IST 664 Natural Language Processing Homework 2

Harper He
xhe128@syr.edu

Part 1

Email Patterns

Epattern 1

Expression

([A-Za-z.]+)\sat\s(([A-Za-z.]+)\.([A-Za-z]+))\.EDU

Description

This regular expression matches strings in the format: (1or more characters) + whitespace + word “at” + whitespace +(1or more characters + punctuation dot “.”+ 1or more characters) + punctuation dot “.”+ word “EDU”.
The first parentheses here is used to capture “someone” while the second is to capture “somewhere”

Examples

This regular expression matches “uma at cs.Stanford.EDU” in the file “cheriton”.

Epattern 2

Expression

([A-Za-z]+)\sWHERE\s([A-Za-z]+)\sDOM\sedu

Description

This regular expression matches strings in the format: (1or more characters) + whitespace + word “WHERE” + whitespace +(1or more characters) + whitespace + word “DOM”+ whitespace + word “edu”.
The first parentheses here is used to capture “someone” while the second is to capture “somewhere”

Examples

This regular expression matches “engler WHERE stanford DOM edu” in the file “engler”.

Epattern 3

Expression

([A-Za-z]+\.[A-Za-z]+)\s\([A-Za-z]+\s[A-Za-z]+\s"@([A-Za-z]+)\.edu

Description

This regular expression matches strings in the format: (1or more characters + punctuation dot “.”+ 1or more characters) + whitespace + left parenthesis + 1or more characters + whitespace+ 1or more characters + whitespace + punctuation quote ““” + punctuation @ + (1or more characters) + punctuation dot “.”+ word “edu”.
The first parentheses here is used to capture “someone” while the second is to capture “somewhere”

Examples

This regular expression matches “teresa.lynn (followed by "@stanford.edu")” in the file “ouster”.

Epattern 4

Expression

([A-Za-z]+)\sAT\s([A-Za-z]+)\sDOT\sedu

Description
This regular expression matches strings in the format: (1or more characters) + whitespace + word “AT” + whitespace + (1or more characters) + word “DOT” + whitespace + word “edu”.
The first parentheses here is used to capture “someone” while the second is to capture “somewhere”

Examples
This regular expression matches “subh AT stanford DOT edu” in the file “subh”.

Epattern 5

Expression
([A-Za-z]+)\sat\s([A-Za-z]+\.[A-Za-z]+)\.edu

Description
This regular expression matches strings in the format: (1or more characters) + whitespace + word “at” + whitespace +(1or more characters + punctuation dot “.”+ 1or more characters) + punctuation dot “.”+ word “edu”.
The first parentheses here is used to capture “someone” while the second is to capture “somewhere”

Examples
This regular expression matches “lam at cs.stanford.edu” in the file “lam”.

Phone Number Patterns

Ppattern 1

Expression
\+\d\s(\d{3})\s(\d{3})\s(\d{4})

Description
This regular expression matches strings in the format: punctuation plus “+” + one digit + whitespace + (3 digits) + whitespace + (3 digits) + whitespace + (4 digits).
The first parentheses here is used to capture “area code”, the second is to capture “exchange part” while the third one is to capture “number part”.

Examples
One example that this regular expression matches is the number “+1 650 723 5666” and “+1 650 725 9046” in the file “jurafsky”.

Ppattern 2

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

Description
This regular expression matches strings in the format: punctuation left bracket “[” + (3 digits) + punctuation right bracket “]” +whitespace + (3 digits) + punctuation hyphen “-” + (4 digits).
The first parentheses here is used to capture “area code”, the second is to capture “exchange part” while the third one is to capture “number part”.

Examples
One example that this regular expression matches is the number “[650] 723-5499” and “[650] 725-2472” in the file “nass”.

Ppattern 3

Expression

\+\d\s\((\d{3})\)\s(\d{3})-(\d{4})

Description

This regular expression matches strings in the format: punctuation plus “+” + one digit + whitespace + left parenthesis “(“ + (3 digits) + right parenthesis “)” + whitespace + (3 digits) + punctuation hyphen “-” + (4 digits).
The first parentheses here is used to capture “area code”, the second is to capture “exchange part” while the third one is to capture “number part”.

Examples

One example that this regular expression matches is the number “+1 (650) 723-7683” and “+1 (650) 725-1449” in the file “manning”.

Ppattern 4

Expression

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

Description

This regular expression matches strings in the format: left parenthesis “(“ + (3 digits) + right parenthesis “)”+ (3 digits) + punctuation hyphen “-” + (4 digits).
The first parentheses here is used to capture “area code”, the second is to capture “exchange part” while the third one is to capture “number part”.

Examples

One example that this regular expression matches is the number “(650)814-1478” and “(650)723-1614” in the file “ashishg”.

Ppattern 5

Expression

'\((\d{3})\)\s(\d{3})-(\d{4})')

Description

This regular expression matches strings in the format: left parenthesis “(“ + (3 digits) + right parenthesis “)”+whitespace+(3 digits) + punctuation hyphen “-” + (4 digits).
The first parentheses here is used to capture “area code”, the second is to capture “exchange part” while the third one is to capture “number part”.

Examples

One example that this regular expression matches is the number “(650) 724-6354” and “(650) 723-4539” in the file “bgirod”.

Ppattern 6

Expression

\+\d\s(\d{3})\s(\d{3})-(\d{4})

Description

This regular expression matches strings in the format: punctuation plus “+” + one digit + whitespace + (3 digits) + whitespace + (3 digits) + punctuation hyphen “-” + (4 digits).

The first parentheses here is used to capture “area code”, the second is to capture “exchange part” while the third one is to capture “number part”.

Examples

One example that this regular expression matches is the number “+1 650 723-3432” and “+1 650 725-1449” in the file “shoham”.

Result

The results before and after I added the expressions are listed below. The final output is 99 TP, 2 FP and 18 FN.

Expression Added (Cumulated)	Result		
	TP	FP	FN
Example Expression: ([A-Za-z.]+)@([A-Za-z.]+)\.edu ([A-Za-z.]+)\s@\s([A-Za-z.]+)\.edu (\d{3})-(\d{3})-(\d{4})	41	0	76
Epattern 1	42	0	75
Epattern 2	43	0	74
Epattern 3	44	0	73
Epattern 4	45	0	72
Epattern 5	46	2	71
Pattern 1	48	2	69
Pattern 2	50	2	67
Pattern 3	53	2	64
Pattern 4	61	2	56
Pattern 5	97	2	20
Pattern 6	99	2	18

Part 2

a.

As the result shown above, there are 20 examples that I cannot match using regular expressions and all of them are email address.

Some examples can be matched by the regular expression on regex online tester but failed to be extracted. I list all the 18 FN examples with some regular expressions I tried but failed as well as my assumption of the reasons. For those regular expressions that didn’t work, I have no idea of the reasons because there is no extracted information for my reference.

File Name	Email Displayed	Standard Form Result	Regular Expression Failed	FP Result	Possible reasons
dabo	dabo @ cs.stanford.edu	dabo@cs.stanford.edu	([A-Za-z]+)\s@\s([A-Za-z]+\.[A-Za-z]+)\.edu		
dlwh	d-l-w-h-@-s-t-a-n-f-o-r-d-.-e-d-u	dlwh@stanford.edu	([A-Za-z]+\-[A-Za-z]+\-[A-Za-z]+\-[A-Za-z]+)\-@(\-[A-Za-z]+\-[A-Za-z]+\-[A-Za-z]+\-[A-Za-z]+\-[A-Za-z]+\-[A-Za-z]\-[A-Za-z]\-[A-Za-z]\-[A-Za-z])\-\.\-e\-d\-u	('dlwh', 'e', 'd-l-w-h@-s-t-a-n-f-o-r-d.edu')	The - between characters is regared as a part of the "someone part"
hager	hager at cs dot jhu dot edu	hager@cs.jhu.edu	([A-Za-z]+)\sat\s(([A-Za-z]+)\s([A-Za-z]+)\s([A-Za-z]+))\sdot\sedu	('hager', 'e', 'hager@cs dot jhu.edu')	The word "dot" between "cs" and "jhu" is regared as a part of the "somewhere part"
jks	jks at robotics;stanford;edu	jks@robotics.stanford.edu	([A-Za-z]+)\sat\s([A-Za-z]+\;[A-Za-z]+)\;edu	('jks', 'e', 'jks@robotics;stanford.edu')	The punctuation ";" is regared as a part of the "somewhere part"
jurafsky	NA	jurafsky@stanford.edu			
latombe	asandra@cs.stanford.edu (with strikethrough from @ to the end of the address)	asandra@cs.stanford.edu	([A-Za-z]+)@([A-Za-z]+\.[A-Za-z]+)\.edu		The strikethrough obstructs the espression to match it.
latombe	latombe@cs.stanford.edu (with strikethrough from @ to the end of the address)	latombe@cs.stanford.edu	Same as above		Same as above
latombe	liliana@cs.stanford.edu (with strikethrough from @ to the end of the address)	liliana@cs.stanford.edu	Same as above		Same as above
levoy	ada@graphics.stanford.edu	ada@graphics.stanford.edu	Same as above		
levoy	melissa@graphics.stanford.edu	melissa@graphics.stanford.edu	Same as above		
manning	dbarros cs.stanford.edu	dbarros@cs.stanford.edu	([A-Za-z]+)\s([A-Za-z]+\.[A-Za-z]+)\.edu		
manning	manning cs.stanford.edu	manning@cs.stanford.edu	Same as above		
ouster	ouster (followed by "@cs.stanford.edu")	ouster@cs.stanford.edu	([A-Za-z]+)\s\([A-Za-z]+\s[A-Za-z]+\s.@([A-Za-z]+\.[A-Za-z]+)\.edu		
pal	pal at cs stanford edu	pal@cs.stanford.edu	([A-Za-z]+)\sat\s([A-Za-z]+)\s([A-Za-z]+)\sedu	('pal', 'e', 'pal@cs.edu')	
			([A-Za-z]+)\sat\s([A-Za-z]+\s[A-Za-z]+)\sedu	('pal', 'e', 'pal@cs stanford.edu')	The whitespace between "cs" and "stanford" is regared as a part of the "somewhere part"
serafim	serafim at cs dot stanford dot edu	serafim@cs.stanford.edu	Same as hager	('serafim', 'e', 'serafim@cs dot stanford.edu')	The word "dot" between "cs" and "stanford" is regared as a part of the "somewhere part"
subh	uma at cs dot stanford dot edu	uma@cs.stanford.edu	Same as hager	('subh', 'e', 'uma@cs dot stanford.edu')	The word "dot" between "cs" and "stanford" is regared as a part of the "somewhere part"
ullman	support at gradiance dt com	support@gradiance.com			The email address does not end with edu
vladlen	vladlen at stanford dot edu	vladlen@stanford.edu	([A-Za-z]+)\sat\s([A-Za-z]+)\sdot\sedu		

(Please click the link for the source file of this screenshot:
https://drive.google.com/file/d/1JUiQ1MVWYRvabXOBlKBa_Ftr-sipEVAu/view?usp=sharing)

There are also 2 False Positives examples, they are “Server at infolab.stanford.edu” and “Server at cs.stanford.edu”, I used the regular expression “([A-Za-z]+)\sat\s([A-Za-z]+\.[A-Za-z]+)\.edu” but failed to get TP results. I had no idea why they cannot match because the regular expression matched them on regex online tester, my assumption is that italic may affect the matching process, just like the strikethrough examples.

b

Based on the FN examples, we can see some characteristics of these example, for example, using word/punctuation/formatting to obscure the email addresses. Below is the email address I designed beads on my observation:
nlp dot ~~c-o_u[r]s;e~~ @ is;ch-oo!>l dot s-%~~y~~-*r* dot edu
(nlp.course@ischool.syr.edu)