Ankur Pandey Assignment 2 SUID: 408067486

1. (Required) Write more regular expressions that correct false positives or fit false negatives. Do as best as you can using the epatterns and ppatterns lists in the program.

=>

I have used following epatterns to match emails without using custom patterns or replace for part 1.

- Emails with optional space in them and edu as domain
  epatterns.append('([A-Za-z.]+)\s?@\s?([A-Za-z.]+)\.edu')
  eg. balaji@stanford.edu , dabo@cs.stanford.edu

- Emails with capital .EDU as domain name and with option space between them
  epatterns.append('([A-Za-z.]+)\s@\s([A-Za-z.]+)\.EDU')
  eg. uma@cs.stanford.EDU

- Cheriton file :
  epatterns.append('([A-Za-z.]+)\s?at\s?([A-Za-z.]+)\.EDU')
  eg. uma at cs.Stanford.EDU

- Engler file
  epatterns.append('([A-Za-z.]+)\s+WHERE+\s([A-Za-z.]+)+\s+DOM+\sedu')
  eg. engler WHERE stanford DOM edu

- Levoy file
  epatterns.append('([A-Za-z.]+)&#x40;([A-Za-z.]+)\.edu')
  eg. melissa&#x40;graphics.stanford.edu
  ada&#x40;graphics.stanford.edu

- ouster file
  epatterns.append('([A-Za-z.]+)\s+\([A-Za-z.\s&]+;@([A-Za-z.]+)\.edu[A-Za-z.\s&]+;\)')
  eg. ouster (followed by &ldquo;@cs.stanford.edu&rdquo;)

- epatterns.append('([A-Za-z.]+)\s+\([A-Za-z.\s&]+"@([A-Za-z.]+)\.edu"\)')
  eg. teresa.lynn (followed by "@stanford.edu")

- epatterns.append('([A-Za-z.]+)\s+at+\s([A-Za-z.\s]+)\s+dot+\sedu')
  eg. uma at cs dot stanford dot edu **false positive**
  eg. uma at cs dot stanford . edu **false positive**
  eg. serafim at cs dot stanford dot edu **false positive**
  eg. serafim at cs dot stanford . edu **false positive**
  eg. hager at cs dot jhu dot edu **false positive**
  eg. hager at cs dot jhu . edu **false positive**

- epatterns.append('([A-Za-z]+)\sat\s([A-Za-z;]+)\;\edu')
  eg. jks at robotics;stanford;edu **false positive**

- epatterns.append('([A-Za-z.]+)\s+AT+\s([A-Za-z.]+)+\s+DOT+\sedu')
  eg. subh AT stanford DOT edu
- epatterns.append('([A-Za-z]+)\sat\s([A-Za-z\s]+)\s\edu')
  eg. pal at cs stanford edu, **false positive**

- epatterns.append('([A-Za-z\-]+)@([A-Za-z\-]+)\.\-e\-d\-u')
  eg. d-l-w-h-@-s-t-a-n-f-o-r-d-.-e-d-u **false positive**

- epatterns.append('([A-Za-z.]+[^Server])\s+at+\s([A-Za-z.]+)\.edu')
  eg. lam at cs.stanford.edu
  Note: This false positive : server@cs.stanford.edu : is fixed by using ^Server in epatterns

- epatterns.append('([A-Za-z.]+)\sat\s<[A-Za-z\s!-]+>\s([A-Za-z\s]+)\s<[A-Za-z\s!-]+>\sdot\s<[A-Za-z\s!-]+>\sedu')
  eg. vladlen at <!-- die!--> stanford <!-- spam pigs!--> dot <!-- die!--> edu

I have used following ppatterns to match phone numbers without using custom patterns or replace for part 1.

- ppatterns.append('(\d{3})-(\d{3})-(\d{4})')
  like XXX-YYY-ZZZZ eg. 650-723-1614

- ppatterns.append('\((\d{3})\)(\d{3})-(\d{4})')
  like (XXX)YYY-ZZZZ eg. (650)723-1614

- ppatterns.append('\((\d{3})\)\s(\d{3})-(\d{4})')
  like (XXX) YYY-ZZZZ eg. (650) 724-6354

- ppatterns.append('(\d{3})\s(\d{3})\s(\d{4})')
  like XXX YYY ZZZZ eg. 650 723 5666

- ppatterns.append('\[(\d{3})\]\s?(\d{3})-(\d{4})')
  like [XXX] YYY-ZZZZ and [XXX]YYY-ZZZZ eg. [650] 723-5499 [650]723-5499

- ppatterns.append('(\d{3})\s(\d{3})-(\d{4})')
  XXX YYY-ZZZZ eg. 650 723-3432

I have found 109 true positive, 9 false positive and 8 false negative without using custom patterns or replace process_file() function for part 1. And choose option 3 to correct it by using custom patterns and replace in process_file() function.

<u>OUTPUT:</u>

Summary for Part 1: tp=109, fp=9, fn=8

False Positives (9):

```
set([('dlwh', 'e', 'd-l-w-h-@-s-t-a-n-f-o-r-d-.edu'),
    ('hager', 'e', 'hager@cs dot jhu dot.edu'),
    ('hager', 'e', 'hager@cs dot jhu.edu'),
    ('jks', 'e', 'jks@robotics;stanford.edu'),
    ('pal', 'e', 'pal@cs stanford.edu'),
    ('serafim', 'e', 'serafim@cs dot stanford dot.edu'),
    ('serafim', 'e', 'serafim@cs dot stanford.edu'),
    ('subh', 'e', 'uma@cs dot stanford dot.edu'),
    ('subh', 'e', 'uma@cs dot stanford.edu')])
```

False Negatives (8):

```
set([('dlwh', 'e', 'dlwh@stanford.edu'),
    ('hager', 'e', 'hager@cs.jhu.edu'),
    ('jks', 'e', 'jks@robotics.stanford.edu'),
    ('jurafsky', 'e', 'jurafsky@stanford.edu'),
    ('pal', 'e', 'pal@cs.stanford.edu'),
    ('serafim', 'e', 'serafim@cs.stanford.edu'),
    ('subh', 'e', 'uma@cs.stanford.edu'),
    ('ullman', 'e', 'support@gradiance.com')])
```

True Positives (109):

```
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'),
    ('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', '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'),
    ('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'),
    ('manning', 'e', 'dbarros@cs.stanford.edu'),
    ('manning', 'e', 'manning@cs.stanford.edu'),
    ('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'),
    ('ouster', 'e', 'ouster@cs.stanford.edu'),
    ('ouster', 'e', 'teresa.lynn@stanford.edu'),
    ('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', 'e', 'subh@stanford.edu'),
    ('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', '650-725-3938'),
    ('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'),
    ('vladlen', 'e', 'vladlen@stanford.edu'),
    ('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')])
```

3. (Option) Python programming: Continue working on the regular expressions to match more examples by having other lists of patterns. For example, you may want to have patterns that will match three parts of the email addresses, or you may want to make a list of patterns for email addresses that end in .com. For each list, you will need to add a part to process_filename that matches that list and puts its parts into a standard format answer. Write a section in the report that describes your approach and gives examples of the extended regular expressions or processing that match some of the emails.

=>

To solve 9 False Positive and 8 False Negatives, I have used two custom patterns and replace function in epatterns.

//[1] First one for obfuscate('stanford.edu','jurafsky'); => jurafsky@stanford.edu
eg. obfuscate('stanford.edu','jurafsky'); => jurafsky@stanford.edu

- epatterns_custom = []
  epatterns_custom.append('obfuscate\(\'([A-Za-z]+.edu)\',\'([A-Za-z]+)\'\)')

CODE:

```
for epat_custom in epatterns_custom:
        # each epat has 2 sets of parentheses so each match will have 2 items in a list
        matches = re.findall(epat_custom,line)
        for m in matches:
                # string formatting operator % takes elements of list m
                #   and inserts them in place of each %s in the result string
                email = '%s@%s' % m
                ename = email.split('@',1)[1]
                edomain = email.split('@',1)[0]
                res.append((name,'e',ename+'@'+edomain))
```

//[2] Second to match .com
eg. support at gradiance dt com   => support@gradiance.com

- compatterns = []
  compatterns.append('([A-Za-z.]+)\s+at+\s([A-Za-z.]+)+\sdt+\scom')
  eg.

CODE:

```
for compat in compatterns:
        # each epat has 2 sets of parentheses so each match will have 2 items in a list
        matches = re.findall(compat,line)
        for m in matches:
                # string formatting operator % takes elements of list m
                #   and inserts them in place of each %s in the result string
```

```
                    email = '%s@%s.com' % m
                    res.append((name,'e',email))
```

//[3] Replace function in epatterns

- email = email.replace('-','')
  eg. dlwh 'd-l-w-h-@-s-t-a-n-f-o-r-d-.edu'

- email = email.replace(' dot ','.')
- email = email.replace('dot.','')
  ```
  eg. ('hager', 'e', 'hager@cs.dot.jhu.dot.edu'),
      ('hager', 'e', 'hager@cs.dot.jhu.edu'),
      ('serafim', 'e', 'serafim@cs.dot.stanford.d
      ('serafim', 'e', 'serafim@cs.dot.stanford.e
      ('subh', 'e', 'uma@cs.dot.stanford.dot.edu'
      ('subh', 'e', 'uma@cs.dot.stanford.edu')])
  ```

- email = email.replace(';','.')
  eg. set([('jks', 'e', 'jks@robotics;stanford.edu')])

- email = email.replace(' ','.')
  ```
  eg. [('hager', 'e', 'hager@cs.jhu edu'),
       ('pal', 'e', 'pal@cs stanford.edu'),
       ('serafim', 'e', 'serafim@cs.stanford ed
       ('subh', 'e', 'uma@cs.stanford edu')])
  ```

CODE:

```
for epat in epatterns:
        # each epat has 2 sets of parentheses so each match will have 2 items in a list
        matches = re.findall(epat,line)
        for m in matches:
                # string formatting operator % takes elements of list m
            #   and inserts them in place of each %s in the result string
                email = '%s@%s.edu' % m
                # dlwh 'd-l-w-h-@-s-t-a-n-f-o-r-d-.edu'
                email = email.replace('-','')
                '''
                        ('hager', 'e', 'hager@cs.dot.jhu.dot.edu'),
                        ('hager', 'e', 'hager@cs.dot.jhu.edu'),
                        ('serafim', 'e', 'serafim@cs.dot.stanford.d
                        ('serafim', 'e', 'serafim@cs.dot.stanford.e
                        ('subh', 'e', 'uma@cs.dot.stanford.dot.edu'
                        ('subh', 'e', 'uma@cs.dot.stanford.edu')])
                '''
                email = email.replace(' dot ','.')
```

```
                email = email.replace('dot.','')
                # set([('jks', 'e', 'jks@robotics;stanford.edu')])
                email = email.replace(';','.')
                '''
                [('hager', 'e', 'hager@cs.jhu edu'),
                ('pal', 'e', 'pal@cs stanford.edu'),
                ('serafim', 'e', 'serafim@cs.stanford ed
                ('subh', 'e', 'uma@cs.stanford edu')])
                '''
                email = email.replace(' ','.')
                res.append((name,'e',email))
```

OUTPUT:

Summary: tp=117, fp=0, fn=0

False Positives (0):

```
set([])
```

False Negatives (0):

```
set([])
```

True Positives (117):

```
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'),
    ('manning', 'e', 'dbarros@cs.stanford.edu'),
    ('manning', 'e', 'manning@cs.stanford.edu'),
    ('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'),
    ('ouster', 'e', 'ouster@cs.stanford.edu'),
    ('ouster', 'e', 'teresa.lynn@stanford.edu'),
    ('pal', 'e', 'pal@cs.stanford.edu'),
    ('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', 'e', 'serafim@cs.stanford.edu'),
    ('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', 'e', 'subh@stanford.edu'),
    ('subh', 'e', 'uma@cs.stanford.edu'),
    ('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', '650-725-3938'),
    ('tim', 'p', '650-724-9147'),
    ('tim', 'p', '650-725-2340'),
    ('tim', 'p', '650-725-4671'),
    ('ullman', 'e', 'support@gradiance.com'),
    ('ullman', 'e', 'ullman@cs.stanford.edu'),
    ('ullman', 'p', '650-494-8016'),
    ('ullman', 'p', '650-725-2588'),
    ('ullman', 'p', '650-725-4802'),
    ('vladlen', 'e', 'vladlen@stanford.edu'),
    ('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')])
```