

fetch_emails

July 11, 2021

Script for 'fetch_emails' program in Python Licensed under the Apache License, Version 2.0
<http://www.apache.org/licenses/LICENSE-2.0>

The program takes as input an URL and prints the list of emails addresses contained in it.

```
[1]: import re
import urllib.request
```

The function 'emails_list' takes as input an URL and returns the list of emails addresses contained in it:

```
[2]: def emails_list(urlname):

    # Opens urlname and store its content, in utf-8 format, in the string data
    # Does so by passing a fake user agent to the website to avoid the possible
    →error message:
    # 'urllib.error.HTTPError: HTTP Error 403: Forbidden'
    header = {'User-Agent': 'Chrome'}
    url = urllib.request.Request(url=urlname, headers=header)
    data = urllib.request.urlopen(url).read().decode('utf-8')

    # Finds in data the email-type strings including '(at)' or '@' and not
    →finishing by 'html'
    # Discards possible repetitions with set() and orders the results with
    →sorted()
    emails1 = re.findall(r'[\w.-]+\s(at\s)[\w.-]+\s\.(?!html)[a-zA-Z]+' , data)
    emails2 = re.findall(r'[\w.-]+\s@\s[\w.-]+\s\.(?!html)[a-zA-Z]+' , data)
    emails = emails1 + emails2
    return sorted(set(emails))
```

The function 'main' takes as input an URL and returns the list of email addresses found in it:

```
[3]: def main():
    urlname = input("Enter URL where to look for email addresses: ")
    emails = emails_list(urlname)
    print('')
    if emails:
        print('The email addresses in', '"' + urlname + '"', 'are:\n')
        for email in emails:
```

```
        print('-', email)
    else:
        print('No email address found in', '"' + urlname + '".')
```

```
[4]: if __name__ == '__main__':
      main()
```

Enter URL where to look for email addresses:
https://en.wikipedia.org/wiki/Email_address

The email addresses in "https://en.wikipedia.org/wiki/Email_address" are:

- JSmith@example.com
- John..Doe@example.com
- Pelé@example.com
- allowed@example.com
- bar@example.com
- c@example.com
- example-indeed@strange-example.com
- example.com@example.org
- example@s.example
- extension@pobox.com
- foobar@example.com
- fully-qualified-domain@example.com
- i_like_underscore@but_its_not_allowed_in_this_part.example.com
- joeuser@example.com
- john.smith@example.com
- john.smith@example.org
- jsmith@example.com
- l@example.com
- other.email-with-hyphen@example.com
- right@example.com
- simple@example.com
- sorting@example.com
- symbol@example.com
- tag@domain.com
- tag@example.com
- test@test.com
- user-@example.org
- user.name@example.com
- user@example.com
- username@example.org
- very.common@example.com
- x@example.com

```
[ ]:
```