String Manipulation and Regular Expressions

Snorre Ralund, Ph.D Fellow, SoDaS, UCPH

KØBENHAVNS UNIVERSITET

It all comes down to a string of text

university_ranking.csv

```
world_rank,university_name,country,total_score,num_students,international_students
1,Harvard University,United States of America,96.1,"20,152",25%
2,California Institute of Technology,United States of America,96.0,"2,243",27%
3,Massachusetts Institute of Technology,United States of America,95.6,"11,074",33%
4,Stanford University,United States of America,94.3,"15,596",22%
5,Princeton University,United States of America,94.2,"7,929",27%
6,University of Cambridge,United Kingdom,91.2,"18,812",34%
6,University of Oxford,United Kingdom,91.2,"19,919",34%
8,"University of California, Berkeley",United States of America,91.1,"36,186",15%
9,Imperial College London,United Kingdom,90.6,"15,060",51%
10,Yale University,United States of America,89.5,"11,751",20%
11,"University of California, Los Angeles",United States of America,87.7,"38,206",15%
12,University of Chicago,United States of America,86.9,"14,221",21%
13,Johns Hopkins University,United States of America,86.4,"15,128",23%
14,Cornell University,United States of America,83.9,"21,424",19%
15,ETH Zurich – Swiss Federal Institute of Technology Zurich,Switzerland,83.4,"18,178",37%
15,University of Michigan,United States of America,83.4,"41,786",16%
17,University of Toronto,Canada,82.0,"66,198",15%
18,Columbia University,United States of America,81.0,"25,055",28%
19,University of Pennsylvania,United States of America,79.5,"20,376",20%
20,Carnegie Mellon University,United States of America,79.3,"11,885",35%
21,University of Hong Kong,Hong Kong,79.2,"19,835",38%
22,University College London,United Kingdom,78.4,"26,607",46%
23,University of Washington,United States of America,78.0,"44,020",13%
24,Duke University,United States of America,76.5,"15,172",17%
25,Northwestern University,United States of America,75.9,"18,334",15%
26,University of Tokyo,Japan,75.6,"26,199",10%
27,Georgia Institute of Technology,United States of America,75.3,"19,967",26%
```

university_ranking.csv 8:24 (1, 14) col# 3, country LF UTF-8 CSV 0 files

It all comes down to a string of text

Python string operations can get you along way.

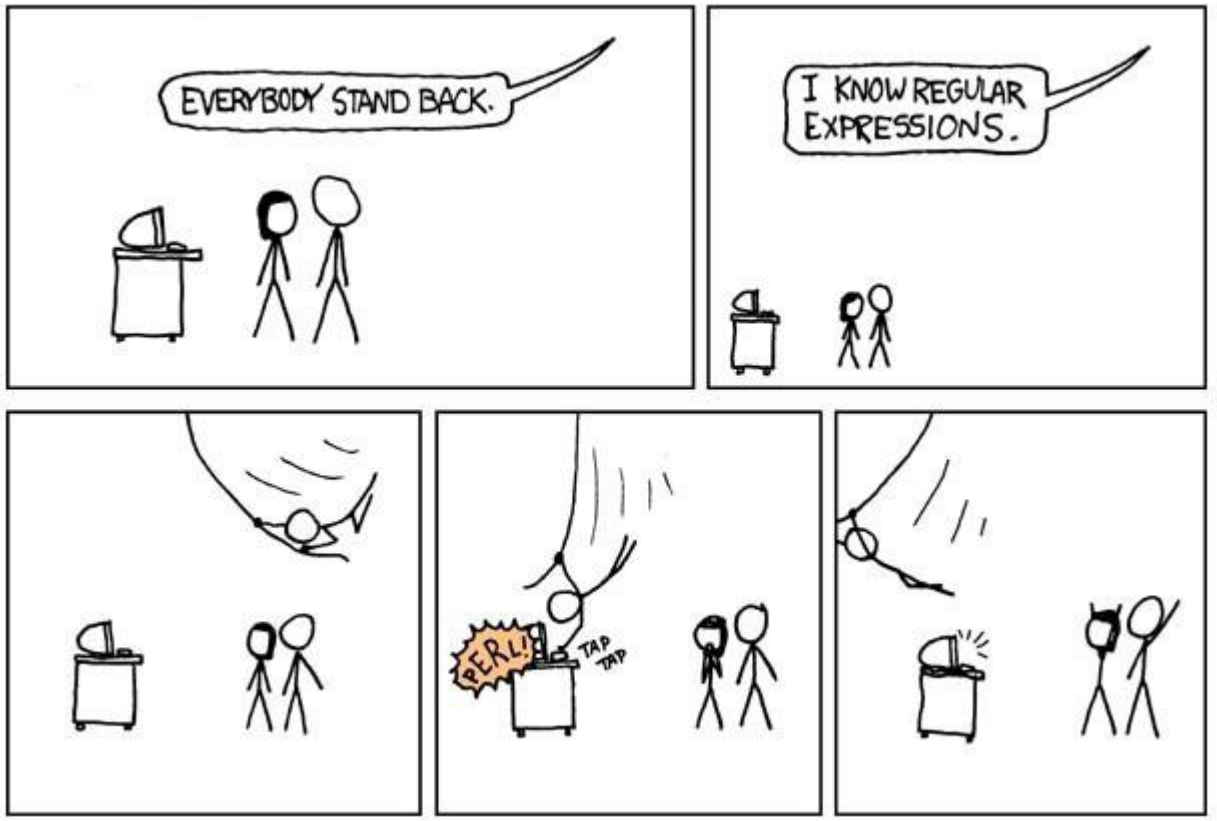
- .split(pattern)
  - E.g. split “Hello word” by whitespace into [‘Hello’,’word’]
- .replace(pattern1,pattern2)
  - E.g. replace all ‘.’ with ‘,’.
- .strip(characters)
  - Greedy removal of characters starting from both ends.
  - E.g. ‘Hello word’.strip(‘Herld’) become ‘o wo’
- .count(pattern)
  - E.g. ‘Hello word’.count(‘o’) is 2

Python string manipulation

```
def simple_csv_reader(filename):
    s = open(filename,'r').read()
    ## remove commas inside ""
    l = s.split('"')
    dat = []
    special_chr = '\x00' # replace commas with specialchr
    for num,i in enumerate(l):
        if num%2==0:
            dat.append(i)
        else:
            dat.append(i.replace(',',special_chr))
    s = ''.join(dat)
    # split by new line
    dat = s.split('\n')[0:-1]
    # split by comma, and replace special character with comma
    dat = [[j.replace(special_chr,',').strip('"') for j in i.split(',')] for i in dat]
    return dat
%timeit simple_csv_reader(filename)
%timeit pd.read_csv(filename)
```

```
535 ms ± 13 ms per loop (mean ± std. dev. of 7 runs, 1 loop each)
339 ms ± 233 µs per loop (mean ± std. dev. of 7 runs, 1 loop each)
```

Regular expressions are **not** avoidable



Regular expressions are **not** avoidable

- **Central to datacleaning and basic information extraction**
- **Central to Understanding Tokenization of documents into words.**
- **Examples:**
  - Extract currency and amount from raw text: $ 20, 10.000 dollars 10,000 £
  - email addresses: here you want to design a pattern (as above), that captures only the uses of @ within an email.
  - urls. Here you are trying to define all the different ways of writing urls (https, http, no http).
  - Dates. Again many variations: 17th of June 2017, 06/17/17 or 17. June 17
  - addresses,
  - phone numbers: 8888888 or 88 88 88 88 or +45 88 88 88 88,
  - emojiies in text. Capturing all the different ways of expressing smiley faces with one regular expression.

Some regular expressions are to be avoided

- pattern = '(?:(?:\r\n)?[\t])*(?:(?:(?:[^()<>@,;:\\".\[\] \000-\031]+(?:(?:(?:\r\n)?[\t])+|\Z|(?=[\["()<>@,;:\\".\[\]]))|"(?:[^\"\r\\]|\\.|(?:(?:\r\n)?[\t]))*"(?:(?:\r\n)?[\t])*)(?:\.(?:(?:\r\n)?[\t])*(?:[^()<>@,;:\\".\[\] \000-\031]+(?:(?:(?:\r\n)?[\t])+|\Z|(?=[\["()<>@,;:\\".\[\]]))|"(?:[^\"\r\\]|\\.|(?:(?:\r\n)?[\t]))*"(?:(?:\r\n)?[\t])*)*@(?:(?:\r\n)?[\t])*(?:[^()<>@,;:\\".\[\] \000-\031]+(?:(?:(?:\r\n)?[\t])+|\Z|(?=[\["()<>@,;:\\".\[\]]))|\[([^\[\]\r\\]|\\.)*\](?:(?:\r\n)?[\t])*)(?:\.(?:(?:\r\n)?[\t])*(?:[^()<>@,;:\\".\[\] \000-\031]+(?:(?:(?:\r\n)?[\t])+|\Z|(?=[\["()<>@,;:\\".\[\]]))|\[([^\[\]\r\\]|\\.)*\](?:(?:\r\n)?[\t])*))*|(?:[^()<>@,;:\\".\[\] \000-\031]+(?:(?:(?:\r\n)?[\t])+|\Z|(?=[\["()<>@,;:\\".\[\]]))|"(?:[^\"\r\\]|\\.|(?:(?:\r\n)?[\t]))*"(?:(?:\r\n)?[\t])*)*\<(?:(?:\r\n)?[\t])*(?:@(?:[^()<>@,;:\\".\[\] \000-\031]+(?:(?:(?:\r\n)?[\t])+|\Z|(?=[\["()<>@,;:\\".\[\]]))|\[([^\[\]\r\\]|\\.)*\](?:(?:\r\n)?[\t])*)(?:\.(?:(?:\r\n)?[\t])*(?:[^()<>@,;:\\".\[\] \000-\031]+(?:(?:(?:\r\n)?[\t])+|\Z|(?=[\["()<>@,;:\\".\[\]]))|\[([^\[\]\r\\]|\\.)*\](?:(?:\r\n)?[\t])*))*(?:,@(?:(?:\r\n)?[\t])*(?:[^()<>@,;:\\".\[\] \000-\031]+(?:(?:(?:\r\n)?[\t])+|\Z|(?=[\["()<>@,;:\\".\[\]]))|\[([^\[\]\r\\]|\\.)*\](?:(?:\r\n)?[\t])*)(?:\.(?:(?:\r\n)?[\t])*(?:[^()<>@,;:\\".\[\] \000-\031]+(?:(?:(?:\r\n)?[\t])+|\Z|(?=[\["()<>@,;:\\".\[\]]))|\[([^\[\]\r\\]|\\.)*\](?:(?:\r\n)?[\t])*))*:(?:(?:\r\n)?[\t])*)?(?:[^()<>@,;:\\".\[\] \000-\031]+(?:(?:(?:\r\n)?[\t])+|\Z|(?=[\["()<>@,;:\\".\[\]]))|"(?:[^\"\r\\]|\\.|(?:(?:\r\n)?[\t]))*"(?:(?:\r\n)?[\t])*)(?:\.(?:(?:\r\n)?[\t])*(?:[^()<>@,;:\\".\[\] \000-\031]+(?:(?:(?:\r\n)?[\t])+|\Z|(?=[\["()<>@,;:\\".\[\]]))|"(?:[^\"\r\\]|\\.|(?:(?:\r\n)?[\t]))*"(?:(?:\r\n)?[\t])*))*@(?:(?:\r\n)?[\t])*(?:[^()<>@,;:\\".\[\] \000-\031]+(?:(?:(?:\r\n)?[\t])+|\Z|(?=[\["()<>@,;:\\".\[\]]))|\[([^\[\]\r\\]|\\.)*\](?:(?:\r\n)?[\t])*)(?:\.(?:(?:\r\n)?[\t])*(?:[^()<>@,;:\\".\[\] \000-\031]+(?:(?:(?:\r\n)?[\t])+|\Z|(?=[\["()<>@,;:\\".\[\]]))|\[([^\[\]\r\\]|\\.)*\](?:(?:\r\n)?[\t])*)*\>(?:(?:\r\n)?[\t])*)|(?:[^()<>@,;:\\".\[\] \000-\031]+(?:(?:(?:\r\n)?[\t])+|\Z|(?=[\["()<>@,;:\\".\[\]]))|"(?:[^\"\r\\]|\\.|(?:(?:\r\n)?[\t]))*"(?:(?:\r\n)?[\t])*)*:(?:(?:\r\n)?[\t])*(?:(?:[^()<>@,;:\\".\[\] \000-\031]+(?:(?:(?:\r\n)?[\t])+|\Z|(?=[\["()<>@,;:\\".\[\]]))|"(?:[^\"\r\\]|\\.|(?:(?:\r\n)?[\t]))*"(?:(?:\r\n)?[\t])*)(?:\.(?:(?:\r\n)?[\t])*(?:[^()<>@,;:\\".\[\] \000-\031]+(?:(?:(?:\r\n)?[\t])+|\Z|(?=[\["()<>@,;:\\".\[\]]))|\[([^\[\]\r\\]|\\.)*\](?:(?:\r\n)?[\t])*))*@(?:(?:\r\n)?[\t])*(?:[^()<>@,;:\\".\[\] \000-\031]+(?:(?:(?:\r\n)?[\t])+|\Z|(?=[\["()<>@,;:\\".\[\]]))|\[([^\[\]\r\\]|\\.)*\](?:(?:\r\n)?[\t])*)(?:\.(?:(?:\r\n)?[\t])*(?:[^()<>@,;:\\".\[\] \000-\031]+(?:(?:(?:\r\n)?[\t])+|\Z|(?=[\["()<>@,;:\\".\[\]]))|\[([^\[\]\r\\]|\\.)*\](?:(?:\r\n)?[\t])*)*\<(?:(?:\r\n)?[\t])*(?:@(?:[^()<>@,;:\\".\[\] \000-\031]+(?:(?:(?:\r\n)?[\t])+|\Z|(?=[\["()<>@,;:\\".\[\]]))|\[([^\[\]\r\\]|\\.)*\](?:(?:\r\n)?[\t])*)(?:\.(?:(?:\r\n)?[\t])*(?:[^()<>@,;:\\".\[\] \000-\031]+(?:(?:(?:\r\n)?[\t])+|\Z|(?=[\["()<>@,;:\\".\[\]]))|\[([^\[\]\r\\]|\\.)*\](?:(?:\r\n)?[\t])*))*(?:,@(?:(?:\r\n)?[\t])*(?:[^()<>@,;:\\".\[\] \000-\031]+(?:(?:(?:\r\n)?[\t])+|\Z|(?=[\["()<>@,;:\\".\[\]]))|\[([^\[\]\r\\]|\\.)*\](?:(?:\r\n)?[\t])*)(?:\.(?:(?:\r\n)?[\t])*(?:[^()<>@,;:\\".\[\] \000-\031]+(?:(?:(?:\r\n)?[\t])+|\Z|(?=[\["()<>@,;:\\".\[\]]))|\[([^\[\]\r\\]|\\.)*\](?:(?:\r\n)?[\t])*))*:(?:(?:\r\n)?[\t])*)?(?:[^()<>@,;:\\".\[\] \000-\031]+(?:(?:(?:\r\n)?[\t])+|\Z|(?=[\["()<>@,;:\\".\[\]]))|"(?:[^\"\r\\]|\\.|(?:(?:\r\n)?[\t]))*"(?:(?:\r\n)?[\t])*)(?:\.(?:(?:\r\n)?[\t])*(?:[^()<>@,;:\\".\[\] \000-\031]+(?:(?:(?:\r\n)?[\t])+|\Z|(?=[\["()<>@,;:\\".\[\]]))|"(?:[^\"\r\\]|\\.|(?:(?:\r\n)?[\t]))*"(?:(?:\r\n)?[\t])*))*@(?:(?:\r\n)?[\t])*(?:[^()<>@,;:\\".\[\] \000-\031]+(?:(?:(?:\r\n)?[\t])+|\Z|(?=[\["()<>@,;:\\".\[\]]))|\[([^\[\]\r\\]|\\.)*\](?:(?:\r\n)?[\t])*)(?:\.(?:(?:\r\n)?[\t])*(?:[^()<>@,;:\\".\[\] \000-\031]+(?:(?:(?:\r\n)?[\t])+|\Z|(?=[\["()<>@,;:\\".\[\]]))|\[([^\[\]\r\\]|\\.)*\](?:(?:\r\n)?[\t])*)*\>(?:(?:\r\n)?[\t])*)(?:,\s*(?:(?:[^()<>@,;:\\".\[\] \000-\031]+(?:(?:(?:\r\n)?[\t])+|\Z|(?=[\["()<>@,;:\\".\[\]]))|"(?:[^\"\r\\]|\\.|(?:(?:\r\n)?[\t]))*"(?:(?:\r\n)?[\t])*)(?:\.(?:(?:\r\n)?[\t])*(?:[^()<>@,;:\\".\[\] \000-\031]+(?:(?:(?:\r\n)?[\t])+|\Z|(?=[\["()<>@,;:\\".\[\]]))|"(?:[^\"\r\\]|\\.|(?:(?:\r\n)?[\t]))*"(?:(?:\r\n)?[\t])*))*@(?:(?:\r\n)?[\t])*(?:[^()<>@,;:\\".\[\] \000-\031]+(?:(?:(?:\r\n)?[\t])+|\Z|(?=[\["()<>@,;:\\".\[\]]))|\[([^\[\]\r\\]|\\.)*\](?:(?:\r\n)?[\t])*)(?:\.(?:(?:\r\n)?[\t])*(?:[^()<>@,;:\\".\[\] \000-\031]+(?:(?:(?:\r\n)?[\t])+|\Z|(?=[\["()<>@,;:\\".\[\]]))|\[([^\[\]\r\\]|\\.)*\](?:(?:\r\n)?[\t])*)*\<(?:(?:\r\n)?[\t])*(?:@(?:[^()<>@,;:\\".\[\] \000-\031]+(?:(?:(?:\r\n)?[\t])+|\Z|(?=[\["()<>@,;:\\".\[\]]))|\[([^\[\]\r\\]|\\.)*\](?:(?:\r\n)?[\t])*)(?:\.(?:(?:\r\n)?[\t])*(?:[^()<>@,;:\\".\[\] \000-\031]+(?:(?:(?:\r\n)?[\t])+|\Z|(?=[\["()<>@,;:\\".\[\]]))|\[([^\[\]\r\\]|\\.)*\](?:(?:\r\n)?[\t])*))*(?:,@(?:(?:\r\n)?[\t])*(?:[^()<>@,;:\\".\[\] \000-\031]+(?:(?:(?:\r\n)?[\t])+|\Z|(?=[\["()<>@,;:\\".\[\]]))|\[([^\[\]\r\\]|\\.)*\](?:(?:\r\n)?[\t])*)(?:\.(?:(?:\r\n)?[\t])*(?:[^()<>@,;:\\".\[\] \000-\031]+(?:(?:(?:\r\n)?[\t])+|\Z|(?=[\["()<>@,;:\\".\[\]]))|\[([^\[\]\r\\]|\\.)*\](?:(?:\r\n)?[\t])*))*:(?:(?:\r\n)?[\t])*)?(?:[^()<>@,;:\\".\[\] \000-\031]+(?:(?:(?:\r\n)?[\t])+|\Z|(?=[\["()<>@,;:\\".\[\]]))|"(?:[^\"\r\\]|\\.|(?:(?:\r\n)?[\t]))*"(?:(?:\r\n)?[\t])*)(?:\.(?:(?:\r\n)?[\t])*(?:[^()<>@,;:\\".\[\] \000-\031]+(?:(?:(?:\r\n)?[\t])+|\Z|(?=[\["()<>@,;:\\".\[\]]))|"(?:[^\"\r\\]|\\.|(?:(?:\r\n)?[\t]))*"(?:(?:\r\n)?[\t])*))*@(?:(?:\r\n)?[\t])*(?:[^()<>@,;:\\".\[\] \000-\031]+(?:(?:(?:\r\n)?[\t])+|\Z|(?=[\["()<>@,;:\\".\[\]]))|\[([^\[\]\r\\]|\\.)*\](?:(?:\r\n)?[\t])*)(?:\.(?:(?:\r\n)?[\t])*(?:[^()<>@,;:\\".\[\] \000-\031]+(?:(?:(?:\r\n)?[\t])+|\Z|(?=[\["()<>@,;:\\".\[\]]))|\[([^\[\]\r\\]|\\.)*\](?:(?:\r\n)?[\t])*)*\>(?:(?:\r\n)?[\t])*))?;\s*)'`

Regular Expressions: Best practices

- Identification: Defining what identifies that pattern, and what does not identify it.
  - E.g. $ identifies a Price.
- Capturing: Defining what we want to match / capture.
  - E.g. All digits around a $ sign. '$[0-9]'

Regular Expression syntax (1)

+ = 1 or more times -- e.g. "a+" will match: "a", and "aaa"

* = 0 or more times -- e.g. "ba*" will match: "b", and "ba", and "baaa"

{3} = exactly three times --- e.g. "ba{3}" will match "baaa", but not "baa"

? = once or none

\ = escape character, used to find characters that has special meaning with regex: e.g. + *

[] = allows you to define a set of characters

^ = applied within a set, it becomes the inverse of the set defined. Applied outside a set it entails the beginning of a string. $ entails the end of a string. '[^0-9]'

. = any characters except line break

| = or statement. -- e.g. a|b means find characters a or b.

\d = digits

\D = any-non-digits.

\s = whitespace-separator

Regular Expression syntax (2)

Sequences

(?:) = Defines a Non-capturing group. -- e.g. "(?:abc)+", will match "abc" and "abcabcabc", but not "aabbcc"

- (?=) = Positive lookahead - only match a certain pattern if a certain pattern comes after it.
- (?!) = Negative lookahead - only match a certain pattern if **not** a certain pattern comes after it.
- (?<=) = Positive lookbehind - only match a certain pattern if a certain pattern precedes it.
- (?<!) = Negative lookbehind - only match a certain pattern if **not** a certain pattern precedes it.

Example: Locating links in text
Defining **Identifiers**

```
# # first we search broadly for www and http
pattern = 'http'
link_dev.explore_pattern(pattern,context=30)
```

```
------ Pattern: http     Matched 21 patterns -----
Match: http     Context:y .. get a cheap one! Look at http://www.radarbusters.com/ or ww
Match: http     Context:ase consider trying to help:

http://www.wm3.org/live/caseinfo/i
Match: http     Context:t download your own drivers - http://www.compgeeks.com/drivers/K
Match: http     Context: and paste this web address:

http://www.arwar.org/wm6caseinform
lick to scroll output; double click to hide ext:sting the web address below:

http://www.wm3.org/live/caseintrod
Match: http     Context:d Catskill's own product at: http://sunnykitchen.com/catskillcr
Match: http     Context: would have. PN is 499432-01. https://servicesales.sel.sony.co .
Match: http     Context:heck out hundreds of posts at http://forums.tivo.com/pe/action/f
Match: http     Context:org/live/caseinfo/index.php
```

Example: Further specification

```
pattern = 'https?://'
link_dev.explore_pattern(pattern,context=30)
```

```
------ Pattern: https?://        Matched 21 patterns -----
Match: https:// Context: would have. PN is 499432401. https://servicesa...s.sel.sony.co ... b
Match: http://  Context:n more about The Yoga Zone at http://www.yogazone.com/. Information
Match: http://  Context:sting the web address below:

http://www.wm3.org/live/caseintroduct
Match: http://  Context:s Rain series you can look at http://wolfsrain.animechain.net. Do n
Match: http://  Context:heck out hundreds of posts at http://forums.tivo.com/pe/action/foru
Match: http://  Context:nd preserve my hearing.  See  http://www.dangerousdecibels.org/hear
Match: http://  Context: it online. Here's the link
```

to scroll output; double click to hide

```
http://www.useandcaremanuals.com/pdf/
Match: http://  Context:ase consider trying to help:

http://www.wm3.org/live/caseinfo/inde
Match: http://  Context:t download your own drivers - http://www.compgeeks.com/drivers/KBGe
Match: http://  Context:as never divorced, ie see
    http://www.columbo-site.freeuk.com/mr
```

Example: Discovery of new identifier

```
------ Pattern: www      Matched 53 patterns -----
Match: www      Context:al human being.
Denise Comeau
www.icefortune.co "The Devil's Pl
Match: www      Context:he web address below:

http://www.wm3.org/live/caseintroduction
Match: www      Context:and seeing he had a web-site (www.ultimatewarrior.com) I eagerl
Match: www      Context:ine. Here's the link

http://www.useandcaremanuals.com/pdf/HS2
Match: www      Context:then holding the W sound like wwwwww. Though most of the sounds
Match: www      Context:entery berries and more! It's Awwwwwwwwesome.Also See The Thing
Match: www      Context:re in amazon.com link:
http://www.amazon.com/gp/product/B000GPQ
Match: www      Context:t a cheap one! Look at http://www.radarbusters.com/ or www.007r
Match: www      Context:s of the page:

http://h20000.www2.hp.com/bc/docs/support/Suppo
Match: www      Context:s coma-inducing phenomenon at www.amazon.com/exec/obidos/tg/lis
```

Example

```
pattern = '\.com'
link_dev.explore_pattern(pattern,context=30)
```

```
------ Pattern: \.com    Matched 176 patterns -----
Match: .com     Context:with all purchases from Amazon.com.  We have only had one articl
Match: .com     Context:ind!
Check eBay, check Amazon.com, check Google "Froogle",
che
Match: .com     Context:dness to come.

P.S: To Amazon.com  Michael McDonald isnt in thi
Match: .com     Context: I am not paid being by AMAZON.com to say all this.  What I say
Match: .com     Context: with only that mocking Amazon.com, 2010 date shown (which basic
Match: .com     Context:hallenges.

~TheRebeccaReview.com One of my favorite seasons. T
Match: .com     Context:day! I just received my Amazon.com..., a Waring FS800 Pro Electr
Match: .com     Context:e film's photo section at imdb.com, and unless it has since been
Match: .com     Context: this DVD recently from Amazon.com, because I injured my back wh
Match: .com     Context:er written a review for Amazon.com before, but I feel compelled
```

Example: Join the identifiers to make on pattern

```
patterns = ['www','http','\.com']
pattern = '|'.join(patterns)
link_dev.explore_pattern(pattern,context=30)
```

```
------ Pattern: www|http|\.com   Matched 250 patterns -----
Match: .com     Context:s depressing. I checked AMAZON.com for "shrimp deveiners" and th
Match: .com     Context:of posts at http://forums.tivo.com/pe/action/forums/displaysingl
Match: http     Context:in without it.

Check it out: http://www.rifftrax.co The memory
Match: .com     Context:official website www.antec-inc.com instead of amazon. this is on
Match: .com     Context:dness to come.

P.S: To Amazon.com  Michael McDonald isnt in thi
Match: www      Context: 911 in Plane Site. Check out www.911inplanesite.com. The movie
Match: .com     Context:. Check out www.911inplanesite.com. The movie shows suppressed n
Match: .com     Context: web-site (www.ultimatewarrior.com) I eagerly logged on to see w
Match: .com     Context:t hesitate to shop with Amazon.com again.

Sandy from W Sit back
Match: www      Context:is on sale at a Friday-sale.  www.amazon.com/fridaysale I wante
```

Example: Get the full link

```
identifiers = ['www','https?://','\.com']
pattern = '|'.join(identifiers)
link_re = '(?:%s)'%pattern # add non-caption group
link_text_re = '[a-zA-Z.]+' # define set of matching characters.
pattern = link_re+link_text_re # add identifier expression with the fulltext expression.
link_dev.explore_pattern(pattern,context=30)
```

```
------ Pattern: (?:www|https?://|\.com)[a-zA-Z.]+        Matched 80 patterns -----
Match: .com.    Context:. Check out www.911inplanesite.com. The movie shows suppressed ne
Match: http://www.useandcaremanuals.com Context: it online. Here's the link

http://www.useandcaremanuals.com/pdf/HS277637.pd Unlike the ov
Match: http://www.dangerousdecibels.org Context:nd preserve my hearing.  See  http://www.dang
erousdecibels.org/hearingloss.cfm.    The Sony A
Match: www.midnitcafe.blogspot.com      Context:ic.

Like this review?  Go to www.midnitcafe.blogspot.com for more The TV stand require
Match: .com.    Context:eviews on both amazon and cnet.com.

The quick installation guide
Match: http://www.arwar.org     Context:org/live/caseinfo/index.php
```

Example: Get the full link

```
# # first we search broadly for www and http
identifiers = ['www','https?://','\.com']
id_pattern = '|'.join(identifiers)
link_re = '(%s)'%id_pattern # add non-caption group
link_text_re = '[^-\s]+' # define the negative set of whitespace
pattern = link_re+link_text_re # add identifier expression with the fulltext expression.
link_dev.explore_pattern(pattern,context=30)
```

```
------ Pattern: (www|https?://|\.com)[^-\s]+     Matched 110 patterns -----
Match: .com..., Context:day! I just received my Amazon.com..., a Waring FS800 Pro Electric F
Match: www.sustworks.com/site/news_usb_ethernet.html,   Context: a driver for OSX 10.3.9 from www.sustworks.com/site/
news_usb_ethernet.html, and the Linksys USB200M works
Match: .com,    Context: with only that mocking Amazon.com, 2010 date shown (which basica
Match: www.AntennaWeb.org       Context: than they really are.  Go to www.AntennaWeb.org to determine what stations b
r
Match: www.wm3.org      Context:the word... tell people about www.wm3.org  tell people about wm3 awaren
Match: www.cutleryandmore.com   Context:t amazon ,but it's cheaper at www.cutleryandmore.com $9.95,so is the $89.00 w
ustho
Match: www.SBrittonReviews.com  Context:nnheiser PC30 is the answer.

www.SBrittonReviews.com Wonderfully done and clear ex
Match: .com,    Context: Bown sight unseen from Amazon.com, and I absolutely love the ric
Match: .completing      Context:ercises being a minute long....completing each one seems like a doable
Match: .com.    Context:1080p) when it comes to Amazon.com. It around $3000.00 After movi
```

Add on: More top-level domain specifiers

- Scrape: https://en.wikipedia.org/wiki/List_of_Internet_top-level_domains

Name	Entity	Administrator[5]	Notes	IDN	DNSSEC	SLD	IPv6
.com	commercial	Verisign	This is an open TLD; any person or entity is permitted to register. Though originally intended for use by for-profit business entities, for a number of reasons it became the "main" TLD for domain names and is currently used by all types of entities including nonprofits, schools, and private individuals. Domain name registrations may be successfully challenged if the holder cannot prove an outside relation justifying reservation of the name,[*citation needed*] to prevent "squatting". It was originally administered by the United States Department of Defense.	Yes	Yes	Yes	Yes
.org	organization	Public Interest Registry	This is an open TLD; any person or entity is permitted to register. Originally intended for use by non-profit organizations, and still primarily used by some.	Yes	Yes	Yes	Yes
.net	network	Verisign	This is an open TLD; any person or entity is permitted to register. Originally intended for use by domains pointing to a distributed network of computers, or "umbrella" sites that act as the portal to a set of smaller websites.	Yes	Yes	Yes	Yes
.int	international organizations	Internet Assigned Numbers Authority	The .int TLD is strictly limited to organizations, offices, and programs endorsed by a treaty between two or more nations. However, a few grandfathered domains do not meet these criteria.	No	Yes	Yes	Yes
.edu	education	Educause (via Verisign)	The .edu TLD is limited to specific higher educational institutions such as, but not limited to, trade schools and universities. In the U.S., its use was restricted in 2001 to post-secondary institutions accredited by an agency on the list of nationally recognized accrediting agencies. This	No	Yes	Yes	Yes

Regular expressions: Precision / Recall

Precision

Percentage of matches correct

- Inspect a sample, and count correct.

Recall

Percentage of positive cases that we identify. E.g. '[;:]-[)(]' will match :-) ;-(but miss the :-D emoticon.

- <- ***harder to ensure, only a thourough search strategy can help establish credibility.***

Regular expressions: Search Strategy

Define-inspect-refine

- **Broad search**
  - Start with simplest and most greedy identifier. E.g. search for all digits '[0-9]' when building a matcher for "Prices".
- **Many and varied inspections**
  - Get many different examples including context to cover the variety.
  - Use negative matching strategies to exclude classes of examples, when searching for different expressions of a pattern.. E.g.
- **Inspect differences between patterns to get a sense of Recall**
  - E.g. compare pattern to initial greedy pattern. E.g. explore_regex.explore_difference('[0-9]','\$[0-9]')

Example: Emoticons

```
emoji_dev = explore_regex.ExploreRegex(text)
```

```
pattern = ':' # start by searching broadly for the eyes
emoji_dev.explore_pattern(':',context=20)
```

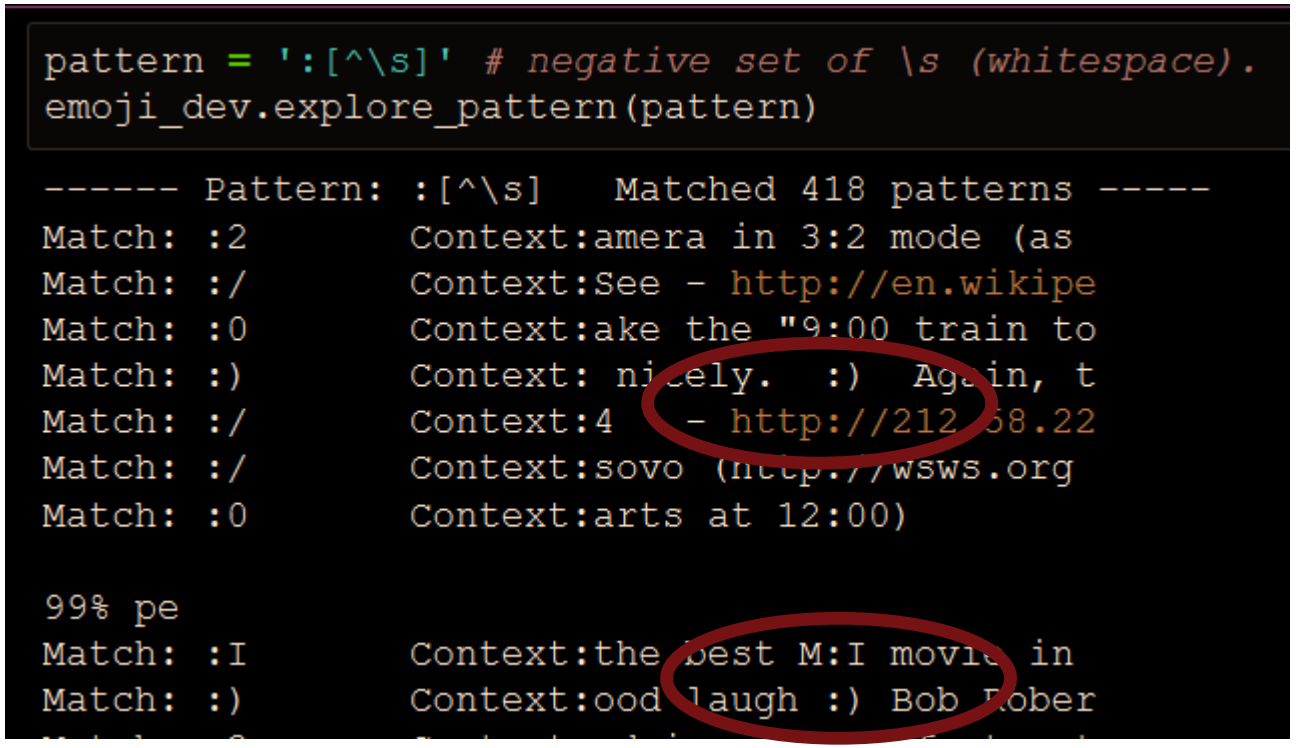
```
------ Pattern: :        Matched 3783 patterns -----
Match: :        Context:
Here's what happens: Han-ki (Jae-hyeon J
Match: :        Context:Introduction to Yoga: Gentle, for Extreme
Match: :        Context:ishwasher safe

Cons:
1.  Do not be foole
Match: :        Context: The Third Dimension: This DVD had some g
Match: :        Context: of two of the leads: one is called Doug
Match: :        Context:arch the Internet on:

Linksys WUSB54GC W
Match: :        Context:rs I have used it on: a dedicated offline
Match: :        Context:tes of the album are:

"Must Be You"
"R
Match: :        Context:nts about this mount:
```

Example: Systematic search for different expressions



Example: Negative match to get different examples

```
mouth_pattern = '[^a-zA-Z\s"!.?)0-9:]-[)(]'
dev.explore_pattern(mouth_pattern,context=30)
```

```
------ Pattern: [^a-zA-Z\s21"!.?)0-9:]-[)(]      Matched 7 patterns -----
Match: ;-)      Context:VCR, though, knows otherwise. ;-)


**********************
EP:
Match: ;-)      Context:that I'm biased or anything.  ;-) I have bought 4 of this model
Match: ;-)      Context:to my Kitchen Aid stand mixer ;-) I really like this dvd overal
Match: ;-)      Context:he child will love it as well.;-)
Elmo is a great master of cer
Match: ;-)      Context:if he can pull that off.......;-) Necesito saber si esta serie
Match: =-(      Context:d buy something else instead. =-( Shame on talented actors like
Match: ;-)      Context: money if they feel like it!) ;-)
```

Example: Structuring the expression

```
EMOTICONS_START = [    r'>:',    r':',    r'=',    r';',    ]
EMOTICONS_MID = [    r'-',    r',',    r'^',    u'\'',    u'\"',    ]
EMOTICONS_END = [    r'D',    r'd',    r'p',    r'P',    r'v',    r')',    r'o',    r'O',    r'(',
r'3(?![0-9])', # Added negative lookahead    r'/',    r'|',    u'\\',    ]
```

Regular expression: Reliability by reporting process

```
import explore_regex
dollar_dev = explore_regex.ExploreRegex(text[0:50000])
```

```
pattern = '[0-9]+' # search
pattern = '[0-9]\.[0-9]+' # narrow search to decimal numbers
pattern = '[0-9]\.[0-9]+ ' # narrow search to add space
pattern = '\$' # broad search for $
pattern = '\$[^0-9]' # negative search for dollar with no digit after
pattern = '\$ ' # narrow search for dollar with space
pattern = 'dollar' # broad search dollar
pattern = '\$ ?[0-9]+' # Specifify capture content
pattern = '\$ ?[0-9]+\.[0-9]+' # Specifify capture content to include decimals
pattern = '\$ ?[0-9]+(?:\.[0-9]+)?' # make the decimals conditional and not defining
dollar_dev.explore_pattern(pattern,context=20)
```

```
------ Pattern: \$ ?[0-9]+(?:\.[0-9]+)?  Matched 5 patterns -----
Match: $110     Context: that sells for the $110 MSRP on Amazon, thi
Match: $60      Context: same item, but for $60 less. A great deal
Match: $100     Context:rdized test. Around $100 more u can get a co
Match: $3.00    Context:ed so much from the $3.00 cups of coffee at S
Match: $25      Context: free shipping over $25 per order After put
```

Regular expression: Reliability by reporting process

```
dollar_dev.report(method='soft')
```

```
------ Pattern: [0-9]+   Matched 117 patterns -----
------ Pattern: [0-9]\.[0-9]+    Matched 7 patterns -----
------ Pattern: [0-9]\.[0-9]+    Matched 5 patterns -----
------ Pattern: \$       Matched 7 patterns -----
------ Pattern: \$[^0-9]         Matched 1 patterns -----
------ Pattern: \$ [^0-9]        Matched 1 patterns -----
------ Pattern: \$       Matched 1 patterns -----
------ Pattern: dollar   Matched 1 patterns -----
------ Pattern: \$ ?[0-9]+       Matched 5 patterns -----
------ Pattern: \$ ?[0-9]+\.[0-9]+       Matched 1 patterns -----
------ Pattern: \$ ?[0-9]+(?:\.[0-9]+)   Matched 1 patterns -----
------ Pattern: \$ ?[0-9]+(?:\.[0-9]+)?  Matched 5 patterns -----
```

Regular expression: Reliability by reporting process