Lab 10: Regular Expression

**References:**

1. Regular Expressions: The Complete Tutorial, by Jan Goyvaerts, 2007.
2. Speech and Language Processing, by Dan Jurafsky and James H. Martin. Prentice Hall Series in Artificial Intelligence, 2008.
3. Natural Language Processing with Python, by Steven Bird, Ewan Klein and Edward Loper, 2014.

Quick Review

A regular expression (regex or regexp for short) is a special text string for describing a search pattern. You can think of regular expressions as wildcards on steroids. You are probably familiar with wildcard notations such as \*.txt to find all text files in a file manager. The regex equivalent is «.\*\.txt». Basically, a regular expression is a pattern describing a certain amount of text.

A regular expression “engine” is a piece of software that can process regular expressions, trying to match the pattern to the given string. Usually, the engine is part of a larger application and you do not access the engine directly. Rather, the application will invoke it for you when needed, making sure the right regular expression is applied to the right file or data.

Practices

Go to online regular expression tool in your web browser: <https://www.regexpal.com/>

Enter the following text below, into the test string box (pay attention to lowercase and uppercase letters):

I knew some good games we could play,

said the cat.

I know some new tricks,

said the cat in the hat.

A lot of good tricks.

I will show them to you.

Your mother will not mind at all if I do.

We looked!

Then we saw him step in on the mat!

We looked!

And we saw him!

The Cat in the Hat!

And he said to us,

Why do you sit there like that?

There is 'the' on lines 2, 4, 9, and 12, but not on other lines.

Oooooh! We have your books in the library.

Color is US and colour is British.

Letters, Characters, and Digits

Use the patterns below in the regular expression pattern box, one by one, and discuss the results.

a

know

some

the

The

we

We

\d

\d\d

\D

\D\D

\D\D *(add a space at the bginning of pattern)*

\D\D *(add a space at the bginning and at the end of pattern)*

Disjunction and Ranges

Use the patterns below in the regular expression pattern box, one by one, and discuss the results.

[nm!]

[12]

[0-9]

[A-Z]

[n|m|!]

Write a pattern to specify ‘we’ and ‘We’.

Repeat it for ‘cat’ and ‘Cat’.

Negation

Use the patterns below in the regular expression pattern box, one by one, and discuss the results.

[^t]

oo

[^gl]oo

Optional Characters

Use the patterns below in the regular expression pattern box, one by one, and discuss the results.

.

wi..

\.

your?

ooo\*

ooo+

kn.w

Write a pattern to recognize both American and British spelling of color.

Line Break

Use the patterns below in the regular expression pattern box, one by one, and discuss the results.

^I

^.

^[A-Z]

^[a-z]

.$

\.$

[.]$

[^.]$