Introduction to Regular Expressions

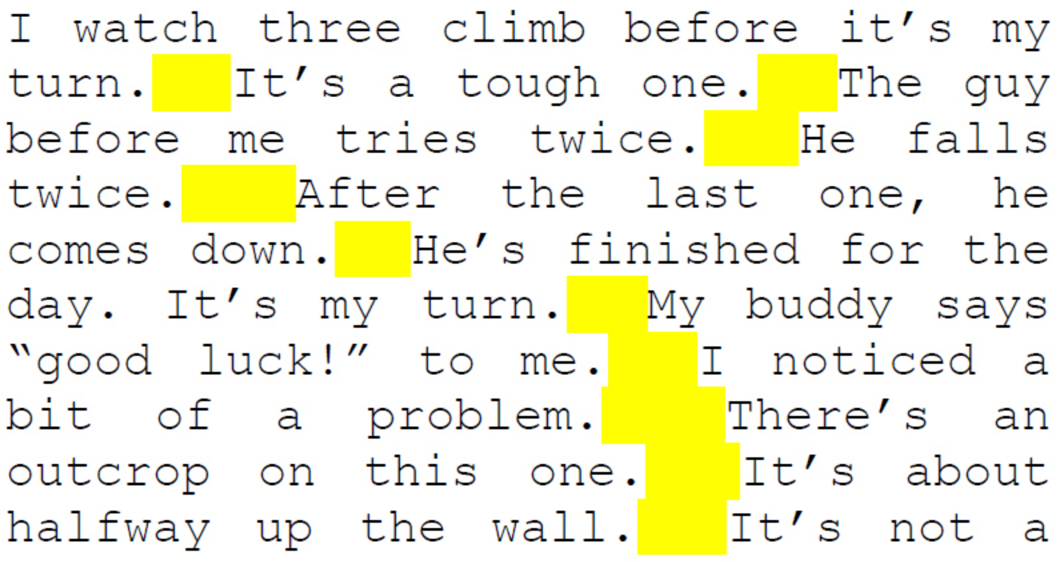
Christopher Ohge, Martin Steer

Riga Technical University, September 2019

Regular expressions

- Sequence of characters that define a search pattern

```
(?<=\.) {2,}(?=[A-Z])
```



I watch three climb before it’s my turn. It’s a tough one. The guy before me tries twice. He falls twice. After the last one, he comes down. He’s finished for the day. It’s my turn. My buddy says “good luck!” to me. I noticed a bit of a problem. There’s an outcrop on this one. It’s about halfway up the wall. It’s not a

https://en.wikipedia.org/wiki/Regular_expression

Regular expressions

[GC]reeting   matches *Greeting* and *Creeting*

Regular expressions

[0123456789]   matches any number

Regular expressions

[0-9]   matches any number

Regular expressions

[0-9I]   matches any number or *I*

So [0-9I]+   matches *1717* or *I7I7*

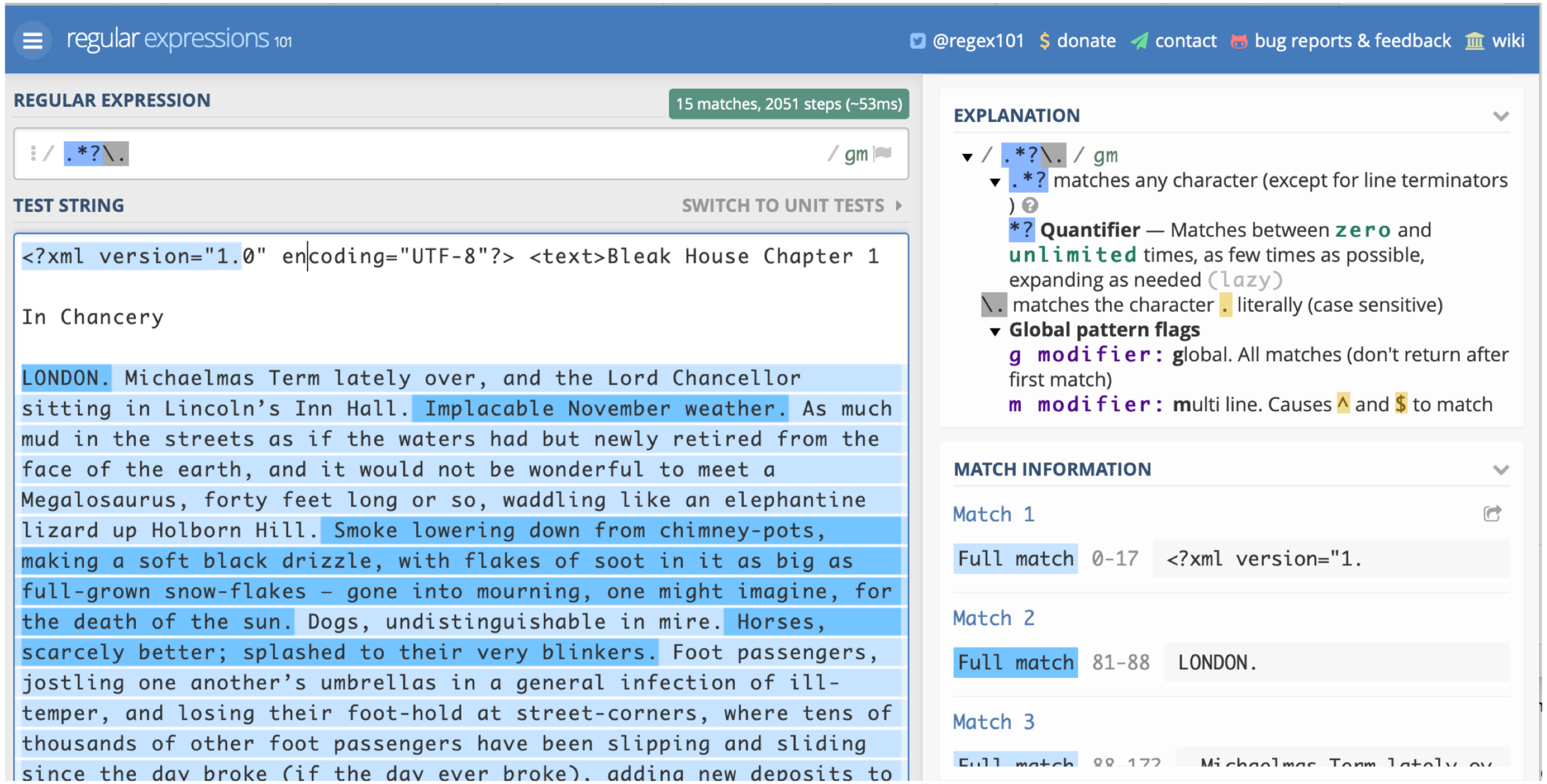
Regular expressions

What does [a-z]+ match?

Demo

https://regex101.com

http://txti.es/bleak-housexml



15 matches, 2051 steps (~53ms)

```
/ .*?\. / gm
```

```
<?xml version="1.0" encoding="UTF-8"?> <text>Bleak House Chapter 1

In Chancery

LONDON. Michaelmas Term lately over, and the Lord Chancellor
sitting in Lincoln’s Inn Hall. Implacable November weather. As much
mud in the streets as if the waters had but newly retired from the
face of the earth, and it would not be wonderful to meet a
Megalosaurus, forty feet long or so, waddling like an elephantine
lizard up Holborn Hill. Smoke lowering down from chimney-pots,
making a soft black drizzle, with flakes of soot in it as big as
full-grown snow-flakes – gone into mourning, one might imagine, for
the death of the sun. Dogs, undistinguishable in mire. Horses,
scarcely better; splashed to their very blinkers. Foot passengers,
jostling one another’s umbrellas in a general infection of ill-
temper, and losing their foot-hold at street-corners, where tens of
thousands of other foot passengers have been slipping and sliding
since the day broke (if the day ever broke), adding new deposits to
```

- / .*?\. / gm
  - .*? matches any character (except for line terminators)
    - *? **Quantifier** — Matches between zero and unlimited times, as few times as possible, expanding as needed (lazy)
  - \. matches the character . literally (case sensitive)
  - **Global pattern flags**
    - g modifier: global. All matches (don't return after first match)
    - m modifier: multi line. Causes ^ and $ to match

Match 1

Full match	0-17	<?xml version="1.

Match 2

Full match	81-88	LONDON.

Match 3

Full match	88-172	Michaelmas Term lately ov

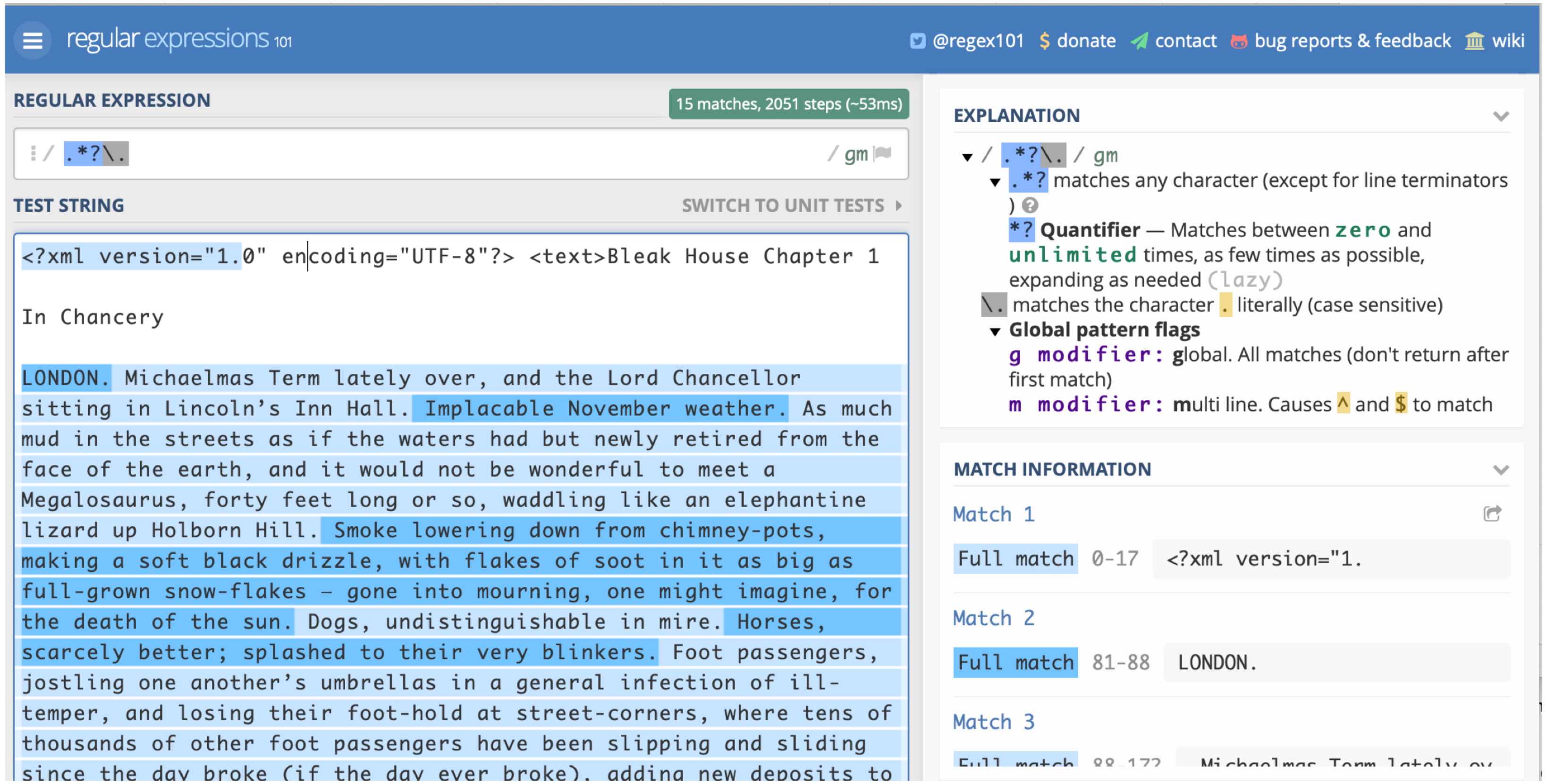
Examples

https://regex101.com

http://txti.es/bleak-housexml

Regular expressions to:

- Match all sentences
- Match all words
- Match all words at end of sentence
- Match all words start with ‘s’
- Match all words ending with ‘s’
- Match all Proper Nouns



REGULAR EXPRESSION

15 matches, 2051 steps (~53ms)

```
/ .*?\. / gm
```

TEST STRING

```
<?xml version="1.0" encoding="UTF-8"?> <text>Bleak House Chapter 1

In Chancery

LONDON. Michaelmas Term lately over, and the Lord Chancellor
sitting in Lincoln’s Inn Hall. Implacable November weather. As much
mud in the streets as if the waters had but newly retired from the
face of the earth, and it would not be wonderful to meet a
Megalosaurus, forty feet long or so, waddling like an elephantine
lizard up Holborn Hill. Smoke lowering down from chimney-pots,
making a soft black drizzle, with flakes of soot in it as big as
full-grown snow-flakes — gone into mourning, one might imagine, for
the death of the sun. Dogs, undistinguishable in mire. Horses,
scarcely better; splashed to their very blinkers. Foot passengers,
jostling one another’s umbrellas in a general infection of ill-
temper, and losing their foot-hold at street-corners, where tens of
thousands of other foot passengers have been slipping and sliding
since the day broke (if the day ever broke), adding new deposits to
```

EXPLANATION

- / .*?\. / gm
  - .*? matches any character (except for line terminators)
    - *? **Quantifier** — Matches between zero and unlimited times, as few times as possible, expanding as needed (lazy)
  - \. matches the character . literally (case sensitive)
  - **Global pattern flags**
    - g modifier: global. All matches (don't return after first match)
    - m modifier: multi line. Causes ^ and $ to match

MATCH INFORMATION

Match 1

Full match	0-17	<?xml version="1.

Match 2

Full match	81-88	LONDON.

Match 3

Full match	88-172	Michaelmas Term lately ov

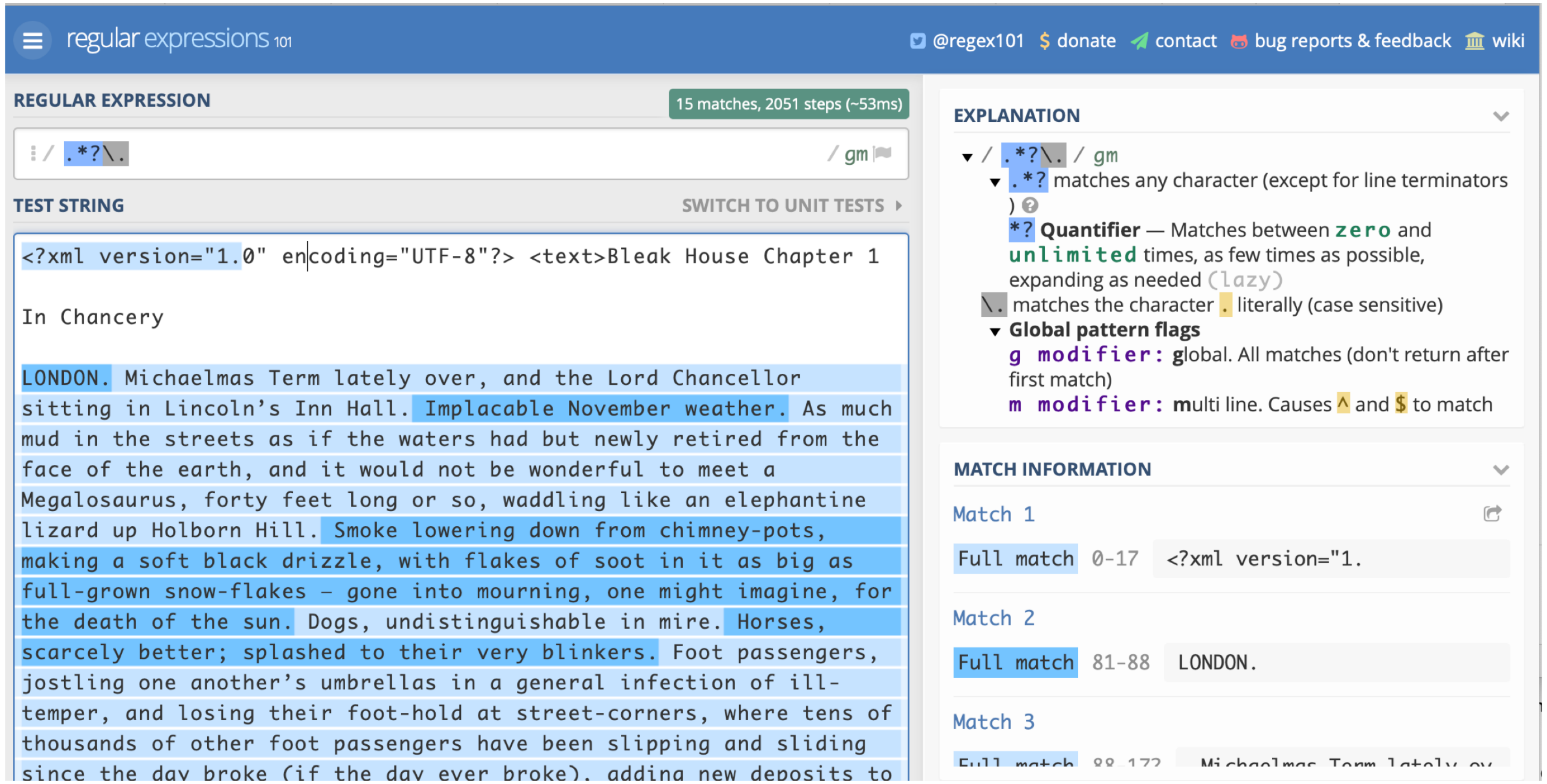
Solutions

https://regex101.com

http://txti.es/bleak-housexml

Regular expressions to:

- Match all sentences
- Match all words
- Match all words at end of sentence
- Match all words start with ‘s’
- Match all words ending with ‘s’
- Match all Proper Nouns



REGULAR EXPRESSION

15 matches, 2051 steps (~53ms)

```
/ .*?\. / gm
```

TEST STRING

SWITCH TO UNIT TESTS

```
<?xml version="1.0" encoding="UTF-8"?> <text>Bleak House Chapter 1

In Chancery

LONDON. Michaelmas Term lately over, and the Lord Chancellor
sitting in Lincoln’s Inn Hall. Implacable November weather. As much
mud in the streets as if the waters had but newly retired from the
face of the earth, and it would not be wonderful to meet a
Megalosaurus, forty feet long or so, waddling like an elephantine
lizard up Holborn Hill. Smoke lowering down from chimney-pots,
making a soft black drizzle, with flakes of soot in it as big as
full-grown snow-flakes – gone into mourning, one might imagine, for
the death of the sun. Dogs, undistinguishable in mire. Horses,
scarcely better; splashed to their very blinkers. Foot passengers,
jostling one another’s umbrellas in a general infection of ill-
temper, and losing their foot-hold at street-corners, where tens of
thousands of other foot passengers have been slipping and sliding
since the day broke (if the day ever broke), adding new deposits to
```

EXPLANATION

- / .*?\. / gm
  - .*? matches any character (except for line terminators)
    - *? **Quantifier** — Matches between zero and unlimited times, as few times as possible, expanding as needed (lazy)
  - \. matches the character . literally (case sensitive)
  - **Global pattern flags**
    - g modifier: global. All matches (don't return after first match)
    - m modifier: multi line. Causes ^ and $ to match

MATCH INFORMATION

Match 1

Full match	0-17	<?xml version="1.

Match 2

Full match	81-88	LONDON.

Match 3

Full match	88-172	Michaelmas Term lately ov

.*?\.   \b.*?\b   \w*\.   \bs.*?\b   [A-Za-z]+s   [A-Z][a-z]*?\b

Regex practice

- https://regexone.com/

Notes

- ar – search for string you want to find
- a.l - wildcards or character classes, the dot
- a.l\. – to find a fullstop (escape with backslash)
- .{5} - quantifiers – number of repetitions, not overlapping
- l{2}
- .* or .+ - all or one or more
- m[aeiou] – any one character, say vowels, or digits [0-9]
- m[0-9]+
- [A-Z][aeiuo]
- \w – any alphanumeric (or \W non alpha)
- \d – any digit (\D non digit)
- \s – any whitespace (space, tab, newline) (or \S non whitespace)
- [a-z]+?s\W
- t.*s - all words begin with T and end with S, then introduce greedy ?
- \b[Tt]\S+?s\b– word break, start with T, end with S, word break