Introduction
Tools
Practice Problem

SCHOOL OF INFORMATION
UNIVERSITY OF MICHIGAN

Discussion 8: Regular Expressions

SI 206: Data-Oriented Programming

Instructor: Dr. Barbara (Barb) Ericson
GSI: Kexuan (Michael) Huang
IA: Cristina & Jade

School of Information
University of Michigan

Fall 2023

Introduction
Tools
Practice Problem

Deadlines

- Homework 5 due this Friday (Oct 27th)
- Midterm 2 on week 10 (now we are at week 9!)

Introduction

Tools

Practice Problem

Table of Contents

1. Introduction
2. Tools
3. Practice Problem

Table of Contents

1. Introduction
2. Tools
3. Practice Problem

Why Regular Expressions?

Validate or search for data with pattern, for example:

- Email
- Address
- Phone number

Introduction
Tools
Practice Problem

Table of Contents

1 Introduction

2 Tools

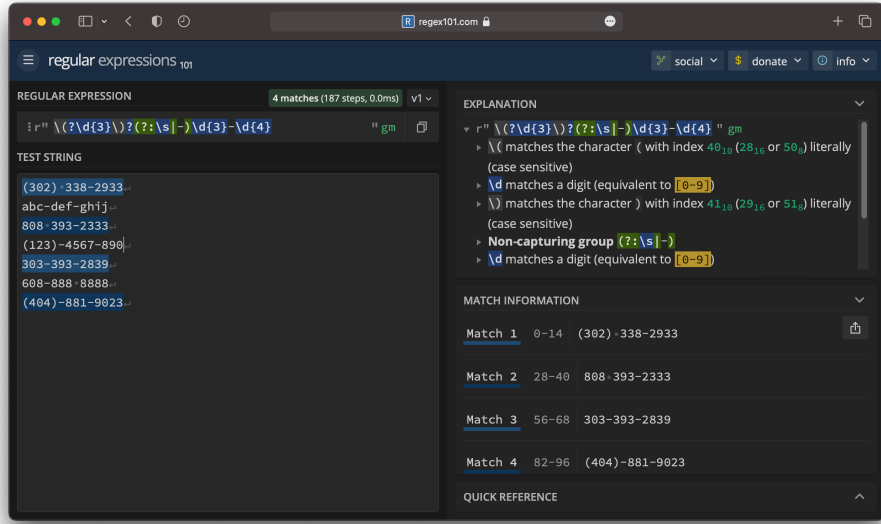
3 Practice Problem

Introduction
Tools
Practice Problem

How to get started?

Start with regex101.com

Use Regex101



Use Regex101 with Cheat-sheet!

SCHOOL OF INFORMATION
UNIVERSITY OF MICHIGAN

Regular Expression Basics	
.	Any character except newline
a	The character a
ab	The string ab
a|b	a or b
a*	0 or more a's
\	Escapes a special character

Regular Expression Quantifiers	
*	0 or more
+	1 or more
?	0 or 1
{2}	Exactly 2
{2, 5}	Between 2 and 5
{2,}	2 or more
{,5}	Up to 5

Default is greedy. Append ? for reluctant.

Regular Expression Groups	
(...)	Capturing group
(?P<Y>...)	Capturing group named Y
(?:...)	Non-capturing group
\Y	Match the Y'th captured group
(?P=Y)	Match the named group Y
(?#...)	Comment

Regular Expression Character Classes	
[ab-d]	One character of: a, b, c, d
[^ab-d]	One character except: a, b, c, d
[\b]	Backspace character
\d	One digit
\D	One non-digit
\s	One whitespace
\S	One non-whitespace
\w	One word character
\W	One non-word character

Regular Expression Assertions	
^	Start of string
\A	Start of string, ignores m flag
$	End of string
\Z	End of string, ignores m flag
\b	Word boundary
\B	Non-word boundary
(?=...)	Positive lookahead
(?!...)	Negative lookahead
(?<=...)	Positive lookbehind
(?<!...)	Negative lookbehind
(?())	Conditional

Regular Expression Flags	
i	Ignore case
m	^ and $ match start and end of line
s	. matches newline as well
x	Allow spaces and comments
L	Locale character classes
u	Unicode character classes
(?iLmsux)	Set flags within regex

Regular Expression Special Characters	
\n	Newline
\r	Carriage return
\t	Tab
\YYY	Octal character YYY
\xYY	Hexadecimal character YY

Regular Expression Replacement	
\g<0>	Insert entire match
\g<Y>	Insert match Y (name or number)
\Y	Insert group numbered Y

https://www.debuggex.com/cheatsheet/regex/python

SCHOOL OF INFORMATION
UNIVERSITY OF MICHIGAN

Use Regex101 with Cheat-sheet!

1. Think of all possible versions of the thing you'd like to retrieve
2. Try it out on regex101.com
3. Code with Python

Introduction

Tools

Practice Problem

Table of Contents

1 Introduction

2 Tools

3 Practice Problem

Introduction
Tools
Practice Problem

Practice Problem

Background and Data

You and your partner are detectives working on a case. You've intercepted a series of fake emails between the killer and victim. Your task is to extract relevant information using regular expressions to aid in your investigation.

```
Sender: Vinny Su (vinnysu@gmail.com)
To: nicolelam@gmail.com
Subject: Re: Dinner Plans

Sure, dinner sounds great! Why not we meet at Michigan Union at 7pm?
The location is: 530 S State St, Ann Arbor, MI 48109

Vinny
666-555-1111
```

Introduction
Tools
Practice Problem

Practice Problems

Go to Canvas → Assignment → Discussion 8 and clone the GitHub Repo.

Your tasks

- Implement `get_email_count()`: returns a dictionary that counts email address
- Implement `get_phone_list()`: returns a list of phone numbers
- Implement `get_address_list()`: returns a list of tuples, each tuple contains state, city, zip code, street name and street number

```
{'mikex@gmail.com': 3, 'nicolelam@gmail.com': 4, 'vinnysu@gmail.com': 2}

['(555)765-4321', '666-555-1111', '(888)765 4321', '608 901 2345']

[('MI', 'Ann Arbor', '48104', 'E William St', '516'),
 ('MI', 'Ann Arbor', '48109', 'S State St', '530'),
 ('MI', 'Ann Arbor', '48104', 'Maynard St', '347'),
 ('MI', 'Ann Arbor', '48109', 'Duffield St', '1931')]
```