

Psychoinformatics - Week 12 (Exercises)

by 徐舒庭 (b11705018@ntu.edu.tw)

1 進一步研究 Regular Expressions (4 points)

用urllib+re在臺大資工系網頁搜集教師emails.

```
In [ ]: import urllib.request
import re
import base64

url = "https://www.csie.ntu.edu.tw/zh_tw/member/Faculty"

response = urllib.request.urlopen(url)
html_content = response.read().decode('utf-8')

professor_pattern = re.compile(r'<div class="i-member-item col-md-6">(.*?)</div>', re.DOTALL)
professor_matches = re.findall(professor_pattern, html_content)

for professor_match in professor_matches:

    name_pattern = re.compile(r'<span class="i-member-title member-data-title-name">姓名:</span>\s*<span class="i-member-value member-data-value-name"><a href="mailto:
email_pattern = re.compile(r'var e = atob\("(.*?)"\);', re.DOTALL)

    name_match = re.search(name_pattern, professor_match)
    email_match = re.search(email_pattern, professor_match)

    if name_match and email_match:
        name = name_match.group(1)
        encoded_email = email_match.group(1)

        email = base64.b64decode(encoded_email).decode('utf-8')
        print(f"姓名: {name}\nemail: {email}\n")

姓名: 陳祝嵩 (Chu-Song Chen)
email: chusong@csie.ntu.edu.tw

姓名: 鄭卜王 (PJ Cheng)
email: pjcheng@csie.ntu.edu.tw

姓名: 鄭文皇 (Wen-Huang Cheng)
email: wenhuang@csie.ntu.edu.tw

姓名: 林忠緯 (CW Lin)
email: cwlin@csie.ntu.edu.tw

姓名: 張瑞峰 (RF Chang)
email: rfchang@csie.ntu.edu.tw

姓名: 趙坤茂 (KM Chao)
email: kmchao@csie.ntu.edu.tw

姓名: 陳炳宇 (BY Chen)
email: robin@ntu.edu.tw

姓名: 陳健輝 (GH Chen)
email: ghchen@csie.ntu.edu.tw

姓名: 陳信希 (HH Chen)
email: hhchen@csie.ntu.edu.tw

姓名: 陳尚澤 (Shang-Tse Chen)
email: stchen@csie.ntu.edu.tw

姓名: 陳文進 (WC Chen)
email: wcchen@csie.ntu.edu.tw

姓名: 陳彥仰 (Mike Y. Chen)
email: mikechen@csie.ntu.edu.tw

姓名: 陳繼儂 (Yun-Nung Chen)
email: yvchen@csie.ntu.edu.tw

姓名: 鄭龍礪 (Lung-Pan Cheng)
email: lung-pan.cheng@csie.ntu.edu.tw

姓名: 周承復 (CF Chou)
email: ccf@csie.ntu.edu.tw

姓名: 莊永裕 (YY Chuang)
email: cyy@csie.ntu.edu.tw

姓名: 傅立成 (LC Fu)
email: lichen@csie.ntu.edu.tw

姓名: 傅楸善 (CS Fuh)
email: fuh@csie.ntu.edu.tw

姓名: 項潔 (Jieh Hsiang)
email: hsiang@csie.ntu.edu.tw

姓名: 蕭旭君 (Hsu-Chun Hsiao)
email: hchsiao@csie.ntu.edu.tw

姓名: 許永真 (YJ Hsu)
email: yjhsu@csie.ntu.edu.tw

姓名: 徐宏民 (Winston Hsu)
email: whsu@ntu.edu.tw

姓名: 薛智文 (CW Hsueh)
email: cwhsueh@csie.ntu.edu.tw

姓名: 洪一平 (YP Hung)
email: hung@csie.ntu.edu.tw

姓名: 洪士灝 (SH Hung)
email: hungsh@csie.ntu.edu.tw

姓名: 張智星 (JS Jang)
email: jang@csie.ntu.edu.tw

姓名: 郭大維 (TW Kuo)
email: ktw@csie.ntu.edu.tw

姓名: 李允中 (J Lee)
email: jlee@csie.ntu.edu.tw

姓名: 李明穗 (MS Lee)
email: mslee@csie.ntu.edu.tw

姓名: 李彥寰 (Yen-Huan Li)
email: yenhuan.li@csie.ntu.edu.tw

姓名: 黎士璋 (Shih-Wei Li)
email: shihwei@csie.ntu.edu.tw

姓名: 廖世偉 (Shih-wei Liao)
email: liao@csie.ntu.edu.tw

姓名: 林智仁 (CJ Lin)
email: cjlin@csie.ntu.edu.tw

姓名: 林風 (P Lin)
email: plin@csie.ntu.edu.tw

姓名: 林守德 (SD Lin)
email: sdlin@csie.ntu.edu.tw

姓名: 林軒田 (HT Lin)
email: htlin@csie.ntu.edu.tw

姓名: 劉邦鋒 (PF Liu)
email: pangfeng@csie.ntu.edu.tw

姓名: 呂學一 (HI Lu)
email: hil@csie.ntu.edu.tw

姓名: 呂育道 (YD Lyuu)
email: lyuu@csie.ntu.edu.tw

姓名: 歐陽彥正 (YJ Oyang)
email: yjoyang@csie.ntu.edu.tw

姓名: 逢愛君 (AC Pang)
email: acpang@csie.ntu.edu.tw

姓名: 施吉昇 (CS Shih)
email: cshih@csie.ntu.edu.tw

姓名: 蔡欣穆 (HM Tsai)
email: hsinmu@csie.ntu.edu.tw

姓名: 曾宇鳳 (YJ Tseng)
email: yjtseng@csie.ntu.edu.tw

姓名: 吳家麟 (JL Wu)
email: wjl@cmlab.csie.ntu.edu.tw

姓名: 巫芳璟 (Fang-Jing Wu)
email: fangjing@csie.ntu.edu.tw

姓名: 楊佳玲 (CL Yang)
email: yangc@csie.ntu.edu.tw
```

2 進一步研究 Topic Modeling (4 points)

對2.5 Topic Modeling的範例程式文本先過濾掉stop words, 使5個主題能夠更加明確

```
In [ ]: # Please write your codes here
```