Bobby Tung <bobbytung@wanderer.tw>,
 Kang-Hao Lu <kanghaol@oupeng.com>
To: =?utf-8?Q?=22public-html-ig-zh=40w3=2Eorg_=E4=B8=AD=E6=96=87?=
 =?utf-8?Q?=E8=88=88=E8=B6=A3=E5=B0=8F=E7=B5=84=22?= <public-html-ig-zh@w3.org>,
 public-digipub-ig@w3.org
Mime-Version: 1.0 (Apple Message framework v1283)
X-Mailer: Apple Mail (2.1283)
Received-SPF: pass client-ip=59.151.113.198; envelope-from=zibin@oupeng.com; helo=mail.oupeng.com
X-W3C-Hub-Spam-Status: No, score=-0.7
X-W3C-Hub-Spam-Report: HTML_MESSAGE=0.001, RCVD_IN_DNSWL_LOW=-0.7, RP_MATCHES_RCVD=-0.001, SPF_PASS=-0.001
X-W3C-Scan-Sig: lisa.w3.org 1VdbFJ-0005If-Kz df780bd28f39fd2bb8cb9bcd65a725aa
X-Original-To: public-html-ig-zh@w3.org
Subject: Requesting a joint meeting between the Digital Publishing IG and HTML5 Chinese IG
Archived-At: <http://www.w3.org/mid/1DCE3F6D-7C7C-4913-9A61-7ED9F110B09C@oupeng.com>
```



```
In light of the upcoming TPAC, I'd like to suggest a joint meeting =
between the two IGs.=20

There have been some discussion in the Chinese IG on things related to =
publishing and I thought it will be nice for us to catch up with the =
Digital Publishing IG.

Agenda

0. Mutual introduction
1. CSS3 text (some discussion on our side)
2. digital publishing requirement for chinese language (Bobby has =
written a spec/requirement and it'd be nice to know how everyone thinks)
3. anything else?

This discussion won't be an exhaustive one, rather it is to put names to =
faces, discuss the agendas, and hopefully drive future online =
discussions.

If everyone is cool, maybe we can do a 90 minute on Thursday during =
TPAC?=20

---
Zi Bin Cheah
HTML5 Chinese IG chair



--Apple-Mail=_F245343E-CC3B-4914-83C5-B0E99CF8B46A
Content-Transfer-Encoding: quoted-printable
Content-Type: text/html;
	charset=us-ascii

<html><head></head><body style=3D"word-wrap: break-word; =
-webkit-nbsp-mode: space; -webkit-line-break: after-white-space; ">Hi =
friends,<br><br>In light of the upcoming TPAC, I'd like to suggest a =
joint meeting between the two IGs. <br><br>There have been some =
discussion in the Chinese IG on things related to publishing and I =
thought it will be nice for us to catch up with the Digital Publishing =
IG.<br><br>Agenda<br><br>0. Mutual introduction<br>1. CSS3 text (some =
discussion on our side)<br>2. digital publishing requirement for chinese =
language (Bobby has written a spec/requirement and it'd be nice to know =
how everyone thinks)<br>3. anything else?<br><br>This discussion won't =
be an exhaustive one, rather it is to put names to faces, discuss the =
```



```
be an exhaustive one, rather it is to put names to faces, discuss the =
agendas, and hopefully drive future online discussions.<br><br>If =
everyone is cool, maybe we can do a 90 minute on Thursday during TPAC? =
<br><br>---<br>Zi Bin Cheah<br>HTML5 Chinese IG chair<br><br><div =
apple-content-edited=3D"true"><span class=3D"Apple-style-span" =
style=3D"border-collapse: separate; color: rgb(0, 0, 0); font-family: =
Helvetica; font-style: normal; font-variant: normal; font-weight: =
normal; letter-spacing: normal; line-height: normal; orphans: 2; =
text-align: -webkit-auto; text-indent: 0px; text-transform: none; =
white-space: normal; widows: 2; word-spacing: 0px; =
-webkit-border-horizontal-spacing: 0px; -webkit-border-vertical-spacing: =
0px; -webkit-text-decorations-in-effect: none; -webkit-text-size-adjust: =
auto; -webkit-text-stroke-width: 0px; font-size: medium; "><span =
class=3D"Apple-style-span" style=3D"border-collapse: separate; color: =
rgb(0, 0, 0); font-family: Helvetica; font-style: normal; font-variant: =
normal; font-weight: normal; letter-spacing: normal; line-height: =
normal; orphans: 2; text-align: -webkit-auto; text-indent: 0px; =
text-transform: none; white-space: normal; widows: 2; word-spacing: 0px; =
-webkit-border-horizontal-spacing: 0px; -webkit-border-vertical-spacing: =
0px; -webkit-text-decorations-in-effect: none; -webkit-text-size-adjust: =
auto; -webkit-text-stroke-width: 0px; font-size: medium; "><div =
style=3D"word-wrap: break-word; -webkit-nbsp-mode: space; =
-webkit-line-break: after-white-space; "><span class=3D"Apple-style-span" =
style=3D"border-collapse: separate; color: rgb(0, 0, 0); font-family: =
Helvetica; font-style: normal; font-variant: normal; font-weight: =
normal; letter-spacing: normal; line-height: normal; orphans: 2; =
text-align: -webkit-auto; text-indent: 0px; text-transform: none; =
white-space: normal; widows: 2; word-spacing: 0px; =
-webkit-border-horizontal-spacing: 0px; -webkit-border-vertical-spacing: =
0px; -webkit-text-decorations-in-effect: none; -webkit-text-size-adjust: =
auto; -webkit-text-stroke-width: 0px; font-size: medium; "><div =
style=3D"word-wrap: break-word; -webkit-nbsp-mode: space; =
-webkit-line-break: after-white-space; =
"><div><div><div><br></div></div></div></div></span></div></span></span></=
div></body></html>=

--Apple-Mail=_F245343E-CC3B-4914-83C5-B0E99CF8B46A--

From zibin@oupeng.com Tue Nov 05 07:44:33 2013
Received: from maggie.w3.org ([128.30.52.39])
        by frink.w3.org with esmtp (Exim 4.72)
        (envelope-from <zibin@oupeng.com>)
        id 1VdbJR-00025N-5B
```

思路

- 分离成单个邮件
- 提取name list
- 用Regex提取name之后的所有字符
- 过滤: 通过限制每条信息的字符长度在一定范围内(300)
- 使用分词, BeatifulSoup4 等工具进行深入提取

坑来了....

分离邮件

- mbox每封邮件的boundary为: “- - boundary- - ”. 具体boundary信息可以在MIME头找到.
- 但是!
  - - - - - - - - - boundary - -
  - <- - !<>……<>- ->
  - boundary里不仅仅是字母和数字, 还有**下划线**, 比如: “- - sfjsdf123123_ - -“
- 所以, 采用了 “From ….@......” 来分离
  - 简单粗暴……

提取name list

- Why?
  - 发件人姓名跟签名档姓名不一致, 比如: 发件人是 Tom Chen, 签名档是.......陈真... 疯不?
  - 尽管是少数例子, 但逐一处理起来还是很耗时

- 做法:
    - 先通过MIME头提取所有发件人姓名,
    - 挨个简单搜索一下, 确认与signature 相关的name 是否正确
    - 如果不正确, 把正确的添加到name list里.
    - **Remark**: 如果数据集name太多, 那就得放弃这么深入的观察, 直接根据每封邮件的发件人姓名去提取signature, 牺牲掉一部分数据. 具体就看实际需求了.

提取Signature的体会

- Signature形式各异, 想用通用模式全部提取不现实.
- 想尽可能提取就需要花时间观察数据, 但同时也要考虑止损时间.
- 观察数据的时间真的很多.
- filter(None, signature_list)

结构化表示

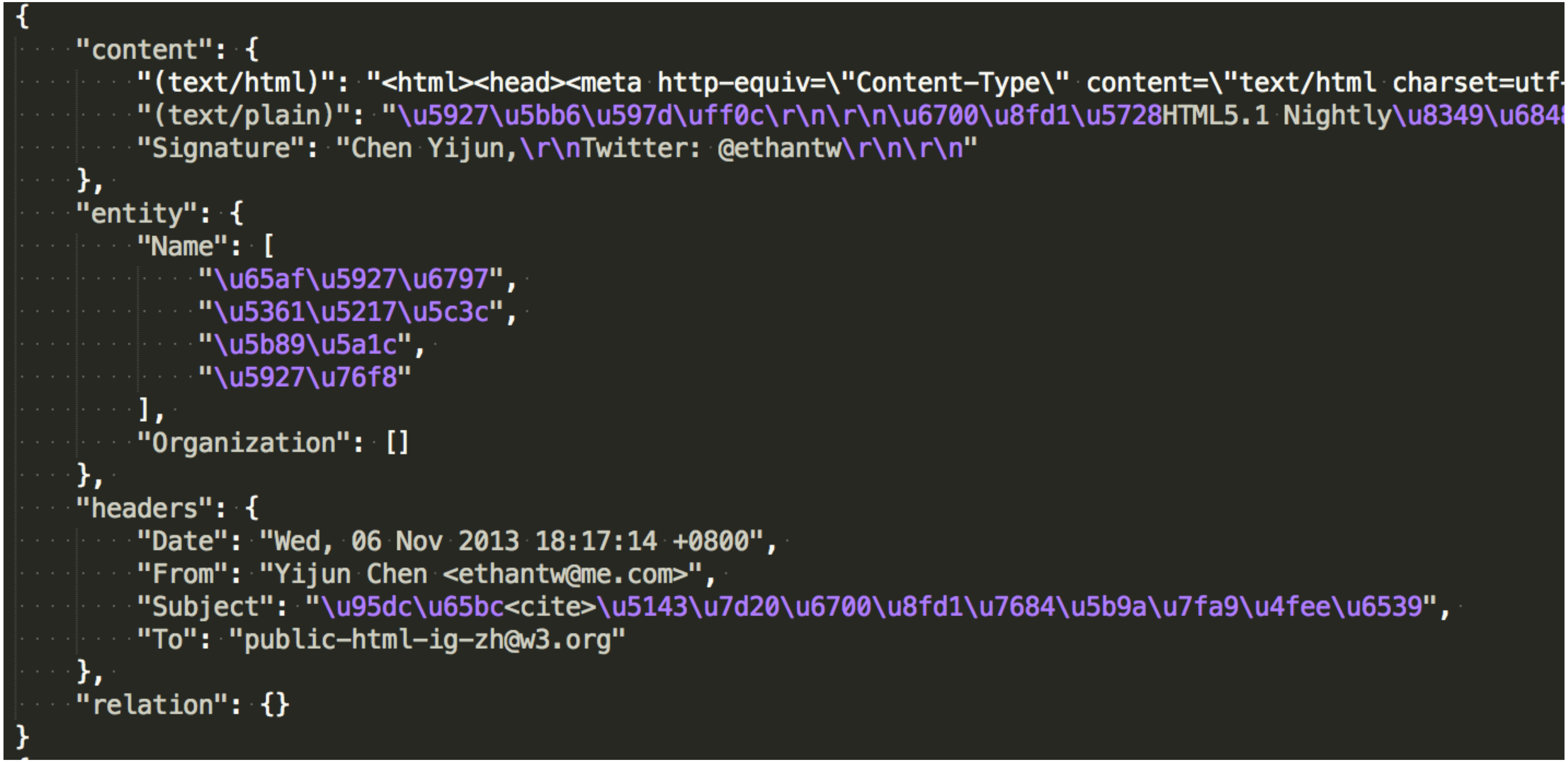
W3C —> JSON

- 将邮件的存储格式分为以下几部分
  - **Headers**: 这部分包含从MIME头提取的信息, 如'From', 'Receive', 'Send_time', ‘Subject'等等
  - **Content**: 这部分会包含所有的内容信息, 如'Body_text', 'Body_html', 'Recite', 'Attachment', ‘Signature'等等
  - **Entities**: 包含邮件中提到的各种实体, 如'Name', 'Organization', 'Time', 'Position', ‘Tel'
  - **Relation**: 包含邮件内的各种关系, 如邮件之间的关系, 邮件内容的语义关系等.

思路

- 层层递进, 由简单到复杂
    - 从MIME头直接提取
    - 到Regex, NLTK 等工具的使用
    - 到关系分析(需要进一步学习…)

结果



```
{
    "content": {
        "(text/html)": "<html><head><meta http-equiv=\"Content-Type\" content=\"text/html charset=utf-
        "(text/plain)": "\u5927\u5bb6\u597d\uff0c\r\n\r\n\u6700\u8fd1\u5728HTML5.1 Nightly\u8349\u684
        "Signature": "Chen Yijun,\r\nTwitter: @ethantw\r\n\r\n"
    },
    "entity": {
        "Name": [
            "\u65af\u5927\u6797",
            "\u5361\u5217\u5c3c",
            "\u5b89\u5a1c",
            "\u5927\u76f8"
        ],
        "Organization": []
    },
    "headers": {
        "Date": "Wed, 06 Nov 2013 18:17:14 +0800",
        "From": "Yijun Chen <ethantw@me.com>",
        "Subject": "\u95dc\u65bc<cite>\u5143\u7d20\u6700\u8fd1\u7684\u5b9a\u7fa9\u4fee\u6539",
        "To": "public-html-ig-zh@w3.org"
    },
    "relation": {}
}
```

- 未解决的问题
    - 邮件内容分段
    - 关系分析
    - NER中organization基本提取不出
    - ……

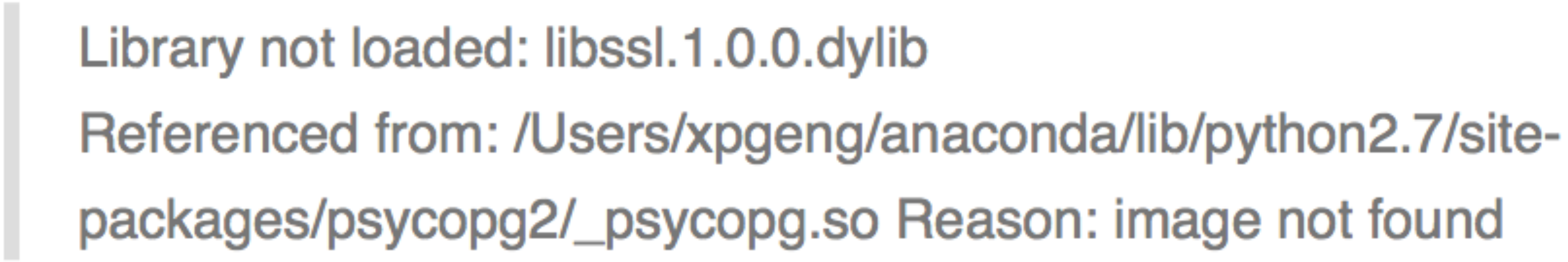
- 收获
    - 感觉分解, 分析整个邮件集, 并以合适的结构呈现出来的过程, 像提炼石油...

咱们工人有力量!

储存W3C-JSON

psycopg2

- 在调用过程中出现了如下错误



> Library not loaded: libssl.1.0.0.dylib
> Referenced from: /Users/xpgeng/anaconda/lib/python2.7/site-packages/psycopg2/_psycopg.so Reason: image not found

- 原因是不能加载 libssl.1.0.0.dylib

- 解决办法: 在`~/.bash_profile`中添加如下代码.
  - export DYLD_FALLBACK_LIBRARY_PATH=$HOME/anaconda/lib/:$DYLD_FALLBACK_LIBRARY_PATH

PostgreSQL References

- PostgreSQL官方文档(这官方文档...呵呵了)
  - [PostgreSql Official Documentation](https://www.postgresql.org/docs/9.5/static/index.html)
  - [Install PostgeSQL on Mac OSX](https://www.postgresql.org/download/macosx/)
  - [GUI Tools of PostgreSQL.app](http://postgresapp.com/documentation/gui-tools.html)
  - [PostgreSql Official Documentation--8.14. JSON Types](https://www.postgresql.org/docs/current/static/datatype-json.html)
  - [PostgreSql Official Documentation--9.15. JSON Functions and Operators](https://www.postgresql.org/docs/current/static/functions-json.html)

- psycopg2
    - [psycopg2.extras](http://initd.org/psycopg/docs/extras.html)
        - Very Useful!!
    - [Frequently Asked Questions](http://initd.org/psycopg/docs/faq.html)
    - [**Querying JSON in Postgres**](http://schinckel.net/2014/05/25/querying-json-in-postgres/)
        - 详细的介绍了querying的语法
- [Psycopg2 Json](http://wiki.zoomquiet.io/pythonic/PsycopgJson)
    - Zoom.Quiet (大妈)写的关于psycopg2与json的文档.

Week4 知识检索

还在学习ing....

Summary

- 从各种工具到知识图谱相关的知识, 掌握的都比较基础, 未来还需要很长时间去学习, 实践.
- 希望接下来参加学习小组的同学务必拿出时间折腾homework
  - 光学习不实践就背离了**从工程角度出发的初衷**.
- 当脏数据在自己手中变得越来越干净, 信息越来越清晰的时候
  - 有快感!

谢谢！