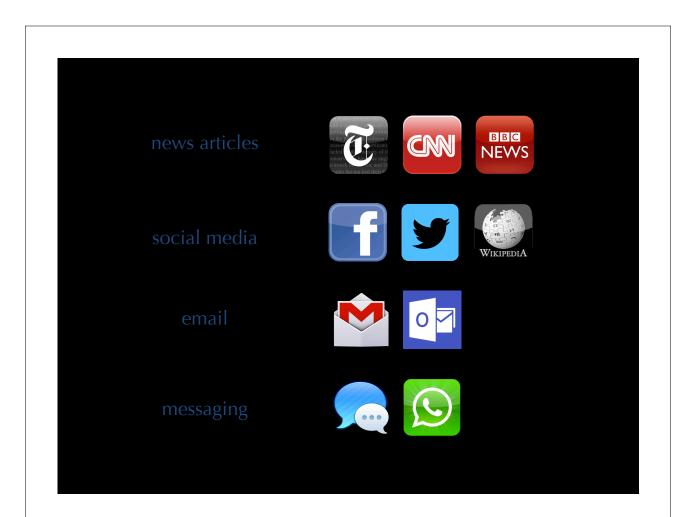
regular expressions Fernando Diaz a lot of data exists as text



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 avgjor

Info - world hockey championships

Isles / Hockey Ramblings

Ioe 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 25

baseball in Spanish

hockey playoff pool: LAST CHANCE!

stats for hockey poo

wanted: mail order hockey equipmen

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

Champion Hockey Team

Sad day for hockey

Selfish hockey fans

Some baseball trivis

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

vanted: 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

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 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

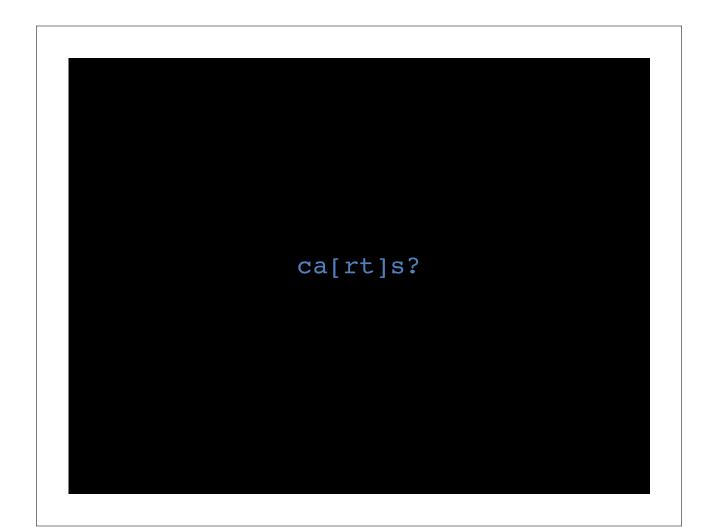
[Hh]ockey

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?

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

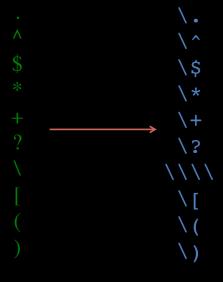
cars

cas

literals: characters matched exactly as written.



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.

<u>\</u>[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

<u>^c</u>at <u>^.</u>t <u>^f</u>r.

\$ match character at the end of the line

* match character zero or more times

+ match character one or more times

* 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<u>.+</u>t c<u>aasda</u>t c<u>aasdat</u>

? match character zero or one times

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

cloalt leal+t$ f[re]{3}

cot     et     free
    cat     eaet     feer
    eoat     t     frare
```

A|B matches either pattern A or
B

cloalt|^wo*l dog|cat

cot cat
wool dog
wol squirrel
eatwol
cat

(A) groups a regular expression

(dog|cat)+ b(ea|oa)t

dog beat

cat boat

dogcatdogcatcat bat

built-in character sets

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

*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

<pre>\$ echo "hello world" egrep "world"</pre>	world



common flags

egrep -c
egrep -v
lines not matching lines
lines not matching the
pattern
lines rease
print n lines after each
matching line
print n lines before each
matching line
egrep -Bn
print n lines before each
matching line
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.

```
replace o with 0.
```

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

repeat oo or ll 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.