Name: Teng Siong (T.S) Yeap
IST 664: Homework 2

1. **First attempt:**
   epatterns: ([A-Za-z]+)@([A-Za-z]+)\.edu

   The regular expression in this attempt is to catch the basic format of email. For example, someone@domain.edu. For "someone" the first letter can be a capitalized or uncapitalized letter, followed by "Kleene plus" which take one or more characters that are designed in the previous way. Then, we have an "@" symbol. I express the "domain" portion with the same regular expression as the "someone" portion. Next, I end the expression with "\.edu". It means the ".edu" is strictly required.

   ppatterns: (\d{3})-(\d{3})-(\d{4})

   The regular expression of phone pattern is using a digit class, specifying 3 digits in the first portion of the format, then a hyphen, another 3-digits class, a hyphen, and a 4-digits class to fit the general format, xxx-xxx-xxxx that we often see in our daily life.

   These regular expressions match 23 true positives, 1 false positive and 94 false negative. They match (part of the examples is listed below):

Obscured	Matched Output
balaji@stanford.edu	'balaji', 'e', 'balaji@stanford.edu'
nass@stanford.edu	'nass', 'e', 'nass@stanford.edu'
650-723-3642	'eroberts', 'p', '650-723-3642'
650-723-4377	'rajeev', 'p', '650-723-4377'

Second attempt:
epatterns: ([A-Za-z.]+)@([A-Za-z.]+)\.edu

This pattern tries to match any letters (capitalized and uncapitalized) and a "." in the "someone" and "domain" portion of the email.

ppatterns: \((\d{3})\)\s*(\d{3})-(\d{4})

For this attempt, I randomly choose some false negative and find the pattern. I find out some very similar patterns in xxx-xxx-xxxx or (xxx)xxx-xxxx, with or without a white space after the first three digits. I set "\(" and "\)" because "(" and ")" are special symbol and they require a "\". Then I add an optional white space "/s*" after the first three digits.

These regular expressions match 85 true positives, 0 false positive and 32 false negative. They match (part of the examples is listed below):

Obscured	Matched Output
patrick.young@stanford.edu	'psyoung', 'e', 'patrick.young@stanford.edu'
engler@lcs.mit.edu	'engler', 'e', 'engler@lcs.mit.edu'
(650)814-1478	'ashishg', 'p', '650-723-1614'
(650)725-3707	'horowitz', 'p', '650-725-3707'

Third attempt:
epatterns: ([A-Za-z.]+)\s*@\s*([A-Za-z.]+)\.edu

This pattern tries to match any letters (capitalized and uncapitalized) and a "." in the "someone" and "domain" portion of the email. It also matches zero or more spaces by using "\s*" before and after the "@" sign.

ppatterns: \[(\d{3})\]\s(\d{3})-(\d{4})

This pattern tries to match nass's phone number which is in the format of [xxx] xxx-xxxx

These regular expressions match 91 true positives, 0 false positive and 26 false negative. They match (part of the examples is listed below):

Obscured	Matched Output
ashishg @ stanford.edu	'ashishg', 'e', 'ashishg@stanford.edu'
[650] 723-5499	'nass', 'p', '650-723-5499'

Forth attempt:
epatterns: ([A-Za-z.]+)\s*@\s*([A-Za-z.]+)\.EDU

In addition to the pattern in the previous expression, I add another expression to match cheriton's email which is uma@cs.stanford.EDU. I make the "edu" into big letters.

ppatterns: \+1\s(\d{3})\s?-?(\d{3})\s?-?(\d{4})

This is the last attempt on phone pattern to match the last few. juraksky's phone number (+1 650 723 5666) and shoham's (+1 650 723-3432) have the similar pattern. To match these, I make an escape for "+" because it is a special symbol, followed by a digit 1, having an optional white space in between digit sets and an optional hyphen in between the last two-digit sets. All the expressions mentioned above have matched all the phone patterns in contact finder.

These regular expressions match 96 true positives, 0 false positive and 21 false negative. They match (part of the examples is listed below):

Obscured	Matched Output
uma@cs.stanford.EDU	'cheriton', 'e', 'uma@cs.stanford.edu'
+1 650 723-3432	'shoham', 'p', '650-723-3432'

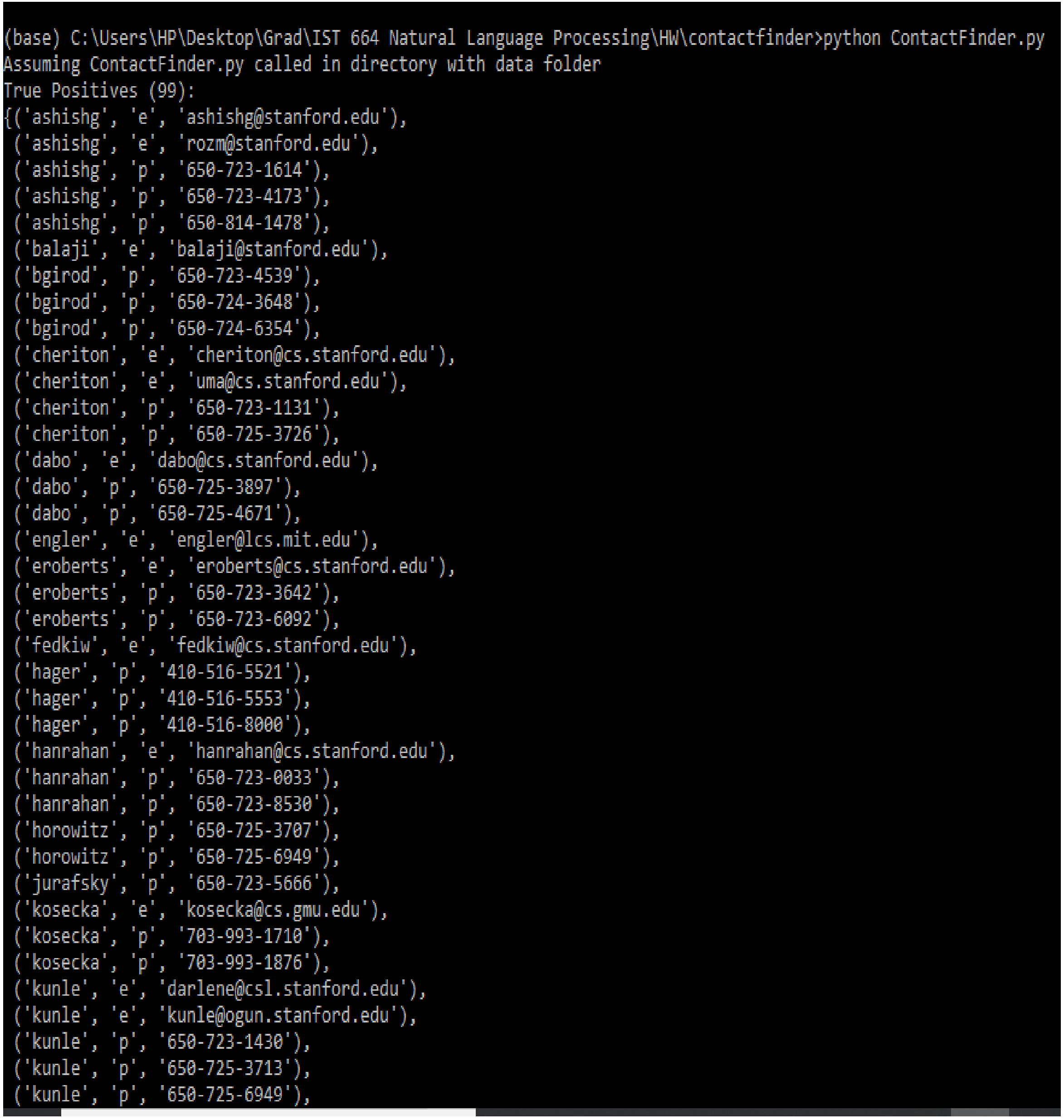
Fifth attempt:
epatterns: ([A-Za-z.]+)<del>@([A-Za-z.]+)\.edu

The last email that I manage to match is lathombe. The email address contains "<del>". SO, I add "<del>" in the regular expression.

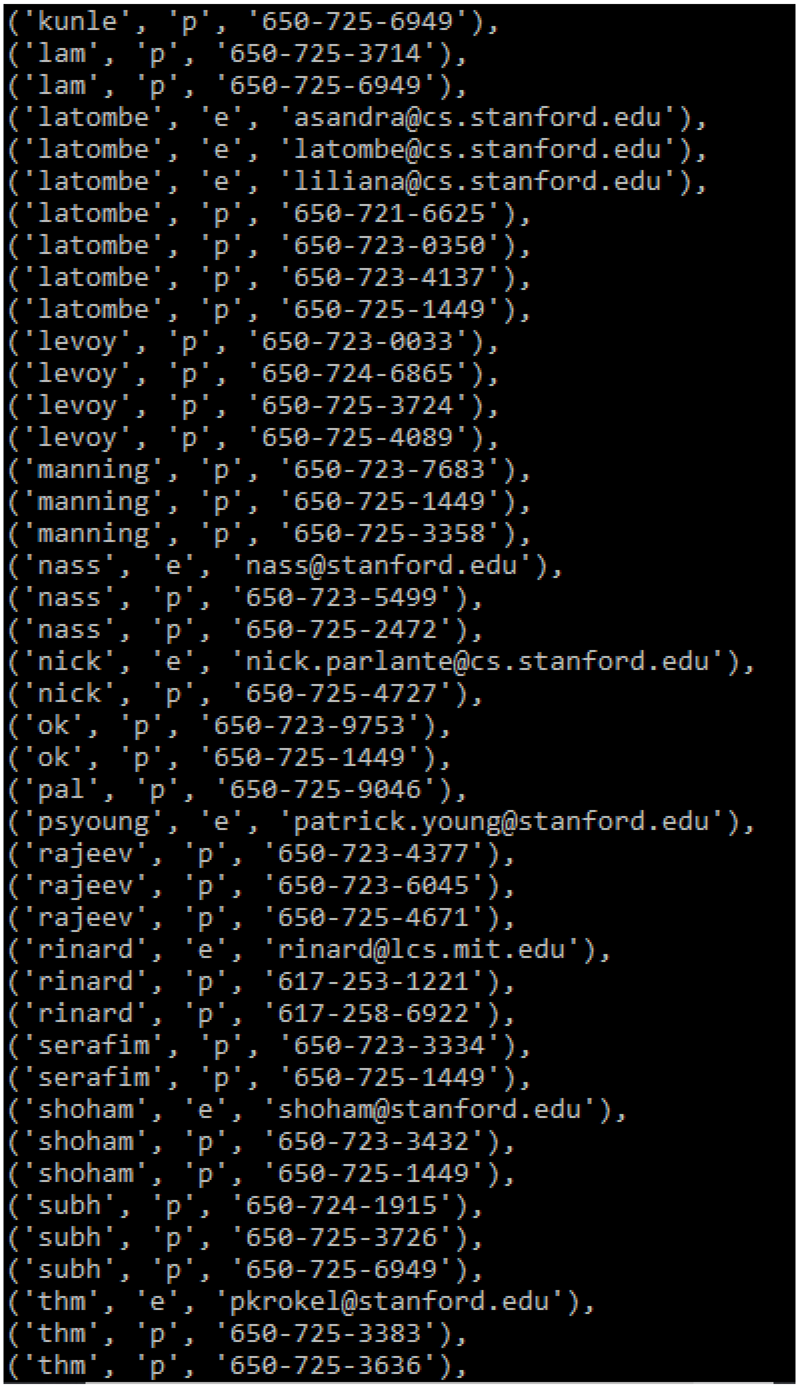
These regular expressions match 99 true positives, 0 false positive and 18 false negative. They match (part of the examples is listed below):

Obscured	Matched Output
latombe<del>@cs.stanford.edu	latombe', 'e', 'latombe@cs.stanford.edu'

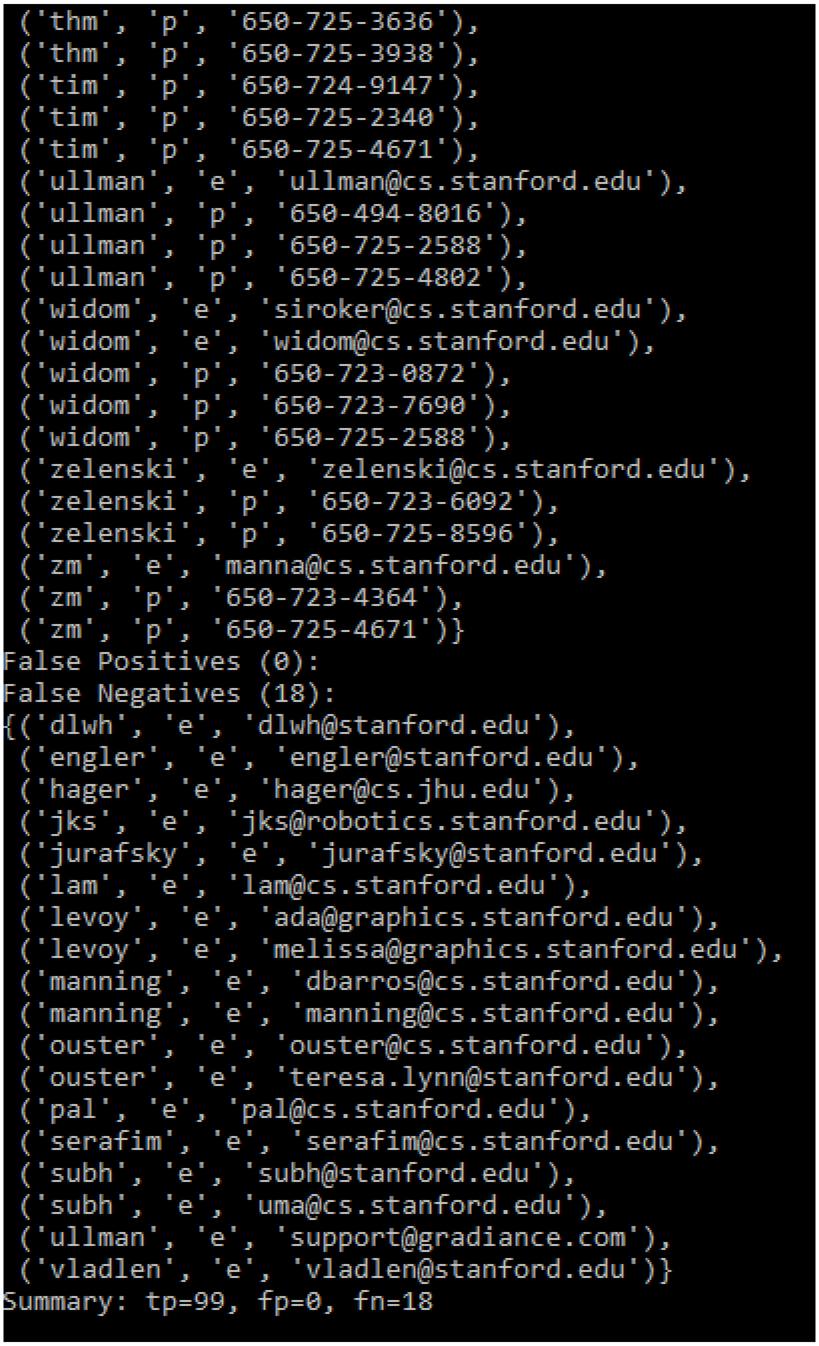
Output of the program:



```
(base) C:\Users\HP\Desktop\Grad\IST 664 Natural Language Processing\HW\contactfinder>python ContactFinder.py
Assuming ContactFinder.py called in directory with data folder
True Positives (99):
{('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'),
 ('engler', 'e', 'engler@lcs.mit.edu'),
 ('eroberts', 'e', 'eroberts@cs.stanford.edu'),
 ('eroberts', 'p', '650-723-3642'),
 ('eroberts', 'p', '650-723-6092'),
 ('fedkiw', 'e', 'fedkiw@cs.stanford.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'),
 ('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'),
```



```
('kunle', 'p', '650-725-6949'),
('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', 'p', '650-723-0033'),
('levoy', 'p', '650-724-6865'),
('levoy', 'p', '650-725-3724'),
('levoy', 'p', '650-725-4089'),
('manning', 'p', '650-723-7683'),
('manning', 'p', '650-725-1449'),
('manning', 'p', '650-725-3358'),
('nass', 'e', 'nass@stanford.edu'),
('nass', 'p', '650-723-5499'),
('nass', 'p', '650-725-2472'),
('nick', 'e', 'nick.parlante@cs.stanford.edu'),
('nick', 'p', '650-725-4727'),
('ok', 'p', '650-723-9753'),
('ok', 'p', '650-725-1449'),
('pal', 'p', '650-725-9046'),
('psyoung', 'e', 'patrick.young@stanford.edu'),
('rajeev', 'p', '650-723-4377'),
('rajeev', 'p', '650-723-6045'),
('rajeev', 'p', '650-725-4671'),
('rinard', 'e', 'rinard@lcs.mit.edu'),
('rinard', 'p', '617-253-1221'),
('rinard', 'p', '617-258-6922'),
('serafim', 'p', '650-723-3334'),
('serafim', 'p', '650-725-1449'),
('shoham', 'e', 'shoham@stanford.edu'),
('shoham', 'p', '650-723-3432'),
('shoham', 'p', '650-725-1449'),
('subh', 'p', '650-724-1915'),
('subh', 'p', '650-725-3726'),
('subh', 'p', '650-725-6949'),
('thm', 'e', 'pkrokel@stanford.edu'),
('thm', 'p', '650-725-3383'),
('thm', 'p', '650-725-3636'),
```



```
 ('thm', 'p', '650-725-3636'),
 ('thm', 'p', '650-725-3938'),
 ('tim', 'p', '650-724-9147'),
 ('tim', 'p', '650-725-2340'),
 ('tim', 'p', '650-725-4671'),
 ('ullman', 'e', 'ullman@cs.stanford.edu'),
 ('ullman', 'p', '650-494-8016'),
 ('ullman', 'p', '650-725-2588'),
 ('ullman', 'p', '650-725-4802'),
 ('widom', 'e', 'siroker@cs.stanford.edu'),
 ('widom', 'e', 'widom@cs.stanford.edu'),
 ('widom', 'p', '650-723-0872'),
 ('widom', 'p', '650-723-7690'),
 ('widom', 'p', '650-725-2588'),
 ('zelenski', 'e', 'zelenski@cs.stanford.edu'),
 ('zelenski', 'p', '650-723-6092'),
 ('zelenski', 'p', '650-725-8596'),
 ('zm', 'e', 'manna@cs.stanford.edu'),
 ('zm', 'p', '650-723-4364'),
 ('zm', 'p', '650-725-4671')}
False Positives (0):
False Negatives (18):
{('dlwh', 'e', 'dlwh@stanford.edu'),
 ('engler', 'e', 'engler@stanford.edu'),
 ('hager', 'e', 'hager@cs.jhu.edu'),
 ('jks', 'e', 'jks@robotics.stanford.edu'),
 ('jurafsky', 'e', 'jurafsky@stanford.edu'),
 ('lam', 'e', 'lam@cs.stanford.edu'),
 ('levoy', 'e', 'ada@graphics.stanford.edu'),
 ('levoy', 'e', 'melissa@graphics.stanford.edu'),
 ('manning', 'e', 'dbarros@cs.stanford.edu'),
 ('manning', 'e', 'manning@cs.stanford.edu'),
 ('ouster', 'e', 'ouster@cs.stanford.edu'),
 ('ouster', 'e', 'teresa.lynn@stanford.edu'),
 ('pal', 'e', 'pal@cs.stanford.edu'),
 ('serafim', 'e', 'serafim@cs.stanford.edu'),
 ('subh', 'e', 'subh@stanford.edu'),
 ('subh', 'e', 'uma@cs.stanford.edu'),
 ('ullman', 'e', 'support@gradiance.com'),
 ('vladlen', 'e', 'vladlen@stanford.edu')}
Summary: tp=99, fp=0, fn=18
```

2

a. Below is the list of e-mails that I can’t match.

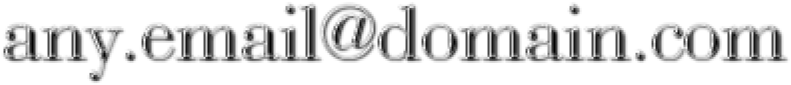
filename	Obscured	regex	Comment
dlwh	d-l-w-h-@-s-t-a-n-f-o-r-d-.-e-d-u	[A-Za-z-]*@[-][A-Za-z-]*\.[-][A-Za-z-]*	I am able to match this but I can’t print it because of the output format.
engler	engler WHERE stanford DOM edu	NA	Wanted to replace “WHERE” and “DOM” with “@” and “.” but it did not work
hager	hager at cs dot jhu dot edu	NA	Wanted to replace “at” and “dot” with “@” and “.” but it did not work
jks	jks at robotics;stanford;edu	NA	Wanted to replace “at” and “;” with “@” and “.” but it did not work
jurafsky	function obfuscate(domain, name) { document.write('<a href="mai' + 'lto:' + name + '@' + domain + '">' + name + '@' + domain + '</' + 'a>'); } obfuscate('stanford.edu','jurafsky');	NA	Used a function in the obsfucate e-mail and there is no way to match it.
lam	lam at cs.stanford.edu	NA	Wanted to replace “at” with “@” but it did not work
levoy	ada&#x40;graphics.stanford.edu melissa&#x40;graphics.stanford.edu	NA	&#x40; is the hex code for “@”
manning	dbarros <at symbol> cs.stanford.edu manning <at symbol> cs.stanford.edu	NA	Wanted to replace “<at symbol>” with “@” but it did not work
ouster	ouster (followed by &ldquo;@cs.stanford.edu&rdquo;) teresa.lynn (followed by "@stanford.edu")	NA	Has to strip and extract the “@” to the “edu” portion
pal	pal at cs stanford edu	NA	Wanted to replace “at” and whitespace with “@” and “.” but it did not work
serafim	serafim at cs dot stanford dot edu	NA	Wanted to replace “at” and “dot” with “@” and “.” but it did not work

subh	subh AT stanford DOT edu uma at cs dot stanford dot edu	NA	Wanted to replace “AT”, “at” with “@” and “DOT”, “dot” with “.” but it did not work
ullman	support at gradiance dt com	NA	Wanted to replace “at” with “@” and “dt” with “.” but it did not work
vladlen	vladlen at <!-- die!--> stanford <!-- spam pigs!--> dot <!-- die!--> edu	NA	This is hard to match as it contains some markup tags that are hard to remove

NOTE: My intention is to replace to words and use the established patterns to match them. Unfortunately, my approach did not give me any fruitful results.

b. – One of the ways that people as use to obscure the e-mail address and phone numbers is to convert these values into hex code. It was shown in one of the example above. Website: http://thenetweb.co.uk/obfuscate-hide-and-obscure-e-mail-addresses-telephone-numbers-and-text, also provides a tool and examples to show how this works. When everything is in hex code, there is no pattern to match using regex because everything is the same.

- Another way of obscure the e-mail address and phone number will be creating a function which returns the e-mail address and phone number.
- The last example I am providing here is hiding your e-mail address and phone number in an image. For example,



any.email@domain.com