Grading Rubric, NLP
Homework 2: ContactFinder

The parts of the report submission:

1. 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. List the regular expressions that you write, examples of email or phone numbers that match each pattern, and a short English explanation of what expressions the pattern matches. Give the text that matches, i.e. the obfuscated example, not only the results! When you are done with as many as you can do, give the output of the program.

Option 2. a. List the examples that you found you could not match with the current regular expressions with two extracted parts, ending in .edu. For each example, or set of examples that fit the same pattern, explain briefly why it won't work. If you can make expressions in regexpal that match, but don't work in the program to extract the email or phone numbers, list those here.

b. Then search the web and find a couple of additional examples of obscured email addresses or phone numbers and report on them, or try to design a way to obscure an email address that would be extremely difficult for a spamlord to match with a regular expression. For the latter, try to have something more specific than things like "To send me email, try the simplest address that makes sense."

Option 3. 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.

The report and submission will be graded according to the following rubric:

	Level of Achievement		
	1 (50%)	**2 (75%)**	**3 (100%)**
reproducibility 20%	ContactFinder program is not provided, or is not provided as a .py program Or the development process is not clear with critical processes missing.	ContactFinder program output is not clear, but program is submitted. The development process is clear in general with some details missing.	ContactFinder output is given in report and program is also submitted. Development process is clear.
correct analysis 40%	Part 1 regular expressions do not extract sufficient items. Regular expressions may not be clear, overly complicated or too specific.	Part 1 regular expressions are complete and extract a significant number of the items. Some regular expressions may not be clear, overly complicated or too specific.	Part 1 regular expressions are complete and extract a significant number of the items. Regular expressions are clear, not overly complicated or too specific.
writing clarity and convincing conclusions 40%	The report is confusing with important information missing as to how the regular expression match and extract text. Either part 2 or 3 fails in some aspect or is not complete.	The report is clear in general with some missing information as to how regular expressions match and extract text. Either 2) there is not a good descript ion of further capabilities of matching and not sufficient ideas of obfuscation OR 3) Good use of text processing and regular expression functionality to further extract emails, but without sufficient documentation.	The report is clear with no confusion. The explanations of rules demonstrate good understanding of how regular expressions match text. Either 2) Good understanding of further capabilities of matching and good ideas of obfuscation to foil regular expressions OR 3) Good use of text processing and regular expression functionality to further extract emails with sufficient documentation.

Interpretation of numeric grades as letter grades:

90 – 100 A
85 – 89.9 A-
80 – 84.9 B+
75 – 79.9 B
70 – 74.9 B-
below 70 has similar interpretation in the C and lower range.

Late assignment submissions will be accepted, but will be penalized in a sliding scale with these interpolation points:
1 Week late- 10 points taken off (2/3 letter grade)
2 Weeks or more late- 20 points taken off (1 1/3 letter grade)
Lateness of assignments may possibly be excused by emailing the instructor with an appropriate excuse.