CSCE-689, Programming Assignment #1 SpamLord

Sambartika Guha

UIN: 127003382

Instructions to run the code:

1. Unzip the folder
2. Go inside the folder PA1 from python console
3. Type the following command:
   python SpamLord.py data_dev/dev data_dev/devGOLD

Results and Analysis:

The program SpamLord.py is written to extract phone numbers and email addresses from HTML files or text files or word files or other kind of files. The email addresses and phone numbers may not be written in the same format in every document. It can be written in different formats, it can be a part of a line or it can be embedded in HTML tags. I have written regular expressions which will be able to extract them. The regular expressions can handle cases given below and many other test cases for email addresses

1. huangrh@cse.tamu.edu
2. huangrh(at)cse.tamu.edu
3. huangrh at cse dot tamu dot edu
4. <script type="text/javascript">obfuscate('cse.tamu.edu','huangrh')</script>
5. huangrh WHERE tamu DOM edu
6. d-l-w-h-@-s-t-a-n-f-o-r-d-.-e-d-u

The regular expression can handle cases given below and many other test cases for phone number:

1. Phone: (979) 862-2908
2. Tel (+1): 979-862-2908
3. <a href="contact.html">TEL</a> +1&thinsp;979&thinsp;862&thinsp;2908

The program is able to reduce the number of false positives and the number of false negatives to zero in the given test files. So, it is able to achieve 100% accuracy for the given test scenarios. I have also tried different other test cases for email addresses and phone numbers to improve the performance of the program.

The output of the program is given below:

```
True Positives (59):

set([('ashishg', 'e', 'ashishg@stanford.edu'),
('ashishg', 'e', 'rozm@stanford.edu'),
('ashishg', 'p', '650-723-1614'),
('ashishg', 'p', '650-723-4173'),
('ashishg', 'p', '650-814-1478'),
('balaji', 'e', 'balaji@stanford.edu'),
('bgirod', 'p', '650-723-4539'),
('bgirod', 'p', '650-724-3648'),
('bgirod', 'p', '650-724-6354'),
('cheriton', 'e', 'cheriton@cs.stanford.edu'),
('cheriton', 'e', 'uma@cs.stanford.edu'),
('cheriton', 'p', '650-723-1131'),
('cheriton', 'p', '650-725-3726'),
('dabo', 'e', 'dabo@cs.stanford.edu'),
('dabo', 'p', '650-725-3897'),
('dabo', 'p', '650-725-4671'),
('dlwh', 'e', 'dlwh@stanford.edu'),
('engler', 'e', 'engler@lcs.mit.edu'),
('engler', 'e', 'engler@stanford.edu'),
('eroberts', 'e', 'eroberts@cs.stanford.edu'),
('eroberts', 'p', '650-723-3642'),
('eroberts', 'p', '650-723-6092'),
('fedkiw', 'e', 'fedkiw@cs.stanford.edu'),
('hager', 'e', 'hager@cs.jhu.edu'),
('hager', 'p', '410-516-5521'),
('hager', 'p', '410-516-5553'),
('hager', 'p', '410-516-8000'),
('hanrahan', 'e', 'hanrahan@cs.stanford.edu'),

('hanrahan', 'p', '650-723-0033'),

('hanrahan', 'p', '650-723-8530'),

('horowitz', 'p', '650-725-3707'),

('horowitz', 'p', '650-725-6949'),

('jks', 'e', 'jks@robotics.stanford.edu'),

('jurafsky', 'e', 'jurafsky@stanford.edu'),

('jurafsky', 'p', '650-723-5666'),

('kosecka', 'e', 'kosecka@cs.gmu.edu'),

('kosecka', 'p', '703-993-1710'),

('kosecka', 'p', '703-993-1876'),

('kunle', 'e', 'darlene@csl.stanford.edu'),

('kunle', 'e', 'kunle@ogun.stanford.edu'),

('kunle', 'p', '650-723-1430'),

('kunle', 'p', '650-725-3713'),

('kunle', 'p', '650-725-6949'),

('lam', 'e', 'lam@cs.stanford.edu'),

('lam', 'p', '650-725-3714'),

('lam', 'p', '650-725-6949'),

('latombe', 'e', 'asandra@cs.stanford.edu'),

('latombe', 'e', 'latombe@cs.stanford.edu'),

('latombe', 'e', 'liliana@cs.stanford.edu'),

('latombe', 'p', '650-721-6625'),

('latombe', 'p', '650-723-0350'),

('latombe', 'p', '650-723-4137'),

('latombe', 'p', '650-725-1449'),
    ('levoy', 'e', 'ada@graphics.stanford.edu'),

    ('levoy', 'e', 'melissa@graphics.stanford.edu'),

    ('levoy', 'p', '650-723-0033'),

    ('levoy', 'p', '650-724-6865'),

    ('levoy', 'p', '650-725-3724'),

    ('levoy', 'p', '650-725-4089')])

False Positives (0):

set([])

False Negatives (0):

set([])

Summary: tp=59, fp=0, fn=0
```

Drawbacks:

In some cases the program may fail to distinguish between actual email address and non email address. For example,

<address> Services at cs.stanford.edu Port 80</address>
will output mail id services@cs.stanford.edu which is actually not an email address.
<address> sambartika at cs.stanford.edu Port 80</address>
will output mail address sambartika@cs.stanford.edu which is an email address.