Regular Expression

- RegEx là chuỗi ký tự đặc biệt để so khớp hoặc so sánh chuỗi thỏa điều kiện nào đó.
- Ví dụ:
  - ^a...s$
  - [0-9]{2,4}
- Để sử dụng thư viện RegEx : **import re**
  - re.findall(r"regex",string)
  - re.split(r"regex",string)
  - re.sub(r"regex",new,string)
- Các hàm hỗ trợ RegEx trong pandas.str: extract, split, match,...

Regular Expression

- Một số ký hiệu:
  - Hoặc : |
  - Nhóm : ()
  - Số lượng ký tự : ?*+{m,n}
  - Ký tự đánh dấu: ^ $
  - Ký tự meta : . [] [-][^]
  - Ký tự : \d\D\w\W...
- Ví dụ:

```
re.findall(r"User\d","The winnners are: User3, User5, User9")
```

```
['User3', 'User5', 'User9']
```

Regular Expression

- Số lượng ký tự: ?*+{m,n}
- Ví dụ:
  - “colou?r” ~ “colour” or “color”
  - “94*9” ~ “99” or “9449” or “944449”
  - “36+40” ~ “3640” or “366640”
  - “go{2,3}gle” ~ “google” or “gooogle”
  - “9{3}” ~ “999”
  - “s{2,} ~”ss” or “sss” or “sssss” ....

Regular Expression

- Ký tự đánh dấu : ^ $
- Ví dụ:
  - "^object" ~ "object" or "object-oriented" ...
  - "^2020" ~ "2020" or "2020/01/05" ...
  - "er$" ~ "driver" or "programer" ...
  - "2019$" ~ "2019" or "05/01/2019" ...

Regular Expression

- Ký tự meta : . [] [-][^]
- Ví dụ:
  - “87.1” ~ “8721” or “8731” or “8751”
  - “[xyz]” ~ “x” or “y” or “z”
  - “[a-zA-Z]” -> tất cả ký tự (chữ hoa, chữ thường)
  - “[^0-9]” -> Không lấy các ký số từ 0-9

Regular Expression

- Ký tự : \d\D\w\W...
- Ví dụ:
  - \d : ký số [0-9]
  - \D : không phải ký số
  - \s : ký tự đơn là tab(\t), newline (\n), khoảng trắng (\v)
  - \w : ký tự [a-zA-Z0-9]
  - \w+ : 1 hoặc nhiều ký tự [a-zA-Z0-9]

```
re.findall(r"\W\d+","The book is on sale, just only $10 today")
```

```
['$20']
```

Regular Expression

```
# Giới tính -
df['female'] = df['data'].str.extract('(\d)', expand=True)
df
```

	data	female
0	Arizona 1 2014-12-23 3242.0	1
1	Iowa 1 2010-02-23 3453.7	1
2	Oregon 0 2014-06-20 2123.0	0
3	Maryland 0 2014-03-14 1123.6	0
4	Florida 1 2013-01-15 2134.0	1
5	Georgia 0 2012-07-14 2345.6	0

```
# Nơi đăng ký
df['state'] = df['data'].str.extract('([A-Z]\w{0,})', expand=True)
df
```

	data	female	date	score	state
0	Arizona 1 2014-12-23 3242.0	1	2014-12-23	3242.0	Arizona
1	Iowa 1 2010-02-23 3453.7	1	2010-02-23	3453.7	Iowa
2	Oregon 0 2014-06-20 2123.0	0	2014-06-20	2123.0	Oregon
3	Maryland 0 2014-03-14 1123.6	0	2014-03-14	1123.6	Maryland
4	Florida 1 2013-01-15 2134.0	1	2013-01-15	2134.0	Florida
5	Georgia 0 2012-07-14 2345.6	0	2012-07-14	2345.6	Georgia

	data
0	Arizona 1 2014-12-23 3242.0
1	Iowa 1 2010-02-23 3453.7
2	Oregon 0 2014-06-20 2123.0
3	Maryland 0 2014-03-14 1123.6
4	Florida 1 2013-01-15 2134.0
5	Georgia 0 2012-07-14 2345.6