In [1]:

```
import spacy
from spacy.lang.en.stop_words import STOP_WORDS
```

In [2]:

```
len(STOP_WORDS)
```

Out[2]:

```
326
```

How to print out the stop words and how to take out the remaning word

In [3]:

```
nlp = spacy.load("en_core_web_sm")
doc = nlp ("We just opened our wings, the flying part is coming soon")

for token in doc:
    if token.is_stop:
        print(token)
```

```
We
just
our
the
part
is
```

Common practise to keep all other non stop words out

In [4]:

```
def preprocess(text):
    doc = nlp (text)
    no_stop_words = [token.text for token in doc if not token.is_stop and not toke
    return " ".join (no_stop_words)
# we used join function above to convert list into string in python
    #for token in doc:
        #if not token.is_stop:
            # now we have to store it in somekind of array - perfect will be list,
            # not token.is_punct - for punctuation marks
```

- Now we can use function preprocess to work on any doc

In [5]:

```
preprocess("We just opened our wings, the flying part is coming so")
```

Out[5]:

```
'opened wings flying coming'
```

In [6]:

```
preprocess("Musk wants time to prepare for a trial over his")
```

Out[6]:

```
'Musk wants time prepare trial'
```

In [7]:

```
preprocess("The other is not other but your divine brother")
```

Out[7]:

```
'divine brother'
```

Remove stop words from pandas dataframe text column

In [8]:

```
import pandas as pd
court_data = pd.read_json(r"C:\Users\s323\Desktop\Gatherings\Data Science\NLP\comb
# line = True means there is one line per JSON object
```

In [9]:
```
court_data.head(5)
```

Out[9]:

	id	title	contents	date	topics	components
0	None	Convicted Bomb Plotter Sentenced to 30 Years	PORTLAND, Oregon. – Mohamed Osman Mohamud, 23,...	2014-10-01T00:00:00-04:00	[]	[National Security Division (NSD)]
1	12-919	$1 Million in Restitution Payments Announced t...	WASHINGTON – North Carolina’s Waccamaw River...	2012-07-25T00:00:00-04:00	[]	[Environment and Natural Resources Division]
2	11-1002	$1 Million Settlement Reached for Natural Reso...	BOSTON– A $1-million settlement has been...	2011-08-03T00:00:00-04:00	[]	[Environment and Natural Resources Division]
3	10-015	10 Las Vegas Men Indicted \r\nfor Falsifying V...	WASHINGTON—A federal grand jury in Las Vegas...	2010-01-08T00:00:00-05:00	[]	[Environment and Natural Resources Division]
4	18-898	$100 Million Settlement Will Speed Cleanup Wor...	The U.S. Department of Justice, the U.S. Envir...	2018-07-09T00:00:00-04:00	[Environment]	[Environment and Natural Resources Division]

In [10]:
```
court_data.shape
```

Out[10]:
```
(13087, 6)
```

Filter out those rows that do not have any topics associated with the case

In [11]:
```
type(court_data.topics[0])
```

Out[11]:
```
list
```

In [12]:
```
court_data.info
```

Out[12]:

```
<bound method DataFrame.info of             id  
title  \
0         None       Convicted Bomb Plotter Sentenced to 30 Years
1      12-919   $1 Million in Restitution Payments Announced t...
2      11-1002  $1 Million Settlement Reached for Natural Reso...
3       10-015  10 Las Vegas Men Indicted \r\nfor Falsifying V...
4       18-898  $100 Million Settlement Will Speed Cleanup Wor...
...        ...                                                 ...
13082   16-735  Yuengling to Upgrade Environmental Measures to...
13083   10-473  Zarein Ahmedzay Pleads Guilty to Terror Violat...
13084   17-045  Zimmer Biomet Holdings Inc. Agrees to Pay $17....
13085   17-252  ZTE Corporation Agrees to Plead Guilty and Pay...
13086   17-304  ZTE Corporation Pleads Guilty for  Violating U...

                                                contents  \
0      PORTLAND, Oregon. – Mohamed Osman Mohamud, 23,...
1        WASHINGTON – North Carolina’s Waccamaw River...
2            BOSTON– A $1-million settlement has been...
3        WASHINGTON—A federal grand jury in Las Vegas...
4      The U.S. Department of Justice, the U.S. Envir...
...                                                  ...
13082  The Department of Justice and the U.S. Environ...
13083    The Justice Department announced that Zarein...
13084  Subsidiary Agrees to Plead Guilty to Violating...
13085  ZTE Corporation has agreed to enter a guilty p...
13086  ZTE Corporation pleaded guilty today to conspi...

                            date  \
0      2014-10-01T00:00:00-04:00
1      2012-07-25T00:00:00-04:00
2      2011-08-03T00:00:00-04:00
3      2010-01-08T00:00:00-05:00
4      2018-07-09T00:00:00-04:00
...                          ...
13082  2016-06-23T00:00:00-04:00
13083  2010-04-23T00:00:00-04:00
13084  2017-01-12T00:00:00-05:00
13085  2017-03-07T00:00:00-05:00
13086  2017-03-22T00:00:00-04:00

                                                  topics  \
0                                                     []
1                                                     []
2                                                     []
3                                                     []
4                                          [Environment]
...                                                  ...
13082                                      [Environment]
13083                                                 []
13084                               [Foreign Corruption]
13085  [Asset Forfeiture, Counterintelligence and Exp...
13086           [Counterintelligence and Export Control]

                                              components
0                     [National Security Division (NSD)]
1           [Environment and Natural Resources Division]
2           [Environment and Natural Resources Division]
3           [Environment and Natural Resources Division]
4           [Environment and Natural Resources Division]
...                                                  ...
13082       [Environment and Natural Resources Division]
13083                   [Office of the Attorney General]
13084  [Criminal Division, Criminal - Criminal Fraud ...
13085  [National Security Division (NSD), USAO - Texa...
```

```
13086  [National Security Division (NSD), USAO - Texa...

[13087 rows x 6 columns]>
```

In [13]:
```
court_data.describe()
```

Out[13]:

	id	title	contents	date	topics	components
count	12810	13087	13087	13087	13087	13087
unique	12672	12887	13080	2400	253	810
top	13-526	Northern California Real Estate Investor Agree...	WASHINGTON – ING Bank N.V., a financial inst...	2018-04-13T00:00:00-04:00	[]	[Criminal Division]
freq	3	8	2	20	8399	2680

In [14]:
```
court_data = court_data[court_data["topics"].str.len() != 0]
court_data.head(5)
```

Out[14]:

	id	title	contents	date	topics	components
4	18-898	$100 Million Settlement Will Speed Cleanup Wor...	The U.S. Department of Justice, the U.S. Envir...	2018-07-09T00:00:00-04:00	[Environment]	[Environment and Natural Resources Division]
7	14-1412	14 Indicted in Connection with New England Com...	A 131-count criminal indictment was unsealed t...	2014-12-17T00:00:00-05:00	[Consumer Protection]	[Civil Division]
19	17-1419	2017 Southeast Regional Animal Cruelty Prosecu...	The United States Attorney's Office for the Mi...	2017-12-14T00:00:00-05:00	[Environment]	[Environment and Natural Resources Division, U...
22	15-1562	21st Century Oncology to Pay $19.75 Million to...	21st Century Oncology LLC, has agreed to pay $...	2015-12-18T00:00:00-05:00	[False Claims Act, Health Care Fraud]	[Civil Division]
23	17-1404	21st Century Oncology to Pay $26 Million to Se...	21st Century Oncology Inc. and certain of its ...	2017-12-12T00:00:00-05:00	[Health Care Fraud, False Claims Act]	[Civil Division, USAO - Florida, Middle]

In [15]:
```
court_data.shape
```

Out[15]:
```
(4688, 6)
```

We will show data preprocessing on 1st 100 rows of my pandas data frame below:

In [16]:
```
court_data.head(100)
```

Out[16]:

	id	title	contents	date	topics	components
4	18-898	$100 Million Settlement Will Speed Cleanup Wor...	The U.S. Department of Justice, the U.S. Envir...	2018-07-09T00:00:00-04:00	[Environment]	[Environment and Natural Resources Division]
7	14-1412	14 Indicted in Connection with New England Com...	A 131-count criminal indictment was unsealed t...	2014-12-17T00:00:00-05:00	[Consumer Protection]	[Civil Division]
19	17-1419	2017 Southeast Regional Animal Cruelty Prosecu...	The United States Attorney's Office for the Mi...	2017-12-14T00:00:00-05:00	[Environment]	[Environment and Natural Resources Division, U...
22	15-1562	21st Century Oncology to Pay $19.75 Million to...	21st Century Oncology LLC, has agreed to pay $...	2015-12-18T00:00:00-05:00	[False Claims Act, Health Care Fraud]	[Civil Division]
23	17-1404	21st Century Oncology to Pay $26 Million to Se...	21st Century Oncology Inc. and certain of its ...	2017-12-12T00:00:00-05:00	[Health Care Fraud, False Claims Act]	[Civil Division, USAO - Florida, Middle]
...	...	...	...	...	...	...
316	15-1359	Alaska Plastic Surgeon Convicted of Wire Fraud...	Doctor Hid Millions in Secret Accounts in Pana...	2015-11-04T00:00:00-05:00	[Tax]	[Tax Division]
318	16-396	Alaska Plastic Surgeon Sentenced to Prison for...	Defendant Concealed Bank Accounts in Panama an...	2016-04-04T00:00:00-04:00	[Tax]	[Tax Division]
321	17-736	Alaskan Commercial Fishing Couple Charged with...	An Alaskan couple was charged in federal court...	2017-07-26T00:00:00-04:00	[Tax]	[Tax Division, USAO - Alaska]
322	18-717	Alaskan Husband And Wife Plead Guilty To Willf...	A husband and wife pleaded guilty yesterday to...	2018-06-01T00:00:00-04:00	[Tax]	[Tax Division]
324	16-1345	Alaskan Oncologist Indicted for Tax Evasion	A resident of Big Lake, Alaska was indicted on...	2016-11-17T00:00:00-05:00	[Tax]	[Tax Division]

100 rows × 6 columns

In [17]:

```
court_data["contents"]
```

Out[17]:

```
4        The U.S. Department of Justice, the U.S. Envir...
7        A 131-count criminal indictment was unsealed t...
19       The United States Attorney’s Office for the Mi...
22       21st Century Oncology LLC, has agreed to pay $...
23       21st Century Oncology Inc. and certain of its ...
                                ...                        
13081    Anthony Merrell Tyler, 34, of Yuba City, Calif...
13082    The Department of Justice and the U.S. Environ...
13084    Subsidiary Agrees to Plead Guilty to Violating...
13085    ZTE Corporation has agreed to enter a guilty p...
13086    ZTE Corporation pleaded guilty today to conspi...
Name: contents, Length: 4688, dtype: object
```

In [18]:

```
court_data["contents"].iloc[4]
```

```
Out[18]: '21st Century Oncology Inc. and certain of its subsidiaries and affiliates have ag
         reed to pay $26 million to the government to resolve a self-disclosure relating to
         the submission of false attestations regarding the company’s use of electronic hea
         lth records software and separate allegations that they violated the False Claims
         Act by submitting, or causing the submission of, claims for certain services provi
         ded pursuant to referrals from physicians with whom they had improper financial re
         lationships. \xa0 “The Justice Department is committed to zealously investigating
         improper financial relationships that have the potential to compromise physicians’
         medical judgment,” said Acting Assistant Attorney General Chad A. Readler of the J
         ustice Department’s Civil Division.\xa0 “However, we will work with companies that
         accept responsibility for their past compliance failures and promptly take correct
         ive action.”  \xa0 21st Century Oncology, which is headquartered in Fort Myers, Fl
         orida, owns and operates subsidiaries and affiliates throughout the United States
         that provide integrated cancer care.\xa0 As part of its business, 21st Century Onc
         ology’s subsidiaries and affiliates employ physicians in specialty fields such as
         radiation oncology, medical oncology, and urology.\xa0  \xa0 The settlement announ
         ced today resolves conduct that was self-disclosed by the company regarding paymen
         ts made by the government as part of the Medicare Electronic Health Records (EHR)
         Incentive Program.\xa0 Under the Medicare EHR Incentive Program, physicians who at
         test to their meaningful use of certified EHR technology may receive incentive pay
         ments and avoid downward adjustments to certain Medicare claims.\xa0 As part of it
         s self-disclosure, 21st Century Oncology reported that it knowingly submitted, or
         caused the submission of, false attestations to CMS concerning employed physician
         s’ use of EHR software.\xa0 The company further reported that, in support of the a
         ttestations, its employees falsified data regarding the company’s use of EHR softw
         are, fabricated software utilization reports, and superimposed EHR vendor logos on
         to the reports to make them look legitimate.\xa0  \xa0 “This settlement represents
         our office’s continued commitment to ensuring compliance with important federal he
         alth care laws,” said Acting U.S. Attorney Stephen Muldrow of the Middle District
         of Florida.\xa0 “We appreciate that 21st Century Oncology self-reported a major fr
         aud affecting Medicare, and we are also pleased that the company has agreed to acc
         ept financial responsibility for past compliance failures.” \xa0 The settlement al
         so resolves the government’s allegations regarding violations of the physician sel
         f-referral law (commonly referred to as the “Stark Law.”)\xa0 The Stark Law prohib
         its an entity from submitting claims to Medicare for designated health services pe
         rformed pursuant to referrals from physicians with whom the entity has a financial
         relationship unless certain designated exceptions apply.\xa0 The government allege
         d that 21st Century Oncology and certain of its subsidiaries and affiliates violat
         ed the FCA by submitting, or causing the submission of, claims for services perfor
         med pursuant to referrals from physicians whose compensation did not satisfy any e
         xception to the Stark Law. \xa0 The Stark Law allegations were originally brought
         in a lawsuit filed by Matthew Moore, 21st Century Oncology’s former Interim Vice P
         resident of Financial Planning, under the qui tam provisions of the False Claims A
         ct.\xa0 Under the Act, private parties may bring suit on behalf of the government
         and share in any recovery.\xa0 Mr. Moore will receive $2,000,000 as his share of t
         he recovery associated with the Stark Law allegations. \xa0 In addition to the civ
         il settlement, 21st Century Oncology has entered into a new five-year Corporate In
         tegrity Agreement with the Office of Inspector General of the United States Depart
         ment of Health and Human Services (HHS-OIG), which obligates 21st Century Oncology
         to undertake substantial internal compliance reforms, including hiring independent
         review organizations to conduct annual claims and arrangements reviews.\xa0 \xa0
         “21st Century Oncology admitted to causing violation of the meaningful use regulat
         ions in order to fund an electronic health records system, as well as falsifying r
         ecords to cover up those actions,” said Shimon R. Richmond, Special Agent in Charg
         e for the Office of Inspector General of the U.S. Department of Health and Human S
         ervices.\xa0 “Separately, the government alleged that same company, through its af
         filiates and subsidiaries, caused certain physicians to enter into illegal financi
         al arrangements.\xa0 Providers engaging in similar behavior should expect attentio
         n from OIG.”\xa0\xa0\xa0 \xa0 The government’s resolution of this matter illustrat
         es the government’s emphasis on combating health care fraud.\xa0 Tips and complain
         ts from all sources about potential fraud, waste, abuse, and mismanagement can be
         reported to the Department of Health and Human Services at 900-HHS-TIPS (800-447-8
         477). \xa0 The investigation was handled by the Civil Division’s Commercial Litiga
         tion Branch and the Fort Myers Division of\xa0the U.S. Attorney’s Office for the M
```

```
iddle District of Florida, with assistance from the U.S. Attorney’s Office for the
Southern District of New York and HHS-OIG.\xa0 The claims resolved by this settlem
ent are allegations only; there has been no determination of liability.\xa0 The ca
se is captioned United States ex rel. Moore v. 21st Century Oncology, LLC, No. 2:1
6-cv-99 (M.D. Fl.).'
```

In [19]:

```python
len(court_data["contents"].iloc[4])
```

Out[19]:

```
5504
```

- Our goal is to remove all stop words from 5504 and make effictive and small nlp application

In [24]:

```python
court_data["contents_new"] = court_data.contents.apply(preprocess)
```

```
C:\Users\s323\AppData\Local\Temp\ipykernel_211572\2370100651.py:1: SettingWithCopy
Warning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stabl
e/user_guide/indexing.html#returning-a-view-versus-a-copy
  court_data["contents_new"] = court_data.contents.apply(preprocess)
```

In [20]:

```python
# we will create a new column of content
# court_data["contents_new"]= court_data.contents.apply(preprocess)
```

In [21]:

```python
# court_data.head(5)
```

In [25]:

```python
 len(court_data.contents[4])
```

Out[25]:

```
6286
```

In [27]:

```python
len(court_data.contents_new[4])
```

Out[27]:

```
4574
```

In [29]:

```python
court_data.contents[4][:300]
```

Out[29]:

```
'The U.S. Department of Justice, the U.S. Environmental Protection Agency (EPA), a
nd the Rhode Island Department of Environmental Management (RIDEM) announced today
that two subsidiaries of Stanley Black & Decker Inc.—Emhart Industries Inc. and Bl
ack & Decker Inc.—have agreed to clean up dioxin conta'
```

Examples where removing stop words can create a problem

(1) Sentiment detection: Not always but in some cases, based on your dataset it can change the sentiment of a sentence if you remove stop words

In [30]:

```python
preprocess("this is a good movie")
```

Out[30]:

```
'good movie'
```

In [31]:

```python
preprocess("this is not a good movie")
```

Out[31]:

```
'good movie'
```

(2) Language translation: Say you want to translate following sentence from english to telugu. Before actual translation if you remove stop words and then translate, it will produce horrible result

```
In [32]: preprocess("how are you doing dhaval?")
```

```
Out[32]: 'dhaval'
```

(3) Chat bot or any Q&A system

```
In [33]: preprocess("I don't find yoga mat on your website. Can you help?")
```

```
Out[33]: 'find yoga mat website help'
```