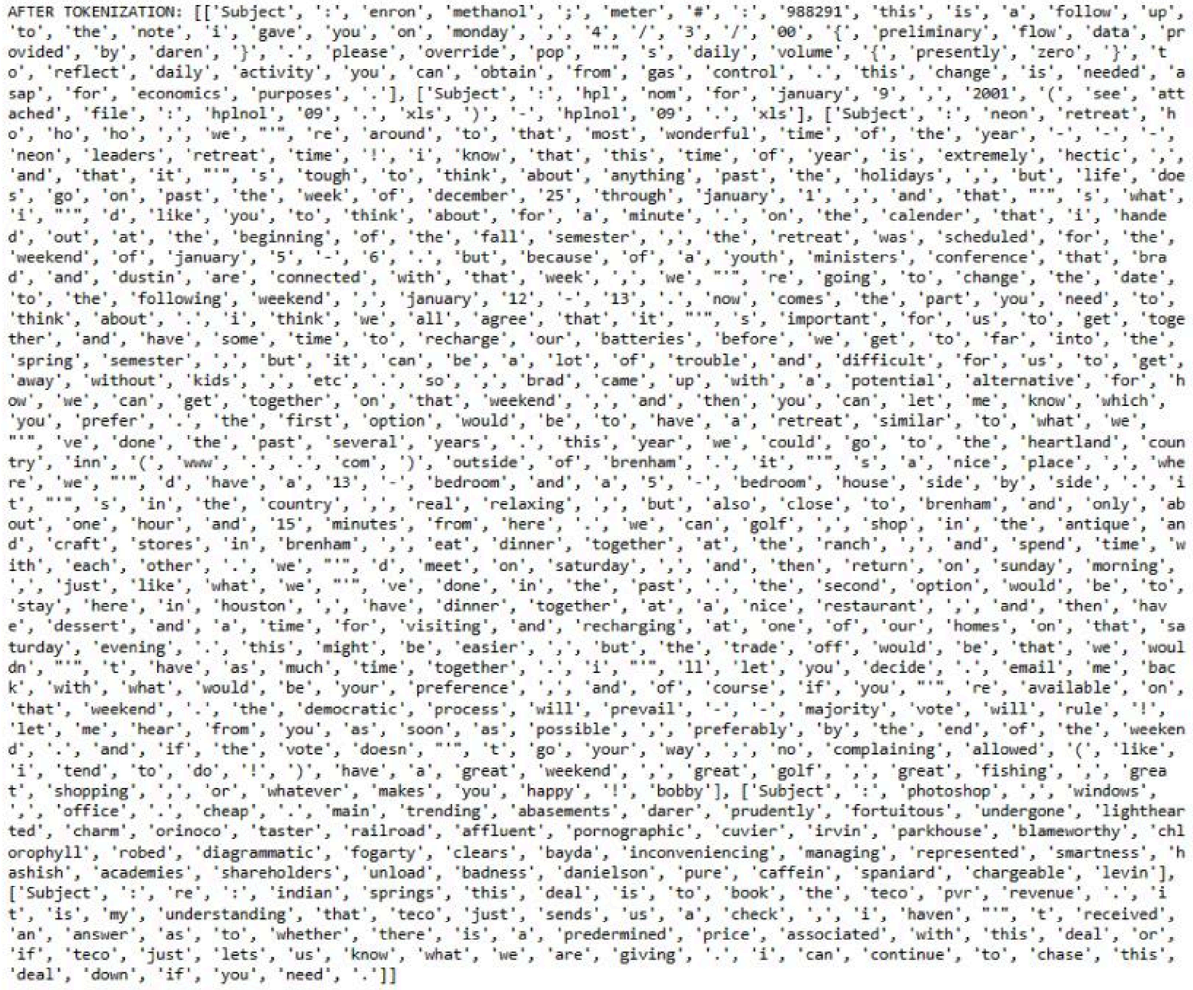
PHASE 3: BUILIDING THE MODEL AND DEVELOPING PREPROCESSING STEPS

TOPIC: BUILDING A SMARTER AI-POWERED SPAM CLASSIFIER

STEP1:TOKENIZATION

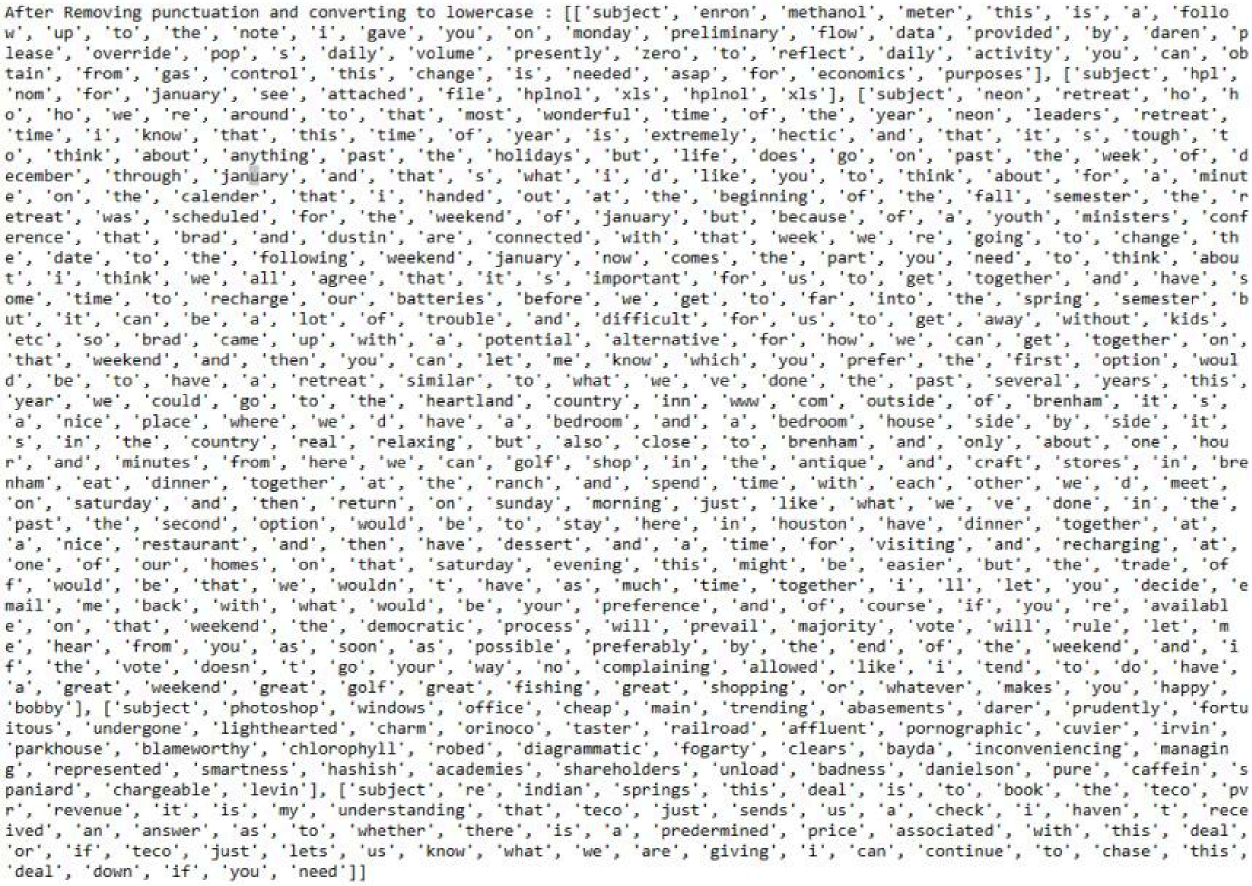
The first step of our model is gathering the dataset as a sentence and breaking it into understandable parts (words) and printing the same for first few sentences of our dataset using head().



```
AFTER TOKENIZATION: [['Subject', ':', 'enron', 'methanol', ';', 'meter', '#', ':', '988291', 'this', 'is', 'a', 'follow', 'up',
'to', 'the', 'note', 'i', 'gave', 'you', 'on', 'monday', ',', '4', '/', '3', '/', '00', '{', 'preliminary', 'flow', 'data', 'pr
ovided', 'by', 'daren', '}', '.', 'please', 'override', 'pop', "'", 's', 'daily', 'volume', '{', 'presently', 'zero', '}', 't
o', 'reflect', 'daily', 'activity', 'you', 'can', 'obtain', 'from', 'gas', 'control', '.', 'this', 'change', 'is', 'needed', 'a
sap', 'for', 'economics', 'purposes', '.'], ['Subject', ':', 'hpl', 'nom', 'for', 'january', '9', ',', '2001', '(', 'see', 'att
ached', 'file', ':', 'hplnol', '09', '.', 'xls', ')', '-', 'hplnol', '09', '.', 'xls'], ['Subject', ':', 'neon', 'retreat', 'h
o', 'ho', 'ho', ',', 'we', "'", 're', 'around', 'to', 'that', 'most', 'wonderful', 'time', 'of', 'the', 'year', '-', '-', '-',
'neon', 'leaders', 'retreat', 'time', '!', 'i', 'know', 'that', 'this', 'time', 'of', 'year', 'is', 'extremely', 'hectic', ',',
'and', 'that', 'it', "'", 's', 'tough', 'to', 'think', 'about', 'anything', 'past', 'the', 'holidays', ',', 'but', 'life', 'doe
s', 'go', 'on', 'past', 'the', 'week', 'of', 'december', '25', 'through', 'january', '1', ',', 'and', 'that', "'", 's', 'what',
'i', "'", 'd', 'like', 'you', 'to', 'think', 'about', 'for', 'a', 'minute', '.', 'on', 'the', 'calender', 'that', 'i', 'hande
d', 'out', 'at', 'the', 'beginning', 'of', 'the', 'fall', 'semester', ',', 'the', 'retreat', 'was', 'scheduled', 'for', 'the',
'weekend', 'of', 'january', '5', '-', '6', '.', 'but', 'because', 'of', 'a', 'youth', 'ministers', 'conference', 'that', 'bra
d', 'and', 'dustin', 'are', 'connected', 'with', 'that', 'week', ',', 'we', "'", 're', 'going', 'to', 'change', 'the', 'date',
'to', 'the', 'following', 'weekend', ',', 'january', '12', '-', '13', '.', 'now', 'comes', 'the', 'part', 'you', 'need', 'to',
'think', 'about', '.', 'i', 'think', 'we', 'all', 'agree', 'that', 'it', "'", 's', 'important', 'for', 'us', 'to', 'get', 'toge
ther', 'and', 'have', 'some', 'time', 'to', 'recharge', 'our', 'batteries', 'before', 'we', 'get', 'to', 'far', 'into', 'the',
'spring', 'semester', ',', 'but', 'it', 'can', 'be', 'a', 'lot', 'of', 'trouble', 'and', 'difficult', 'for', 'us', 'to', 'get',
'away', 'without', 'kids', ',', 'etc', '.', 'so', ',', 'brad', 'came', 'up', 'with', 'a', 'potential', 'alternative', 'for', 'h
ow', 'we', 'can', 'get', 'together', 'on', 'that', 'weekend', ',', 'and', 'then', 'you', 'can', 'let', 'me', 'know', 'which',
'you', 'prefer', '.', 'the', 'first', 'option', 'would', 'be', 'to', 'have', 'a', 'retreat', 'similar', 'to', 'what', 'we',
"'", 've', 'done', 'the', 'past', 'several', 'years', '.', 'this', 'year', 'we', 'could', 'go', 'to', 'the', 'heartland', 'coun
try', 'inn', '(', 'www', '.', '.', 'com', ')', 'outside', 'of', 'brenham', '.', 'it', "'", 's', 'a', 'nice', 'place', ',', 'whe
re', 'we', "'", 'd', 'have', 'a', '13', '-', 'bedroom', 'and', 'a', '5', '-', 'bedroom', 'house', 'side', 'by', 'side', '.', 'i
t', "'", 's', 'in', 'the', 'country', ',', 'real', 'relaxing', ',', 'but', 'also', 'close', 'to', 'brenham', 'and', 'only', 'ab
out', 'one', 'hour', 'and', '15', 'minutes', 'from', 'here', '.', 'we', 'can', 'golf', ',', 'shop', 'in', 'the', 'antique', 'an
d', 'craft', 'stores', 'in', 'brenham', ',', 'eat', 'dinner', 'together', 'at', 'the', 'ranch', ',', 'and', 'spend', 'time', 'w
ith', 'each', 'other', '.', 'we', "'", 'd', 'meet', 'on', 'saturday', ',', 'and', 'then', 'return', 'on', 'sunday', 'morning',
',', 'just', 'like', 'what', 'we', "'", 've', 'done', 'in', 'the', 'past', '.', 'the', 'second', 'option', 'would', 'be', 'to',
'stay', 'here', 'in', 'houston', ',', 'have', 'dinner', 'together', 'at', 'a', 'nice', 'restaurant', ',', 'and', 'then', 'hav
e', 'dessert', 'and', 'a', 'time', 'for', 'visiting', 'and', 'recharging', 'at', 'one', 'of', 'our', 'homes', 'on', 'that', 'sa
turday', 'evening', '.', 'this', 'might', 'be', 'easier', ',', 'but', 'the', 'trade', 'off', 'would', 'be', 'that', 'we', 'woul
dn', "'", 't', 'have', 'as', 'much', 'time', 'together', '.', 'i', "'", 'll', 'let', 'you', 'decide', '.', 'email', 'me', 'bac
k', 'with', 'what', 'would', 'be', 'your', 'preference', ',', 'and', 'of', 'course', 'if', 'you', "'", 're', 'available', 'on',
'that', 'weekend', '.', 'the', 'democratic', 'process', 'will', 'prevail', '-', '-', 'majority', 'vote', 'will', 'rule', '!',
'let', 'me', 'hear', 'from', 'you', 'as', 'soon', 'as', 'possible', ',', 'preferably', 'by', 'the', 'end', 'of', 'the', 'weeken
d', '.', 'and', 'if', 'the', 'vote', 'doesn', "'", 't', 'go', 'your', 'way', ',', 'no', 'complaining', 'allowed', '(', 'like',
'i', 'tend', 'to', 'do', '!', ')', 'have', 'a', 'great', 'weekend', ',', 'great', 'golf', ',', 'great', 'fishing', ',', 'grea
t', 'shopping', ',', 'or', 'whatever', 'makes', 'you', 'happy', '!', 'bobby'], ['Subject', ':', 'photoshop', ',', 'windows',
',', 'office', '.', 'cheap', '.', 'main', 'trending', 'abasements', 'darer', 'prudently', 'fortuitous', 'undergone', 'lighthear
ted', 'charm', 'orinoco', 'taster', 'railroad', 'affluent', 'pornographic', 'cuvier', 'irvin', 'parkhouse', 'blameworthy', 'chl
orophyll', 'robed', 'diagrammatic', 'fogarty', 'clears', 'bayda', 'inconveniencing', 'managing', 'represented', 'smartness', 'h
ashish', 'academies', 'shareholders', 'unload', 'badness', 'danielson', 'pure', 'caffein', 'spaniard', 'chargeable', 'levin'],
['Subject', ':', 're', ':', 'indian', 'springs', 'this', 'deal', 'is', 'to', 'book', 'the', 'teco', 'pvr', 'revenue', '.', 'i
t', 'is', 'my', 'understanding', 'that', 'teco', 'just', 'sends', 'us', 'a', 'check', ',', 'i', 'haven', "'", 't', 'received',
'an', 'answer', 'as', 'to', 'whether', 'there', 'is', 'a', 'predermined', 'price', 'associated', 'with', 'this', 'deal', 'or',
'if', 'teco', 'just', 'lets', 'us', 'know', 'what', 'we', 'are', 'giving', '.', 'i', 'can', 'continue', 'to', 'chase', 'this',
'deal', 'down', 'if', 'you', 'need', '.']]
```

STEP 2: REMOVING PUNCTUATIONS AND CASE CONVERSION:

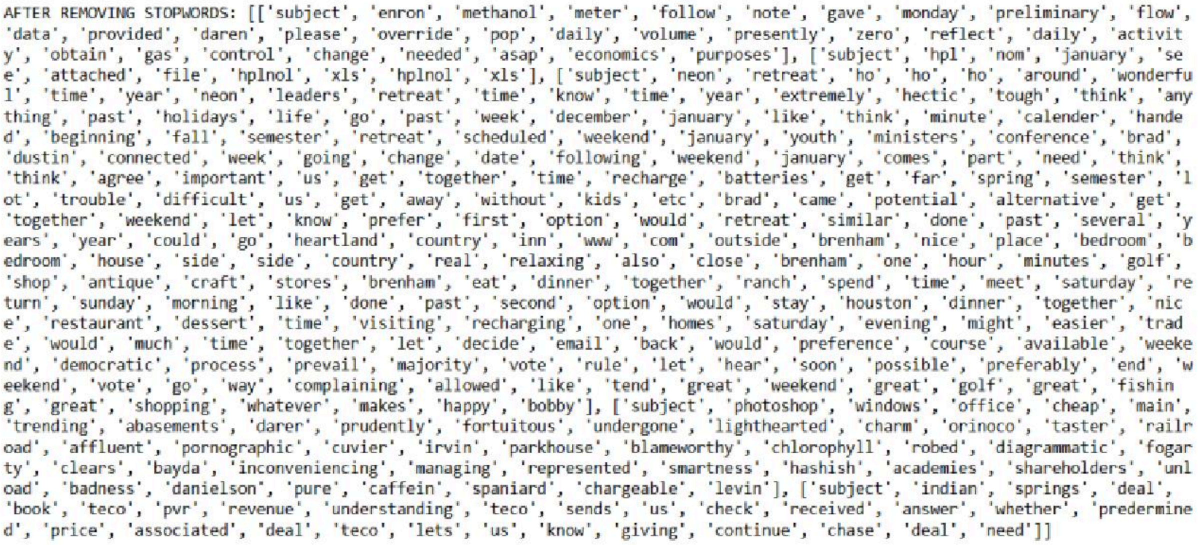
The next step of our model is gathering tokenized words and removing the punctuations and converting the tokens to the lowercase for the learning ability of our model and printing the same.



```
After Removing punctuation and converting to lowercase : [['subject', 'enron', 'methanol', 'meter', 'this', 'is', 'a', 'follo
w', 'up', 'to', 'the', 'note', 'i', 'gave', 'you', 'on', 'monday', 'preliminary', 'flow', 'data', 'provided', 'by', 'daren', 'p
lease', 'override', 'pop', 's', 'daily', 'volume', 'presently', 'zero', 'to', 'reflect', 'daily', 'activity', 'you', 'can', 'ob
tain', 'from', 'gas', 'control', 'this', 'change', 'is', 'needed', 'asap', 'for', 'economics', 'purposes'], ['subject', 'hpl',
'nom', 'for', 'january', 'see', 'attached', 'file', 'hplnol', 'xls', 'hplnol', 'xls'], ['subject', 'neon', 'retreat', 'ho', 'h
o', 'ho', 'we', 're', 'around', 'to', 'that', 'most', 'wonderful', 'time', 'of', 'the', 'year', 'neon', 'leaders', 'retreat',
'time', 'i', 'know', 'that', 'this', 'time', 'of', 'year', 'is', 'extremely', 'hectic', 'and', 'that', 'it', 's', 'tough', 't
o', 'think', 'about', 'anything', 'past', 'the', 'holidays', 'but', 'life', 'does', 'go', 'on', 'past', 'the', 'week', 'of', 'd
ecember', 'through', 'january', 'and', 'that', 's', 'what', 'i', 'd', 'like', 'you', 'to', 'think', 'about', 'for', 'a', 'minut
e', 'on', 'the', 'calender', 'that', 'i', 'handed', 'out', 'at', 'the', 'beginning', 'of', 'the', 'fall', 'semester', 'the', 'r
etreat', 'was', 'scheduled', 'for', 'the', 'weekend', 'of', 'january', 'but', 'because', 'of', 'a', 'youth', 'ministers', 'conf
erence', 'that', 'brad', 'and', 'dustin', 'are', 'connected', 'with', 'that', 'week', 'we', 're', 'going', 'to', 'change', 'th
e', 'date', 'to', 'the', 'following', 'weekend', 'january', 'now', 'comes', 'the', 'part', 'you', 'need', 'to', 'think', 'abou
t', 'i', 'think', 'we', 'all', 'agree', 'that', 'it', 's', 'important', 'for', 'us', 'to', 'get', 'together', 'and', 'have', 's
ome', 'time', 'to', 'recharge', 'our', 'batteries', 'before', 'we', 'get', 'to', 'far', 'into', 'the', 'spring', 'semester', 'b
ut', 'it', 'can', 'be', 'a', 'lot', 'of', 'trouble', 'and', 'difficult', 'for', 'us', 'to', 'get', 'away', 'without', 'kids',
'etc', 'so', 'brad', 'came', 'up', 'with', 'a', 'potential', 'alternative', 'for', 'how', 'we', 'can', 'get', 'together', 'on',
'that', 'weekend', 'and', 'then', 'you', 'can', 'let', 'me', 'know', 'which', 'you', 'prefer', 'the', 'first', 'option', 'woul
d', 'be', 'to', 'have', 'a', 'retreat', 'similar', 'to', 'what', 'we', 've', 'done', 'the', 'past', 'several', 'years', 'this',
'year', 'we', 'could', 'go', 'to', 'the', 'heartland', 'country', 'inn', 'www', 'com', 'outside', 'of', 'brenham', 'it', 's',
'a', 'nice', 'place', 'where', 'we', 'd', 'have', 'a', 'bedroom', 'and', 'a', 'bedroom', 'house', 'side', 'by', 'side', 'it',
's', 'in', 'the', 'country', 'real', 'relaxing', 'but', 'also', 'close', 'to', 'brenham', 'and', 'only', 'about', 'one', 'hou
r', 'and', 'minutes', 'from', 'here', 'we', 'can', 'golf', 'shop', 'in', 'the', 'antique', 'and', 'craft', 'stores', 'in', 'bre
nham', 'eat', 'dinner', 'together', 'at', 'the', 'ranch', 'and', 'spend', 'time', 'with', 'each', 'other', 'we', 'd', 'meet',
'on', 'saturday', 'and', 'then', 'return', 'on', 'sunday', 'morning', 'just', 'like', 'what', 'we', 've', 'done', 'in', 'the',
'past', 'the', 'second', 'option', 'would', 'be', 'to', 'stay', 'here', 'in', 'houston', 'have', 'dinner', 'together', 'at',
'a', 'nice', 'restaurant', 'and', 'then', 'have', 'dessert', 'and', 'a', 'time', 'for', 'visiting', 'and', 'recharging', 'at',
'one', 'of', 'our', 'homes', 'on', 'that', 'saturday', 'evening', 'this', 'might', 'be', 'easier', 'but', 'the', 'trade', 'of
f', 'would', 'be', 'that', 'we', 'wouldn', 't', 'have', 'as', 'much', 'time', 'together', 'i', 'll', 'let', 'you', 'decide', 'e
mail', 'me', 'back', 'with', 'what', 'would', 'be', 'your', 'preference', 'and', 'of', 'course', 'if', 'you', 're', 'availabl
e', 'on', 'that', 'weekend', 'the', 'democratic', 'process', 'will', 'prevail', 'majority', 'vote', 'will', 'rule', 'let', 'm
e', 'hear', 'from', 'you', 'as', 'soon', 'as', 'possible', 'preferably', 'by', 'the', 'end', 'of', 'the', 'weekend', 'and', 'i
f', 'the', 'vote', 'doesn', 't', 'go', 'your', 'way', 'no', 'complaining', 'allowed', 'like', 'i', 'tend', 'to', 'do', 'have',
'a', 'great', 'weekend', 'great', 'golf', 'great', 'fishing', 'great', 'shopping', 'or', 'whatever', 'makes', 'you', 'happy',
'bobby'], ['subject', 'photoshop', 'windows', 'office', 'cheap', 'main', 'trending', 'abasements', 'darer', 'prudently', 'fortu
itous', 'undergone', 'lighthearted', 'charm', 'orinoco', 'taster', 'railroad', 'affluent', 'pornographic', 'cuvier', 'irvin',
'parkhouse', 'blameworthy', 'chlorophyll', 'robed', 'diagrammatic', 'fogarty', 'clears', 'bayda', 'inconveniencing', 'managin
g', 'represented', 'smartness', 'hashish', 'academies', 'shareholders', 'unload', 'badness', 'danielson', 'pure', 'caffein', 's
paniard', 'chargeable', 'levin'], ['subject', 're', 'indian', 'springs', 'this', 'deal', 'is', 'to', 'book', 'the', 'teco', 'pv
r', 'revenue', 'it', 'is', 'my', 'understanding', 'that', 'teco', 'just', 'sends', 'us', 'a', 'check', 'i', 'haven', 't', 'rece
ived', 'an', 'answer', 'as', 'to', 'whether', 'there', 'is', 'a', 'predermined', 'price', 'associated', 'with', 'this', 'deal',
'or', 'if', 'teco', 'just', 'lets', 'us', 'know', 'what', 'we', 'are', 'giving', 'i', 'can', 'continue', 'to', 'chase', 'this',
'deal', 'down', 'if', 'you', 'need']]
```

STEP 3:STOPWORDS REMOVAL:

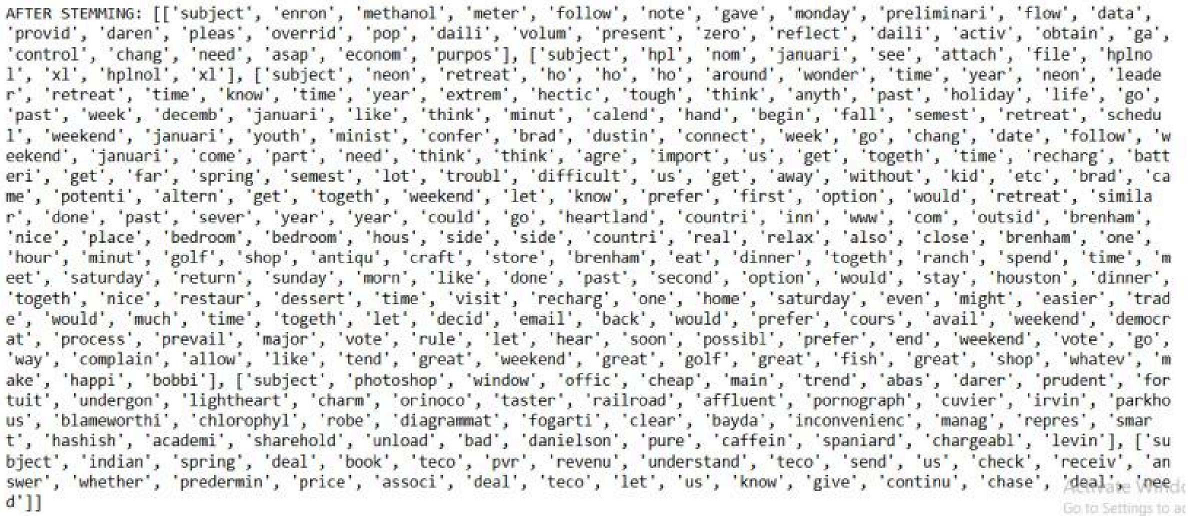
After removing the punctuations and case conversion stop words has been removed from the same and printing the same.



```
AFTER REMOVING STOPWORDS: [['subject', 'enron', 'methanol', 'meter', 'follow', 'note', 'gave', 'monday', 'preliminary', 'flow',
'data', 'provided', 'daren', 'please', 'override', 'pop', 'daily', 'volume', 'presently', 'zero', 'reflect', 'daily', 'activit
y', 'obtain', 'gas', 'control', 'change', 'needed', 'asap', 'economics', 'purposes'], ['subject', 'hpl', 'nom', 'january', 'se
e', 'attached', 'file', 'hplnol', 'xls', 'hplnol', 'xls'], ['subject', 'neon', 'retreat', 'ho', 'ho', 'ho', 'around', 'wonderfu
l', 'time', 'year', 'neon', 'leaders', 'retreat', 'time', 'know', 'time', 'year', 'extremely', 'hectic', 'tough', 'think', 'any
thing', 'past', 'holidays', 'life', 'go', 'past', 'week', 'december', 'january', 'like', 'think', 'minute', 'calender', 'hande
d', 'beginning', 'fall', 'semester', 'retreat', 'scheduled', 'weekend', 'january', 'youth', 'ministers', 'conference', 'brad',
'dustin', 'connected', 'week', 'going', 'change', 'date', 'following', 'weekend', 'january', 'comes', 'part', 'need', 'think',
'think', 'agree', 'important', 'us', 'get', 'together', 'time', 'recharge', 'batteries', 'get', 'far', 'spring', 'semester', 'l
ot', 'trouble', 'difficult', 'us', 'get', 'away', 'without', 'kids', 'etc', 'brad', 'came', 'potential', 'alternative', 'get',
'together', 'weekend', 'let', 'know', 'prefer', 'first', 'option', 'would', 'retreat', 'similar', 'done', 'past', 'several', 'y
ears', 'year', 'could', 'go', 'heartland', 'country', 'inn', 'www', 'com', 'outside', 'brenham', 'nice', 'place', 'bedroom', 'b
edroom', 'house', 'side', 'side', 'country', 'real', 'relaxing', 'also', 'close', 'brenham', 'one', 'hour', 'minutes', 'golf',
'shop', 'antique', 'craft', 'stores', 'brenham', 'eat', 'dinner', 'together', 'ranch', 'spend', 'time', 'meet', 'saturday', 're
turn', 'sunday', 'morning', 'like', 'done', 'past', 'second', 'option', 'would', 'stay', 'houston', 'dinner', 'together', 'nic
e', 'restaurant', 'dessert', 'time', 'visiting', 'recharging', 'one', 'homes', 'saturday', 'evening', 'might', 'easier', 'trad
e', 'would', 'much', 'time', 'together', 'let', 'decide', 'email', 'back', 'would', 'preference', 'course', 'available', 'weeke
nd', 'democratic', 'process', 'prevail', 'majority', 'vote', 'rule', 'let', 'hear', 'soon', 'possible', 'preferably', 'end', 'w
eekend', 'vote', 'go', 'way', 'complaining', 'allowed', 'like', 'tend', 'great', 'weekend', 'great', 'golf', 'great', 'fishin
g', 'great', 'shopping', 'whatever', 'makes', 'happy', 'bobby'], ['subject', 'photoshop', 'windows', 'office', 'cheap', 'main',
'trending', 'abasements', 'darer', 'prudently', 'fortuitous', 'undergone', 'lighthearted', 'charm', 'orinoco', 'taster', 'railr
oad', 'affluent', 'pornographic', 'cuvier', 'irvin', 'parkhouse', 'blameworthy', 'chlorophyll', 'robed', 'diagrammatic', 'fogar
ty', 'clears', 'bayda', 'inconveniencing', 'managing', 'represented', 'smartness', 'hashish', 'academies', 'shareholders', 'unl
oad', 'badness', 'danielson', 'pure', 'caffein', 'spaniard', 'chargeable', 'levin'], ['subject', 'indian', 'springs', 'deal',
'book', 'teco', 'pvr', 'revenue', 'understanding', 'teco', 'sends', 'us', 'check', 'received', 'answer', 'whether', 'predermine
d', 'price', 'associated', 'deal', 'teco', 'lets', 'us', 'know', 'giving', 'continue', 'chase', 'deal', 'need']]
```

STEP 4:STEMMING:

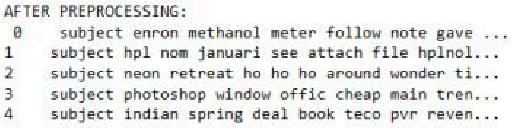
After removing the unnecessary tokens from the data it has been convert to its root forms using stemming and printing the same.



```
AFTER STEMMING: [['subject', 'enron', 'methanol', 'meter', 'follow', 'note', 'gave', 'monday', 'preliminari', 'flow', 'data',
'provid', 'daren', 'pleas', 'overrid', 'pop', 'daili', 'volum', 'present', 'zero', 'reflect', 'daili', 'activ', 'obtain', 'ga',
'control', 'chang', 'need', 'asap', 'econom', 'purpos'], ['subject', 'hpl', 'nom', 'januari', 'see', 'attach', 'file', 'hplno
l', 'xl', 'hplnol', 'xl'], ['subject', 'neon', 'retreat', 'ho', 'ho', 'ho', 'around', 'wonder', 'time', 'year', 'neon', 'leade
r', 'retreat', 'time', 'know', 'time', 'year', 'extrem', 'hectic', 'tough', 'think', 'anyth', 'past', 'holiday', 'life', 'go',
'past', 'week', 'decemb', 'januari', 'like', 'think', 'minut', 'calend', 'hand', 'begin', 'fall', 'semest', 'retreat', 'schedu
l', 'weekend', 'januari', 'youth', 'minist', 'confer', 'brad', 'dustin', 'connect', 'week', 'go', 'chang', 'date', 'follow', 'w
eekend', 'januari', 'come', 'part', 'need', 'think', 'think', 'agre', 'import', 'us', 'get', 'togeth', 'time', 'recharg', 'batt
eri', 'get', 'far', 'spring', 'semest', 'lot', 'troubl', 'difficult', 'us', 'get', 'away', 'without', 'kid', 'etc', 'brad', 'ca
me', 'potenti', 'altern', 'get', 'togeth', 'weekend', 'let', 'know', 'prefer', 'first', 'option', 'would', 'retreat', 'simila
r', 'done', 'past', 'sever', 'year', 'year', 'could', 'go', 'heartland', 'countri', 'inn', 'www', 'com', 'outsid', 'brenham',
'nice', 'place', 'bedroom', 'bedroom', 'hous', 'side', 'side', 'countri', 'real', 'relax', 'also', 'close', 'brenham', 'one',
'hour', 'minut', 'golf', 'shop', 'antiqu', 'craft', 'store', 'brenham', 'eat', 'dinner', 'togeth', 'ranch', 'spend', 'time', 'm
eet', 'saturday', 'return', 'sunday', 'morn', 'like', 'done', 'past', 'second', 'option', 'would', 'stay', 'houston', 'dinner',
'togeth', 'nice', 'restaur', 'dessert', 'time', 'visit', 'recharg', 'one', 'home', 'saturday', 'even', 'might', 'easier', 'trad
e', 'would', 'much', 'time', 'togeth', 'let', 'decid', 'email', 'back', 'would', 'prefer', 'cours', 'avail', 'weekend', 'democr
at', 'process', 'prevail', 'major', 'vote', 'rule', 'let', 'hear', 'soon', 'possibl', 'prefer', 'end', 'weekend', 'vote', 'go',
'way', 'complain', 'allow', 'like', 'tend', 'great', 'weekend', 'great', 'golf', 'great', 'fish', 'great', 'shop', 'whatev', 'm
ake', 'happi', 'bobbi'], ['subject', 'photoshop', 'window', 'offic', 'cheap', 'main', 'trend', 'abas', 'darer', 'prudent', 'for
tuit', 'undergon', 'lightheart', 'charm', 'orinoco', 'taster', 'railroad', 'affluent', 'pornograph', 'cuvier', 'irvin', 'parkho
us', 'blameworthi', 'chlorophyl', 'robe', 'diagrammat', 'fogarti', 'clear', 'bayda', 'inconvenienc', 'manag', 'repres', 'smar
t', 'hashish', 'academi', 'sharehold', 'unload', 'bad', 'danielson', 'pure', 'caffein', 'spaniard', 'chargeabl', 'levin'], ['su
bject', 'indian', 'spring', 'deal', 'book', 'teco', 'pvr', 'revenu', 'understand', 'teco', 'send', 'us', 'check', 'receiv', 'an
swer', 'whether', 'predermin', 'price', 'associ', 'deal', 'teco', 'let', 'us', 'know', 'give', 'continu', 'chase', 'deal', 'nee
d']]
```

FINAL STEP:

After doing all the pre-processing steps the final pre-processed output has been printed.

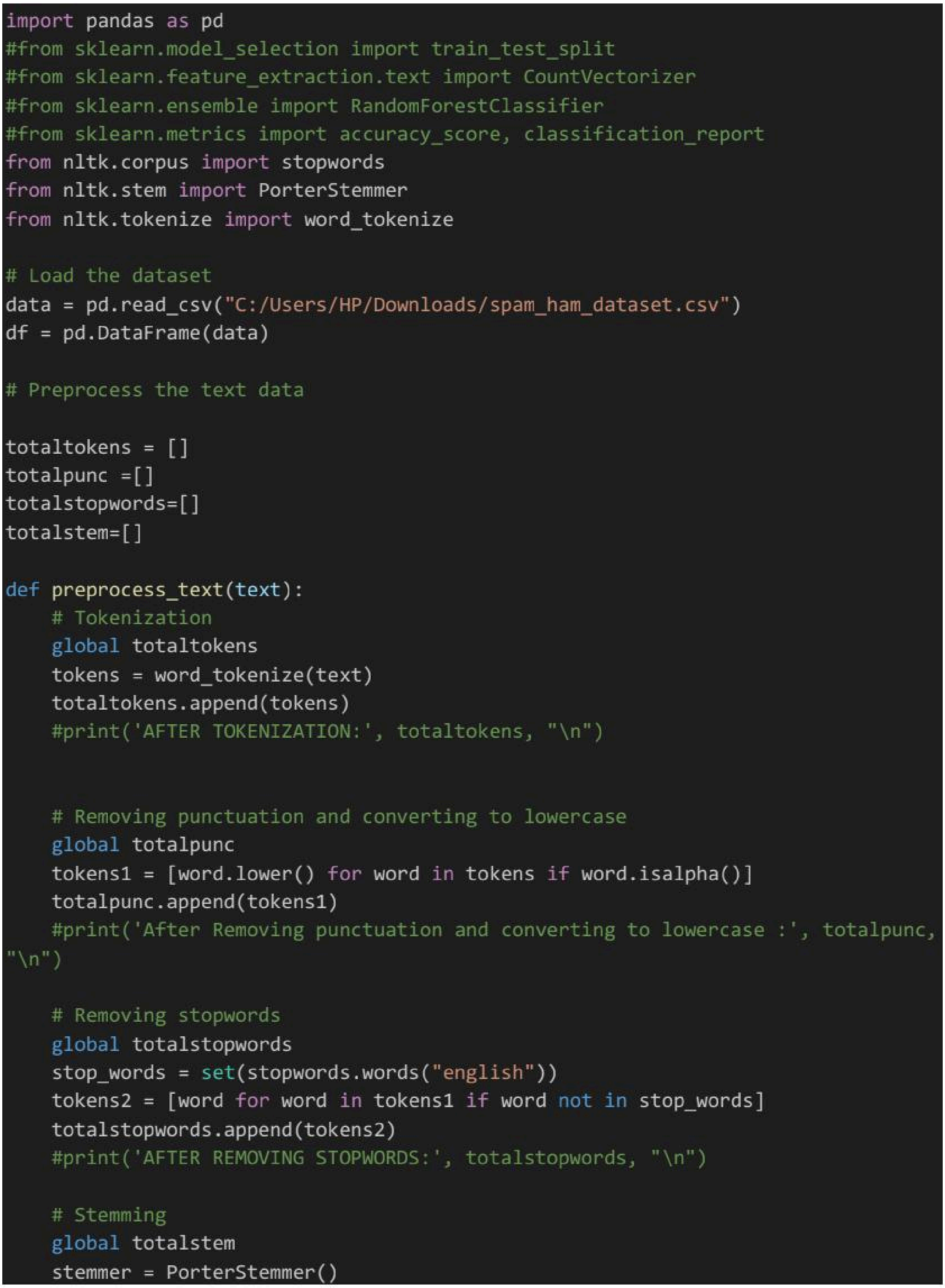


```
AFTER PREPROCESSING:
 0     subject enron methanol meter follow note gave ...
1     subject hpl nom januari see attach file hplnol...
2     subject neon retreat ho ho ho around wonder ti...
3     subject photoshop window offic cheap main tren...
4     subject indian spring deal book teco pvr reven...
```

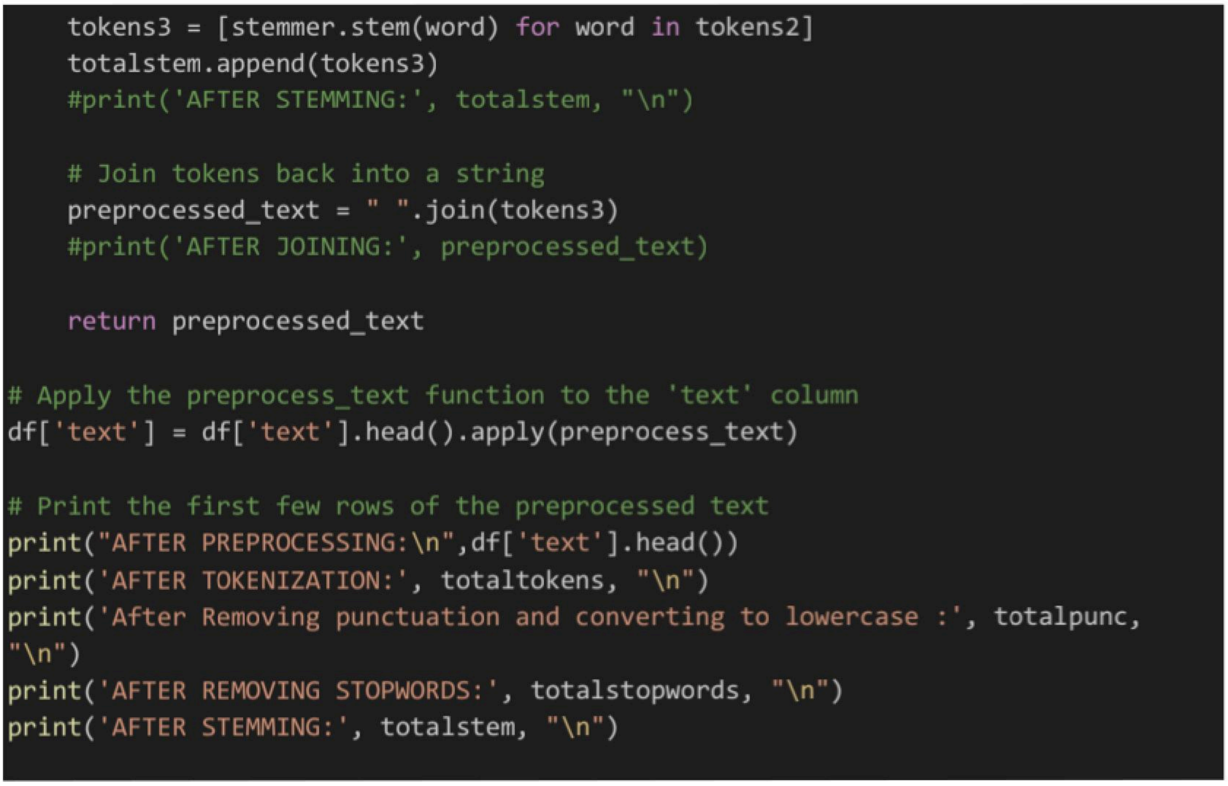
PREPROCESSING CODE:

LANGUAGE: PYTHON

PLATFORM: JUPYTER NOTEBOOK



```
import pandas as pd
#from sklearn.model_selection import train_test_split
#from sklearn.feature_extraction.text import CountVectorizer
#from sklearn.ensemble import RandomForestClassifier
#from sklearn.metrics import accuracy_score, classification_report
from nltk.corpus import stopwords
from nltk.stem import PorterStemmer
from nltk.tokenize import word_tokenize

# Load the dataset
data = pd.read_csv("C:/Users/HP/Downloads/spam_ham_dataset.csv")
df = pd.DataFrame(data)

# Preprocess the text data

totaltokens = []
totalpunc =[]
totalstopwords=[]
totalstem=[]

def preprocess_text(text):
    # Tokenization
    global totaltokens
    tokens = word_tokenize(text)
    totaltokens.append(tokens)
    #print('AFTER TOKENIZATION:', totaltokens, "\n")


    # Removing punctuation and converting to lowercase
    global totalpunc
    tokens1 = [word.lower() for word in tokens if word.isalpha()]
    totalpunc.append(tokens1)
    #print('After Removing punctuation and converting to lowercase :', totalpunc,
"\n")

    # Removing stopwords
    global totalstopwords
    stop_words = set(stopwords.words("english"))
    tokens2 = [word for word in tokens1 if word not in stop_words]
    totalstopwords.append(tokens2)
    #print('AFTER REMOVING STOPWORDS:', totalstopwords, "\n")

    # Stemming
    global totalstem
    stemmer = PorterStemmer()
```



```
    tokens3 = [stemmer.stem(word) for word in tokens2]
    totalstem.append(tokens3)
    #print('AFTER STEMMING:', totalstem, "\n")

    # Join tokens back into a string
    preprocessed_text = " ".join(tokens3)
    #print('AFTER JOINING:', preprocessed_text)

    return preprocessed_text

# Apply the preprocess_text function to the 'text' column
df['text'] = df['text'].head().apply(preprocess_text)

# Print the first few rows of the preprocessed text
print("AFTER PREPROCESSING:\n",df['text'].head())
print('AFTER TOKENIZATION:', totaltokens, "\n")
print('After Removing punctuation and converting to lowercase :', totalpunc,
"\n")
print('AFTER REMOVING STOPWORDS:', totalstopwords, "\n")
print('AFTER STEMMING:', totalstem, "\n")
```