Assignment 2 - Regular expressions

Basic practice questions

Give appropriate command to achieve following objectives. (These are practice questions to prepare you for assignment questions. These questions will not be evaluated.).

You can use (`/usr/share/dict/words` file available on most Red-Hat / Cent-OS / Fedora distributions to test your expressions

1. Search for all lines in the file that contain pattern 'hello' in them.
2. Search for all lines that contain digit '4' somewhere in the line
3. Search for all lines that contain comma ',' somewhere in the line
4. Search for all lines that contain any digit somewhere in the line
5. Search for all lines that contain word with pattern 'oo' immediately followed by a vowel (a, e, i, o or u)
6. Search for all lines that contain word 'Yahoo' or 'yahoo' somewhere between them.

Assignment Questions

Give appropriate command to achieve following objectives. These are compulsary questions that will be evaluated as part of assignment

Note: You are expected to explain the step-by-step construction and logic behind the regular expression that you have used in detail, to help with evaluation.

1. Write regular expressions using grep to achieve following ouput: (You can use (`/usr/share/dict/words` file available on most Red-Hat / Cent-OS / Fedora distributions to test your expressions
   (a) Search for all non-blank lines that contain only capital letters.

(b) Search all lines that end with pattern 'point'.

(c) Search for all lines that contains repetition of single alphabet, digit or special character three times. For example any line that contains AAA, ZZZ, ooo, @@@, etc. should get grepped.

2. Write regular expression that can extract all tags like '<html>', '<body style="color:black">', '</head>' etc. from a HTML file.

3. Write shell command that can extract all HTML tags that have 'style=' attribute set inside tag.

4. Write regular expression that extracts only absolutely correct IPv4 addresses from text file. Criteria that correct IPv4 address must satisfy are:

   (a) It must consist of four integers, say, A, B, C, D in form A.B.C.D

   (b) All A,B,C,D must be between 0 and 255 both inclusive

5. Write regular expression that extracts real numbers from text file. Real numbers could be both positive or negative. They can optionally have decimal and any number of decimal digits. Criteria that real numbers satisfy are:

   (a) Real number have optional '+' or '-' as first character to indicate positive or negative.

   (b) There is at least one digit between 0 to 9. There can be any number of digits.

   (c) There can be optional decimal point (.) anywhere between number. Decimal point should occur only once.

   (d) If there is a decimal point there must be at least one digit after decimal point.

6. Write regular expression that extracts proper email address from text file. Email addresses can be assumed to satisfy following criteria:

   (a) Email username can consists of digits ([0-9]), alphabets ([a-z], [A-Z]), hypen (-), underscore(_) and dot.

   (b) Usernames do not start or end with hypen (-), underscore(_) or dot(.)

(c) There can be multiple hypens, underscore and dot in email address but no two occurences of these characters can be consecutive. That is email address wont have two consecutive hypens(-) or hypen immediately followed by dot(.), etc.

(d) There must be exactly one '@' sign in every email address.

(e) '@' sign must be followed by domain name which contains at least one dot.

(f) Domain name can contain numbers, alphabets, hypen and dot.

(g) Domain name does not have two consecutive dots or hypens.

(h) Domain name can have multiple dots separated by characters

(i) Domain names do not start with dot or hypen.

(j) Final part of domain name, that is, .in in case of iiit.ac.in does not contains number or hypen. Hence domain names like .abc.ac.3in, .abc.uk3, iiit.ac.i-n etc. are not possible.

7. Write sed command that takes set of correct email addresses from text file and seperates usernames and domain names. For example if input file contains `a@b.c`, sed output should contain `Username : a, Domain : b.c`