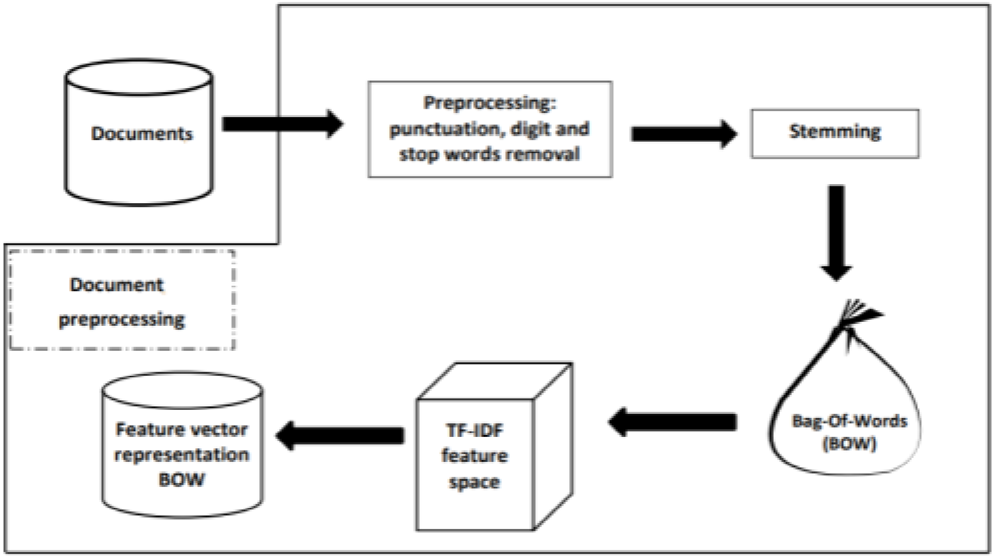
Mojeeb Alrahmaan Hasan: 2141987

Hatem Yousef: 2132159

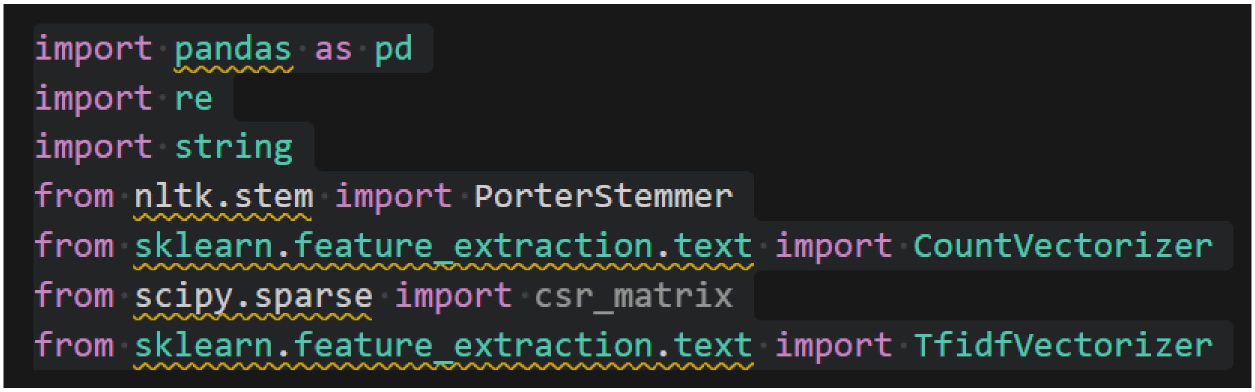
Abdallah Sadeq Elzurba: 2135654

Information retrieval

In this assignment we will take a csv file and conduct some tasks in the following figure and compare between the results using different experiment setups with different options for stemming, tokenizing, weighting, case folding and stop words handling.


Documents
Preprocessing: punctuation, digit and stop words removal
Stemming
Document preprocessing
Bag-Of-Words (BOW)
TF-IDF feature space
Feature vector representation BOW

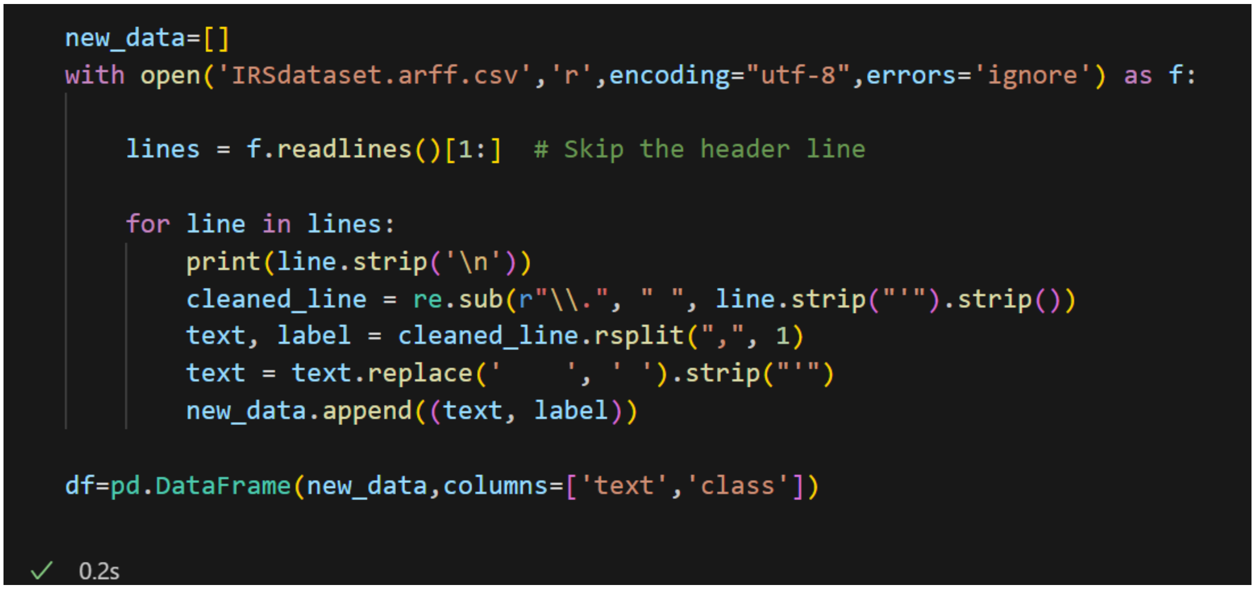
In our tasks we will be using each one of these libraries we imported in Python.



```python
import pandas as pd
import re
import string
from nltk.stem import PorterStemmer
from sklearn.feature_extraction.text import CountVectorizer
from scipy.sparse import csr_matrix
from sklearn.feature_extraction.text import TfidfVectorizer
```

We have imported here pandas, re(regular expression) library, nltk(natural language Toolkit) library for stemming and other libraries we will show in the next slides.

Step1. Preprocessing: punctuation, digit and stop words removal.

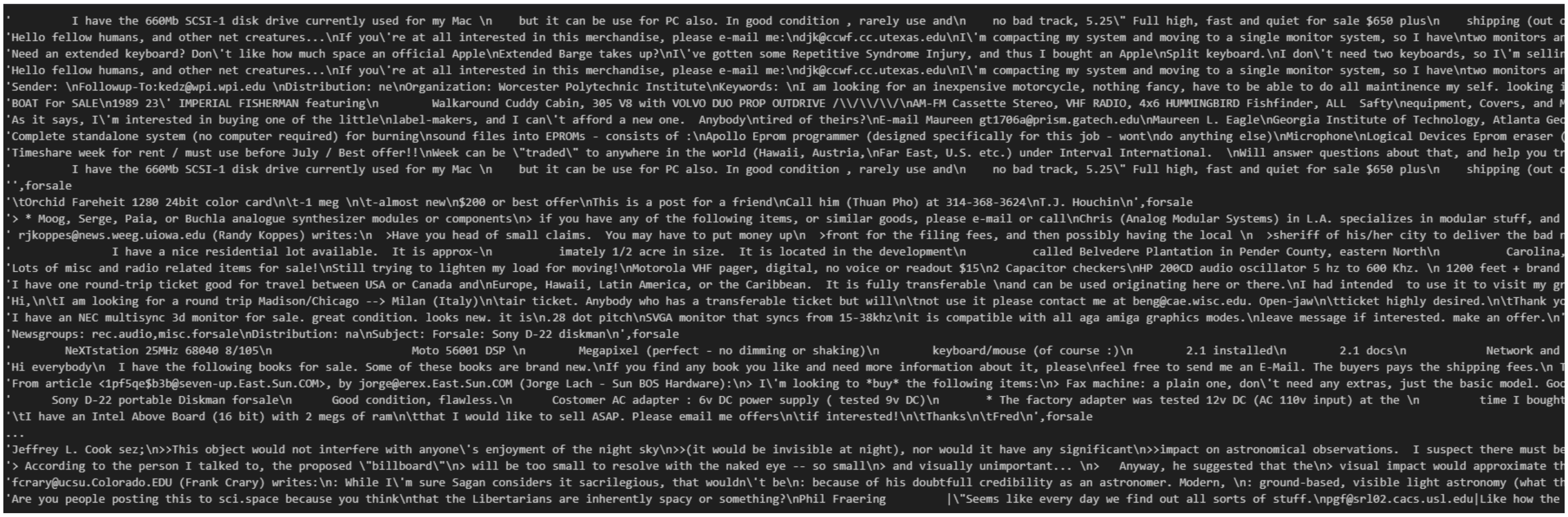


```
new_data=[]
with open('IRSdataset.arff.csv','r',encoding="utf-8",errors='ignore') as f:

    lines = f.readlines()[1:]  # Skip the header line

    for line in lines:
        print(line.strip('\n'))
        cleaned_line = re.sub(r"\\.", " ", line.strip("'").strip())
        text, label = cleaned_line.rsplit(",", 1)
        text = text.replace('     ', ' ').strip("'")
        new_data.append((text, label))

df=pd.DataFrame(new_data,columns=['text','class'])
```

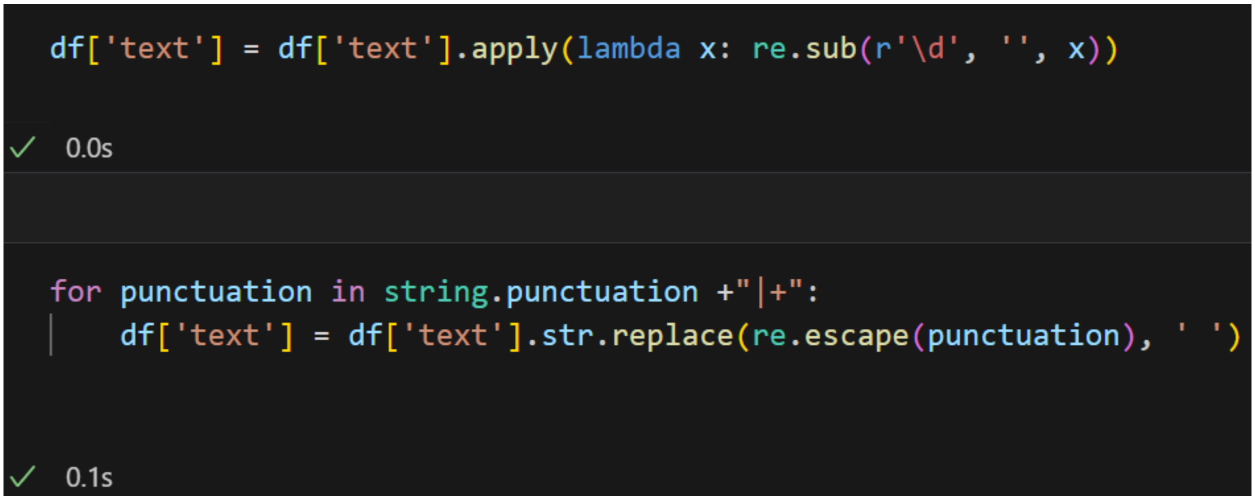
✓ 0.2s

Here we need to read our csv file and create a DataFrame using pandas and after reading we skipped the header then entered a for loop to visit each line, we cleaned the data from spaces and we splitted the lines into (text, label) and appended it into “new_data”.

This is our original data that we are about to process.



```
'         I have the 660Mb SCSI-1 disk drive currently used for my Mac \n    but it can be use for PC also. In good condition , rarely use and\n    no bad track, 5.25\" Full high, fast and quiet for sale $650 plus\n    shipping (out o
'Hello fellow humans, and other net creatures...\nIf you\'re at all interested in this merchandise, please e-mail me:\ndjk@ccwf.cc.utexas.edu\nI\'m compacting my system and moving to a single monitor system, so I have\ntwo monitors an
'Need an extended keyboard? Don\'t like how much space an official Apple\nExtended Barge takes up?\nI\'ve gotten some Repetitive Syndrome Injury, and thus I bought an Apple\nSplit keyboard.\nI don\'t need two keyboards, so I\'m sellin
'Hello fellow humans, and other net creatures...\nIf you\'re at all interested in this merchandise, please e-mail me:\ndjk@ccwf.cc.utexas.edu\nI\'m compacting my system and moving to a single monitor system, so I have\ntwo monitors an
'Sender: \nFollowup-To:kedz@wpi.wpi.edu \nDistribution: ne\nOrganization: Worcester Polytechnic Institute\nKeywords: \nI am looking for an inexpensive motorcycle, nothing fancy, have to be able to do all maintinence my self. looking i
'BOAT For SALE\n1989 23\' IMPERIAL FISHERMAN featuring\n        Walkaround Cuddy Cabin, 305 V8 with VOLVO DUO PROP OUTDRIVE /\\/\\/\\/\nAM-FM Cassette Stereo, VHF RADIO, 4x6 HUMMINGBIRD Fishfinder, ALL  Safty\nequipment, Covers, and M
'As it says, I\'m interested in buying one of the little\nlabel-makers, and I can\'t afford a new one.  Anybody\ntired of theirs?\nE-mail Maureen gt1706a@prism.gatech.edu\nMaureen L. Eagle\nGeorgia Institute of Technology, Atlanta Geo
'Complete standalone system (no computer required) for burning\nsound files into EPROMs - consists of :\nApollo Eprom programmer (designed specifically for this job - wont\ndo anything else)\nMicrophone\nLogical Devices Eprom eraser (
'Timeshare week for rent / must use before July / Best offer!!\nWeek can be \"traded\" to anywhere in the world (Hawaii, Austria,\nFar East, U.S. etc.) under Interval International.  \nWill answer questions about that, and help you tr
'         I have the 660Mb SCSI-1 disk drive currently used for my Mac \n    but it can be use for PC also. In good condition , rarely use and\n    no bad track, 5.25\" Full high, fast and quiet for sale $650 plus\n    shipping (out o
'',forsale
'\tOrchid Fareheit 1280 24bit color card\n\t-1 meg \n\t-almost new\n$200 or best offer\nThis is a post for a friend\nCall him (Thuan Pho) at 314-368-3624\nT.J. Houchin\n',forsale
'> * Moog, Serge, Paia, or Buchla analogue synthesizer modules or components\n> if you have any of the following items, or similar goods, please e-mail or call\nChris (Analog Modular Systems) in L.A. specializes in modular stuff, and 
' rjkoppes@news.weeg.uiowa.edu (Randy Koppes) writes:\n  >Have you head of small claims.  You may have to put money up\n  >front for the filing fees, and then possibly having the local \n  >sheriff of his/her city to deliver the bad n
'               I have a nice residential lot available.  It is approx-\n          imately 1/2 acre in size.  It is located in the development\n          called Belvedere Plantation in Pender County, eastern North\n          Carolina,
'Lots of misc and radio related items for sale!\nStill trying to lighten my load for moving!\nMotorola VHF pager, digital, no voice or readout $15\n2 Capacitor checkers\nHP 200CD audio oscillator 5 hz to 600 Khz. \n 1200 feet + brand 
'I have one round-trip ticket good for travel between USA or Canada and\nEurope, Hawaii, Latin America, or the Caribbean.  It is fully transferable \nand can be used originating here or there.\nI had intended  to use it to visit my gr
'Hi,\n\tI am looking for a round trip Madison/Chicago --> Milan (Italy)\n\tair ticket. Anybody who has a transferable ticket but will\n\tnot use it please contact me at beng@cae.wisc.edu. Open-jaw\n\tticket highly desired.\n\tThank yo
'I have an NEC multisync 3d monitor for sale. great condition. looks new. it is\n.28 dot pitch\nSVGA monitor that syncs from 15-38khz\nit is compatible with all aga amiga graphics modes.\nleave message if interested. make an offer.\n'
'Newsgroups: rec.audio,misc.forsale\nDistribution: na\nSubject: Forsale: Sony D-22 diskman\n',forsale
'        NeXTstation 25MHz 68040 8/105\n                     Moto 56001 DSP \n        Megapixel (perfect - no dimming or shaking)\n        keyboard/mouse (of course :)\n        2.1 installed\n        2.1 docs\n           Network and 
'Hi everybody\n  I have the following books for sale. Some of these books are brand new.\nIf you find any book you like and need more information about it, please\nfeel free to send me an E-Mail. The buyers pays the shipping fees.\n T
'From article <1pf5qe$b3b@seven-up.East.Sun.COM>, by jorge@erex.East.Sun.COM (Jorge Lach - Sun BOS Hardware):\n> I\'m looking to *buy* the following items:\n> Fax machine: a plain one, don\'t need any extras, just the basic model. Goo
'      Sony D-22 portable Diskman forsale\n      Good condition, flawless.\n      Costomer AC adapter : 6v DC power supply ( tested 9v DC)\n      * The factory adapter was tested 12v DC (AC 110v input) at the \n        time I bought
'\tI have an Intel Above Board (16 bit) with 2 megs of ram\n\tthat I would like to sell ASAP. Please email me offers\n\tif interested!\n\tThanks\n\tFred\n',forsale
...
'Jeffrey L. Cook sez;\n>>This object would not interfere with anyone\'s enjoyment of the night sky\n>>(it would be invisible at night), nor would it have any significant\n>>impact on astronomical observations.  I suspect there must be
'> According to the person I talked to, the proposed \"billboard\"\n> will be too small to resolve with the naked eye -- so small\n> and visually unimportant... \n>   Anyway, he suggested that the\n> visual impact would approximate th
'fcrary@ucsu.Colorado.EDU (Frank Crary) writes:\n: While I\'m sure Sagan considers it sacrilegious, that wouldn\'t be\n: because of his doubtfull credibility as an astronomer. Modern, \n: ground-based, visible light astronomy (what th
'Are you people posting this to sci.space because you think\nthat the Libertarians are inherently spacy or something?\nPhil Fraering        |\"Seems like every day we find out all sorts of stuff.\npgf@srl02.cacs.usl.edu|Like how the 
```



```
df['text'] = df['text'].apply(lambda x: re.sub(r'\d', '', x))
```

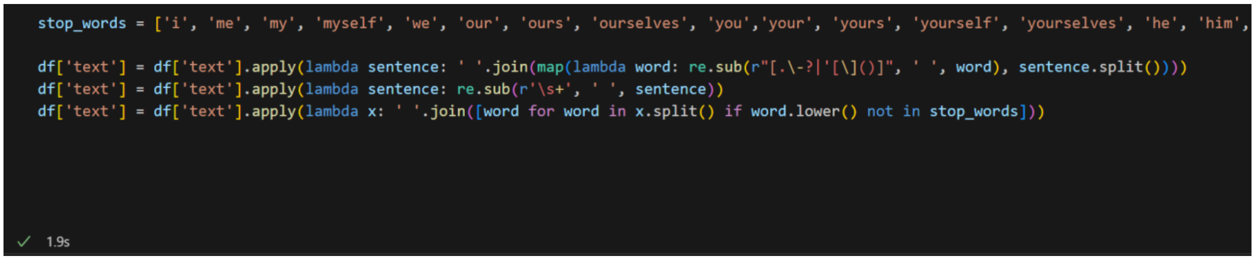
✓ 0.0s

```
for punctuation in string.punctuation +"|+":
    df['text'] = df['text'].str.replace(re.escape(punctuation), ' ')
```

✓ 0.1s

In the first line we cleaned the dataFrame text from digits and replaced it with an empty string then we did the punctuation in a for loop to replace every punctuation character including “ | “ in the text dataFrame with a space.

After we were done with punctuation and digits we needed to work on stop words that are included in the list, so we removed special characters, extra white Spaces and removed stop words that we put in the list with the following code:

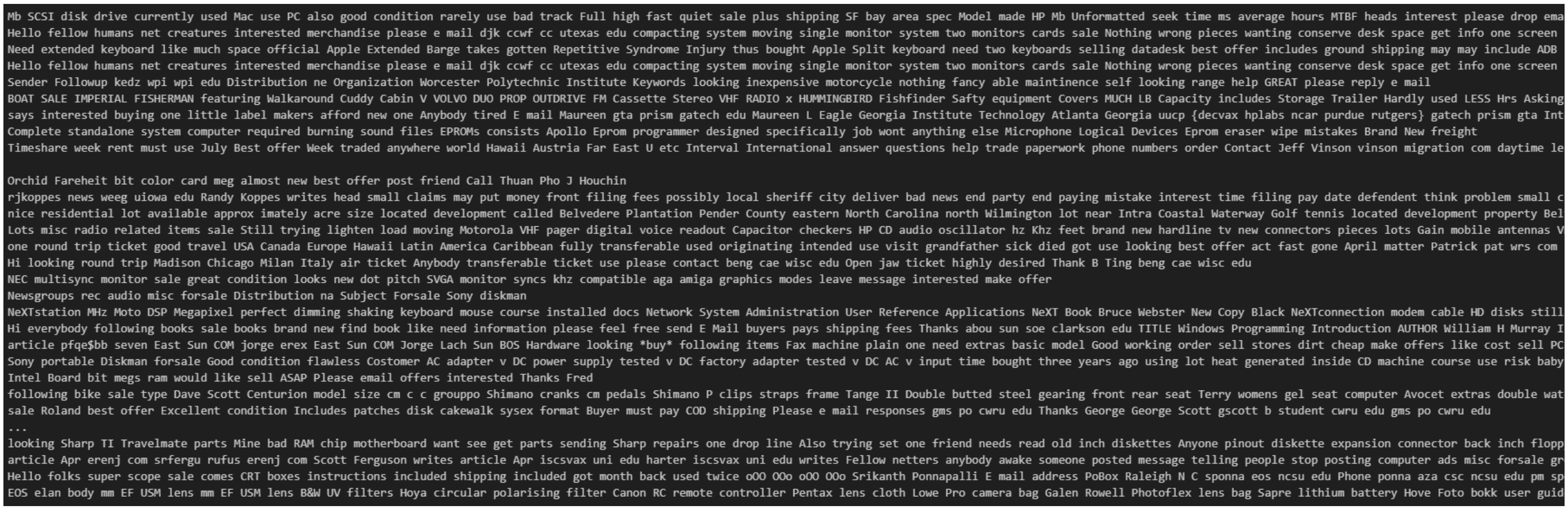


```
stop_words = ['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you','your', 'yours', 'yourself', 'yourselves', 'he', 'him',

df['text'] = df['text'].apply(lambda sentence: ' '.join(map(lambda word: re.sub(r"[.\-?|'[\]()]", ' ', word), sentence.split())))
df['text'] = df['text'].apply(lambda sentence: re.sub(r'\s+', ' ', sentence))
df['text'] = df['text'].apply(lambda x: ' '.join([word for word in x.split() if word.lower() not in stop_words]))
```

✓ 1.9s

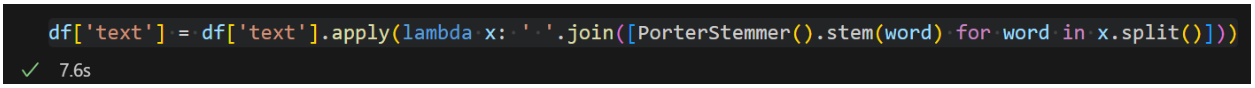
The goal here was to preprocess the data that is in the text dataFrame to filter it from stop words and the results should be cleaned and tokenized text data.

Now we can say we are done with our first step which was preprocessing our data and you can see the results in the next slide



```
Mb SCSI disk drive currently used Mac use PC also good condition rarely use bad track Full high fast quiet sale plus shipping SF bay area spec Model made HP Mb Unformatted seek time ms average hours MTBF heads interest please drop ema
Hello fellow humans net creatures interested merchandise please e mail djk ccwf cc utexas edu compacting system moving single monitor system two monitors cards sale Nothing wrong pieces wanting conserve desk space get info one screen
Need extended keyboard like much space official Apple Extended Barge takes gotten Repetitive Syndrome Injury thus bought Apple Split keyboard need two keyboards selling datadesk best offer includes ground shipping may may include ADB
Hello fellow humans net creatures interested merchandise please e mail djk ccwf cc utexas edu compacting system moving single monitor system two monitors cards sale Nothing wrong pieces wanting conserve desk space get info one screen
Sender Followup kedz wpi wpi edu Distribution ne Organization Worcester Polytechnic Institute Keywords looking inexpensive motorcycle nothing fancy able maintinence self looking range help GREAT please reply e mail
BOAT SALE IMPERIAL FISHERMAN featuring Walkaround Cuddy Cabin V VOLVO DUO PROP OUTDRIVE FM Cassette Stereo VHF RADIO x HUMMINGBIRD Fishfinder Safty equipment Covers MUCH LB Capacity includes Storage Trailer Hardly used LESS Hrs Asking
says interested buying one little label makers afford new one Anybody tired E mail Maureen gta prism gatech edu Maureen L Eagle Georgia Institute Technology Atlanta Georgia uucp {decvax hplabs ncar purdue rutgers} gatech prism gta Int
Complete standalone system computer required burning sound files EPROMs consists Apollo Eprom programmer designed specifically job wont anything else Microphone Logical Devices Eprom eraser wipe mistakes Brand New freight
Timeshare week rent must use July Best offer Week traded anywhere world Hawaii Austria Far East U etc Interval International answer questions help trade paperwork phone numbers order Contact Jeff Vinson vinson migration com daytime le

Orchid Fareheit bit color card meg almost new best offer post friend Call Thuan Pho J Houchin
rjkoppes news weeg uiowa edu Randy Koppes writes head small claims may put money front filing fees possibly local sheriff city deliver bad news end party end paying mistake interest time filing pay date defendent think problem small c
nice residential lot available approx imately acre size located development called Belvedere Plantation Pender County eastern North Carolina north Wilmington lot near Intra Coastal Waterway Golf tennis located development property Bel
Lots misc radio related items sale Still trying lighten load moving Motorola VHF pager digital voice readout Capacitor checkers HP CD audio oscillator hz Khz feet brand new hardline tv new connectors pieces lots Gain mobile antennas V
one round trip ticket good travel USA Canada Europe Hawaii Latin America Caribbean fully transferable used originating intended use visit grandfather sick died got use looking best offer act fast gone April matter Patrick pat wrs com
Hi looking round trip Madison Chicago Milan Italy air ticket Anybody transferable ticket use please contact beng cae wisc edu Open jaw ticket highly desired Thank B Ting beng cae wisc edu
NEC multisync monitor sale great condition looks new dot pitch SVGA monitor syncs khz compatible aga amiga graphics modes leave message interested make offer
Newsgroups rec audio misc forsale Distribution na Subject Forsale Sony diskman
NeXTstation MHz Moto DSP Megapixel perfect dimming shaking keyboard mouse course installed docs Network System Administration User Reference Applications NeXT Book Bruce Webster New Copy Black NeXTconnection modem cable HD disks still
Hi everybody following books sale books brand new find book like need information please feel free send E Mail buyers pays shipping fees Thanks abou sun soe clarkson edu TITLE Windows Programming Introduction AUTHOR William H Murray I
article pfqe$bb seven East Sun COM jorge erex East Sun COM Jorge Lach Sun BOS Hardware looking *buy* following items Fax machine plain one need extras basic model Good working order sell stores dirt cheap make offers like cost sell PC
Sony portable Diskman forsale Good condition flawless Costomer AC adapter v DC power supply tested v DC factory adapter tested v DC AC v input time bought three years ago using lot heat generated inside CD machine course use risk baby
Intel Board bit megs ram would like sell ASAP Please email offers interested Thanks Fred
following bike sale type Dave Scott Centurion model size cm c c grouppo Shimano cranks cm pedals Shimano P clips straps frame Tange II Double butted steel gearing front rear seat Terry womens gel seat computer Avocet extras double wat
sale Roland best offer Excellent condition Includes patches disk cakewalk sysex format Buyer must pay COD shipping Please e mail responses gms po cwru edu Thanks George George Scott gscott b student cwru edu gms po cwru edu
...
looking Sharp TI Travelmate parts Mine bad RAM chip motherboard want see get parts sending Sharp repairs one drop line Also trying set one friend needs read old inch diskettes Anyone pinout diskette expansion connector back inch flopp
article Apr erenj com srfergu rufus erenj com Scott Ferguson writes article Apr iscsvax uni edu harter iscsvax uni edu writes Fellow netters anybody awake someone posted message telling people stop posting computer ads misc forsale gr
Hello folks super scope sale comes CRT boxes instructions included shipping included got month back used twice oOO OOo oOO OOo Srikanth Ponnapalli E mail address PoBox Raleigh N C sponna eos ncsu edu Phone ponna aza csc ncsu edu pm sp
EOS elan body mm EF USM lens mm EF USM lens B&W UV filters Hoya circular polarising filter Canon RC remote controller Pentax lens cloth Lowe Pro camera bag Galen Rowell Photoflex lens bag Sapre lithium battery Hove Foto bokk user guid
```

Step2. Stemming.



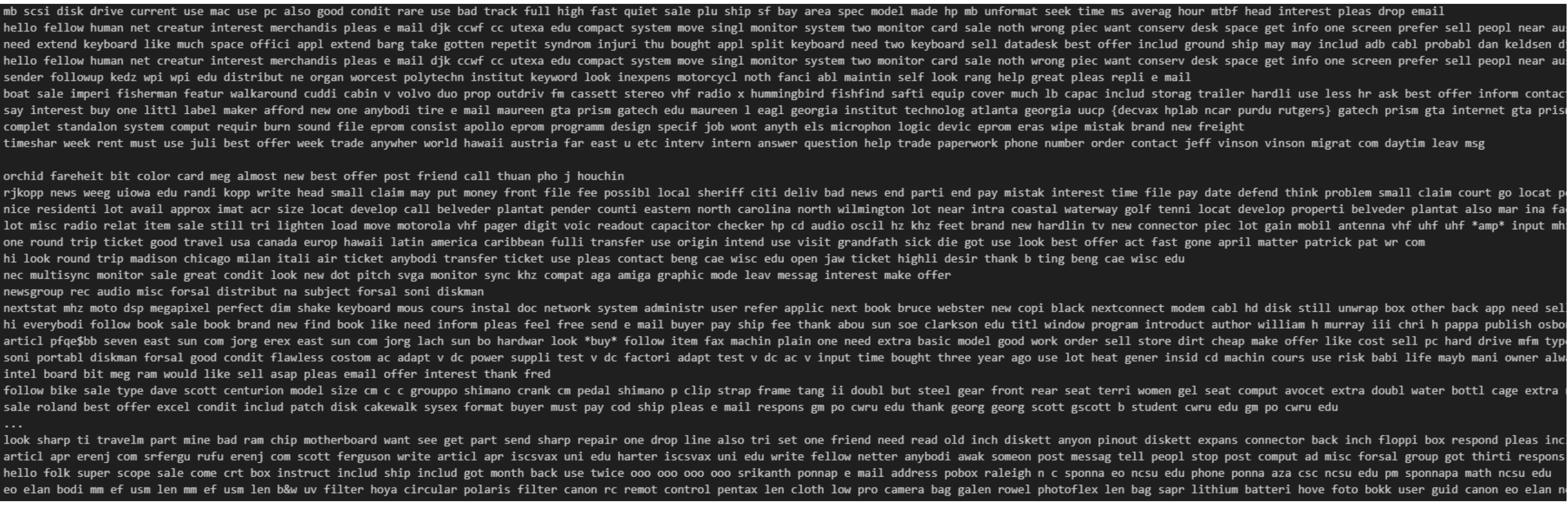
```python
df['text'] = df['text'].apply(lambda x: ' '.join([PorterStemmer().stem(word) for word in x.split()]))
```

✓ 7.6s

What is stemming ?

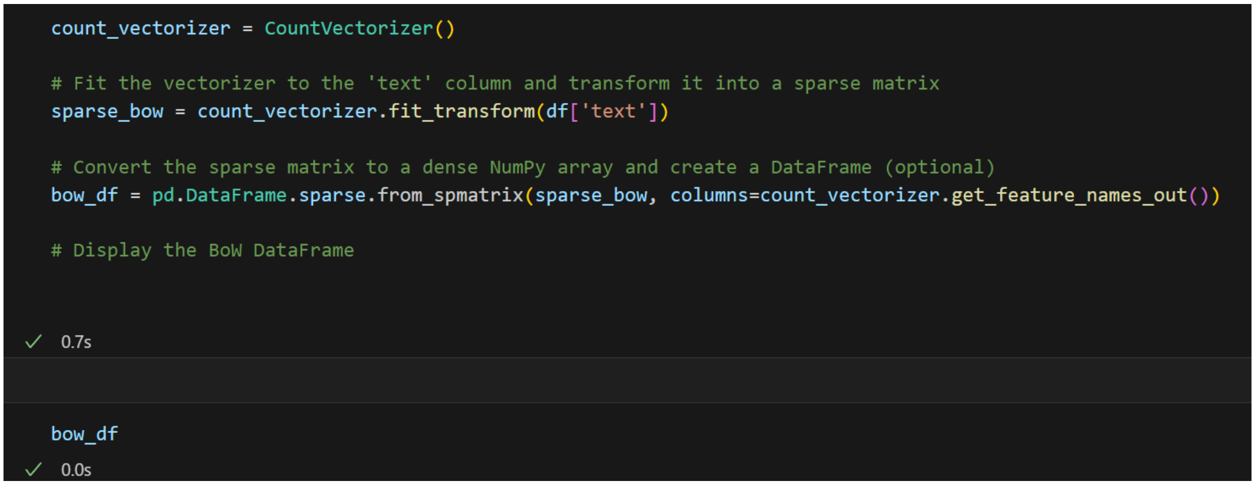
Example : "running quickly in the park," the stemming process might transform it into "run quickli in the park."

We are doing the same on DataFrame text.



```
mb scsi disk drive current use mac use pc also good condit rare use bad track full high fast quiet sale plu ship sf bay area spec model made hp mb unformat seek time ms averag hour mtbf head interest pleas drop email
hello fellow human net creatur interest merchandis pleas e mail djk ccwf cc utexa edu compact system move singl monitor system two monitor card sale noth wrong piec want conserv desk space get info one screen prefer sell peopl near au
need extend keyboard like much space offici appl extend barg take gotten repetit syndrom injuri thu bought appl split keyboard need two keyboard sell datadesk best offer includ ground ship may may includ adb cabl probabl dan keldsen d
hello fellow human net creatur interest merchandis pleas e mail djk ccwf cc utexa edu compact system move singl monitor system two monitor card sale noth wrong piec want conserv desk space get info one screen prefer sell peopl near au
sender followup kedz wpi wpi edu distribut ne organ worcest polytechn institut keyword look inexpens motorcycl noth fanci abl maintin self look rang help great pleas repli e mail
boat sale imperi fisherman featur walkaround cuddi cabin v volvo duo prop outdriv fm cassett stereo vhf radio x hummingbird fishfind safti equip cover much lb capac includ storag trailer hardli use less hr ask best offer inform contac
say interest buy one littl label maker afford new one anybodi tire e mail maureen gta prism gatech edu maureen l eagl georgia institut technolog atlanta georgia uucp {decvax hplab ncar purdu rutgers} gatech prism gta internet gta pris
complet standalon system comput requir burn sound file eprom consist apollo eprom programm design specif job wont anyth els microphon logic devic eprom eras wipe mistak brand new freight
timeshar week rent must use juli best offer week trade anywher world hawaii austria far east u etc interv intern answer question help trade paperwork phone number order contact jeff vinson vinson migrat com daytim leav msg

orchid fareheit bit color card meg almost new best offer post friend call thuan pho j houchin
rjkopp news weeg uiowa edu randi kopp write head small claim may put money front file fee possibl local sheriff citi deliv bad news end parti end pay mistak interest time file pay date defend think problem small claim court go locat p
nice residenti lot avail approx imat acr size locat develop call belveder plantat pender counti eastern north carolina north wilmington lot near intra coastal waterway golf tenni locat develop properti belveder plantat also mar ina fa
lot misc radio relat item sale still tri lighten load move motorola vhf pager digit voic readout capacitor checker hp cd audio oscil hz khz feet brand new hardlin tv new connector piec lot gain mobil antenna vhf uhf uhf *amp* input mh
one round trip ticket good travel usa canada europ hawaii latin america caribbean fulli transfer use origin intend use visit grandfath sick die got use look best offer act fast gone april matter patrick pat wr com
hi look round trip madison chicago milan itali air ticket anybodi transfer ticket use pleas contact beng cae wisc edu open jaw ticket highli desir thank b ting beng cae wisc edu
nec multisync monitor sale great condit look new dot pitch svga monitor sync khz compat aga amiga graphic mode leav messag interest make offer
newsgroup rec audio misc forsal distribut na subject forsal soni diskman
nextstat mhz moto dsp megapixel perfect dim shake keyboard mous cours instal doc network system administr user refer applic next book bruce webster new copi black nextconnect modem cabl hd disk still unwrap box other back app need sel
hi everybodi follow book sale book brand new find book like need inform pleas feel free send e mail buyer pay ship fee thank abou sun soe clarkson edu titl window program introduct author william h murray iii chri h pappa publish osbo
articl pfqe$bb seven east sun com jorg erex east sun com jorg lach sun bo hardwar look *buy* follow item fax machin plain one need extra basic model good work order sell store dirt cheap make offer like cost sell pc hard drive mfm typ
soni portabl diskman forsal good condit flawless costom ac adapt v dc power suppli test v dc factori adapt test v dc ac v input time bought three year ago use lot heat gener insid cd machin cours use risk babi life mayb mani owner alw
intel board bit meg ram would like sell asap pleas email offer interest thank fred
follow bike sale type dave scott centurion model size cm c c grouppo shimano crank cm pedal shimano p clip strap frame tang ii doubl but steel gear front rear seat terri women gel seat comput avocet extra doubl water bottl cage extra
sale roland best offer excel condit includ patch disk cakewalk sysex format buyer must pay cod ship pleas e mail respons gm po cwru edu thank georg georg scott gscott b student cwru edu gm po cwru edu
...
look sharp ti travelm part mine bad ram chip motherboard want see get part send sharp repair one drop line also tri set one friend need read old inch diskett anyon pinout diskett expans connector back inch floppi box respond pleas inc
articl apr erenj com srfergu rufu erenj com scott ferguson write articl apr iscsvax uni edu harter iscsvax uni edu write fellow netter anybodi awak someon post messag tell peopl stop post comput ad misc forsal group got thirti respons
hello folk super scope sale come crt box instruct includ ship includ got month back use twice ooo ooo ooo ooo srikanth ponnap e mail address pobox raleigh n c sponna eo ncsu edu phone ponna aza csc ncsu edu pm sponnapa math ncsu edu
eo elan bodi mm ef usm len mm ef usm len b&w uv filter hoya circular polaris filter canon rc remot control pentax len cloth low pro camera bag galen rowel photoflex len bag sapr lithium batteri hove foto bokk user guid canon eo elan n
```

Step3. BOW(Bag-Of-Words)

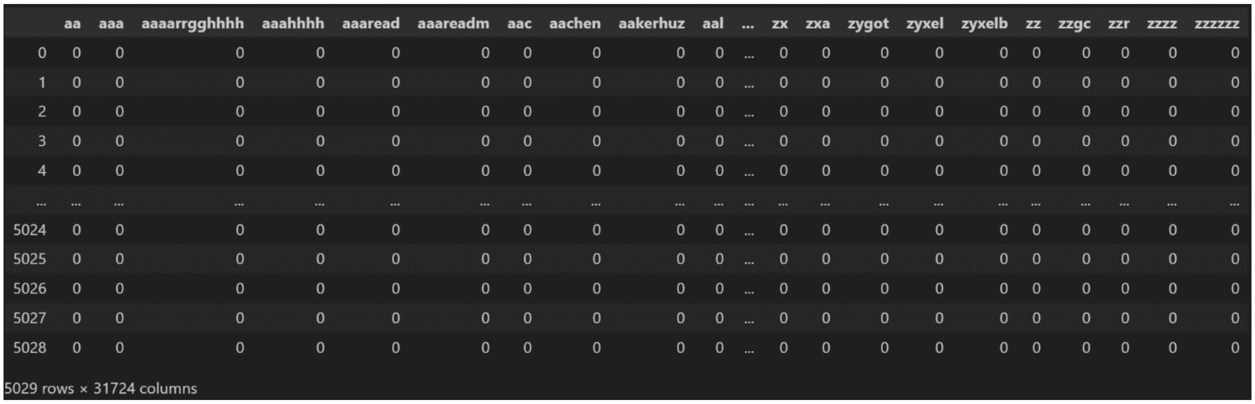


```
count_vectorizer = CountVectorizer()

# Fit the vectorizer to the 'text' column and transform it into a sparse matrix
sparse_bow = count_vectorizer.fit_transform(df['text'])

# Convert the sparse matrix to a dense NumPy array and create a DataFrame (optional)
bow_df = pd.DataFrame.sparse.from_spmatrix(sparse_bow, columns=count_vectorizer.get_feature_names_out())

# Display the BoW DataFrame
```

✓ 0.7s

```
bow_df
```

✓ 0.0s



	aa	aaa	aaaarrgghhhh	aaahhhh	aaaread	aaareadm	aac	aachen	aakerhuz	aal	...	zx	zxa	zygot	zyxel	zyxelb	zz	zzgc	zzr	zzzz	zzzzzz
0	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
5024	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
5025	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
5026	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
5027	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
5028	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0

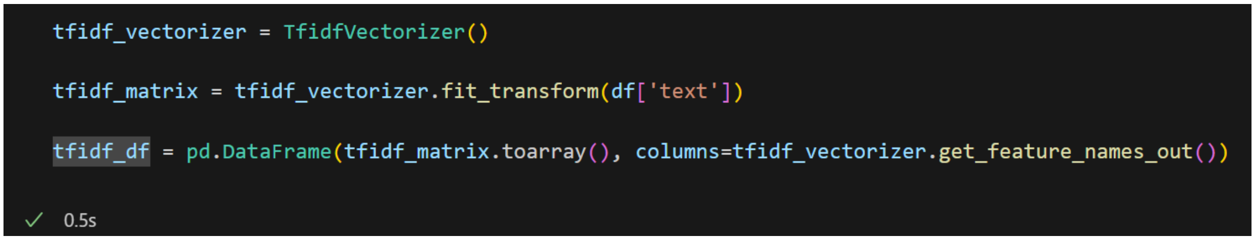
5029 rows × 31724 columns

The values in the cells represent the count of each word in the corresponding document.

Step4. TF-IDF feature space.

(Term Frequency-Inverse Document Frequency).

In the next code we are converting the text column of a DataFrame into a TF-IDF to see how much each word is relevant in our text DataFrame

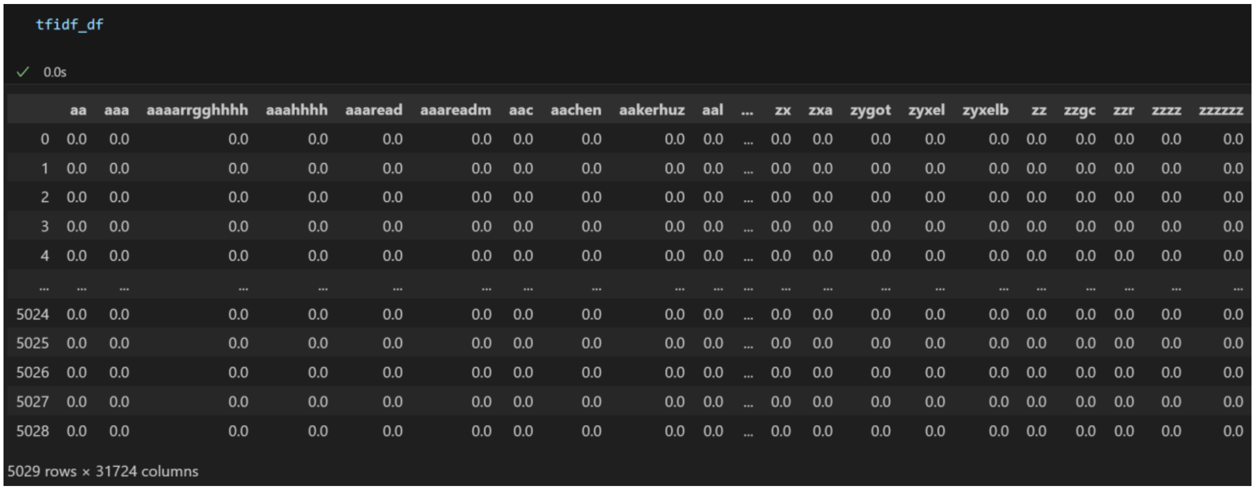


```
tfidf_vectorizer = TfidfVectorizer()

tfidf_matrix = tfidf_vectorizer.fit_transform(df['text'])

tfidf_df = pd.DataFrame(tfidf_matrix.toarray(), columns=tfidf_vectorizer.get_feature_names_out())
```

✓ 0.5s

This code gives us the values of TF-IDF of each word, so basically the higher TF-IDF indicate words that are more important.

So after running our code we will get the following result:



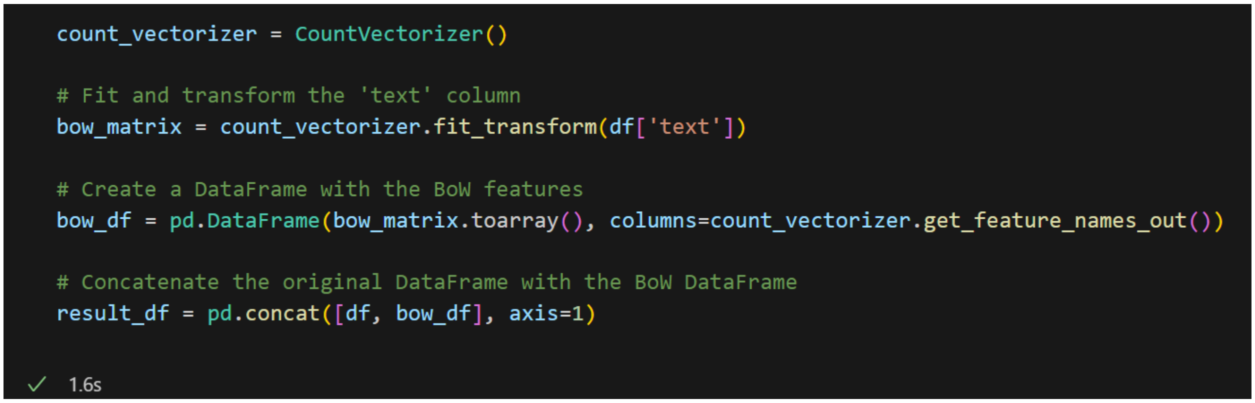
```
tfidf_df
```

✓ 0.0s

	aa	aaa	aaaarrgghhhh	aaahhhh	aaaread	aaareadm	aac	aachen	aakerhuz	aal	...	zx	zxa	zygot	zyxel	zyxelb	zz	zzgc	zzr	zzzz	zzzzzz
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
5024	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
5025	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
5026	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
5027	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
5028	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

5029 rows × 31724 columns

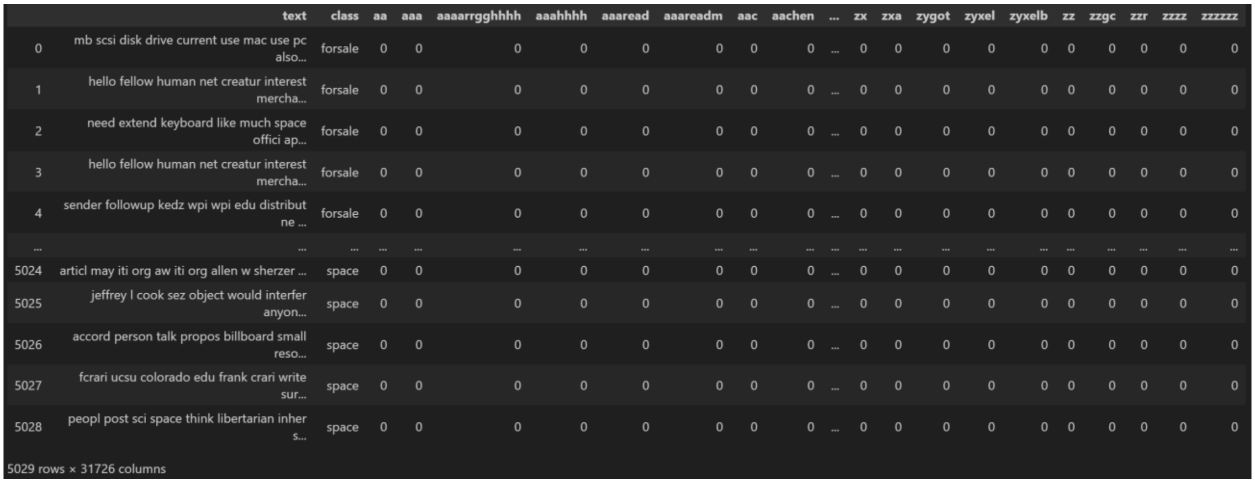
Step5. Feature Vector representation BOW.



```
count_vectorizer = CountVectorizer()

# Fit and transform the 'text' column
bow_matrix = count_vectorizer.fit_transform(df['text'])

# Create a DataFrame with the BoW features
bow_df = pd.DataFrame(bow_matrix.toarray(), columns=count_vectorizer.get_feature_names_out())

# Concatenate the original DataFrame with the BoW DataFrame
result_df = pd.concat([df, bow_df], axis=1)
```

✓ 1.6s

Results after run:



	text	class	aa	aaa	aaaarrgghhhh	aaahhhh	aaaread	aaareadm	aac	aachen	...	zx	zxa	zygot	zyxel	zyxelb	zz	zzgc	zzr	zzzz	zzzzzz
0	mb scsi disk drive current use mac use pc also...	forsale	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
1	hello fellow human net creatur interest mercha...	forsale	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
2	need extend keyboard like much space offici ap...	forsale	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
3	hello fellow human net creatur interest mercha...	forsale	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
4	sender followup kedz wpi wpi edu distribut ne ...	forsale	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
5024	articl may iti org aw iti org allen w sherzer ...	space	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
5025	jeffrey l cook sez object would interfer anyon...	space	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
5026	accord person talk propos billboard small reso...	space	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
5027	fcrari ucsu colorado edu frank crari write sur...	space	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
5028	peopl post sci space think libertarian inher s...	space	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0

5029 rows × 31726 columns