In [1]:
```
# Regular expression - Regex
import re
```

In [2]:
```
text ="""I want to be a rich person with at least 100000 salary and my mobile no is
pattern = "\d{3}\-\d{3}-\d{4}|\d{6}"
matches =re.findall (pattern, text)
matches
```

Out[2]:
```
['100000', '944-567-6421']
```

In [ ]:
```
!pip install spacy
!python -m spacy download en_core_web_sm
!python -m spacy download en
!python -m spacy download en_core_web_sm-2.2.0
```

```
Requirement already satisfied: spacy in c:\users\s323\anaconda3\lib\site-packages
(3.5.0)
Requirement already satisfied: wasabi<1.2.0,>=0.9.1 in c:\users\s323\anaconda3\lib
\site-packages (from spacy) (1.1.1)
Requirement already satisfied: smart-open<7.0.0,>=5.2.1 in c:\users\s323\anaconda3
\lib\site-packages (from spacy) (6.3.0)
Requirement already satisfied: murmurhash<1.1.0,>=0.28.0 in c:\users\s323\anaconda
3\lib\site-packages (from spacy) (1.0.9)
Requirement already satisfied: numpy>=1.15.0 in c:\users\s323\anaconda3\lib\site-p
ackages (from spacy) (1.21.5)
Requirement already satisfied: spacy-loggers<2.0.0,>=1.0.0 in c:\users\s323\anacon
da3\lib\site-packages (from spacy) (1.0.4)
Requirement already satisfied: thinc<8.2.0,>=8.1.0 in c:\users\s323\anaconda3\lib
\site-packages (from spacy) (8.1.7)
Requirement already satisfied: jinja2 in c:\users\s323\anaconda3\lib\site-packages
(from spacy) (2.11.3)
Requirement already satisfied: spacy-legacy<3.1.0,>=3.0.11 in c:\users\s323\anacon
da3\lib\site-packages (from spacy) (3.0.12)
Requirement already satisfied: typer<0.8.0,>=0.3.0 in c:\users\s323\anaconda3\lib
\site-packages (from spacy) (0.7.0)
Requirement already satisfied: pathy>=0.10.0 in c:\users\s323\anaconda3\lib\site-p
ackages (from spacy) (0.10.1)
Requirement already satisfied: srsly<3.0.0,>=2.4.3 in c:\users\s323\anaconda3\lib
\site-packages (from spacy) (2.4.5)
Requirement already satisfied: requests<3.0.0,>=2.13.0 in c:\users\s323\anaconda3
\lib\site-packages (from spacy) (2.27.1)
Requirement already satisfied: langcodes<4.0.0,>=3.2.0 in c:\users\s323\anaconda3
\lib\site-packages (from spacy) (3.3.0)
Requirement already satisfied: packaging>=20.0 in c:\users\s323\anaconda3\lib\site
-packages (from spacy) (21.3)
Requirement already satisfied: catalogue<2.1.0,>=2.0.6 in c:\users\s323\anaconda3
\lib\site-packages (from spacy) (2.0.8)
Requirement already satisfied: cymem<2.1.0,>=2.0.2 in c:\users\s323\anaconda3\lib
\site-packages (from spacy) (2.0.7)
Requirement already satisfied: preshed<3.1.0,>=3.0.2 in c:\users\s323\anaconda3\li
b\site-packages (from spacy) (3.0.8)
Requirement already satisfied: pydantic!=1.8,!=1.8.1,<1.11.0,>=1.7.4 in c:\users\s
323\anaconda3\lib\site-packages (from spacy) (1.10.4)
Requirement already satisfied: setuptools in c:\users\s323\anaconda3\lib\site-pack
ages (from spacy) (61.2.0)
Requirement already satisfied: tqdm<5.0.0,>=4.38.0 in c:\users\s323\anaconda3\lib
\site-packages (from spacy) (4.64.0)
Requirement already satisfied: pyparsing!=3.0.5,>=2.0.2 in c:\users\s323\anaconda3
\lib\site-packages (from packaging>=20.0->spacy) (3.0.4)
Requirement already satisfied: typing-extensions>=4.2.0 in c:\users\s323\anaconda3
\lib\site-packages (from pydantic!=1.8,!=1.8.1,<1.11.0,>=1.7.4->spacy) (4.4.0)
Requirement already satisfied: charset-normalizer~=2.0.0 in c:\users\s323\anaconda
3\lib\site-packages (from requests<3.0.0,>=2.13.0->spacy) (2.0.4)
Requirement already satisfied: urllib3<1.27,>=1.21.1 in c:\users\s323\anaconda3\li
b\site-packages (from requests<3.0.0,>=2.13.0->spacy) (1.26.9)
Requirement already satisfied: certifi>=2017.4.17 in c:\users\s323\anaconda3\lib\s
ite-packages (from requests<3.0.0,>=2.13.0->spacy) (2021.10.8)
Requirement already satisfied: idna<4,>=2.5 in c:\users\s323\anaconda3\lib\site-pa
ckages (from requests<3.0.0,>=2.13.0->spacy) (3.3)
Requirement already satisfied: blis<0.8.0,>=0.7.8 in c:\users\s323\anaconda3\lib\s
ite-packages (from thinc<8.2.0,>=8.1.0->spacy) (0.7.9)
Requirement already satisfied: confection<1.0.0,>=0.0.1 in c:\users\s323\anaconda3
\lib\site-packages (from thinc<8.2.0,>=8.1.0->spacy) (0.0.4)
Requirement already satisfied: colorama in c:\users\s323\anaconda3\lib\site-packag
es (from tqdm<5.0.0,>=4.38.0->spacy) (0.4.6)
Requirement already satisfied: click<9.0.0,>=7.1.1 in c:\users\s323\anaconda3\lib
\site-packages (from typer<0.8.0,>=0.3.0->spacy) (8.0.4)
Requirement already satisfied: MarkupSafe>=0.23 in c:\users\s323\anaconda3\lib\sit
e-packages (from jinja2->spacy) (2.0.1)
```

```
Collecting en-core-web-sm==3.5.0
  Downloading https://github.com/explosion/spacy-models/releases/download/en_core_
web_sm-3.5.0/en_core_web_sm-3.5.0-py3-none-any.whl (12.8 MB)
Requirement already satisfied: spacy<3.6.0,>=3.5.0 in c:\users\s323\anaconda3\lib
\site-packages (from en-core-web-sm==3.5.0) (3.5.0)
Requirement already satisfied: pathy>=0.10.0 in c:\users\s323\anaconda3\lib\site-p
ackages (from spacy<3.6.0,>=3.5.0->en-core-web-sm==3.5.0) (0.10.1)
Requirement already satisfied: smart-open<7.0.0,>=5.2.1 in c:\users\s323\anaconda3
\lib\site-packages (from spacy<3.6.0,>=3.5.0->en-core-web-sm==3.5.0) (6.3.0)
Requirement already satisfied: catalogue<2.1.0,>=2.0.6 in c:\users\s323\anaconda3
\lib\site-packages (from spacy<3.6.0,>=3.5.0->en-core-web-sm==3.5.0) (2.0.8)
Requirement already satisfied: pydantic!=1.8,!=1.8.1,<1.11.0,>=1.7.4 in c:\users\s
323\anaconda3\lib\site-packages (from spacy<3.6.0,>=3.5.0->en-core-web-sm==3.5.0)
(1.10.4)
Requirement already satisfied: numpy>=1.15.0 in c:\users\s323\anaconda3\lib\site-p
ackages (from spacy<3.6.0,>=3.5.0->en-core-web-sm==3.5.0) (1.21.5)
Requirement already satisfied: murmurhash<1.1.0,>=0.28.0 in c:\users\s323\anaconda
3\lib\site-packages (from spacy<3.6.0,>=3.5.0->en-core-web-sm==3.5.0) (1.0.9)
Requirement already satisfied: jinja2 in c:\users\s323\anaconda3\lib\site-packages
(from spacy<3.6.0,>=3.5.0->en-core-web-sm==3.5.0) (2.11.3)
Requirement already satisfied: thinc<8.2.0,>=8.1.0 in c:\users\s323\anaconda3\lib
\site-packages (from spacy<3.6.0,>=3.5.0->en-core-web-sm==3.5.0) (8.1.7)
Requirement already satisfied: requests<3.0.0,>=2.13.0 in c:\users\s323\anaconda3
\lib\site-packages (from spacy<3.6.0,>=3.5.0->en-core-web-sm==3.5.0) (2.27.1)
Requirement already satisfied: wasabi<1.2.0,>=0.9.1 in c:\users\s323\anaconda3\lib
\site-packages (from spacy<3.6.0,>=3.5.0->en-core-web-sm==3.5.0) (1.1.1)
Requirement already satisfied: langcodes<4.0.0,>=3.2.0 in c:\users\s323\anaconda3
\lib\site-packages (from spacy<3.6.0,>=3.5.0->en-core-web-sm==3.5.0) (3.3.0)
Requirement already satisfied: spacy-legacy<3.1.0,>=3.0.11 in c:\users\s323\anacon
da3\lib\site-packages (from spacy<3.6.0,>=3.5.0->en-core-web-sm==3.5.0) (3.0.12)
Requirement already satisfied: spacy-loggers<2.0.0,>=1.0.0 in c:\users\s323\anacon
da3\lib\site-packages (from spacy<3.6.0,>=3.5.0->en-core-web-sm==3.5.0) (1.0.4)
Requirement already satisfied: preshed<3.1.0,>=3.0.2 in c:\users\s323\anaconda3\li
b\site-packages (from spacy<3.6.0,>=3.5.0->en-core-web-sm==3.5.0) (3.0.8)
Requirement already satisfied: typer<0.8.0,>=0.3.0 in c:\users\s323\anaconda3\lib
\site-packages (from spacy<3.6.0,>=3.5.0->en-core-web-sm==3.5.0) (0.7.0)
Requirement already satisfied: srsly<3.0.0,>=2.4.3 in c:\users\s323\anaconda3\lib
\site-packages (from spacy<3.6.0,>=3.5.0->en-core-web-sm==3.5.0) (2.4.5)
Requirement already satisfied: cymem<2.1.0,>=2.0.2 in c:\users\s323\anaconda3\lib
\site-packages (from spacy<3.6.0,>=3.5.0->en-core-web-sm==3.5.0) (2.0.7)
Requirement already satisfied: packaging>=20.0 in c:\users\s323\anaconda3\lib\site
-packages (from spacy<3.6.0,>=3.5.0->en-core-web-sm==3.5.0) (21.3)
Requirement already satisfied: setuptools in c:\users\s323\anaconda3\lib\site-pack
ages (from spacy<3.6.0,>=3.5.0->en-core-web-sm==3.5.0) (61.2.0)
Requirement already satisfied: tqdm<5.0.0,>=4.38.0 in c:\users\s323\anaconda3\lib
\site-packages (from spacy<3.6.0,>=3.5.0->en-core-web-sm==3.5.0) (4.64.0)
Requirement already satisfied: pyparsing!=3.0.5,>=2.0.2 in c:\users\s323\anaconda3
\lib\site-packages (from packaging>=20.0->spacy<3.6.0,>=3.5.0->en-core-web-sm==3.
5.0) (3.0.4)
Requirement already satisfied: typing-extensions>=4.2.0 in c:\users\s323\anaconda3
\lib\site-packages (from pydantic!=1.8,!=1.8.1,<1.11.0,>=1.7.4->spacy<3.6.0,>=3.5.
0->en-core-web-sm==3.5.0) (4.4.0)
Requirement already satisfied: urllib3<1.27,>=1.21.1 in c:\users\s323\anaconda3\li
b\site-packages (from requests<3.0.0,>=2.13.0->spacy<3.6.0,>=3.5.0->en-core-web-sm
==3.5.0) (1.26.9)
Requirement already satisfied: charset-normalizer~=2.0.0 in c:\users\s323\anaconda
3\lib\site-packages (from requests<3.0.0,>=2.13.0->spacy<3.6.0,>=3.5.0->en-core-we
b-sm==3.5.0) (2.0.4)
Requirement already satisfied: certifi>=2017.4.17 in c:\users\s323\anaconda3\lib\s
ite-packages (from requests<3.0.0,>=2.13.0->spacy<3.6.0,>=3.5.0->en-core-web-sm==
3.5.0) (2021.10.8)
Requirement already satisfied: idna<4,>=2.5 in c:\users\s323\anaconda3\lib\site-pa
ckages (from requests<3.0.0,>=2.13.0->spacy<3.6.0,>=3.5.0->en-core-web-sm==3.5.0)
(3.3)
```

```
Requirement already satisfied: confection<1.0.0,>=0.0.1 in c:\users\s323\anaconda3
\lib\site-packages (from thinc<8.2.0,>=8.1.0->spacy<3.6.0,>=3.5.0->en-core-web-sm=
=3.5.0) (0.0.4)
Requirement already satisfied: blis<0.8.0,>=0.7.8 in c:\users\s323\anaconda3\lib\s
ite-packages (from thinc<8.2.0,>=8.1.0->spacy<3.6.0,>=3.5.0->en-core-web-sm==3.5.
0) (0.7.9)
Requirement already satisfied: colorama in c:\users\s323\anaconda3\lib\site-packag
es (from tqdm<5.0.0,>=4.38.0->spacy<3.6.0,>=3.5.0->en-core-web-sm==3.5.0) (0.4.6)
Requirement already satisfied: click<9.0.0,>=7.1.1 in c:\users\s323\anaconda3\lib
\site-packages (from typer<0.8.0,>=0.3.0->spacy<3.6.0,>=3.5.0->en-core-web-sm==3.
5.0) (8.0.4)
Requirement already satisfied: MarkupSafe>=0.23 in c:\users\s323\anaconda3\lib\sit
e-packages (from jinja2->spacy<3.6.0,>=3.5.0->en-core-web-sm==3.5.0) (2.0.1)
[+] Download and installation successful
You can now load the package via spacy.load('en_core_web_sm')
```

In [88]:
```
import spacy
# Create blank language object and tokenize words in a sentence
nlp = spacy.blank("en")
# nlp = spacy.load("en_core_web_sm")
```

Create blank language object and tokenize words in a sentence

In [89]:
```
doc = nlp("Tony gave 2$ to peter.")
for token in doc:
    print(token)
```

```
Tony
gave
2
$
to
peter
.
```

Creating blank language object gives a tokenizer and an empty pipeline. We will look more into language pipelines in next tutorial

Using index to grab tokens

Token attributes

In [84]:
```
token0 = doc[0]
```

In [85]:
```
token0
```

Out[85]:
```
Tony
```

In [86]:
```
dir (token0)
# are the objects
```

Out[86]:
```
['_',
 '__bytes__',
 '__class__',
 '__delattr__',
 '__dir__',
 '__doc__',
 '__eq__',
 '__format__',
 '__ge__',
 '__getattribute__',
 '__gt__',
 '__hash__',
 '__init__',
 '__init_subclass__',
 '__le__',
 '__len__',
 '__lt__',
 '__ne__',
 '__new__',
 '__pyx_vtable__',
 '__reduce__',
 '__reduce_ex__',
 '__repr__',
 '__setattr__',
 '__sizeof__',
 '__str__',
 '__subclasshook__',
 '__unicode__',
 'ancestors',
 'check_flag',
 'children',
 'cluster',
 'conjuncts',
 'dep',
 'dep_',
 'doc',
 'ent_id',
 'ent_id_',
 'ent_iob',
 'ent_iob_',
 'ent_kb_id',
 'ent_kb_id_',
 'ent_type',
 'ent_type_',
 'get_extension',
 'has_dep',
 'has_extension',
 'has_head',
 'has_morph',
 'has_vector',
 'head',
 'i',
 'idx',
 'iob_strings',
 'is_alpha',
 'is_ancestor',
 'is_ascii',
 'is_bracket',
 'is_currency',
 'is_digit',
 'is_left_punct',
 'is_lower',
 'is_oov',
 'is_punct',
```

```
 'is_quote',
 'is_right_punct',
 'is_sent_end',
 'is_sent_start',
 'is_space',
 'is_stop',
 'is_title',
 'is_upper',
 'lang',
 'lang_',
 'left_edge',
 'lefts',
 'lemma',
 'lemma_',
 'lex',
 'lex_id',
 'like_email',
 'like_num',
 'like_url',
 'lower',
 'lower_',
 'morph',
 'n_lefts',
 'n_rights',
 'nbor',
 'norm',
 'norm_',
 'orth',
 'orth_',
 'pos',
 'pos_',
 'prefix',
 'prefix_',
 'prob',
 'rank',
 'remove_extension',
 'right_edge',
 'rights',
 'sent',
 'sent_start',
 'sentiment',
 'set_extension',
 'set_morph',
 'shape',
 'shape_',
 'similarity',
 'subtree',
 'suffix',
 'suffix_',
 'tag',
 'tag_',
 'tensor',
 'text',
 'text_with_ws',
 'vector',
 'vector_norm',
 'vocab',
 'whitespace_']
```

In [48]:
```
type (token0)
```

Out[48]:
```
spacy.tokens.token.Token
```

In [49]:
```
token0.like_num
```

Out[49]:
```
False
```

In [50]:
```
token2 = doc[2]
token2.text
```

Out[50]:
```
'2'
```

In [51]:
```
token2.like_num
```

Out[51]:
```
True
```

In [52]:
```
token3 = doc[3]
token3.text
```

Out[52]:
```
'$'
```

In [53]:
```
token3.is_currency
```

Out[53]:
```
True
```

Span object

In [92]:
```
span = doc[0:5]
span
```

Out[92]:
```
Tony gave 2$ to
```

In [93]:
```
type(span)
```

Out[93]:
```
spacy.tokens.span.Span
```

In [54]:
```
for token in doc:
    print (token,"==>","index: ",token.i,
           "is_alpha: ", token.is_alpha,
           "is_punct: ", token.is_punct,
           "like_num: ",token.like_num,
           "is_currency: ",token.is_currency)
```

```
Tony ==> index:  0 is_alpha:  True is_punct:  False like_num:  False is_currency:
False
gave ==> index:  1 is_alpha:  True is_punct:  False like_num:  False is_currency:
False
2 ==> index:  2 is_alpha:  False is_punct:  False like_num:  True is_currency:  Fa
lse
$ ==> index:  3 is_alpha:  False is_punct:  False like_num:  False is_currency:  T
rue
to ==> index:  4 is_alpha:  True is_punct:  False like_num:  False is_currency:  F
alse
peter ==> index:  5 is_alpha:  True is_punct:  False like_num:  False is_currency:
False
. ==> index:  6 is_alpha:  False is_punct:  True like_num:  False is_currency:  Fa
lse
```

Collecting email ids of students from students information sheet

In [94]:
```
"""# read file in python
with open("student.txt") as f:
    text = f.readlines()
text

# it will read all the texts of the file in the array"""
```

Out[94]:
```
'"# read file in python\nwith open("student.txt") as f:\n    text = f.readlines()
\ntext\n\n# it will read all the texts of the file in the array'
```

In [56]:
```
"""# spacy uses single texts- convert array into big textx
text = " ".join(text)
text"""
```

Out[56]:
```
'"# spacy uses single texts- convert array into big textx\ntext = " ".join(text)\n
text'
```

In [57]:
```
"""doc = nlp(text)
emails =[]
for token in doc:
    if token.like_email:
        emails.append(token.text)
emails"""

# print out all emails from the doc
```

Out[57]:
```
'"doc = nlp(text)\nemails =[]\nfor token in doc:\n    if token.like_email:\n
emails.append(token.text)\nemails'
```

In [58]:
```
### Model in a diffrent language
#we will look for hindi now

nlp = spacy.blank("hi")
doc = nlp("कैसे हो मेरे भै, मेरे को निन्द आ रहि है")
```

In [59]:
```
for token in doc:
    print(token, token.is_currency, token.like_num)
```

```
कैसे False False
हो False False
मेरे False False
भै False False
, False False
मेरे False False
को False False
निन्द False False
आ False False
रहि False False
है False False
```

Customize your tokenizer

In [60]:
```
from spacy.symbols import ORTH

nlp = spacy.blank("en")
doc = nlp("gimme double cheese large healthy pizza")

# instead for token in token sprint, i will try to grab everything in a array so w

tokens = [token.text for token in doc]
```

```
tokens
# list comprehension is a easy way to write for loop so instead of writing for toke
```

Out[60]:
```
['gimme', 'double', 'cheese', 'large', 'healthy', 'pizza']
```

In [61]:
```
# since gimme consists 2 words- customize with spacy
from spacy.symbols import ORTH
nlp.tokenizer.add_special_case("gimme",[
    {ORTH:"gim"},
    {ORTH:"me"}
])
doc = nlp("gimme double cheese large healthy pizza")
tokens = [token.text for token in doc]
tokens
```

Out[61]:
```
['gim', 'me', 'double', 'cheese', 'large', 'healthy', 'pizza']
```

Sentence Tokenization or Segmentation

In [70]:
```
doc = nlp("Dr. Strange loves pav bhaji of mumbai. Hulk loves chat of delhi")
for sentence in doc.sents:
    print("sentence")
```

```
sentence
sentence
```

In [64]:
```
# above error means nlp.pipeline is block
nlp.add_pipe("sentencizer")
```

Out[64]:
```
<spacy.pipeline.sentencizer.Sentencizer at 0x1c733162480>
```

In [65]:
```
nlp.pipe_names
```

Out[65]:
```
['sentencizer']
```

In [66]:
```
doc = nlp("Dr. Strange loves pav bhaji of mumbai. Hulk loves chat of delhi")
for sentence in doc.sents:
    print("sentence")
```

```
sentence
sentence
```

In [67]:
```
doc = nlp("Dr. Strange loves pav bhaji of mumbai. Hulk loves chat of delhi")
for sentence in doc.sents:
    print(sentence)
```

```
Dr. Strange loves pav bhaji of mumbai.
Hulk loves chat of delhi
```

Exercises:

Collecting dataset websites from a book paragraph

In [79]:
```
text='''
Look for data to help you address the question. Governments are good
sources because data from public research is often freely available. Good
places to start include http://www.data.gov/, and http://www.science.
gov/, and in the United Kingdom, http://data.gov.uk/.
Two of my favorite data sets are the General Social Survey at http://www3.norc.org/
```

```
and the European Social Survey at http://www.europeansocialsurvey.org/.
'''
```

In [80]:
```
doc = nlp(text)
```

In [81]:
```
for token in doc:
    if token.like_url:
        print(token.text)
# above one is via Array and below one is conversion into list
#data_websites = [token.text for token in doc if token.like_url ]
#data_websites
```

```
http://www.data.gov/
http://www.science
http://data.gov.uk/.
http://www3.norc.org/gss+website/
http://www.europeansocialsurvey.org/.
```

Figure out all transactions from this text with amount and currency

In [77]:
```
transactions = "Tony gave two $ to Peter, Bruce gave 500 € to Steve"
doc = nlp(transactions)
for token in doc:
    if token.like_num and doc[token.i+1].is_currency:
        print(token.text, doc[token.i+1].text)
```

```
two $
500 €
```