[89]:
```
text = '''
<h1>?sooo yesterday!!!</h1> i taught       the <b>students</b> about <i>nlp preprocessing</i> , yeah yeah like lower  CASE , stop words, and punctuation ... oh wow!! what a  mess mess... there were ab
then some <div> of </div> them said like “sir why we remove punctuation 123 456 789 numbers , html tags <p></p> and spaces spaces   and more  spaces ???” i said because it makes <code>the data clean c
we also found sooo many stopwords like "the" "a" "an" "is" "of" "in" "to" "on" "for" <span>and</span>  "that" "by" oh nooo, and ugh  repeated words everywhere everywhere everywhere !!!
sometimes  the  text  had  BIG  letters, small letters, Missing letters, ...  punctuation,  punctuation,  punctuation!!!  😩😩 lowercase is important, uppercase looks  BAD BAD BAD!!!
and  we  found  multiple  spaces  like  this  one  and  even  worse  ones.  some sentences ended ??? without any reason ... others ended !!!!!!! or ........ or 1234567890 randomly
students were like <a href="#">sir we tired</a> 🤕 but i said “no!! clean the DATA!! remove html tags <img> <body> <head>, remove punctuation!!!!! remove stopwords, remove numbers 000111222333, remov
then one student wrote: REMOVE  STOPWORDS  PLEASE!!!!!! PLEASE  PLEASE!!!! in ALL CAPS 😭😭😭 and forgot half of the text lol lol lol.
we used nltk , regex , and re.sub() and .split() and join() join() join()  functions ... oh so many ( ) ( ) [ ] { } symbols; ; ; ; !!!
btw <html> we </html> also found that lowercase makes the <p>text</p> more readable readable readable readable !!! not LIKE THIS or LIKE THAT .
<article> after </article> finishing <section> class </section> , we all laughed, smiled, cleaned cleaned data, removed numbers(12345 67890), stopwords stopwords stopwords stopwords, and extra  spaces
finally the dataset looked better better better better, before it was a big mess mess mess mess mess full of tags <div> <meta> <h1> </h1> <title> broken html <br> <hr> </hr> punctuation!!!
some text was like:  <b>this.. is.. just.. messy.. text..!!!</b>  others like  what??? why???? no idea... anyway.
cleaning text is so boring boring boring but also fun fun fun!!! it makes model smarter, faster, better, clearer, not slower slower slower 😛😛😛.
and i said again again again, “don’t forget to remove punctuation, stopwords, numbers, emojis, and html tags!!!”
then they all said “yes sir yes sir” three times three times three times.
<footer> and </footer> that’s how preprocessing works works works — lower, clean, trim, remove, filter, normalize normalize normalize normalize!!!'''
```

[90]:
```
import re
```

[91]:
```
#Remove HTML Tags
text = re.sub(r'<.*?>', '', text)
```

[92]:
```
# Change case to Lowercase
text = text.lower()
```

[93]:
```
# Remove numbers
text = re.sub(r'\d+', '', text)
```

[94]:
```
#get list of punctuations
import string
string.punctuation
```

[94]:
```
'!"#$%&\'()*+,-./:;<=>?@[\\]^_`{|}~'
```

[95]:
```
#Remove Punctuations
text = "".join([m for m in text if m not in string.punctuation])
```

[96]:
```
#Remove extra spaces from start and end of words, new line characters, tab characters
text.strip()
```

[96]:
```
'sooo yesterday i taught       the students about nlp preprocessing  yeah yeah like lower  case  stop words and punctuation  oh wow what a  mess mess there were about  or  students  maybe  not sure
😕 anyway  we talked and talked  and talked about commas dots dots more dots  question marks semicolons and other stuff  \nthen some  of  them said like “sir why we remove punctuation  numbers  ht
ml tags  and spaces spaces  and more  spaces ” i said because it makes the data clean clean clean oh yes yes yes  \nwe also found sooo many stopwords like the a an is of in to on for and  that by oh
nooo and ugh  repeated words everywhere everywhere everywhere  \nsometimes  the  text  had  big  letters small letters missing letters  punctuation  punctuation  punctuation  😩😩 lowercase is imp
ortant uppercase looks  bad bad bad  \nand  we  found  multiple  spaces  like  this  one  and  even  worse  ones  some sentences ended  without any reason  others ended  or  or
randomly inserted in middle of text 😐  \nstudents were like sir we tired 🤕 but i said “no clean the data remove html tags  remove punctuation remove stopwords remove numbers  remove emojis 😅
😂🤳 remove extra spaces ”  \nthen one student wrote remove  stopwords  please please  please in all caps 😭😭😭 and forgot half of the text lol lol lol  \nwe used nltk  regex  and resub and split
and join join join  functions  oh so many  symbols  \nbtw  we  also found that lowercase makes the text more readable readable readable readable  not like this or like that  \n a
fter  finishing  class  we all laughed smiled cleaned cleaned data removed numbers  stopwords stopwords stopwords stopwords and extra  spaces  \nfinally the dataset looked better better better bette
r before it was a big mess mess mess mess mess full of tags  broken html  punctuation  \nsome text was like  this is just messy text  others like  what why no idea anyway  \ncleaning text is so
boring boring boring but also fun fun fun it makes model smarter faster better clearer not slower slower slower 😛😛😛  \nand i said again again again “don’t forget to remove punctuation stopwords n
umbers emojis and html tags”  \nthen they all said “yes sir yes sir” three times three times three times  \n and  that’s how preprocessing works works works — lower clean trim remove filter normalize
normalize normalize normalize'
```

[97]:
```
# Remove extra spaces
text = re.sub(r'\s+', ' ', text).strip()
```

[98]:
```
text
```

[98]:
```
'sooo yesterday i taught the students about nlp preprocessing yeah yeah like lower case stop words and punctuation oh wow what a mess mess there were about or students maybe not sure 😕 anyway we tal
ked and talked and talked about commas dots dots more dots question marks semicolons and other stuff then some of them said like “sir why we remove punctuation numbers html tags and spaces spaces and
more spaces ” i said because it makes the data clean clean clean oh yes yes yes we also found sooo many stopwords like the a an is of in to on for and that by oh nooo and ugh repeated words everywher
e everywhere everywhere sometimes the text had big letters small letters missing letters punctuation punctuation punctuation 😩😩 lowercase is important uppercase looks bad bad bad and we found mult
iple spaces like this one and even worse ones some sentences ended without any reason others ended or or randomly inserted in middle of text 😐 students were like sir we tired 🤕 but i said “no clea
n the data remove html tags remove punctuation remove stopwords remove numbers remove emojis 😅😂🤳 remove extra spaces ” then one student wrote remove stopwords please please please please in all caps 😭
😭😭 and forgot half of the text lol lol lol we used nltk regex and resub and split and join join functions oh so many symbols btw we also found that lowercase makes the text more readable read
able readable readable readable not like this or like that after finishing class we all laughed smiled cleaned cleaned data removed numbers stopwords stopwords stopwords stopwords and extra spaces fi
nally the dataset looked better better better better before it was a big mess mess mess mess mess full of tags broken html punctuation some text was like this is just messy text others like what why
no idea anyway cleaning text is so boring boring boring but also fun fun fun it makes model smarter faster better clearer not slower slower slower 😛😛😛 and i said again again again “don’t forget t
o remove punctuation stopwords numbers emojis and html tags” then they all said “yes sir yes sir” three times three times three times and that’s how preprocessing works works works — lower clean trim
remove filter normalize normalize normalize normalize'
```

[99]:
```
import nltk
from nltk.corpus import stopwords
nltk.download('stopwords')
```

```
[nltk_data] Downloading package stopwords to
[nltk_data]     C:\Users\Mushtaq\AppData\Roaming\nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

[99]:
```
True
```

[100]:
```
stop_words = stopwords.words('english')
```

[101]:
```
text = ' '.join([word for word in text.split() if word not in stop_words])
```

[102]:
```
text = re.sub(r'[^a-zA-Z0-9\s\x22\x27\u201c\u201d]', '', text)
```

[103]:
```
text
```

[103]:
```
'sooo yesterday taught students nlp preprocessing yeah yeah like lower case stop words punctuation oh wow mess mess students maybe sure  anyway talked talked talked commas dots dots dots question mar
ks semicolons stuff said like “sir remove punctuation numbers html tags spaces spaces spaces ” said makes data clean clean clean oh yes yes yes also found sooo many stopwords like oh nooo ugh repeate
d words everywhere everywhere everywhere sometimes text big letters small letters missing letters punctuation punctuation punctuation  lowercase important uppercase looks bad bad bad found multiple s
paces like one even worse ones sentences ended without reason others ended randomly inserted middle text  students like sir tired  said “no clean data remove html tags remove punctuation remove stopw
ords remove numbers remove emojis  remove extra spaces ” one student wrote remove stopwords please please please caps  forgot half text lol lol lol used nltk regex resub split join join join function
s oh many symbols btw also found lowercase makes text readable readable readable readable readable like like finishing class laughed smiled cleaned cleaned data removed numbers stopwords stopwords st
opwords stopwords extra spaces finally dataset looked better better better better big mess mess mess mess mess full tags broken html punctuation text like messy text others like idea anyway cleaning
text boring boring boring also fun fun fun makes model smarter faster better clearer slower slower slower  said “dont forget remove punctuation stopwords numbers emojis html tags” said “yes sir yes s
ir” three times three times three times thats preprocessing works works works  lower clean trim remove filter normalize normalize normalize normalize'
```

[104]:
```
text = text.replace(' ”',’”’)
```

[105]:
```
text
```

[105]:
```
'sooo yesterday taught students nlp preprocessing yeah yeah like lower case stop words punctuation oh wow mess mess students maybe sure  anyway talked talked talked commas dots dots dots question marks semicolons stuff said like “sir remove punctuation numbers html tags spaces spaces spaces” said makes data clean clean clean oh yes yes yes also found sooo many stopwords like oh nooo ugh repeated words everywhere everywhere everywhere sometimes text big letters small letters missing letters punctuation punctuation punctuation  lowercase important uppercase looks bad bad bad found multiple spaces like one even worse ones sentences ended without reason others ended randomly inserted middle text  students like sir tired  said “no clean data remove html tags remove punctuation remove stopwords remove numbers remove emojis  remove extra spaces” one student wrote remove stopwords please please please caps  forgot half text lol lol lol used nltk regex resub split join join join functions oh many symbols btw also found lowercase makes text readable readable readable readable readable like like finishing class laughed smiled cleaned cleaned data removed numbers stopwords stopwords stopwords stopwords extra spaces finally dataset looked better better better better big mess mess mess mess mess full tags broken html punctuation text like messy text others like idea anyway cleaning text boring boring boring also fun fun fun makes model smarter faster better clearer slower slower slower  said “dont forget remove punctuation stopwords numbers emojis html tags” said “yes sir yes sir” three times three times three times thats preprocessing works works works  lower clean trim remove filter normalize normalize normalize normalize'
```

[ ]: