

regular expressions

Fernando Diaz

a lot of data exists as text

news articles



social media



email



messaging



regular expression: a language for writing strings that define the patterns to match in text.

A.L. East is best in baseball!
Aargh! Great Hockey Coverage!! (Devils)
Atlanta Hockey Hell!!
Baseball Stats
Baseball spreads?
College Hockey All-Star Roster
ESPN cares less about hockey
European/Russian Hockey team addresses?
Hockey Equip. Recommendations?
Hockey coverage
Hockey on TV in the Bay area, NOT!
Hockeytipset 93 avgjort
Info - world hockey championships
Isles / Hockey Ramblings
Joe Robbie Stadium "NOT FOR BASEBALL"
Looking for: Strategic Boardgame for Baseball
NCAA Hockey Final
Need software for baseball stats
Official Rules of Baseball ISBN
Please join my hockey playoff pool.
Remarks by President Clinton to NCAA Division I
Champion Hockey Team
Sad day for hockey
Selfish hockey fans..
Some baseball trivia
The Bob Dylan Baseball Abstract
Truly a sad day for hockey
UMass Big East hockey underway
Where can I find baseball statistics ??
baseball in Spanish
hockey playoff pool: LAST CHANCE!
stats for hockey pool
wanted: mail order hockey equipment

A.L. East is best in baseball!
Aargh! Great Hockey Coverage!! (Devils)
Atlanta Hockey Hell!!
Baseball Stats
Baseball spreads?
College Hockey All-Star Roster
ESPN cares less about hockey
European/Russian Hockey team addresses?
Hockey Equip. Recommendations?
Hockey coverage
Hockey on TV in the Bay area, NOT!
Hockeytipset 93 avgjort
Info - world hockey championships
Isles / Hockey Ramblings
Joe Robbie Stadium "NOT FOR BASEBALL"
Looking for: Strategic Boardgame for Baseball
NCAA Hockey Final
Need software for baseball stats
Official Rules of Baseball ISBN
Please join my hockey playoff pool.
Remarks by President Clinton to NCAA Division I
Champion Hockey Team
Sad day for hockey
Selfish hockey fans..
Some baseball trivia
The Bob Dylan Baseball Abstract
Truly a sad day for hockey
UMass Big East hockey underway
Where can I find baseball statistics ??
baseball in Spanish
hockey playoff pool: LAST CHANCE!
stats for hockey pool
wanted: mail order hockey equipment

W

A.L. East is best in baseball!
Aargh! Great Hockey Coverage!! (Devils)
Atlanta Hockey Hell!!
Baseball Stats
Baseball spreads?
College Hockey All-Star Roster
ESPN cares less about hockey
European/Russian Hockey team addresses?
Hockey Equip. Recommendations?
Hockey coverage
Hockey on TV in the Bay area, NOT!
Hockeytipset 93 avgjort
Info - world hockey championships
Isles / Hockey Ramblings
Joe Robbie Stadium "NOT FOR BASEBALL"
Looking for: Strategic Boardgame for Baseball
NCAA Hockey Final
Need software for baseball stats
Official Rules of Baseball ISBN
Please join my hockey playoff pool.
Remarks by President Clinton to NCAA Division I
Champion Hockey Team
Sad day for hockey
Selfish hockey fans..
Some baseball trivia
The Bob Dylan Baseball Abstract
Truly a sad day for hockey
UMass Big East hockey underway
Where can I find baseball statistics ??
baseball in Spanish
hockey playoff pool: LAST CHANCE!
stats for hockey pool
wanted: mail order hockey equipment

hockey

A.L. East is best in baseball!
Aargh! Great Hockey Coverage!! (Devils)
Atlanta Hockey Hell!!
Baseball Stats
Baseball spreads?
College Hockey All-Star Roster
ESPN cares less about hockey
European/Russian Hockey team addresses?
Hockey Equip. Recommendations?
Hockey coverage
Hockey on TV in the Bay area, NOT!
Hockeytipset 93 avgjort
Info - world hockey championships
Isles / Hockey Ramblings
Joe Robbie Stadium "NOT FOR BASEBALL"
Looking for: Strategic Boardgame for Baseball
NCAA Hockey Final
Need software for baseball stats
Official Rules of Baseball ISBN
Please join my hockey playoff pool.
Remarks by President Clinton to NCAA Division I
Champion Hockey Team
Sad day for hockey
Selfish hockey fans..
Some baseball trivia
The Bob Dylan Baseball Abstract
Truly a sad day for hockey
UMass Big East hockey underway
Where can I find baseball statistics ??
baseball in Spanish
hockey playoff pool: LAST CHANCE!
stats for hockey pool
wanted: mail order hockey equipment

[Hh]ockey

A.L. East is best in baseball!
Aargh! Great Hockey Coverage!! (Devils)
Atlanta Hockey Hell!!
Baseball Stats
Baseball spreads?
College Hockey All-Star Roster
ESPN cares less about hockey
European/Russian Hockey team addresses?
Hockey Equip. Recommendations?
Hockey coverage
Hockey on TV in the Bay area, NOT!
Hockeytipset 93 avgjort
Info - world hockey championships
Isles / Hockey Ramblings
Joe Robbie Stadium "NOT FOR BASEBALL"
Looking for: Strategic Boardgame for Baseball
NCAA Hockey Final
Need software for baseball stats
Official Rules of Baseball ISBN
Please join my hockey playoff pool.
Remarks by President Clinton to NCAA Division I
Champion Hockey Team
Sad day for hockey
Selfish hockey fans..
Some baseball trivia
The Bob Dylan Baseball Abstract
Truly a sad day for hockey
UMass Big East hockey underway
Where can I find baseball statistics ??
baseball in Spanish
hockey playoff pool: LAST CHANCE!
stats for hockey pool
wanted: mail order hockey equipment

[Hh] . *

A.L. East is best in baseball!
Aargh! Great Hockey Coverage!! (Devils)
Atlanta Hockey Hell!!
Baseball Stats
Baseball spreads?
College Hockey All-Star Roster
ESPN cares less about hockey
European/Russian Hockey team addresses?
Hockey Equip. Recommendations?
Hockey coverage
Hockey on TV in the Bay area, NOT!
Hockeytipset 93 avgjort
Info - world hockey championships
Isles / Hockey Ramblings
Joe Robbie Stadium "NOT FOR BASEBALL"
Looking for: Strategic Boardgame for Baseball
NCAA Hockey Final
Need software for baseball stats
Official Rules of Baseball ISBN
Please join my hockey playoff pool.
Remarks by President Clinton to NCAA Division I
Champion Hockey Team
Sad day for hockey
Selfish hockey fans..
Some baseball trivia
The Bob Dylan Baseball Abstract
Truly a sad day for hockey
UMass Big East hockey underway
Where can I find baseball statistics ??
baseball in Spanish
hockey playoff pool: LAST CHANCE!
stats for hockey pool
wanted: mail order hockey equipment

. *

```
ca[rt]s?
```

```
ca[rt]s?
```

match a single character **c** in the first position

`ca[rt]s?`

match a single character **a** in the second position

`ca[rt]s?`

match a single character “**r** or **t**” in the third position

`ca[rt]s?`

match zero or one character **s** in the fourth position

`ca[rt]s?`

art
car
cat
carts
cars
cas

literals: characters matched exactly as written.

ca[rt]s?
cat



A diagram illustrating a literal match. The pattern 'ca[rt]s?' is shown with 'ca' underlined in red. The text 'cat' is shown with 'cat' underlined in green. Two red arrows point from the underlined 'ca' in the pattern to the underlined 'cat' in the text, indicating a successful match for the literal part of the pattern.

special characters: literals reserved for pattern matching; must be escaped if matching these characters.

.		\.
^		\^
\$		\\$
*		*
+		\+
?	→	\?
\		\\
[\[
(\(
)		\)

special characters: literals reserved for pattern matching; must be escaped if matching these characters.

\[at

[at

- match any character once

.at

b.t

z.r.

cat

bat

zero

hat

bot

zzrz

at

bxxt

zaaro

^ match character at the beginning of the line

^cat

^.t

^fr.

cat

at

fry

category

iinerary

free

seat

vbt

afraid

\$ match character at the end of the line

catt\$

.t\$

^fr..\$

catt

at

fry

cattegory

itinerary

free

scatt

vbt

afraid

- * match character zero or more times

c.*t

e*t\$

^fre*\$

cat

et

free

caasdat

t

fr

ct

ea

afree

- + match character one or more times

c.+t

e+t\$

^fre+t\$

cat

et

free

caasdat

t

fr

ct

ea

afree

* and + operators are greedy

given a string and a pattern, a regular expression will match as much of the string as possible while satisfying the pattern.

c.+t
caasdat
easdat

? match character zero or one times

c <u>.</u> ?t	<u>e</u> ?t\$	^fr <u>e</u> ?\$
cat	et	free
caasdat	t	fr
ct	ea	afree

`{m}` match character exactly m times

`c.{2}t` `e{3}t$` `^fre{1}$`

<u>co</u> at	<u>ee</u> t	<u>fr</u> ee
ea <u>s</u> dat	t	fr <u>e</u>
e <u>t</u>	ea	a <u>f</u> ree

`[a]` matches any character in the set a

`c[oa]t` `[ea]+t$` `f[re]{3}`

<u>c</u> oat	<u>e</u> t	<u>f</u> ree
<u>c</u> at	<u>ea</u> t	<u>f</u> eer
eo <u>a</u> t	t	fr <u>e</u> re

`[^a]` matches any character
not in the set a

`c[^oa]t` `[^ea]+t`

~~ee~~t

e~~t~~

e~~a~~t

it

cet

oat

`[a-b]` matches any character
in the range a-b*

`c[a-z]t` `[0-9]+t$` `F[A-Z]{3}`

cot

0t

FOUR

cat

213t

FREE

coat

aet

Free

*defined by ASCII codes

$A|B$ matches either pattern A or B

$c[\underline{oa}]t|^wo*l$

cot
wool
wol
catwol
cat

$dog|cat$

cat
dog
squirrel

(A) groups a regular expression

$(dog|cat)^+$

dog
cat
dogcatdogcatcat

$b(ea|oa)t$

beat
boat
~~bat~~

built-in character sets

<code>\s</code>	<code>[\t\n\e\f\v]</code>
<code>\S</code>	<code>[^ \t\n\e\f\v]</code>
<code>\d</code>	<code>[0-9]</code>
<code>\D</code>	<code>[^0-9]</code>
<code>\w</code>	<code>[a-zA-Z0-9_]</code>
<code>\W</code>	<code>[^a-zA-Z0-9_]</code>

*may not work on Windows machines

tools for regular expression matching

grep: command line tool for
detecting regular expressions in
lines of text

egrep: grep with extended
regular expression syntax

```
$ echo "hello world" | egrep "world"
```

world

```
$ echo "hello world" | egrep "w.*d"
```

w.*d

```
$ echo "hello world" | egrep "d.*w"
```

d.*w

common flags

`egrep -c` number of matching lines

`egrep -v` lines **not** matching the pattern

`egrep -i` ignore case

`egrep -An` print **n** lines after each matching line

`egrep -Bn` print **n** lines before each matching line

`egrep -f PATH` read patterns from **PATH**
(one pattern per line)

tools for regular expression processing

sed: command line tool for
substituting regular expressions
in lines of text

```
$ echo "hello wood" | sed -E "s/wood/world/g"
```

replace wood with world.

```
$ echo "hello wood" | sed -E "s/o/0/g"
```

replace o with 0.

```
$ echo "hello wood" | sed -E "s/(oo|11)/\1\1/g"
```

repeat oo or 11 twice.

exercises

1. define a regular expression for zip codes.
2. define a regular expression for US phone numbers.
3. define a regular expression for email addresses from US universities.

exercises

download and uncompress the “20 newsgroups” dataset (20news-19997.tar.gz),

<http://qwone.com/~jason/20Newsgroups/>

1. list the subject lines for all messages in the package.
2. count the number of mentions of “baseball” per newsgroup.
3. count the number of mentions of “hockey” per newsgroup.