HW2

Contact Finder: Using Regular Expressions to find email addresses and phone numbers

Bing-Je Wu

IST 664 – Natural Language Processing

Table of Contents

[Introduction 2](#_Toc39668101)

[Analysis and Models 2](#_Toc39668102)

[About the Data 2](#_Toc39668103)

[Data Processing 2](#_Toc39668104)

[Results 10](#_Toc39668105)

[Conclusion 10](#_Toc39668106)

[Reference 10](#_Toc39668107)

# Introduction

In the modern day, people receive spam calls or emails easily because of sharing contact information on the internet, especially with automatically collecting program. To reduce the risk of getting spam calls or emails, some try to make it hard for spammers to automatically search text for email addresses and phone numbers by giving non-standard forms of them, also called obfuscation. For example, name(at)domain.com is non-standard form of email address. With that obscured format, spammers need to spend extra effort to decode the address to get the useful information.

Analysis and Models  
A python program, Contact Finder, originally from Stanford NLP course, used in the assignment was adapted for the course, IST664. The goal is to find the various email addresses and phone numbers in both standard format and obscured format. The ‘regular expression’ library was implemented throughout the program to achieve the goal. Some of python built-in functions were used in aid of collecting correct format of email address and phone numbers.

## About the Data

The data are in html format that includes varies information, such as the bio of faculties or staffs at Stanford. Another file contains a list of correct phone numbers and correct email address corresponding to the html data. Each html file may have one or more email addresses and one or more phone numbers.

Data Processing  
***Table 1****. Email Patterns with Examples and Result*

|  |  |  |  |
| --- | --- | --- | --- |
| # | Pattern | Explain (& obfuscated text) | Result (matched list /TP,FP,FN) |
| 1 | r'([A-Za-z.]+)@([A-Za-z.]+)\.edu' | Explain:   |  | | --- | | Match standard emails with a ‘period’ sign in the ‘name’ group or in the ‘domain’ group |   Obfuscated text from the files:   |  | | --- | | [balaji@stanford.edu](mailto:balaji@stanford.edu), [cheriton@cs.stanford.edu](mailto:cheriton@cs.stanford.edu), [engler@lcs.mit.edu](mailto:engler@lcs.mit.edu), [eroberts@cs.stanford.edu](mailto:eroberts@cs.stanford.edu), [fedkiw@cs.stanford.edu](mailto:fedkiw@cs.stanford.edu), [hanrahan@cs.stanford.edu](mailto:hanrahan@cs.stanford.edu), [kosecka@cs.gmu.edu](mailto:kosecka@cs.gmu.edu),  [kunle@ogun.stanford.edu](mailto:kunle@ogun.stanford.edu),  [darlene@csl.stanford.edu](mailto:darlene@csl.stanford.edu),  [nass@stanford.edu](mailto:nass@stanford.edu),  [nick.parlante@cs.stanford.edu](mailto:nick.parlante@cs.stanford.edu),  [patrick.young@stanford.edu](mailto:patrick.young@stanford.edu),  [rinard@lcs.mit.edu](mailto:rinard@lcs.mit.edu),  [shoham@stanford.edu](mailto:shoham@stanford.edu),  [pkrokel@Stanford.edu](mailto:pkrokel@Stanford.edu),  [widom@cs.stanford.edu](mailto:widom@cs.stanford.edu),  [siroker@cs.stanford.edu](mailto:siroker@cs.stanford.edu),  [zelenski@cs.stanford.edu](mailto:zelenski@cs.stanford.edu),  [manna@cs.stanford.edu](mailto:manna@cs.stanford.edu) | | Before applying the pattern:   |  |  |  | | --- | --- | --- | | TP | FP | FN | | 0 | 0 | 117 |   After applying the pattern:   |  |  |  | | --- | --- | --- | | TP | FP | FN | | 19 | 0 | 98 |   Matched list with the pattern:   |  | | --- | | 'balaji@stanford.edu', 'cheriton@cs.stanford.edu', 'engler@lcs.mit.edu',  ’eroberts@cs.stanford.edu', 'fedkiw@cs.stanford.edu',  'hanrahan@cs.stanford.edu', 'kosecka@cs.gmu.edu', 'darlene@csl.stanford.edu', 'kunle@ogun.stanford.edu', 'nass@stanford.edu', 'nick.parlante@cs.stanford.edu',  'patrick.young@stanford.edu', 'rinard@lcs.mit.edu', 'shoham@stanford.edu', 'pkrokel@stanford.edu', 'siroker@cs.stanford.edu', 'widom@cs.stanford.edu', 'zelenski@cs.stanford.edu', 'manna@cs.stanford.edu' | |
| 2 | r'([A-Za-z.]+)\s{1,3}@\s{1,3}([A-Za-z.]+)\.edu' | Explain:   |  | | --- | | Match emails with ' @ ' instead of '@' |   Obfuscated text from the files:   |  | | --- | | ashishg @ stanford.edu,  rozm @ stanford.edu,  dabo @ cs.stanford.edu,  ullman @ cs.stanford.edu,  zelenski @ cs.stanford.edu | | Before applying the pattern:   |  |  |  | | --- | --- | --- | | TP | FP | FN | | 0 | 0 | 117 |   After applying the pattern:   |  |  |  | | --- | --- | --- | | TP | FP | FN | | 5 | 0 | 112 |   Matched list with the pattern:   |  | | --- | | 'ashishg@stanford.edu', 'rozm@stanford.edu', 'dabo@cs.stanford.edu', 'ullman@cs.stanford.edu', 'zelenski@cs.stanford.edu' | |
| 3 | r'([A-Za-z.]+)@([A-Za-z.]+)\.[EDU]+' | Explain:   |  | | --- | | Match emails with capital 'EDU' |   Obfuscated text from the files:   |  | | --- | | uma@cs.stanford.EDU | | Before applying the pattern:   |  |  |  | | --- | --- | --- | | TP | FP | FN | | 0 | 0 | 117 |   After applying the pattern:   |  |  |  | | --- | --- | --- | | TP | FP | FN | | 1 | 0 | 116 |   Matched list with the pattern:   |  | | --- | | uma@cs.stanford.edu | |
| 4 | r'([a-zA-Z-]+)-@-([a-z-A-Z]+)-\.[eduEDU-]+' | Explain:   |  | | --- | | Match emails address with dash sign within characters |   Obfuscated text from the files:   |  | | --- | | d-l-w-h-@-s-t-a-n-f-o-r-d-.-e-d-u | | Before applying the pattern:   |  |  |  | | --- | --- | --- | | TP | FP | FN | | 0 | 0 | 117 |   After applying the pattern:   |  |  |  | | --- | --- | --- | | TP | FP | FN | | 0 | 1 | 117 |   Matched list with the pattern:   |  | | --- | | 'd-l-w-h@s-t-a-n-f-o-r-d.edu' |   After modifying ‘process\_file’:   |  |  |  | | --- | --- | --- | | TP | FP | FN | | 1 | 0 | 116 |   Matched list with modification:   |  | | --- | | dlwh@stanford.edu | |
| 5 | r'[\s>]([a-zA-Z\s]+)\sat\s([a-zA-Z\s]+)\sedu' | Explain:   |  | | --- | | Match emails with 'dot' instead of '.' ; 'at' instead of '@' |   Obfuscated text from the files:   |  | | --- | | serafim at cs dot stanford dot edu,  hager at cs dot jhu dot edu,  pal at cs stanford edu,  uma at cs dot stanford dot edu | | Before applying the pattern:   |  |  |  | | --- | --- | --- | | TP | FP | FN | | 0 | 0 | 117 |   After applying the pattern:   |  |  |  | | --- | --- | --- | | TP | FP | FN | | 0 | 4 | 117 |   Matched list with the pattern:   |  | | --- | | ‘serafim@cs dot stanford’, ‘dot.edu, hager@cs dot jhu’, ‘dot.edu, pal@cs stanford.edu’, ‘uma@cs dot stanford dot.edu’ |   After modifying ‘process\_file’:   |  |  |  | | --- | --- | --- | | TP | FP | FN | | 4 | 0 | 113 |   Matched list with modification:   |  | | --- | | 'hager@cs.jhu.edu', 'serafim@cs.stanford.edu', 'uma@cs.stanford.edu', 'uma@cs.stanford.edu' | |
| 6 | r'([A-Za-z]+)\s[atAT]+\s([A-Za-z;]+);[edu]+' | Explain:   |  | | --- | | Match emails with ';' sign instead of '.'; ' at ' instead of '@' |   Obfuscated text from the files:   |  | | --- | | jks at robotics;stanford;edu | | Before applying the pattern:   |  |  |  | | --- | --- | --- | | TP | FP | FN | | 0 | 0 | 117 |   After applying the pattern:   |  |  |  | | --- | --- | --- | | TP | FP | FN | | 0 | 1 | 117 |   Matched list with the pattern:   |  | | --- | | 'jks@robotics;stanford.edu' |   After modifying ‘process\_file’:   |  |  |  | | --- | --- | --- | | TP | FP | FN | | 1 | 0 | 116 |   Matched list with modification:   |  | | --- | | 'jks@robotics.stanford.edu' | |
| 7 | r"([a-zA-Z]+)\s\+\s'@'\s\+\s([a-zA-Z]+)" | Explain:   |  | | --- | | Match emails with the format of "name + '@' + domain" |   Obfuscated text from the files:   |  | | --- | | name + '@' + domain | | Before applying the pattern:   |  |  |  | | --- | --- | --- | | TP | FP | FN | | 0 | 0 | 117 |   After applying the pattern:   |  |  |  | | --- | --- | --- | | TP | FP | FN | | 0 | 1 | 117 |   Matched list with the pattern:   |  | | --- | | 'name@domain.edu' |   After modifying ‘process\_file’:   |  |  |  | | --- | --- | --- | | TP | FP | FN | | 1 | 0 | 116 |   Matched list with modification:   |  | | --- | | 'jurafsky@stanford.edu' | |
| 8 | r"\s([a-z]+)\sat\s([a-zA-Z.]+)\.edu" | Explain:   |  | | --- | | Match emails with ' at ' instead of '@' |   Obfuscated text from the files:   |  | | --- | | lam at cs.stanford.edu | | Before applying the pattern:   |  |  |  | | --- | --- | --- | | TP | FP | FN | | 0 | 0 | 117 |   After applying the pattern:   |  |  |  | | --- | --- | --- | | TP | FP | FN | | 1 | 0 | 116 |   Matched list with the pattern:   |  | | --- | | 'lam@cs.stanford.edu' | |
| 9 | r"([a-zA-Z]+)<del>@([a-zA-Z.]+)\.edu" | Explain:   |  | | --- | | Match emails with '<del>@' instead of '@' |   Obfuscated text from the files:   |  | | --- | | latombe<del>@cs.stanford.edu,  asandra<del>@cs.stanford.edu,  liliana<del>@cs.stanford.edu | | Before applying the pattern:   |  |  |  | | --- | --- | --- | | TP | FP | FN | | 0 | 0 | 117 |   After applying the pattern:   |  |  |  | | --- | --- | --- | | TP | FP | FN | | 3 | 0 | 114 |   Matched list with the pattern:   |  | | --- | | 'asandra@cs.stanford.edu',  'latombe@cs.stanford.edu',  'liliana@cs.stanford.edu' | |
| 10 | r"([a-zA-Z]+)&#x40;([a-zA-Z.]+)\.edu" | Explain:   |  | | --- | | Match emails with '&#x40;' instead of '@' |   Obfuscated text from the files:   |  | | --- | | ada&#x40;graphics.stanford.edu,  melissa&#x40;graphics.stanford.edu | | Before applying the pattern:   |  |  |  | | --- | --- | --- | | TP | FP | FN | | 0 | 0 | 117 |   After applying the pattern:   |  |  |  | | --- | --- | --- | | TP | FP | FN | | 2 | 0 | 115 |   Matched list with the pattern:   |  | | --- | | 'ada@graphics.stanford.edu',  'melissa@graphics.stanford.edu' | |
| 11 | r"([a-zA-Z]+)\s<at\ssymbol>\s([a-zA-Z.]+)\.edu" | Explain:   |  | | --- | | Match emails with ' <at symbol> ' instead of '@' |   Obfuscated text from the files:   |  | | --- | | manning <at symbol> cs.stanford.edu,  dbarros <at symbol> cs.stanford.edu | | Before applying the pattern:   |  |  |  | | --- | --- | --- | | TP | FP | FN | | 0 | 0 | 117 |   After applying the pattern:   |  |  |  | | --- | --- | --- | | TP | FP | FN | | 2 | 0 | 115 |   Matched list with the pattern:   |  | | --- | | 'dbarros@cs.stanford.edu',  'manning@cs.stanford.edu' | |
| 12 | r"([a-zA-Z.]+)\s\(followed\sby\s&ldquo;@([a-zA-Z.]+)\.edu" | Explain:   |  | | --- | | Match emails with ' (followed by &ldquo;@' instead of '@' |   Obfuscated text from the files:   |  | | --- | | ouster (followed by &ldquo;@cs.stanford.edu | | Before applying the pattern:   |  |  |  | | --- | --- | --- | | TP | FP | FN | | 0 | 0 | 117 |   After applying the pattern:   |  |  |  | | --- | --- | --- | | TP | FP | FN | | 1 | 0 | 116 |   Matched list with the pattern:   |  | | --- | | 'ouster@cs.stanford.edu' | |
| 13 | r"([a-zA-Z.]+)\s\(followed\sby\s\"@([a-zA-Z.]+)\.edu" | Explain:   |  | | --- | | Match emails with ' (followed by "@' instead of '@' |   Obfuscated text from the files:   |  | | --- | | teresa.lynn (followed by "@stanford.edu | | Before applying the pattern:   |  |  |  | | --- | --- | --- | | TP | FP | FN | | 0 | 0 | 117 |   After applying the pattern:   |  |  |  | | --- | --- | --- | | TP | FP | FN | | 1 | 0 | 116 |   Matched list with the pattern:   |  | | --- | | 'teresa.lynn@stanford.edu' | |
| 14 | r"[\s>]([a-zA-Z\s]+)\sAT\s([a-zA-Z\s]+)\sDOT\sedu" | Explain:   |  | | --- | | Match emails with 'DOT' instead of '.'; 'AT' instead of '@' |   Obfuscated text from the files:   |  | | --- | | subh AT stanford DOT edu | | Before applying the pattern:   |  |  |  | | --- | --- | --- | | TP | FP | FN | | 0 | 0 | 117 |   After applying the pattern:   |  |  |  | | --- | --- | --- | | TP | FP | FN | | 1 | 0 | 116 |   Matched list with the pattern:   |  | | --- | | 'subh@stanford.edu' | |
| 15 | r"([a-zA-Z]+)\sat\s([a-zA-Z]+)\sdt\scom" | Explain:   |  | | --- | | Match emails with 'gradiance dt com' instead of 'stanford.edu' |   Obfuscated text from the files:   |  | | --- | | support at gradiance dt com | | Before applying the pattern:   |  |  |  | | --- | --- | --- | | TP | FP | FN | | 0 | 0 | 117 |   After applying the pattern:   |  |  |  | | --- | --- | --- | | TP | FP | FN | | 0 | 1 | 117 |   Matched list with the pattern:   |  | | --- | | 'support@gradiance.edu' |   After modifying ‘process\_file’:   |  |  |  | | --- | --- | --- | | TP | FP | FN | | 1 | 0 | 117 |   Matched list with modification:   |  | | --- | | 'support@gradiance.com' | |
| 16 | r"([a-zA-Z]+)\sat\s[<die!\s\->]+([a-z]+)\s[a-z<>!\-\s]+\sedu" | Explain:   |  | | --- | | Match emails with 'tags', 'at' and 'dot' |   Obfuscated text from the files:   |  | | --- | | vladlen at <!-- die!--> stanford <!-- spam pigs!--> dot <!-- die!--> edu | | Before applying the pattern:   |  |  |  | | --- | --- | --- | | TP | FP | FN | | 0 | 0 | 117 |   After applying the pattern:   |  |  |  | | --- | --- | --- | | TP | FP | FN | | 1 | 0 | 116 |   Matched list with the pattern:   |  | | --- | | 'vladlen@stanford.edu' | |
| 17 | r"([a-zA-Z]+)\sWHERE\s([a-zA-Z]+)\s\DOM\sedu" | Explain:   |  | | --- | | Match emails with ' WHERE ' and ' DOM ' instead of '@' and '.' |   Obfuscated text from the files:   |  | | --- | | engler WHERE stanford DOM edu | | Before applying the pattern:   |  |  |  | | --- | --- | --- | | TP | FP | FN | | 0 | 0 | 117 |   After applying the pattern:   |  |  |  | | --- | --- | --- | | TP | FP | FN | | 1 | 0 | 116 |   Matched list with the pattern:   |  | | --- | | 'engler@stanford.edu' | |

***Table 2****. Phone Patterns with Examples and Result*

|  |  |  |  |
| --- | --- | --- | --- |
| # | Pattern | Explain (& obfuscated text) | Result (matched list /TP,FP,FN) |
| 18 | r'(\d{3})-(\d{3})-(\d{4})' | Explain:   |  | | --- | | Match phone with format  \_\_\_-\_\_\_-\_\_\_\_ |   Obfuscated text from the files:   |  | | --- | | '650-723-1131', '650-725-3726',  '650-723-3642', '650-723-6092',  '410-516-8000', '650-723-4377',  '650-723-6045', '650-725-4671',  '650-724-1915', '650-725-3726',  '650-725-6949', '650-494-8016',  '650-725-2588', '650-725-4802',  '650-723-0872', '650-723-7690',  '650-725-2588', '650-723-6092',  '650-725-8596' | | Before applying the pattern:   |  |  |  | | --- | --- | --- | | TP | FP | FN | | 0 | 0 | 117 |   After applying the pattern:   |  |  |  | | --- | --- | --- | | TP | FP | FN | | 19 | 0 | 98 |   Matched list with the pattern:   |  | | --- | | '650-723-1131', '650-725-3726', '650-723-3642', '650-723-6092', '410-516-8000', '650-723-4377', '650-723-6045', '650-725-4671', '650-724-1915', '650-725-3726', '650-725-6949', '650-494-8016', '650-725-2588', '650-725-4802', '650-723-0872', '650-723-7690', '650-725-2588', '650-723-6092', '650-725-8596' | |
| 19 | r'\((\d{3})\)(\d{3})-(\d{4})' | Explain:   |  | | --- | | Match phone with format  (\_\_\_)\_\_\_-\_\_\_\_ |   Obfuscated text from the files:   |  | | --- | | '(650)723-1614', '(650)723-4173', '(650)814-1478', '(650)725-3707', '(650)725-6949', '(650)724-9147', '(650)725-2340', '(650)725-4671' | | Before applying the pattern:   |  |  |  | | --- | --- | --- | | TP | FP | FN | | 0 | 0 | 117 |   After applying the pattern:   |  |  |  | | --- | --- | --- | | TP | FP | FN | | 8 | 0 | 109 |   Matched list with the pattern:   |  | | --- | | '650-723-1614', '650-723-4173', '650-814-1478', '650-725-3707', '650-725-6949', '650-724-9147', '650-725-2340', '650-725-4671' | |
| 20 | r'\((\d{3})\)\s(\d{3})-(\d{4})' | Explain:   |  | | --- | | Match phone with format  (\_\_\_) \_\_\_-\_\_\_\_ |   Obfuscated text from the files:   |  | | --- | | '(650) 723-4539', '(650) 724-3648', '(650) 724-6354', '(650) 725-3897', '(650) 725-4671', '(410) 516-5521', '(410) 516-5553', '(650) 723-0033', '(650) 723-8530', '(703) 993-1710',  '(703) 993-1876', '(650) 723-1430', '(650) 725-3713', '(650) 725-6949', '(650) 725-3714', '(650) 725-6949', '(650) 721-6625', '(650) 723-0350', '(650) 723-4137', '(650) 725-1449',  '(650) 723-0033', '(650) 724-6865', '(650) 725-3724', '(650) 725-4089', '(650) 723-7683', '(650) 725-1449', '(650) 725-3358', '(650) 725-4727', '(650) 723-9753', '(650) 725-1449',  '(617) 253-1221', '(617) 258-6922', '(650) 723-3334', '(650) 725-1449', '(650) 725-3383', '(650) 725-3636', '(650) 725-3938', '(650) 723-4364', '(650) 725-4671' | | Before applying the pattern:   |  |  |  | | --- | --- | --- | | TP | FP | FN | | 0 | 0 | 117 |   After applying the pattern:   |  |  |  | | --- | --- | --- | | TP | FP | FN | | 39 | 0 | 78 |   Matched list with the pattern:   |  | | --- | | '650-723-4539', '650-724-3648', '650-724-6354', '650-725-3897', '650-725-4671', '410-516-5521', '410-516-5553', '650-723-0033', '650-723-8530', '703-993-1710',  '703-993-1876', '650-723-1430', '650-725-3713', '650-725-6949', '650-725-3714', '650-725-6949', '650-721-6625', '650-723-0350', '650-723-4137', '650-725-1449',  '650-723-0033', '650-724-6865', '650-725-3724', '650-725-4089', '650-723-7683', '650-725-1449', '650-725-3358', '650-725-4727', '650-723-9753', '650-725-1449',  '617-253-1221', '617-258-6922', '650-723-3334', '650-725-1449', '650-725-3383', '650-725-3636', '650-725-3938', '650-723-4364', '650-725-4671' | |
| 21 | r'\+1\s(\d{3})\s(\d{3})\s(\d{4})' | Explain:   |  | | --- | | Match phone with format  +1 \_\_\_ \_\_\_ \_\_\_\_ |   Obfuscated text from the files:   |  | | --- | | '+1 650 723 5666', '+1 650 725 9046' | | Before applying the pattern:   |  |  |  | | --- | --- | --- | | TP | FP | FN | | 0 | 0 | 117 |   After applying the pattern:   |  |  |  | | --- | --- | --- | | TP | FP | FN | | 2 | 0 | 115 |   Matched list with the pattern:   |  | | --- | | '650-723-5666', '650-725-9046' | |
| 22 | r'\[(\d{3})\]\s(\d{3})-(\d{4})' | Explain:   |  | | --- | | Match phone with format  [\_\_\_] \_\_\_-\_\_\_\_ |   Obfuscated text from the files:   |  | | --- | | '[650] 723-5499', '[650] 725-2472' | | Before applying the pattern:   |  |  |  | | --- | --- | --- | | TP | FP | FN | | 0 | 0 | 117 |   After applying the pattern:   |  |  |  | | --- | --- | --- | | TP | FP | FN | | 2 | 0 | 115 |   Matched list with the pattern:   |  | | --- | | '650-723-5499', '650-725-2472' | |
| 23 | r'\+1\s(\d{3})\s(\d{3})-(\d{4})' | Explain:   |  | | --- | | Match phone with format  +1 \_\_\_ \_\_\_-\_\_\_\_ |   Obfuscated text from the files:   |  | | --- | | '+1 650 725-1449', '+1 650 723-3432' | | Before applying the pattern:   |  |  |  | | --- | --- | --- | | TP | FP | FN | | 0 | 0 | 117 |   After applying the pattern:   |  |  |  | | --- | --- | --- | | TP | FP | FN | | 2 | 0 | 115 |   Matched list with the pattern:   |  | | --- | | '650-725-1449', '650-723-3432' | |

23 patterns were created during the assignment. The approach was used by examining the person file and match the obscured format case by case. Table 1 and Table 2 show 23 patterns with explanation and results for email addresses and phone numbers. In the ‘result’ column, it includes metrics before applying the pattern, metrics after applying the pattern, matched list with pattern. Some of patterns matched with unformal format by comparing with the correct list of email address and phone number. Therefore, some modifications were added to the ‘process\_filename’ function to get the correct data. And the ‘result’ also includes the metrics after having the modification and matched list with modification. The detail modification has been commented in the program, ‘Wu\_ContactFinder.py’.

# Results

***Table 3: Total Patterns with Results***

|  |  |  |
| --- | --- | --- |
| # | Pattern | Result with modification (TP, FN) |
| 1 | r'([A-Za-z.]+)@([A-Za-z.]+)\.edu' | TP: 19, FN: 98 |
| 2 | r'([A-Za-z.]+)\s{1,3}@\s{1,3}([A-Za-z.]+)\.edu' | TP: 23, FN: 94 |
| 3 | r'([A-Za-z.]+)@([A-Za-z.]+)\.[EDU]+' | TP: 24, FN: 93 |
| 4 | r'([a-zA-Z-]+)-@-([a-z-A-Z]+)-\.[eduEDU-]+' | TP: 25, FN: 92 |
| 5 | r'[\s>]([a-zA-Z\s]+)\sat\s([a-zA-Z\s]+)\sedu' | TP: 29, FN: 88 |
| 6 | r'([A-Za-z]+)\s[atAT]+\s([A-Za-z;]+);[edu]+' | TP: 30, FN: 87 |
| 7 | r"([a-zA-Z]+)\s\+\s'@'\s\+\s([a-zA-Z]+)" | TP: 31, FN: 86 |
| 8 | r"\s([a-z]+)\sat\s([a-zA-Z.]+)\.edu" | TP: 32, FN: 85 |
| 9 | r"([a-zA-Z]+)<del>@([a-zA-Z.]+)\.edu" | TP: 35, FN: 82 |
| 10 | r"([a-zA-Z]+)&#x40;([a-zA-Z.]+)\.edu" | TP: 37, FN: 80 |
| 11 | r"([a-zA-Z]+)\s<at\ssymbol>\s([a-zA-Z.]+)\.edu" | TP: 38, FN: 79 |
| 12 | r"([a-zA-Z.]+)\s\(followed\sby\s&ldquo;@([a-zA-Z.]+)\.edu" | TP: 40, FN:77 |
| 13 | r"([a-zA-Z.]+)\s\(followed\sby\s\"@([a-zA-Z.]+)\.edu" | TP: 41, FN:76 |
| 14 | r"[\s>]([a-zA-Z\s]+)\sAT\s([a-zA-Z\s]+)\sDOT\sedu" | TP: 42, FN: 75 |
| 15 | r"([a-zA-Z]+)\sat\s([a-zA-Z]+)\sdt\scom" | TP: 43, FN: 74 |
| 16 | r"([a-zA-Z]+)\sat\s[<die!\s\->]+([a-z]+)\s[a-z<>!\-\s]+\sedu" | TP: 44, FN: 73 |
| 17 | r"([a-zA-Z]+)\sWHERE\s([a-zA-Z]+)\s\DOM\sedu" | TP: 45, FN: 72 |
| 18 | r'(\d{3})-(\d{3})-(\d{4})' | TP: 64, FN: 53 |
| 19 | r'\((\d{3})\)(\d{3})-(\d{4})' | TP: 72, FN: 45 |
| 20 | r'\((\d{3})\)\s(\d{3})-(\d{4})' | TP: 111, FN: 6 |
| 21 | r'\+1\s(\d{3})\s(\d{3})\s(\d{4})' | TP: 113, FN: 4 |
| 22 | r'\[(\d{3})\]\s(\d{3})-(\d{4})' | TP: 115, FN: 2 |
| 23 | r'\+1\s(\d{3})\s(\d{3})-(\d{4})' | TP: 117, FN: 0 |

Table 3 show a list of patterns used throughout the assignment. There were 23 patterns used to match all the data. Each pattern corresponds with the metrics, TP and FN, that was evaluated in the sequences. 117 correct email addresses and phone numbers were matched by using regular expression technique.

# Conclusion

In the world of big data, there are information everywhere. Extracting information such as contact info from the web is not as simple as usual. In order to reduce the risk of getting spam call or emails, people use encoding method on its contact info. To decode those information, regular expression could be a useful tool to extract information from text.

# Reference

1 Regex101. [Web Link](https://regex101.com/) ;

**Attachment:** Modification on ‘process\_file’

|  |
| --- |
| def process\_file(name, f):  res = []  for line in f:  for epat in epatterns:  matches = re.findall(epat,line)  for m in matches:  if '-' in m[0] or '-' in m[1]: ## total +1  ## Substitute the '-' sign with None  ## ex: d-l-w-h-@-s-t-a-n-f-o-r-d-.-e-d-u  ## matched with pattern 4: r'([a-zA-Z-]+)-@-([a-z-A-Z]+)-\.[eduEDU-]+'  email = '{}@{}.edu'.format(re.sub(r'-','',m[0]),re.sub(r'-','',m[1]))  res.append((name,'e',email))  elif 'dot' in m[1]: ## total +3  ## Replace ' dot ' with '.'  ## hager at cs dot jhu dot edu / serafim at cs dot stanford dot edu  ## uma at cs dot stanford dot edu  ## matched with pattern 5: r'[\s>]([a-zA-Z\s]+)\sat\s([a-zA-Z\s]+)\sedu'  domain\_temp = re.sub(r' dot ','.',m[1])  domain= re.sub(r' dot','',domain\_temp)  ename = m[0].strip()  email = '{}@{}.edu'.format(ename,domain)  res.append((name,'e',email))  if m[0] == 'subh':  ## uma at cs dot stanford dot edu / subh AT stanford DOT edu +1  email = '{}@{}.edu'.format(m[0],'cs.'+ m[1])  res.append((name,'e',email))  email2 = '{}@{}.edu'.format(name,m[1])  res.append((name,'e',email2))    elif 'cs ' in m[1]: ## total +1  ## Replace ' ' with '.'  ## ex: pal at cs stanford edu +1  ## matched with pattern 5: r'[\s>]([a-zA-Z\s]+)\sat\s([a-zA-Z\s]+)\sedu'  domain = re.sub(r' ','.',m[1])  email = '{}@{}.edu'.format(m[0], domain)  res.append((name,'e',email))  elif ';' in m[1]: ## total +1  ## Substitute the ';' sign with '.'  ## ex: jks at robotics;stanford;edu  ## matched with pattern 6: r'([A-Za-z]+)\s[atAT]+\s([A-Za-z;]+);[edu]+'  email = '{}@{}.edu'.format(re.sub(r'-','',m[0]),re.sub(r';','.',m[1]))  res.append((name,'e',email))  elif 'domain' in m[1]: ## total +1  ## email assigned with variables instead of traditional obscured email  ## ex: name + '@' + domain  ## matched with pattern 7: r"([a-zA-Z]+)\s\+\s'@'\s\+\s([a-zA-Z]+)"  email = '{}@{}.edu'.format(name,'stanford')  res.append((name,'e',email))    elif 'gradiance' in m[1]: ## total +1  ## save email as '.com' instead of 'edu'  ## ex: support at gradiance dt com  ## matched with pattern 15: r"([a-zA-Z]+)\sat\s([a-zA-Z]+)\sdt\scom"  email = '{}@{}.com'.format(m[0],m[1])  res.append((name,'e',email))  else:  email = '{}@{}.edu'.format(m[0],m[1])  res.append((name,'e',email))  # phone pattern list  for ppat in ppatterns:  matches = re.findall(ppat,line)  for m in matches:  phone = '{}-{}-{}'.format(m[0],m[1],m[2])  res.append((name,'p',phone))  return res |