From Regular Expressions to Regular Languages

a.k.a How to build a regex engine

Phúc Phạm

June 10, 2025

Outline

1. Regex basics
2. Regular Languages
3. Regex tips in practice
4. Introduction to parsing
5. Summary

Regex basics

What is a regular expression?

A regular expression is a sequence of characters that specifies a match pattern in text.

— Wikipedia

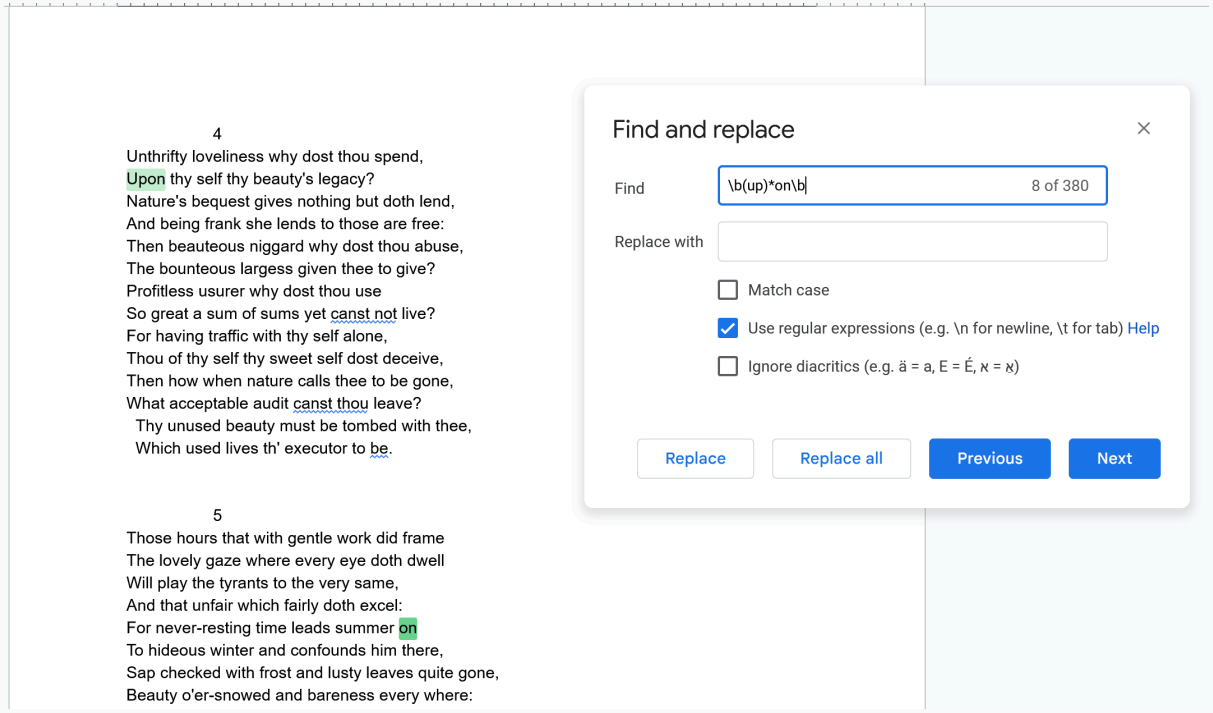
```
What [a-z]{2} [^A-Z]+ (\w*)\?
What is regular expression?
```

Where can I use regex?

Any where there is text search feature, there is a decent chance it support regex

- Text search
- Pattern validation

Word/Google Docs



4

Unthrifty loveliness why dost thou spend,
Upon thy self thy beauty's legacy?
Nature's bequest gives nothing but doth lend,
And being frank she lends to those are free:
Then beauteous niggard why dost thou abuse,
The bounteous largess given thee to give?
Profitless usurer why dost thou use
So great a sum of sums yet canst not live?
For having traffic with thy self alone,
Thou of thy self thy sweet self dost deceive,
Then how when nature calls thee to be gone,
What acceptable audit canst thou leave?
Thy unused beauty must be tombed with thee,
Which used lives th' executor to be.

5

Those hours that with gentle work did frame
The lovely gaze where every eye doth dwell
Will play the tyrants to the very same,
And that unfair which fairly doth excel:
For never-resting time leads summer on
To hideous winter and confounds him there,
Sap checked with frost and lusty leaves quite gone,
Beauty o'er-snowed and bareness every where:

Find and replace

Find: \b(up)*on\b 8 of 380

Replace with:

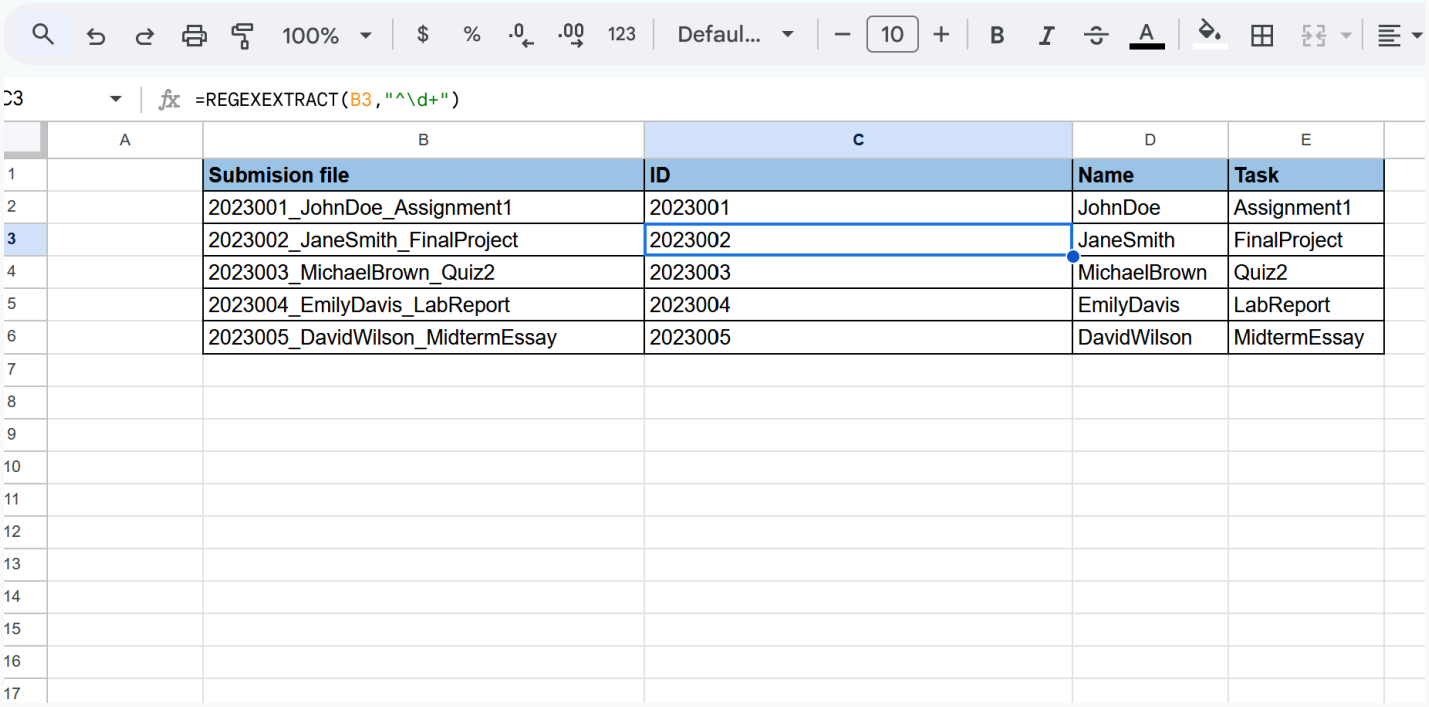
Match case

Use regular expressions (e.g. \n for newline, \t for tab) Help

Ignore diacritics (e.g. ä = a, E = É, א = א)

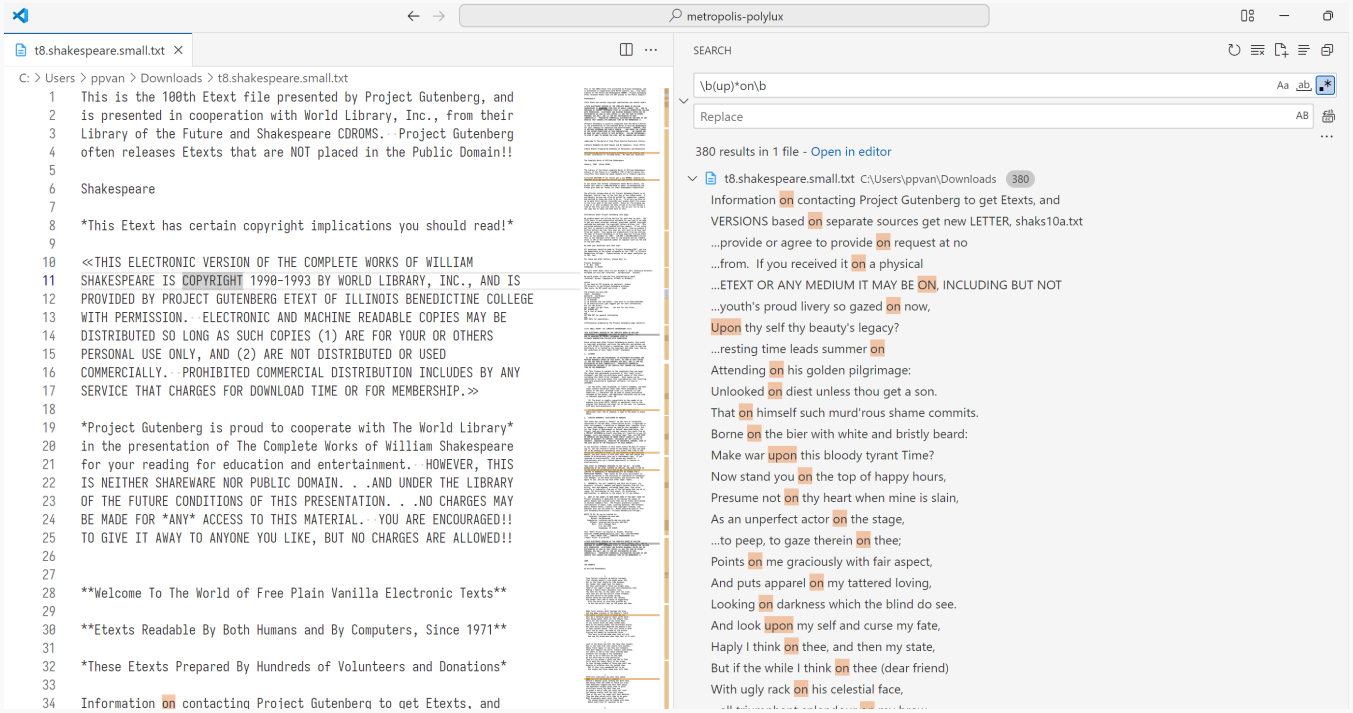
Replace | Replace all | Previous | Next

Excel



	A	B	C	D	E
1		**Submision file**	**ID**	**Name**	**Task**
2		2023001_JohnDoe_Assignment1	2023001	JohnDoe	Assignment1
3		2023002_JaneSmith_FinalProject	2023002	JaneSmith	FinalProject
4		2023003_MichaelBrown_Quiz2	2023003	MichaelBrown	Quiz2
5		2023004_EmilyDavis_LabReport	2023004	EmilyDavis	LabReport
6		2023005_DavidWilson_MidtermEssay	2023005	DavidWilson	MidtermEssay
7					
8					
9					
10					
11					
12					
13					
14					
15					
16					
17					

Text editor


metropolis-polylux
t8.shakespeare.small.txt
C: > Users > ppvan > Downloads > t8.shakespeare.small.txt
1 This is the 100th Etext file presented by Project Gutenberg, and
2 is presented in cooperation with World Library, Inc., from their
3 Library of the Future and Shakespeare CDROMS. Project Gutenberg
4 often releases Etexts that are NOT placed in the Public Domain!!
5
6 Shakespeare
7
8 *This Etext has certain copyright implications you should read!*
9
10 <<THIS ELECTRONIC VERSION OF THE COMPLETE WORKS OF WILLIAM
11 SHAKESPEARE IS COPYRIGHT 1990-1993 BY WORLD LIBRARY, INC., AND IS
12 PROVIDED BY PROJECT GUTENBERG ETEXT OF ILLINOIS BENEDICTINE COLLEGE
13 WITH PERMISSION. ELECTRONIC AND MACHINE READABLE COPIES MAY BE
14 DISTRIBUTED SO LONG AS SUCH COPIES (1) ARE FOR YOUR OR OTHERS
15 PERSONAL USE ONLY, AND (2) ARE NOT DISTRIBUTED OR USED
16 COMMERCIALLY. PROHIBITED COMMERCIAL DISTRIBUTION INCLUDES BY ANY
17 SERVICE THAT CHARGES FOR DOWNLOAD TIME OR FOR MEMBERSHIP.>>
18
19 *Project Gutenberg is proud to cooperate with The World Library*
20 in the presentation of The Complete Works of William Shakespeare
21 for your reading for education and entertainment. HOWEVER, THIS
22 IS NEITHER SHAREWARE NOR PUBLIC DOMAIN. . .AND UNDER THE LIBRARY
23 OF THE FUTURE CONDITIONS OF THIS PRESENTATION. . .NO CHARGES MAY
24 BE MADE FOR *ANY* ACCESS TO THIS MATERIAL. YOU ARE ENCOURAGED!!
25 TO GIVE IT AWAY TO ANYONE YOU LIKE, BUT NO CHARGES ARE ALLOWED!!
26
27
28 **Welcome To The World of Free Plain Vanilla Electronic Texts**
29
30 **Etexts Readable By Both Humans and By Computers, Since 1971**
31
32 *These Etexts Prepared By Hundreds of Volunteers and Donations*
33
34 Information on contacting Project Gutenberg to get Etexts, and
SEARCH
\b(up)*on\b
Replace
380 results in 1 file - Open in editor
t8.shakespeare.small.txt C:\Users\ppvan\Downloads 380
Information on contacting Project Gutenberg to get Etexts, and
VERSIONS based on separate sources get new LETTER, shaks10a.txt
...provide or agree to provide on request at no
...from. If you received it on a physical
...ETEXT OR ANY MEDIUM IT MAY BE ON, INCLUDING BUT NOT
...youth's proud livery so gazed on now,
Upon thy self thy beauty's legacy?
...resting time leads summer on
Attending on his golden pilgrimage:
Unlooked on diest unless thou get a son.
That on himself such murd'rous shame commits.
Borne on the bier with white and bristly beard:
Make war upon this bloody tyrant Time?
Now stand you on the top of happy hours,
Presume not on thy heart when mine is slain,
As an unperfect actor on the stage,
...to peep, to gaze therein on thee;
Points on me graciously with fair aspect,
And puts apparel on my tattered loving,
Looking on darkness which the blind do see.
And look upon my self and curse my fate,
Haply I think on thee, and then my state,
But if the while I think on thee (dear friend)
With ugly rack on his celestial face,

Command line program

```
ppvan at Thinkbook in ~\Downloads
↪rg "\b(up)*on\b" .\t8.shakespeare.small.txt
34:Information on contacting Project Gutenberg to get Etexts, and
48:VERSIONS based on separate sources get new LETTER, shaks10a.txt
167:    (3) You provide or agree to provide on request at no
189:person you received it from.  If you received it on a physical
273:  Thy youth's proud livery so gazed on now,
326:  For never-resting time leads summer on
363:  Attending on his golden pilgrimage:
369:    Unlooked on diest unless thou get a son.
403:    That on himself such murd'rous shame commits.
448:  Borne on the bier with white and bristly beard:
510:  Make war upon this bloody tyrant Time?
513:  Now stand you on the top of happy hours,
623:    Presume not on thy heart when mine is slain,
628:  As an unperfect actor on the stage,
656:  Delights to peep, to gaze therein on thee;
688:  Points on me graciously with fair aspect,
689:  And puts apparel on my tattered loving,
703:  Looking on darkness which the blind do see.
733:  And look upon my self and curse my fate,
739:  Haply I think on thee, and then my state
```

How do I use regex?

Literal (normal character)

- any ascii character

Regex: /a/

Every object will remain at rest or in uniform motion in a straight line unless compelled to change its state by the action of an external force

How do I use regex?

Alternation (pipe)

- a OR b

Regex: `/a|b|e/`

Every object will remain at rest or in uniform motion in a straight line unless compelled to change its state by the action of an external force

How do I use regex?

Characters group

- Match character in []
- [milk] = m|i|l|k

Regex: /[milk]/

Energy cannot be created or destroyed; it can only be transformed from one form to another.

How do I use regex?

Characters group exclude

- Match what's not in the []

Regex: /[^milk]/

Energy cannot be created or destroyed; it can only be transformed from one form to another.

How do I use regex?

Characters group short hand

- [0-9] ≡ [0123456789]
- [a-z] ≡ [abcdefghijklmnopqrstuvwxyz]
- \w ≡ [0-9a-zA-Z_]
- \d ≡ [0-9]
- \W ≡ [^\w]

Regex: /\W/

Energy cannot be created or destroyed; it can only be transformed from one form to another.

How do I use regex?

The dot

- Match every thing
- Escape to match only actual dot (\.)

Regex: /./

Electrons orbit the nucleus only in certain allowed energy levels.

Regex: /\./

Electrons orbit the nucleus only in certain allowed energy levels.

How do I use regex?

Repetitions

- `{3}`: exactly three times
- `{2, 6}`: from 2 to 6 times

Regex: `/w{3}/`

At a constant temperature, the current through a conductor is directly proportional to the voltage across it.

Regex: `/w{2,6}/`

At a constant temperature, the current through a conductor is directly proportional to the voltage across it.

How do I use regex?

Repetitions (unknown times)

- `?`: zezo or one
- `*`: zezo or more
- `+`: one or more

Regex: `/colou?r/`

The colour of light depends on its frequency

Regex: `/\w*e/`

A force can cause an object to move, stop, or change direction

---

Regex: `/s\w+/`

An accelerating object gains speed as time passes

How do I use regex?

Anchor

- ^: start of string
- $: end of string

Regex: /treature/

treature here, treature there,
everywhere treature

Regex: /treature/

treature here, treature there,
everywhere treature

Regex: /treature/

treature here, treature there,
everywhere treature

How do I use regex?

Capture group

- Extract information from match
- Mostly used in code

Regex: `/^(?P<file>.+)\.(?P<ext>.+)$/`

Nắng ấm xa dần.mp3
Hồng nhan.wav
Bạc phận.flac
Faded.mp3

Regular Languages

What is regular languages?

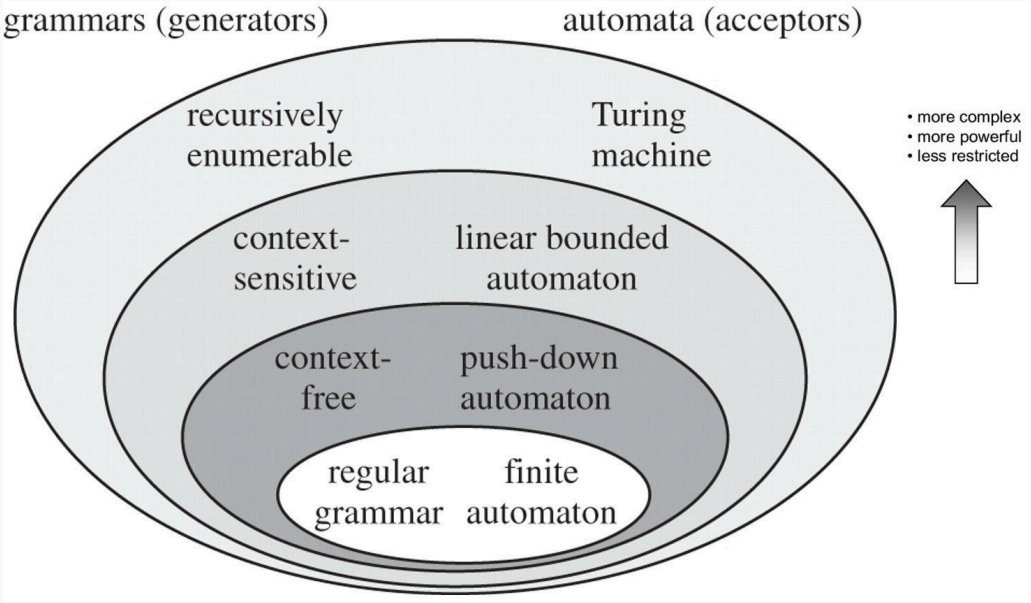
Regular language is a set of string can be recognized by an finite automata (FA).

- Regex is way to describe it

Example:

/a(bb)+a/

→ L = {abba, abbba, abbbbbba..}



Finite Automata?

A thing can “consume” regular language

Regex engine actually compile regex to Non-determistic Finite Automata (NFA)

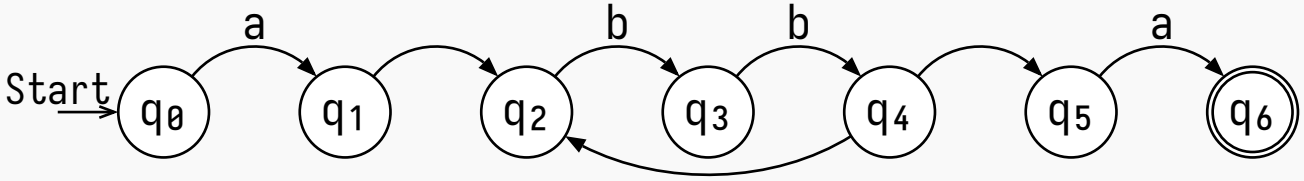


Figure 1: /a(bb)+a/ state machine

NFA properties

- Has a finite number of states
- Doesn’t have auxiliary memory

→ Can’t not consume “context-free” language

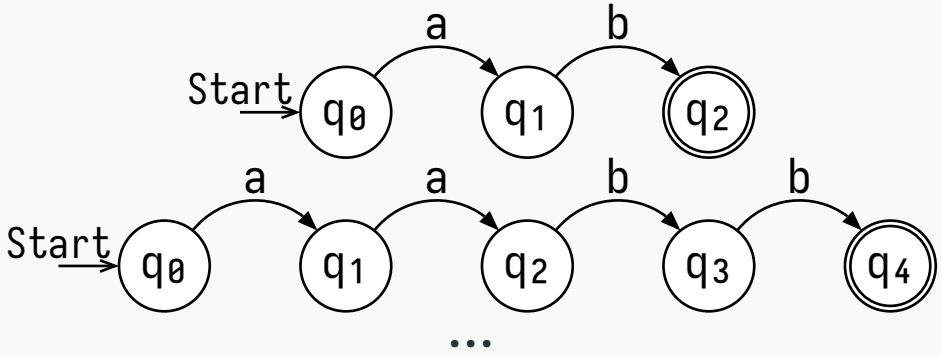


Figure 2: NFA can't represent {ab, aabb, aaabbb, ...}

Backtracking

Regex repetitions (e.g `?*+`) causes backtracking in regex

→ `O(2^n)` complexity

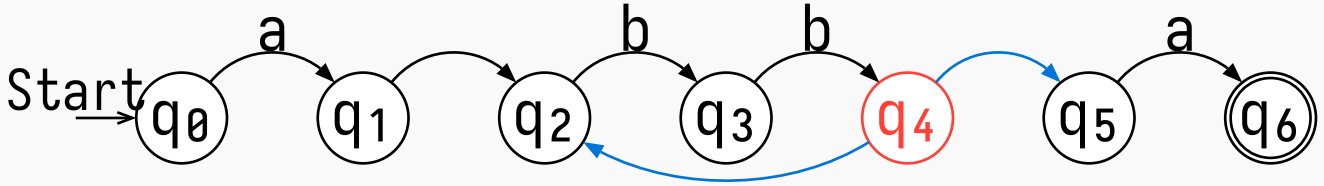


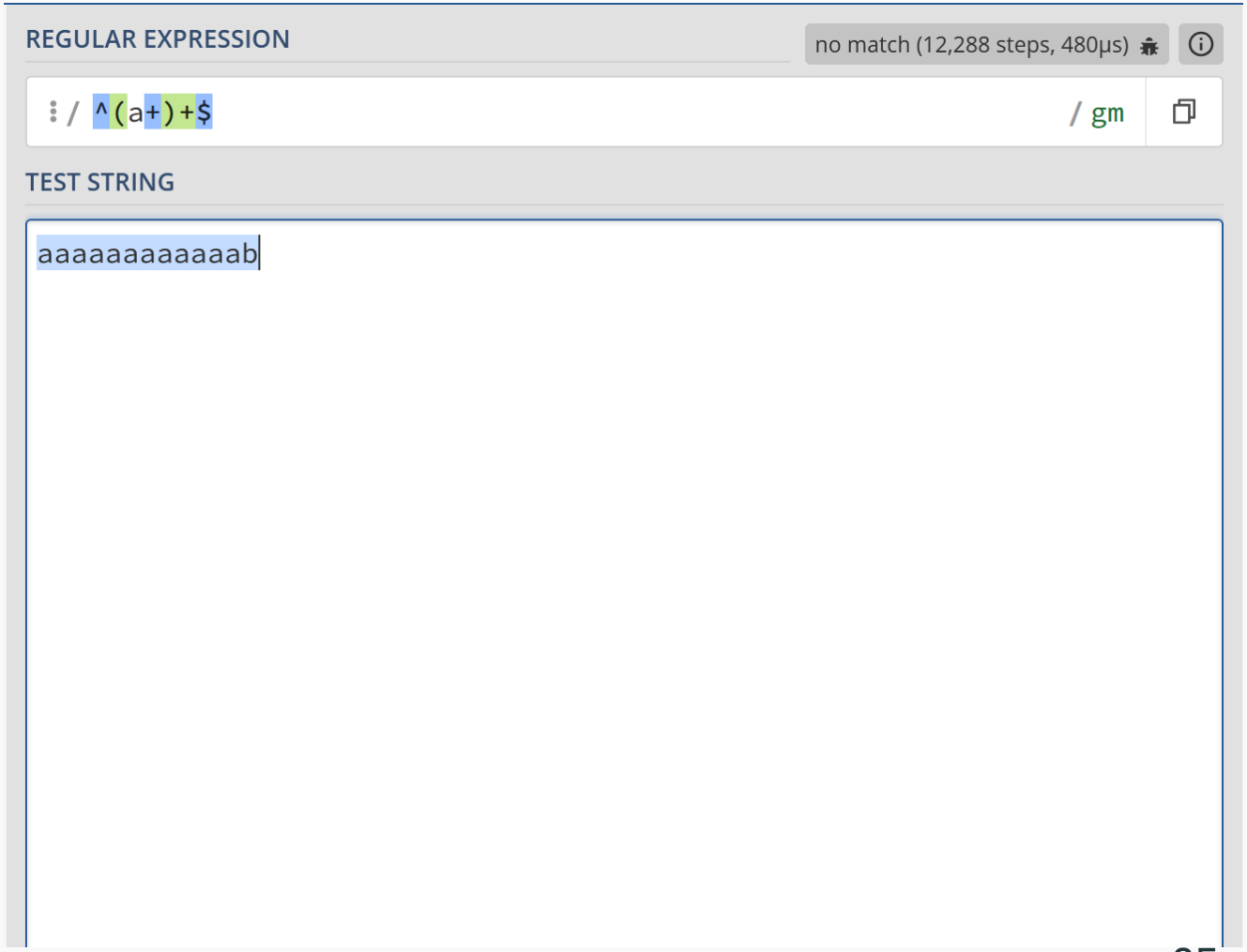
Figure 3: q4 has 2 ways to move

Backtracking

Pattern: `^(a+)+$`

Text: `aaaaaaaaaaaab`

→ ~12.000 steps



Regex tips in practice

Use anchor (^$) when possible

- Make the engine fail early
- Improve accuracy

Regex: /s/

treature here, treature there, everywhere treature

Regex: /treature/

treature here, treature there, everywhere treature

Regex: /treature/

treature here, treature there, everywhere treature

`/.*/` is rarely what you want

Regex: `/user_id="(.*)"/`

```
2024-01-15 user_id="123" action=login error="Invalid password" ip=192.168.1.100 session="abc123"
```

Regex: `/user_id="([^"])"/`

```
2024-01-15 user_id="123" action=login error="Invalid password" ip=192.168.1.100 session="abc123"
```

Regex is code. DO NOT execuate regex from user input

- Regex can check prime number!

Regex: `/^(?!(aa+)\1+$)aa+$/`

aa → 2

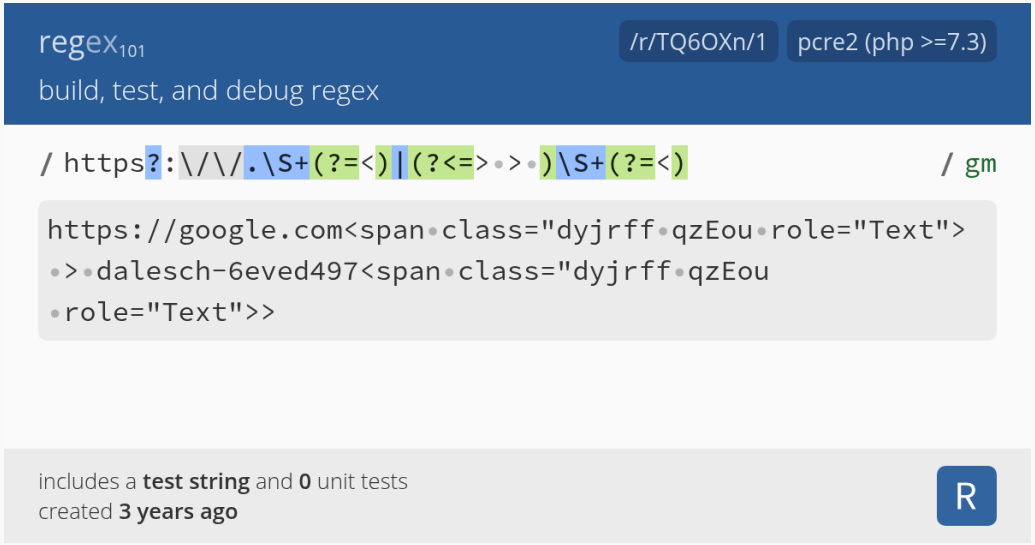
aaa → 3

aaaa → 4

aaaaaaaaaaa → 11

Document your regex

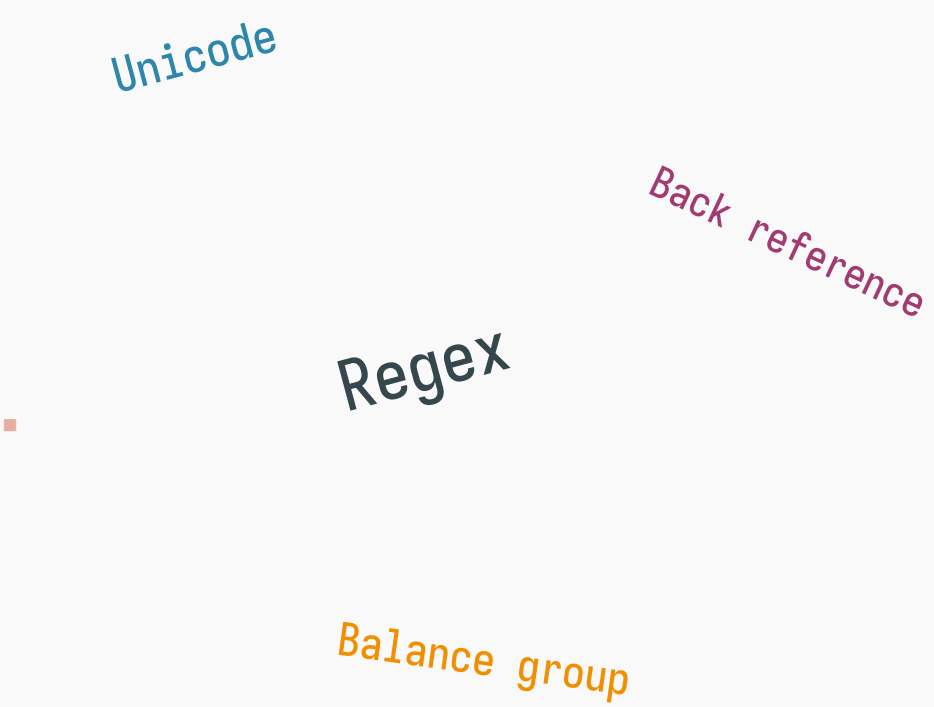
- Always use named group
- Document every thing with EXAMPLES
- Use visualization tools



Only use basic feature

- Advanced features are implementation specific (regex flavors)

→ If thing gets complicated, consider a parser

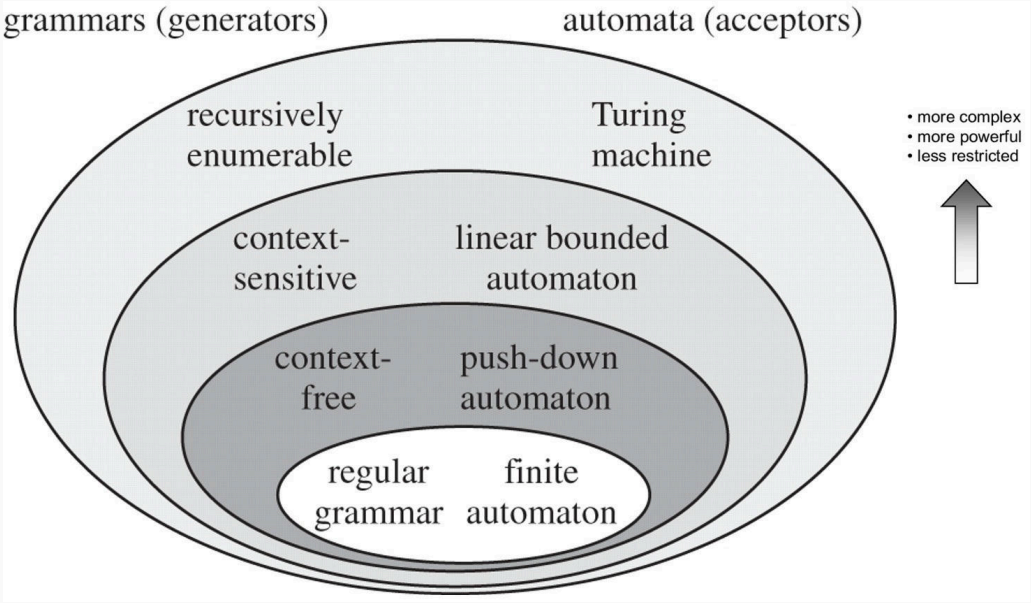


Introduction to parsing

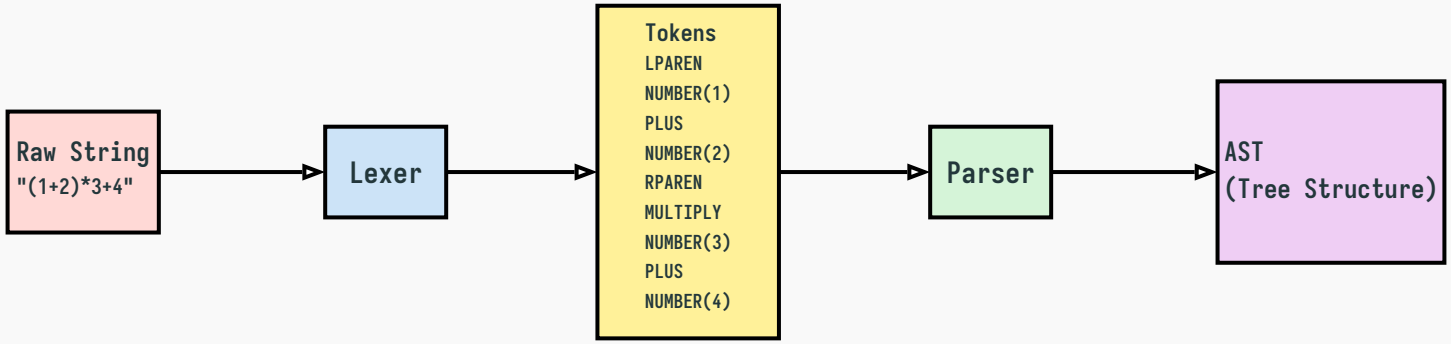
Why we need parser?

Regex can’t handle “context-free” (stuctured) languages

- math expression: `(1+2) * 3 + 4`
- json/xml/html/yaml
- python/javascript



What is a parser?



Compilation Pipeline: Lexical Analysis → Syntax Analysis

Raw String → Lexer → Tokens → Parser → AST

Regex vs AST

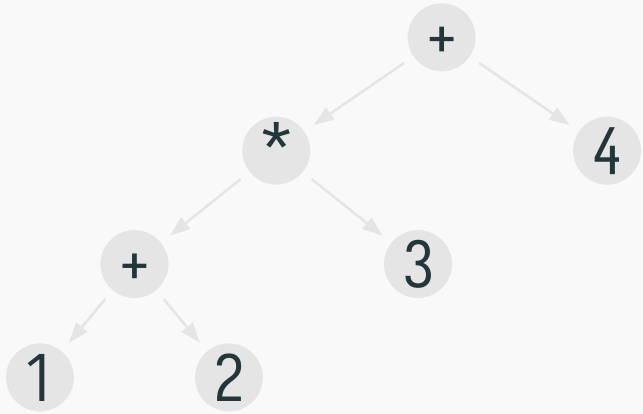
Parsers are more powerful and intuitive

Regex

```
/^\s*(?:\d+|\([^()]*(?:\([^()]*\)
[^()]*)*\))\s*(?:[+\-*\/]\s*(?:\d+|
\([^()]*(?:\([^()]*\)[^()]*)*\))\s*)*$/
```

(1+2) * 3 + 4

AST



Paser learning resource

- Crafting Interpreters - Robert Nystrom
- Compilers: Principles, Techniques, and Tools

Summary

- Regex is a powerful tools to match “flat” structure extremely fast
- Modern regex is complicated and require careful consideration
- When regex become too hard, consider parsing

Questions?