egex-for-information-extraction-1

March 3, 2024

NLP Tutorial: Regular Expressions

(1) Regex in customer support

Retrieve order number

[1]:
```
import re

chat1='codebasics: Hello, I am having an issue with my order # 412889912'

pattern = 'order[^\d]*(\d*)'
matches = re.findall(pattern, chat1)
matches
```

[1]:
```
['412889912']
```

[2]:
```
chat2='codebasics: I have a problem with my order number 412889912'
pattern = 'order[^\d]*(\d*)'
matches = re.findall(pattern, chat2)
matches
```

[2]:
```
['412889912']
```

[4]:
```
chat3='codebasics: My order 412889912 is having an issue, I was charged 300$␣
  ↪when online it says 280$'
pattern = 'order[^\d]*(\d*)'
matches = re.findall(pattern, chat3)
matches
```

[4]:
```
['412889912']
```

[5]:
```
def get_pattern_match(pattern, text):
    matches = re.findall(pattern, text)
    if matches:
        return matches[0]
```

[6]:
```
get_pattern_match('order[^\d]*(\d*)', chat1)
```

[6]:
```
'412889912'
```

Retrieve email id and phone

```
[7]: chat1 = 'codebasics: you ask lot of questions   1235678912, abc@xyz.com'
     chat2 = 'codebasics: here it is: (123)-567-8912, abc@xyz.com'
     chat3 = 'codebasics: yes, phone: 1235678912 email: abc@xyz.com'
```

——Email id——

```
[8]: get_pattern_match('[a-zA-Z0-9_]*@[a-z]*\.[a-zA-Z0-9]*',chat1)
```

```
[8]: 'abc@xyz.com'
```

```
[9]: get_pattern_match('[a-zA-Z0-9_]*@[a-z]*\.[a-zA-Z0-9]*',chat2)
```

```
[9]: 'abc@xyz.com'
```

```
[10]: get_pattern_match('[a-zA-Z0-9_]*@[a-z]*\.[a-zA-Z0-9]*',chat3)
```

```
[10]: 'abc@xyz.com'
```

——Phone number——

```
[11]: get_pattern_match('(\d{10})|(\(\d{3}\)-\d{3}-\d{4})',chat1)
```

```
[11]: ('1235678912', '')
```

```
[12]: get_pattern_match('(\d{10})|(\(\d{3}\)-\d{3}-\d{4})', chat2)
```

```
[12]: ('', '(123)-567-8912')
```

```
[13]: get_pattern_match('(\d{10})|(\(\d{3}\)-\d{3}-\d{4})', chat3)
```

```
[13]: ('1235678912', '')
```

(2) Regex for Information Extraction

```
[14]: text='''
      Born        Elon Reeve Musk
      June 28, 1971 (age 50)
      Pretoria, Transvaal, South Africa
      Citizenship
      South Africa (1971–present)
      Canada (1971–present)
      United States (2002–present)
      Education        University of Pennsylvania (BS, BA)
      Title
      Founder, CEO and Chief Engineer of SpaceX
      CEO and product architect of Tesla, Inc.
      Founder of The Boring Company and X.com (now part of PayPal)
```

```
Co-founder of Neuralink, OpenAI, and Zip2
Spouse(s)
Justine Wilson

(m. 2000; div. 2008)
Talulah Riley

(m. 2010; div. 2012)

(m. 2013; div. 2016)
'''
```

[15]:
```python
get_pattern_match(r'age (\d+)', text)
```

[15]:
```
'50'
```

[16]:
```python
get_pattern_match(r'Born(.*)\n', text).strip()
```

[16]:
```
'Elon Reeve Musk'
```

[17]:
```python
get_pattern_match(r'Born.*\n(.*)\(age', text).strip()
```

[17]:
```
'June 28, 1971'
```

[18]:
```python
get_pattern_match(r'\(age.*\n(.*)', text)
```

[18]:
```
'Pretoria, Transvaal, South Africa'
```

[19]:
```python
def extract_personal_information(text):
    age = get_pattern_match('age (\d+)', text)
    full_name = get_pattern_match('Born(.*)\n', text)
    birth_date = get_pattern_match('Born.*\n(.*)\(age', text)
    birth_place = get_pattern_match('\(age.*\n(.*)', text)
    return {
        'age': int(age),
        'name': full_name.strip(),
        'birth_date': birth_date.strip(),
        'birth_place': birth_place.strip()
    }
```

[20]:
```python
extract_personal_information(text)
```

[20]:
```
{'age': 50,
 'name': 'Elon Reeve Musk',
 'birth_date': 'June 28, 1971',
 'birth_place': 'Pretoria, Transvaal, South Africa'}
```

[21]:

```
text = '''
Born        Mukesh Dhirubhai Ambani
19 April 1957 (age 64)
Aden, Colony of Aden
(present-day Yemen)[1][2]
Nationality        Indian
Alma mater
St. Xavier's College, Mumbai
Institute of Chemical Technology (B.E.)
Stanford University (drop-out)
Occupation        Chairman and MD, Reliance Industries
Spouse(s)        Nita Ambani (m. 1985)[3]
Children        3
Parent(s)
Dhirubhai Ambani (father)
Kokilaben Ambani (mother)
Relatives        Anil Ambani (brother)
Tina Ambani (sister-in-law)
'''
```

[22]:

```
extract_personal_information(text)
```

[22]:

```
{'age': 64,
 'name': 'Mukesh Dhirubhai Ambani',
 'birth_date': '19 April 1957',
 'birth_place': 'Aden, Colony of Aden'}
```

References

Please refer to my videon on python regular expressions to learn more: https://www.youtube.com/watch?v=sHw5hLYFaIw

Here is the code of that video: https://github.com/codebasics/py/blob/master/Advanced/regex/regex_tutorial_p

Exercise

https://github.com/codebasics/nlp-tutorials/blob/main/1_regex/regex_nlp_exercise_questions.ipynb