UE19CS332

Algorithms For Web And Information Retrieval

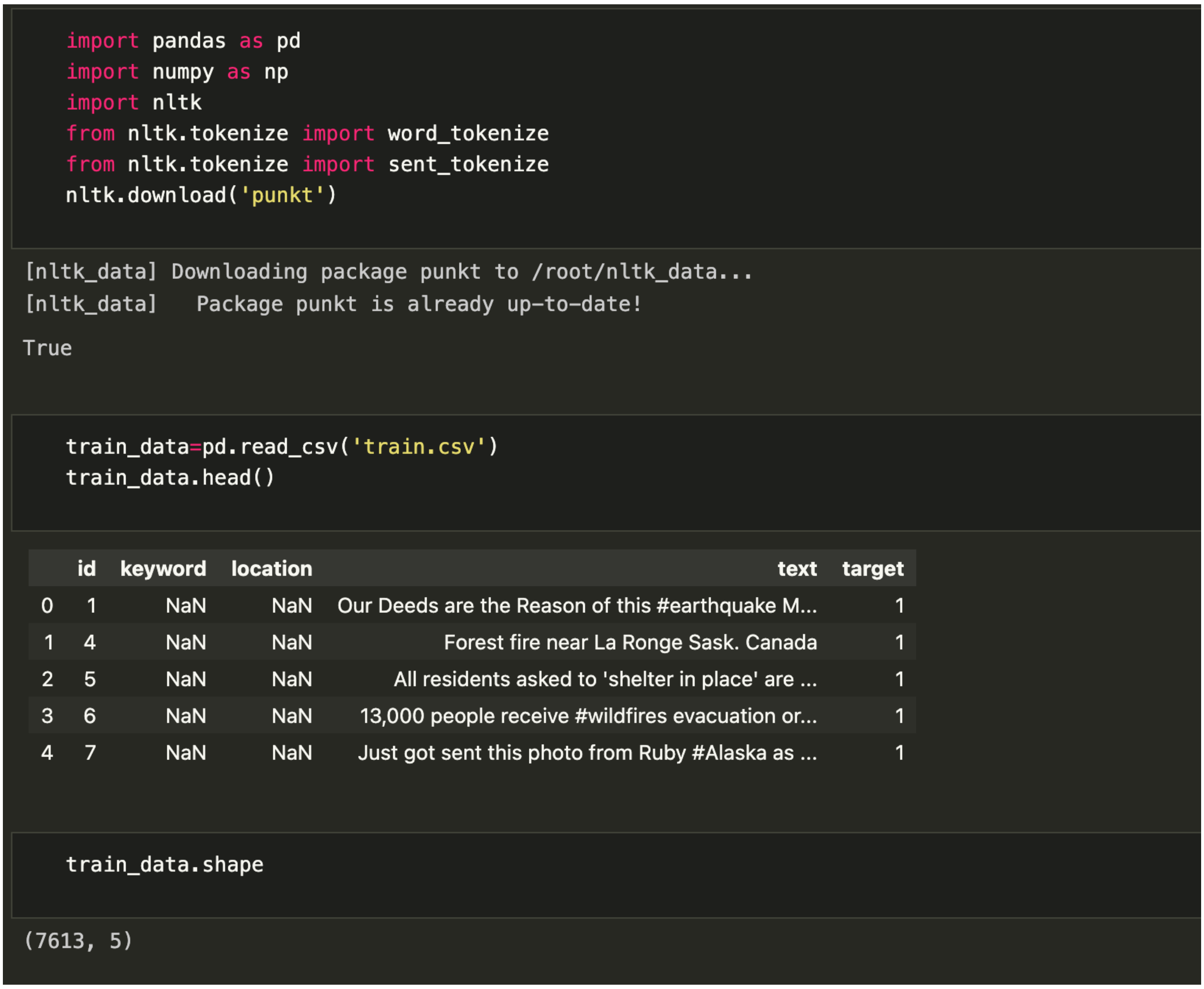
Assignment – 2

NAME : PRIYA MOHATA
SRN : PES2UG19CS301
SECTION: E

Problem Statement :

For a given dataset
A) Tokenize into sentences
B) Tokenize each tweet into words
C) Remove stop words in each tweet

TRAIN.CSV
Importing all libraries + displaying the train dataset



```
import pandas as pd
import numpy as np
import nltk
from nltk.tokenize import word_tokenize
from nltk.tokenize import sent_tokenize
nltk.download('punkt')
```

```
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Package punkt is already up-to-date!

True
```

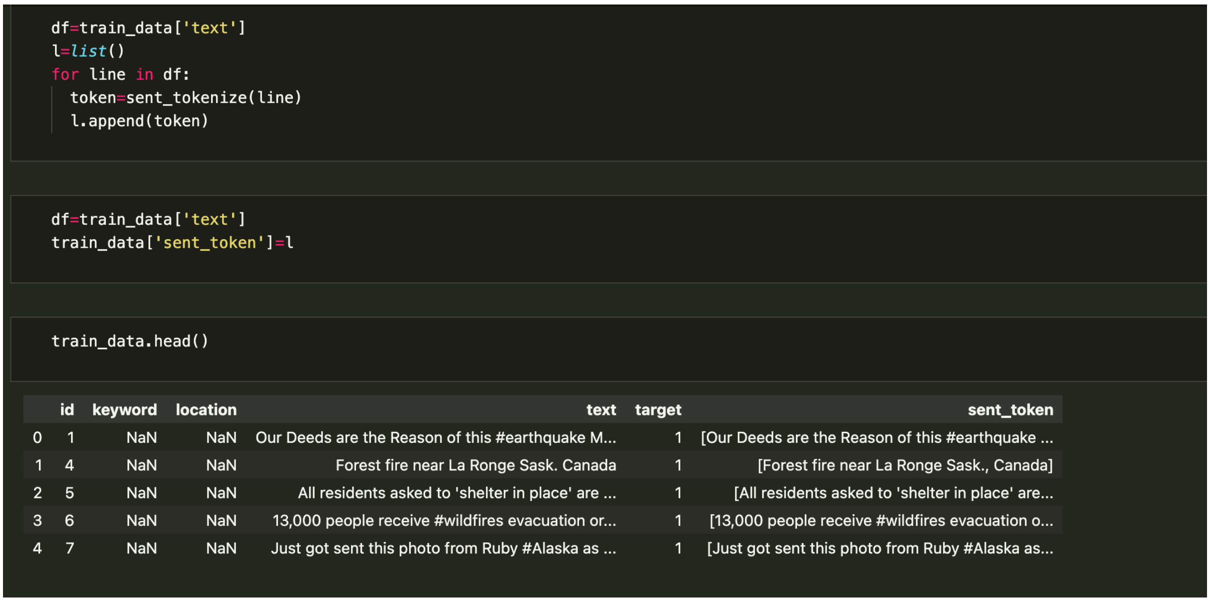
```
train_data=pd.read_csv('train.csv')
train_data.head()
```

	id	keyword	location	text	target
0	1	NaN	NaN	Our Deeds are the Reason of this #earthquake M...	1
1	4	NaN	NaN	Forest fire near La Ronge Sask. Canada	1
2	5	NaN	NaN	All residents asked to 'shelter in place' are ...	1
3	6	NaN	NaN	13,000 people receive #wildfires evacuation or...	1
4	7	NaN	NaN	Just got sent this photo from Ruby #Alaska as ...	1

```
train_data.shape
```

```
(7613, 5)
```

Performing SENT TOKENIZATION :



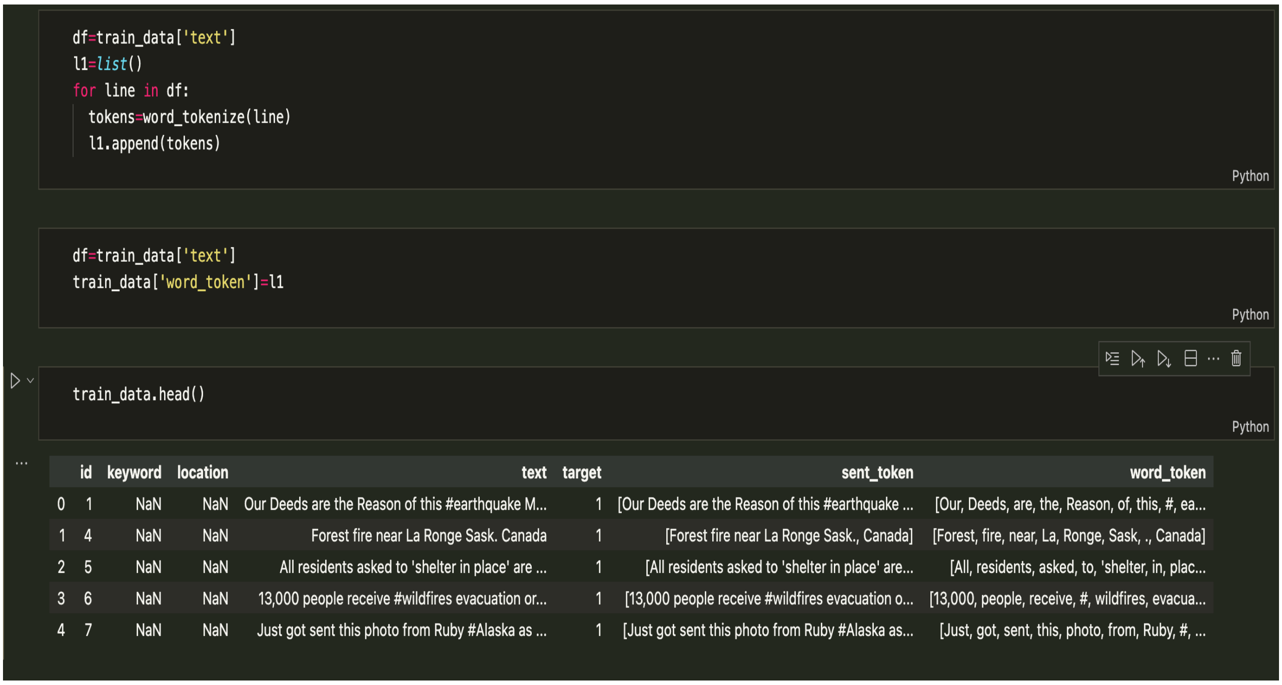
```
df=train_data['text']
l=list()
for line in df:
    token=sent_tokenize(line)
    l.append(token)
```

```
df=train_data['text']
train_data['sent_token']=l
```

```
train_data.head()
```

	id	keyword	location	text	target	sent_token
0	1	NaN	NaN	Our Deeds are the Reason of this #earthquake M...	1	[Our Deeds are the Reason of this #earthquake ...
1	4	NaN	NaN	Forest fire near La Ronge Sask. Canada	1	[Forest fire near La Ronge Sask., Canada]
2	5	NaN	NaN	All residents asked to 'shelter in place' are ...	1	[All residents asked to 'shelter in place' are...
3	6	NaN	NaN	13,000 people receive #wildfires evacuation or...	1	[13,000 people receive #wildfires evacuation o...
4	7	NaN	NaN	Just got sent this photo from Ruby #Alaska as ...	1	[Just got sent this photo from Ruby #Alaska as...

Performing WORD TOKENIZATION :



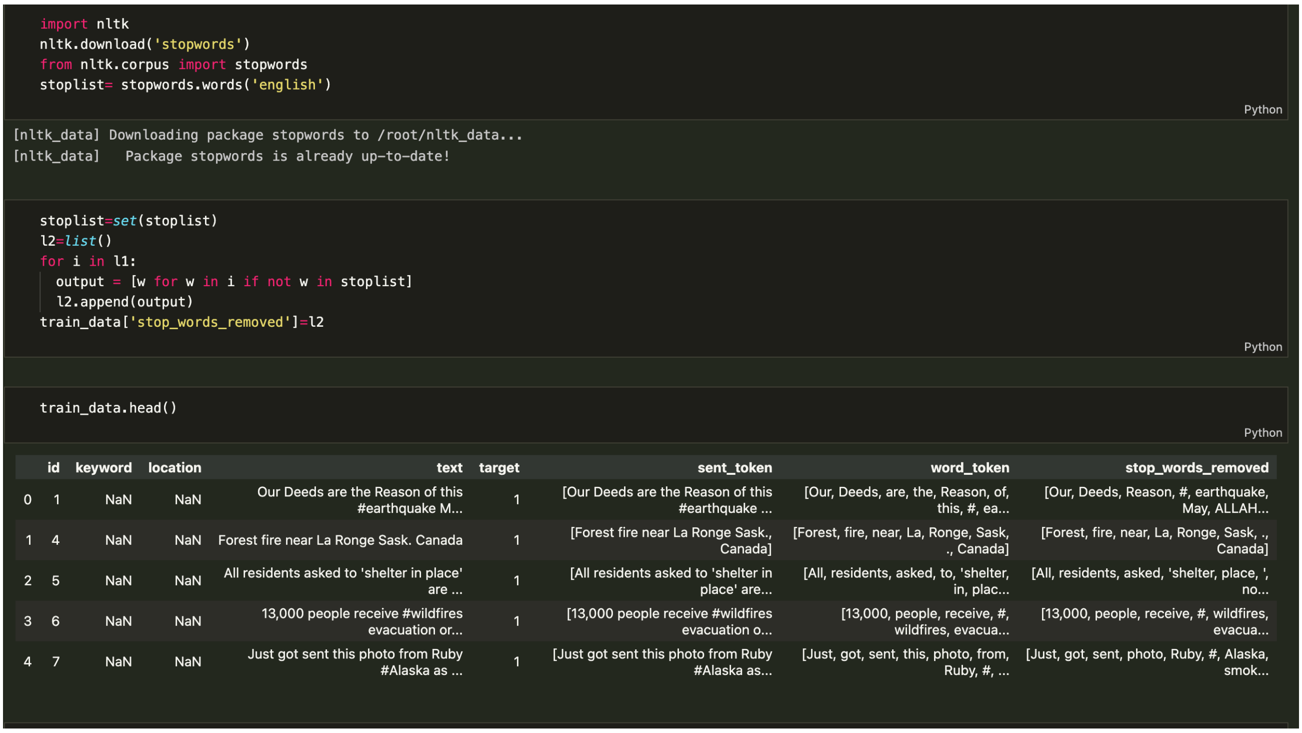
```
df=train_data['text']
l1=list()
for line in df:
    tokens=word_tokenize(line)
    l1.append(tokens)
```

```
df=train_data['text']
train_data['word_token']=l1
```

```
train_data.head()
```

	id	keyword	location	text	target	sent_token	word_token
0	1	NaN	NaN	Our Deeds are the Reason of this #earthquake M...	1	[Our Deeds are the Reason of this #earthquake ...	[Our, Deeds, are, the, Reason, of, this, #, ea...
1	4	NaN	NaN	Forest fire near La Ronge Sask. Canada	1	[Forest fire near La Ronge Sask., Canada]	[Forest, fire, near, La, Ronge, Sask, ., Canada]
2	5	NaN	NaN	All residents asked to 'shelter in place' are ...	1	[All residents asked to 'shelter in place' are...	[All, residents, asked, to, 'shelter, in, plac...
3	6	NaN	NaN	13,000 people receive #wildfires evacuation or...	1	[13,000 people receive #wildfires evacuation o...	[13,000, people, receive, #, wildfires, evacua...
4	7	NaN	NaN	Just got sent this photo from Ruby #Alaska as ...	1	[Just got sent this photo from Ruby #Alaska as...	[Just, got, sent, this, photo, from, Ruby, #, ...

Removing STOP WORDS :



```
import nltk
nltk.download('stopwords')
from nltk.corpus import stopwords
stoplist= stopwords.words('english')
```

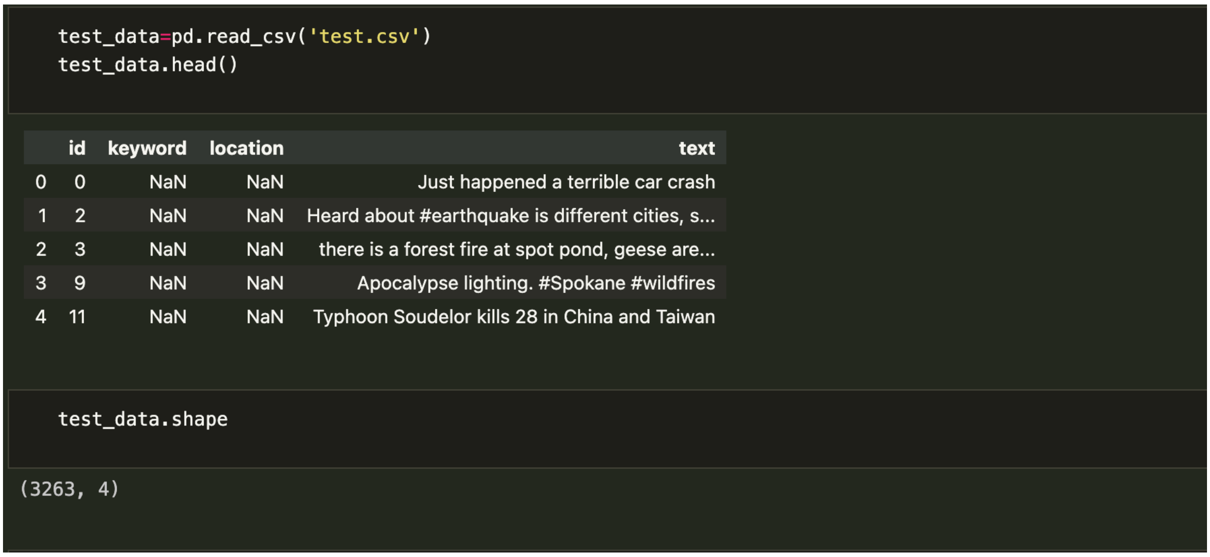
```
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

```
stoplist=set(stoplist)
l2=list()
for i in l1:
    output = [w for w in i if not w in stoplist]
    l2.append(output)
train_data['stop_words_removed']=l2
```

```
train_data.head()
```

	id	keyword	location	text	target	sent_token	word_token	stop_words_removed
0	1	NaN	NaN	Our Deeds are the Reason of this #earthquake M...	1	[Our Deeds are the Reason of this #earthquake ...	[Our, Deeds, are, the, Reason, of, this, #, ea...	[Our, Deeds, Reason, #, earthquake, May, ALLAH...
1	4	NaN	NaN	Forest fire near La Ronge Sask. Canada	1	[Forest fire near La Ronge Sask., Canada]	[Forest, fire, near, La, Ronge, Sask, ., Canada]	[Forest, fire, near, La, Ronge, Sask, ., Canada]
2	5	NaN	NaN	All residents asked to 'shelter in place' are ...	1	[All residents asked to 'shelter in place' are...	[All, residents, asked, to, 'shelter, in, plac...	[All, residents, asked, 'shelter, place, ', no...
3	6	NaN	NaN	13,000 people receive #wildfires evacuation or...	1	[13,000 people receive #wildfires evacuation o...	[13,000, people, receive, #, wildfires, evacua...	[13,000, people, receive, #, wildfires, evacua...
4	7	NaN	NaN	Just got sent this photo from Ruby #Alaska as ...	1	[Just got sent this photo from Ruby #Alaska as...	[Just, got, sent, this, photo, from, Ruby, #, ...	[Just, got, sent, photo, Ruby, #, Alaska, smok...

<u>**TEST.CSV**</u>



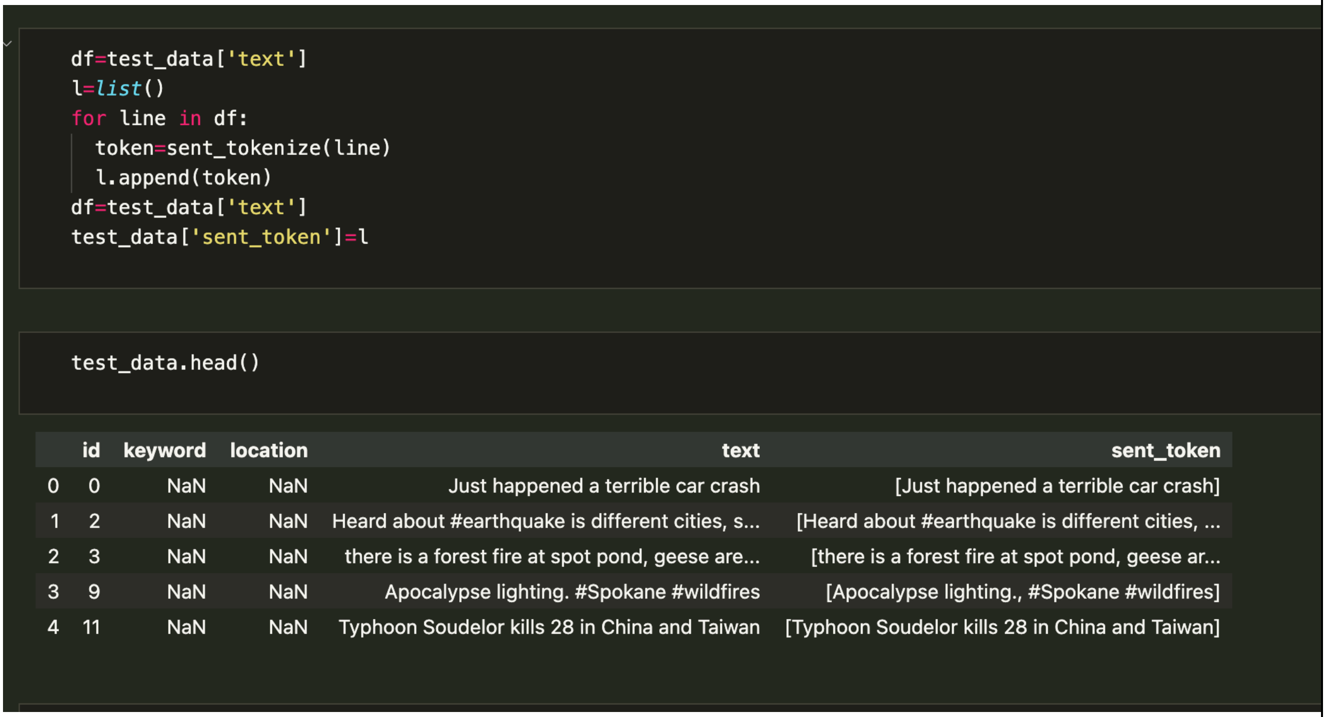
```
test_data=pd.read_csv('test.csv')
test_data.head()
```

	id	keyword	location	text
0	0	NaN	NaN	Just happened a terrible car crash
1	2	NaN	NaN	Heard about #earthquake is different cities, s...
2	3	NaN	NaN	there is a forest fire at spot pond, geese are...
3	9	NaN	NaN	Apocalypse lighting. #Spokane #wildfires
4	11	NaN	NaN	Typhoon Soudelor kills 28 in China and Taiwan

```
test_data.shape
```

```
(3263, 4)
```

Performing SENT TOKENIZATION :

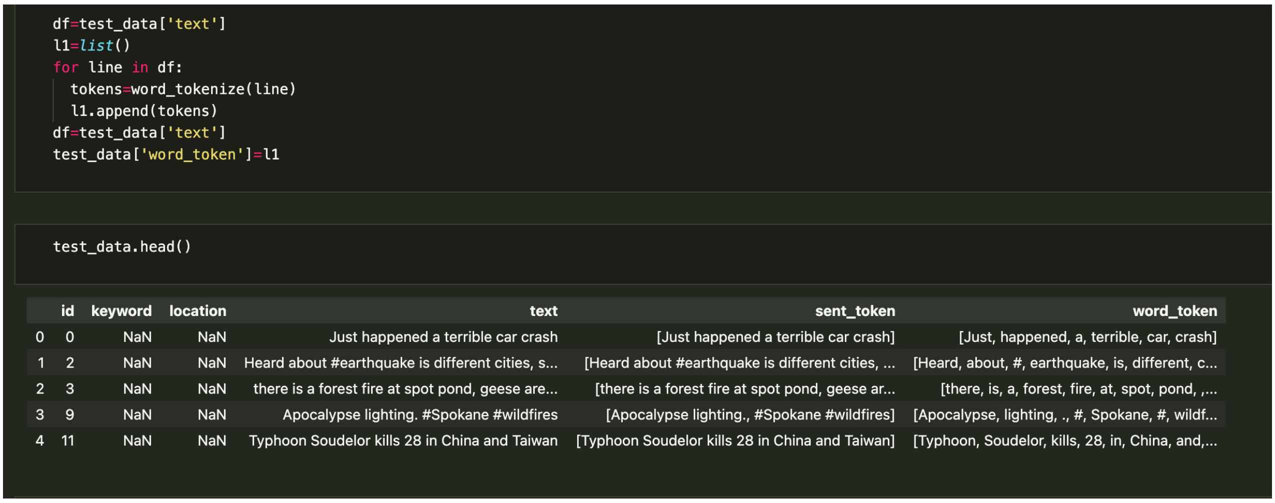


```
df=test_data['text']
l=list()
for line in df:
  token=sent_tokenize(line)
  l.append(token)
df=test_data['text']
test_data['sent_token']=l
```

```
test_data.head()
```

	id	keyword	location	text	sent_token
0	0	NaN	NaN	Just happened a terrible car crash	[Just happened a terrible car crash]
1	2	NaN	NaN	Heard about #earthquake is different cities, s...	[Heard about #earthquake is different cities, ...
2	3	NaN	NaN	there is a forest fire at spot pond, geese are...	[there is a forest fire at spot pond, geese ar...
3	9	NaN	NaN	Apocalypse lighting. #Spokane #wildfires	[Apocalypse lighting., #Spokane #wildfires]
4	11	NaN	NaN	Typhoon Soudelor kills 28 in China and Taiwan	[Typhoon Soudelor kills 28 in China and Taiwan]

Performing WORD TOKENIZATION :

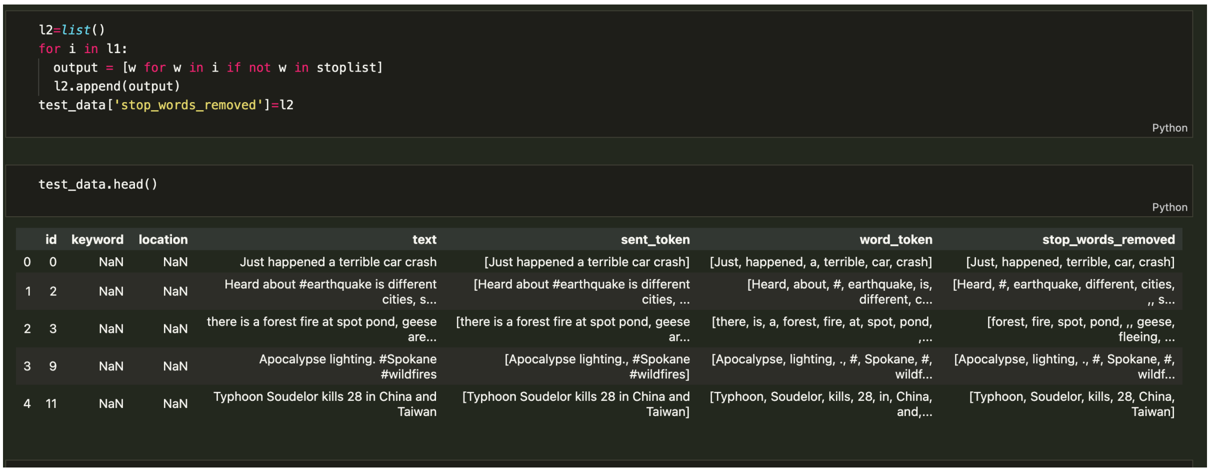


```
df=test_data['text']
l1=list()
for line in df:
  tokens=word_tokenize(line)
  l1.append(tokens)
df=test_data['text']
test_data['word_token']=l1
```

```
test_data.head()
```

	id	keyword	location	text	sent_token	word_token
0	0	NaN	NaN	Just happened a terrible car crash	[Just happened a terrible car crash]	[Just, happened, a, terrible, car, crash]
1	2	NaN	NaN	Heard about #earthquake is different cities, s...	[Heard about #earthquake is different cities, ...	[Heard, about, #, earthquake, is, different, c...
2	3	NaN	NaN	there is a forest fire at spot pond, geese are...	[there is a forest fire at spot pond, geese ar...	[there, is, a, forest, fire, at, spot, pond, ,...
3	9	NaN	NaN	Apocalypse lighting. #Spokane #wildfires	[Apocalypse lighting., #Spokane #wildfires]	[Apocalypse, lighting, ., #, Spokane, #, wildf...
4	11	NaN	NaN	Typhoon Soudelor kills 28 in China and Taiwan	[Typhoon Soudelor kills 28 in China and Taiwan]	[Typhoon, Soudelor, kills, 28, in, China, and,...

Removing STOP WORDS :



```python
l2=list()
for i in l1:
  output = [w for w in i if not w in stoplist]
  l2.append(output)
test_data['stop_words_removed']=l2
```

Python

```python
test_data.head()
```

Python

	id	keyword	location	text	sent_token	word_token	stop_words_removed
0	0	NaN	NaN	Just happened a terrible car crash	[Just happened a terrible car crash]	[Just, happened, a, terrible, car, crash]	[Just, happened, terrible, car, crash]
1	2	NaN	NaN	Heard about #earthquake is different cities, s...	[Heard about #earthquake is different cities, ...	[Heard, about, #, earthquake, is, different, c...	[Heard, #, earthquake, different, cities, ,, s...
2	3	NaN	NaN	there is a forest fire at spot pond, geese are...	[there is a forest fire at spot pond, geese ar...	[there, is, a, forest, fire, at, spot, pond, ,...	[forest, fire, spot, pond, ,, geese, fleeing, ...
3	9	NaN	NaN	Apocalypse lighting. #Spokane #wildfires	[Apocalypse lighting., #Spokane #wildfires]	[Apocalypse, lighting, ., #, Spokane, #, wildf...	[Apocalypse, lighting, ., #, Spokane, #, wildf...
4	11	NaN	NaN	Typhoon Soudelor kills 28 in China and Taiwan	[Typhoon Soudelor kills 28 in China and Taiwan]	[Typhoon, Soudelor, kills, 28, in, China, and,...	[Typhoon, Soudelor, kills, 28, China, Taiwan]