INFORMATION RETRIEVAL

Assignment -1

Anshuman Uniyal (MT 23018)

● os : os is used for importing dataset folders from directory.

● NLTK : The Natural Language Toolkit (nltk) is used for performing tasks such as: Tokenization, Part-of-speech tagging, Sentiment Analysis Stemming and Lemmatization etc.

● BeautifulSoup : It is used for web scraping and parsing HTML and XML documents, providing a convenient and flexible way to extract and manipulate data from web pages.

● pickle : The pickle module in Python is used for serializing and deserializing Python objects.

● re : Provides operations of detecting patterns with the help of regular expressions.

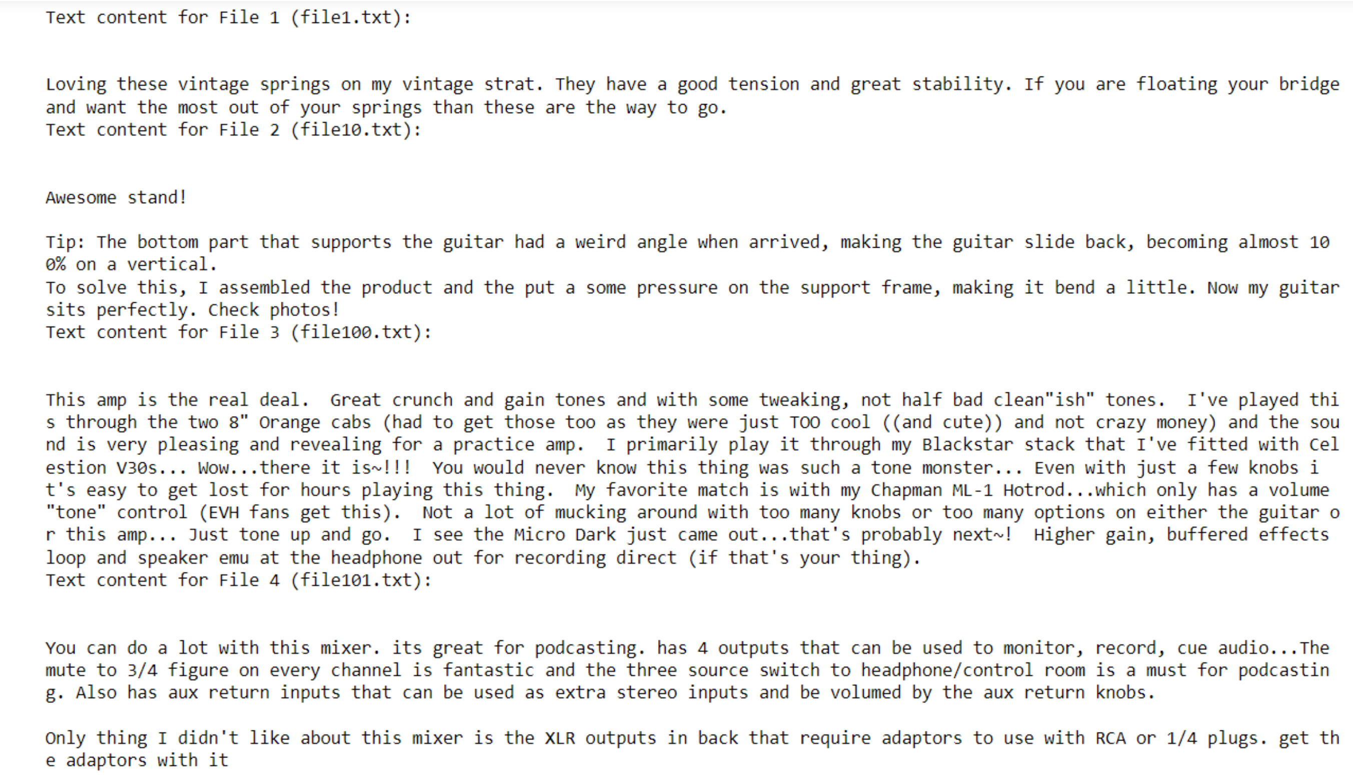
● string : For working with text data present in dataset files.

QUESTION 1: Data Pre-processing

Pre-processing

1. Read data using os.listdir(“location”).
2. Checking the total count of files in the folder. The number of files in the folder are 999.
3. A dictionary is created with docId name.

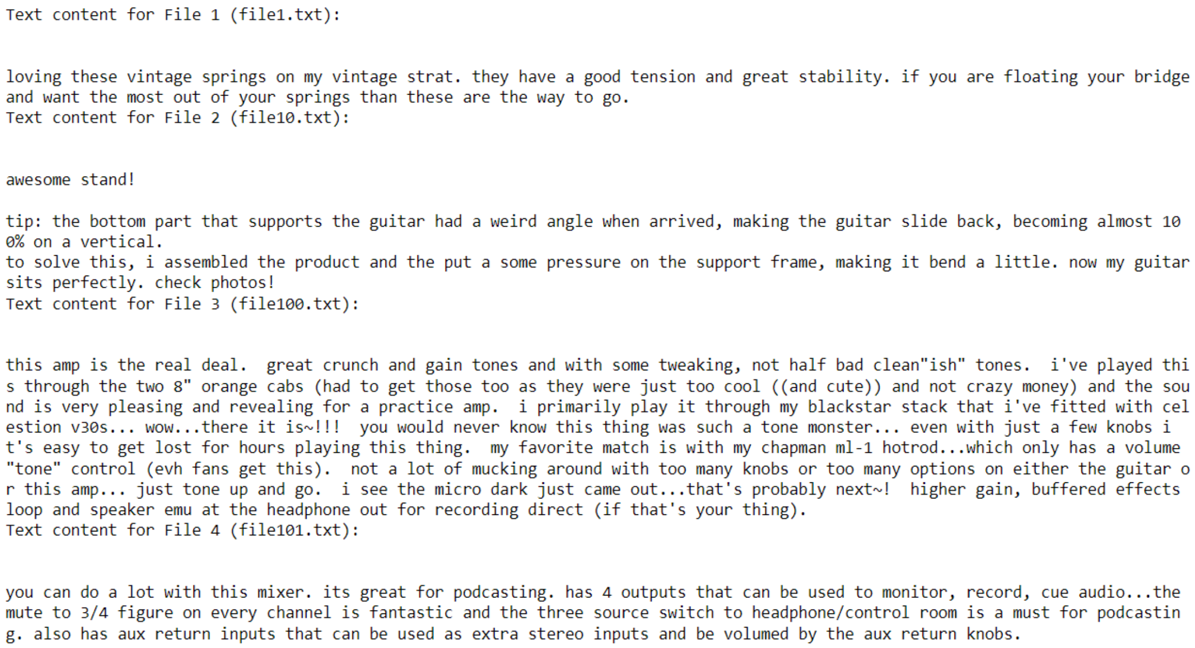
The key and value of the dictionary are taken and a dataframe is created for storing the filenames.



```
Text content for File 1 (file1.txt):


Loving these vintage springs on my vintage strat. They have a good tension and great stability. If you are floating your bridge
and want the most out of your springs than these are the way to go.
Text content for File 2 (file10.txt):


Awesome stand!

Tip: The bottom part that supports the guitar had a weird angle when arrived, making the guitar slide back, becoming almost 10
0% on a vertical.
To solve this, I assembled the product and the put a some pressure on the support frame, making it bend a little. Now my guitar
sits perfectly. Check photos!
Text content for File 3 (file100.txt):


This amp is the real deal.  Great crunch and gain tones and with some tweaking, not half bad clean"ish" tones.  I've played thi
s through the two 8" Orange cabs (had to get those too as they were just TOO cool ((and cute)) and not crazy money) and the sou
nd is very pleasing and revealing for a practice amp.  I primarily play it through my Blackstar stack that I've fitted with Cel
estion V30s... Wow...there it is~!!!  You would never know this thing was such a tone monster... Even with just a few knobs i
t's easy to get lost for hours playing this thing.  My favorite match is with my Chapman ML-1 Hotrod...which only has a volume
"tone" control (EVH fans get this).  Not a lot of mucking around with too many knobs or too many options on either the guitar o
r this amp... Just tone up and go.  I see the Micro Dark just came out...that's probably next~!  Higher gain, buffered effects
loop and speaker emu at the headphone out for recording direct (if that's your thing).
Text content for File 4 (file101.txt):


You can do a lot with this mixer. its great for podcasting. has 4 outputs that can be used to monitor, record, cue audio...The
mute to 3/4 figure on every channel is fantastic and the three source switch to headphone/control room is a must for podcastin
g. Also has aux return inputs that can be used as extra stereo inputs and be volumed by the aux return knobs.

Only thing I didn't like about this mixer is the XLR outputs in back that require adaptors to use with RCA or 1/4 plugs. get th
e adaptors with it
```

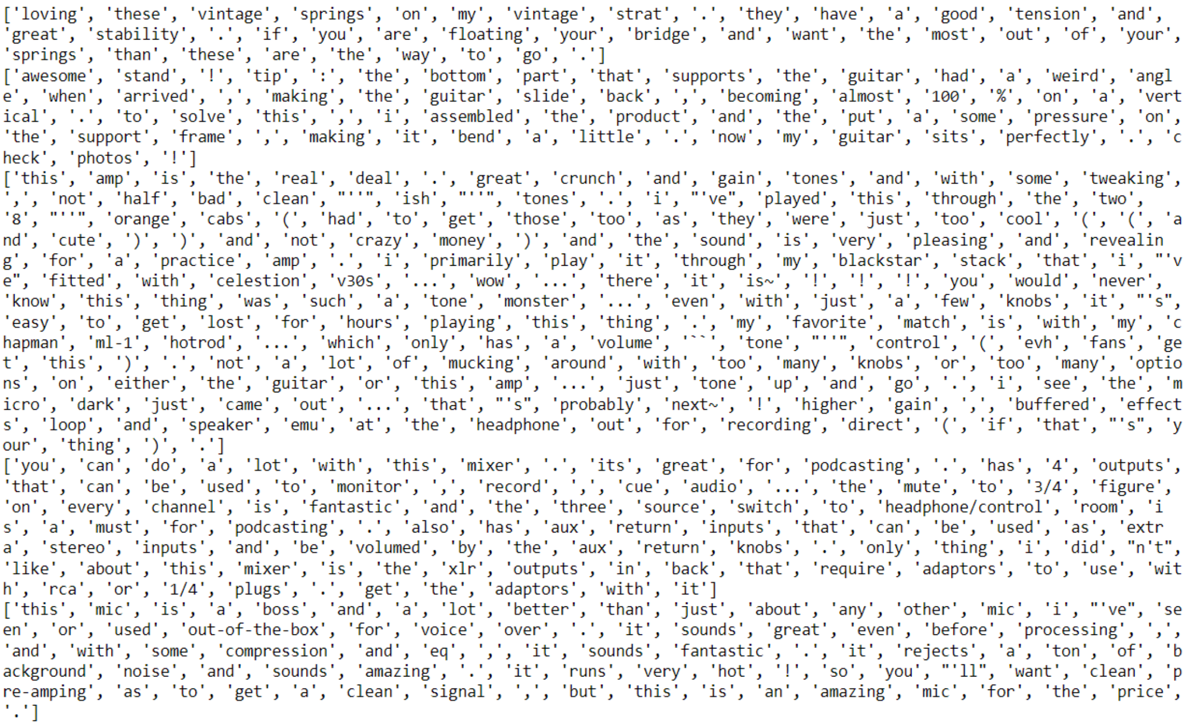
Original Data

After Converting to lowercase



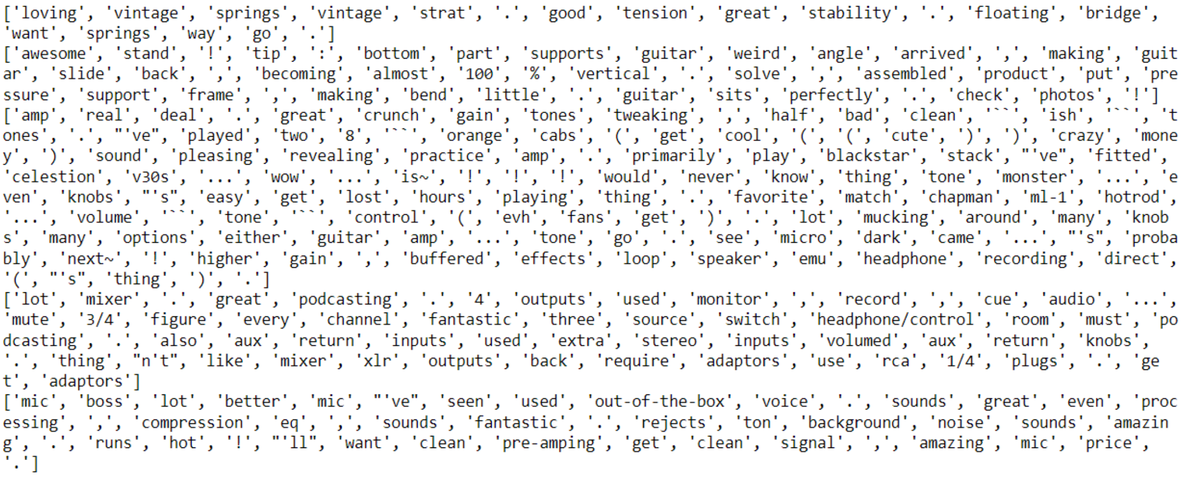
```
Text content for File 1 (file1.txt):


loving these vintage springs on my vintage strat. they have a good tension and great stability. if you are floating your bridge
and want the most out of your springs than these are the way to go.
Text content for File 2 (file10.txt):


awesome stand!

tip: the bottom part that supports the guitar had a weird angle when arrived, making the guitar slide back, becoming almost 10
0% on a vertical.
to solve this, i assembled the product and the put a some pressure on the support frame, making it bend a little. now my guitar
sits perfectly. check photos!
Text content for File 3 (file100.txt):


this amp is the real deal.  great crunch and gain tones and with some tweaking, not half bad clean"ish" tones.  i've played thi
s through the two 8" orange cabs (had to get those too as they were just too cool ((and cute)) and not crazy money) and the sou
nd is very pleasing and revealing for a practice amp.  i primarily play it through my blackstar stack that i've fitted with cel
estion v30s... wow...there it is~!!!  you would never know this thing was such a tone monster... even with just a few knobs i
t's easy to get lost for hours playing this thing.  my favorite match is with my chapman ml-1 hotrod...which only has a volume
"tone" control (evh fans get this).  not a lot of mucking around with too many knobs or too many options on either the guitar o
r this amp... just tone up and go.  i see the micro dark just came out...that's probably next~!  higher gain, buffered effects
loop and speaker emu at the headphone out for recording direct (if that's your thing).
Text content for File 4 (file101.txt):


you can do a lot with this mixer. its great for podcasting. has 4 outputs that can be used to monitor, record, cue audio...the
mute to 3/4 figure on every channel is fantastic and the three source switch to headphone/control room is a must for podcastin
g. also has aux return inputs that can be used as extra stereo inputs and be volumed by the aux return knobs.
```

After Tokenization



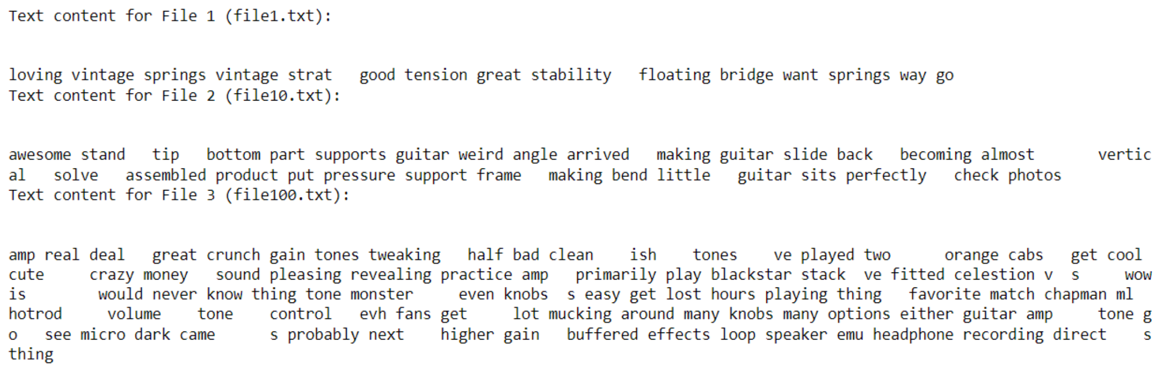
```
['loving', 'these', 'vintage', 'springs', 'on', 'my', 'vintage', 'strat', '.', 'they', 'have', 'a', 'good', 'tension', 'and',
'great', 'stability', '.', 'if', 'you', 'are', 'floating', 'your', 'bridge', 'and', 'want', 'the', 'most', 'out', 'of', 'your',
'springs', 'than', 'these', 'are', 'the', 'way', 'to', 'go', '.']
['awesome', 'stand', '!', 'tip', ':', 'the', 'bottom', 'part', 'that', 'supports', 'the', 'guitar', 'had', 'a', 'weird', 'angl
e', 'when', 'arrived', ',', 'making', 'the', 'guitar', 'slide', 'back', ',', 'becoming', 'almost', '100', '%', 'on', 'a', 'vert
ical', '.', 'to', 'solve', 'this', ',', 'i', 'assembled', 'the', 'product', 'and', 'the', 'put', 'a', 'some', 'pressure', 'on',
'the', 'support', 'frame', ',', 'making', 'it', 'bend', 'a', 'little', '.', 'now', 'my', 'guitar', 'sits', 'perfectly', '.', 'c
heck', 'photos', '!']
['this', 'amp', 'is', 'the', 'real', 'deal', '.', 'great', 'crunch', 'and', 'gain', 'tones', 'and', 'with', 'some', 'tweaking',
',', 'not', 'half', 'bad', 'clean', "''", 'ish', "''", 'tones', '.', 'i', "'ve", 'played', 'this', 'through', 'the', 'two',
'8', "''", 'orange', 'cabs', '(', 'had', 'to', 'get', 'those', 'too', 'as', 'they', 'were', 'just', 'too', 'cool', '(', '(', 'a
nd', 'cute', ')', ')', 'and', 'not', 'crazy', 'money', ')', 'and', 'the', 'sound', 'is', 'very', 'pleasing', 'and', 'revealin
g', 'for', 'a', 'practice', 'amp', '.', 'i', 'primarily', 'play', 'it', 'through', 'my', 'blackstar', 'stack', 'that', 'i', "'v
e", 'fitted', 'with', 'celestion', 'v30s', '...', 'wow', '...', 'there', 'it', 'is~', '!', '!', '!', 'you', 'would', 'never',
'know', 'this', 'thing', 'was', 'such', 'a', 'tone', 'monster', '...', 'even', 'with', 'just', 'a', 'few', 'knobs', 'it', "'s",
'easy', 'to', 'get', 'lost', 'for', 'hours', 'playing', 'this', 'thing', '.', 'my', 'favorite', 'match', 'is', 'with', 'my', 'c
hapman', 'ml-1', 'hotrod', '...', 'which', 'only', 'has', 'a', 'volume', '``', 'tone', "''", 'control', '(', 'evh', 'fans', 'ge
t', 'this', ')', '.', 'not', 'a', 'lot', 'of', 'mucking', 'around', 'with', 'too', 'many', 'knobs', 'or', 'too', 'many', 'optio
ns', 'on', 'either', 'the', 'guitar', 'or', 'this', 'amp', '...', 'just', 'tone', 'up', 'and', 'go', '.', 'i', 'see', 'the', 'm
icro', 'dark', 'just', 'came', 'out', '...', 'that', "'s", 'probably', 'next~', '!', 'higher', 'gain', ',', 'buffered', 'effect
s', 'loop', 'and', 'speaker', 'emu', 'at', 'the', 'headphone', 'out', 'for', 'recording', 'direct', '(', 'if', 'that', "'s", 'y
our', 'thing', ')', '.']
['you', 'can', 'do', 'a', 'lot', 'with', 'this', 'mixer', '.', 'its', 'great', 'for', 'podcasting', '.', 'has', '4', 'outputs',
'that', 'can', 'be', 'used', 'to', 'monitor', ',', 'record', ',', 'cue', 'audio', '...', 'the', 'mute', 'to', '3/4', 'figure',
'on', 'every', 'channel', 'is', 'fantastic', 'and', 'the', 'three', 'source', 'switch', 'to', 'headphone/control', 'room', 'i
s', 'a', 'must', 'for', 'podcasting', '.', 'also', 'has', 'aux', 'return', 'inputs', 'that', 'can', 'be', 'used', 'as', 'extr
a', 'stereo', 'inputs', 'and', 'be', 'volumed', 'by', 'the', 'aux', 'return', 'knobs', '.', 'only', 'thing', 'i', 'did', "n't",
'like', 'about', 'this', 'mixer', 'is', 'the', 'xlr', 'outputs', 'in', 'back', 'that', 'require', 'adaptors', 'to', 'use', 'wit
h', 'rca', 'or', '1/4', 'plugs', '.', 'get', 'the', 'adaptors', 'with', 'it']
['this', 'mic', 'is', 'a', 'boss', 'and', 'a', 'lot', 'better', 'than', 'just', 'about', 'any', 'other', 'mic', 'i', "'ve", 'se
en', 'or', 'used', 'out-of-the-box', 'for', 'voice', 'over', '.', 'it', 'sounds', 'great', 'even', 'before', 'processing', ',',
'and', 'with', 'some', 'compression', 'and', 'eq', ',', 'it', 'sounds', 'fantastic', '.', 'it', 'rejects', 'a', 'ton', 'of', 'b
ackground', 'noise', 'and', 'sounds', 'amazing', '.', 'it', 'runs', 'very', 'hot', '!', 'so', 'you', "'ll", 'want', 'clean', 'p
re-amping', 'as', 'to', 'get', 'a', 'clean', 'signal', ',', 'but', 'this', 'is', 'an', 'amazing', 'mic', 'for', 'the', 'price',
'.']
```

After Removing stop words



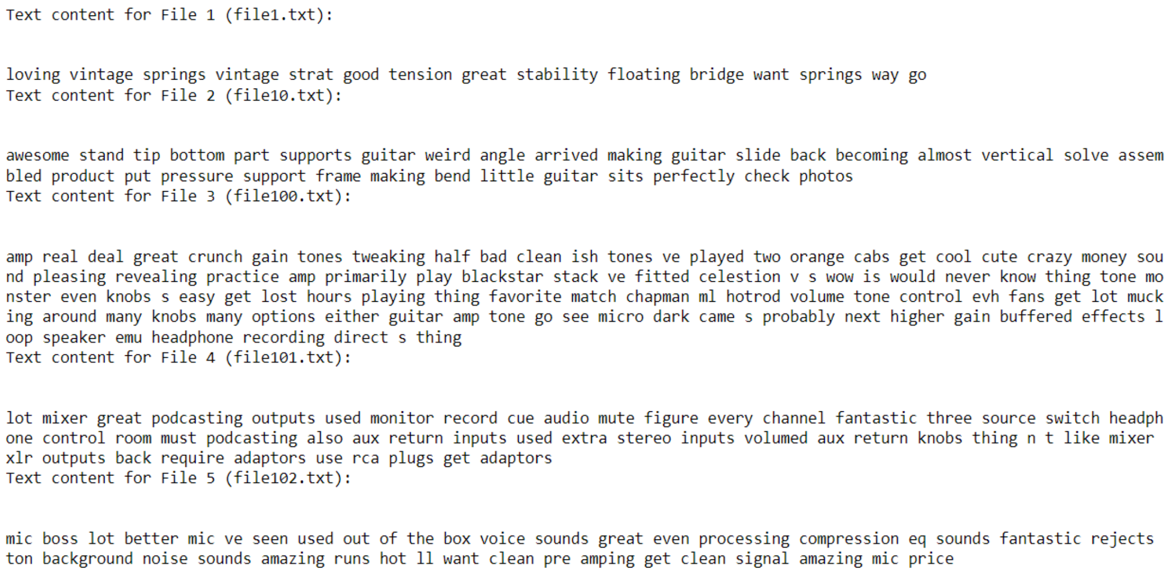
```
['loving', 'vintage', 'springs', 'vintage', 'strat', '.', 'good', 'tension', 'great', 'stability', '.', 'floating', 'bridge',
'want', 'springs', 'way', 'go', '.']
['awesome', 'stand', '!', 'tip', ':', 'bottom', 'part', 'supports', 'guitar', 'weird', 'angle', 'arrived', ',', 'making', 'guit
ar', 'slide', 'back', ',', 'becoming', 'almost', '100', '%', 'vertical', '.', 'solve', ',', 'assembled', 'product', 'put', 'pre
ssure', 'support', 'frame', ',', 'making', 'bend', 'little', '.', 'guitar', 'sits', 'perfectly', '.', 'check', 'photos', '!']
['amp', 'real', 'deal', '.', 'great', 'crunch', 'gain', 'tones', 'tweaking', ',', 'half', 'bad', 'clean', '``', 'ish', '``', 't
ones', '.', "'ve", 'played', 'two', '8', '``', 'orange', 'cabs', '(', 'get', 'cool', '(', '(', 'cute', ')', ')', 'crazy', 'mone
y', ')', 'sound', 'pleasing', 'revealing', 'practice', 'amp', '.', 'primarily', 'play', 'blackstar', 'stack', "'ve", 'fitted',
'celestion', 'v30s', '...', 'wow', '...', 'is~', '!', '!', '!', 'would', 'never', 'know', 'thing', 'tone', 'monster', '...', 'e
ven', 'knobs', "'s", 'easy', 'get', 'lost', 'hours', 'playing', 'thing', '.', 'favorite', 'match', 'chapman', 'ml-1', 'hotrod',
'...', 'volume', '``', 'tone', '``', 'control', '(', 'evh', 'fans', 'get', ')', '.', 'lot', 'mucking', 'around', 'many', 'knob
s', 'many', 'options', 'either', 'guitar', 'amp', '...', 'tone', 'go', '.', 'see', 'micro', 'dark', 'came', '...', "'s", 'proba
bly', 'next~', '!', 'higher', 'gain', ',', 'buffered', 'effects', 'loop', 'speaker', 'emu', 'headphone', 'recording', 'direct',
'(', "'s", 'thing', ')', '.']
['lot', 'mixer', '.', 'great', 'podcasting', '.', '4', 'outputs', 'used', 'monitor', ',', 'record', ',', 'cue', 'audio', '...',
'mute', '3/4', 'figure', 'every', 'channel', 'fantastic', 'three', 'source', 'switch', 'headphone/control', 'room', 'must', 'po
dcasting', '.', 'also', 'aux', 'return', 'inputs', 'used', 'extra', 'stereo', 'inputs', 'volumed', 'aux', 'return', 'knobs',
'.', 'thing', "n't", 'like', 'mixer', 'xlr', 'outputs', 'back', 'require', 'adaptors', 'use', 'rca', '1/4', 'plugs', '.', 'ge
t', 'adaptors']
['mic', 'boss', 'lot', 'better', 'mic', "'ve", 'seen', 'used', 'out-of-the-box', 'voice', '.', 'sounds', 'great', 'even', 'proc
essing', ',', 'compression', 'eq', ',', 'sounds', 'fantastic', '.', 'rejects', 'ton', 'background', 'noise', 'sounds', 'amazin
g', '.', 'runs', 'hot', '!', "'ll", 'want', 'clean', 'pre-amping', 'get', 'clean', 'signal', ',', 'amazing', 'mic', 'price',
'.']
```

After removing Punctuation



```
Text content for File 1 (file1.txt):


loving vintage springs vintage strat   good tension great stability   floating bridge want springs way go
Text content for File 2 (file10.txt):


awesome stand   tip   bottom part supports guitar weird angle arrived   making guitar slide back   becoming almost       vertic
al   solve   assembled product put pressure support frame   making bend little   guitar sits perfectly   check photos
Text content for File 3 (file100.txt):


amp real deal   great crunch gain tones tweaking   half bad clean    ish    tones    ve played two     orange cabs   get cool
cute     crazy money   sound pleasing revealing practice amp   primarily play blackstar stack  ve fitted celestion v  s     wow
is        would never know thing tone monster     even knobs  s easy get lost hours playing thing   favorite match chapman ml
hotrod     volume    tone    control   evh fans get     lot mucking around many knobs many options either guitar amp     tone g
o   see micro dark came      s probably next    higher gain   buffered effects loop speaker emu headphone recording direct    s
thing
```

After Removal of Blank space



```
Text content for File 1 (file1.txt):


loving vintage springs vintage strat good tension great stability floating bridge want springs way go
Text content for File 2 (file10.txt):


awesome stand tip bottom part supports guitar weird angle arrived making guitar slide back becoming almost vertical solve assem
bled product put pressure support frame making bend little guitar sits perfectly check photos
Text content for File 3 (file100.txt):


amp real deal great crunch gain tones tweaking half bad clean ish tones ve played two orange cabs get cool cute crazy money sou
nd pleasing revealing practice amp primarily play blackstar stack ve fitted celestion v s wow is would never know thing tone mo
nster even knobs s easy get lost hours playing thing favorite match chapman ml hotrod volume tone control evh fans get lot muck
ing around many knobs many options either guitar amp tone go see micro dark came s probably next higher gain buffered effects l
oop speaker emu headphone recording direct s thing
Text content for File 4 (file101.txt):


lot mixer great podcasting outputs used monitor record cue audio mute figure every channel fantastic three source switch headph
one control room must podcasting also aux return inputs used extra stereo inputs volumed aux return knobs thing n t like mixer
xlr outputs back require adaptors use rca plugs get adaptors
Text content for File 5 (file102.txt):


mic boss lot better mic ve seen used out of the box voice sounds great even processing compression eq sounds fantastic rejects
ton background noise sounds amazing runs hot ll want clean pre amping get clean signal amazing mic price
```

QUESTION 2:

Inverted Index Creation:

An empty dictionary index is initialized to store the inverted index.

The code processes each file in the specified directory (list1), reads its content, and tokenizes each word using word_tokenize. For each token in the file, it adds the corresponding filename to the postings list in the index dictionary.

The result is a dictionary where each term (token) maps to a list of filenames where that term appears.

Function Definitions for OR,AND,NOT,AND NOT and OR NOT

AND Operator:

When encountering an "AND" operator, it takes the documents that contain the first word (word1) and the documents that contain the second word (word2).

Then, it calculates the intersection of these two sets using the merge_lists function with 'AND' operation. The result (fileNames) will be the documents that contain both words.

OR Operator:

For "OR" operations, it takes the documents containing the first word (word1) and the documents containing the second word (word2).

It calculates the union of these sets using the merge_lists function with 'OR' operation. The result (fileNames) will be the documents that contain either of the words.

AND NOT Operator:

When encountering "AND NOT," it considers the documents containing the first word (word1) and subtracts the documents containing the second word (word2).

It calculates the set difference using the merge_lists function with 'AND NOT' operation. The result (fileNames) will be the documents that contain the first word but not the second.

OR NOT Operator:

For "OR NOT" operations, it takes the documents containing the first word (word1) and adds the documents containing the second word (word2) but not already in word1.

It calculates the set union with set difference using the merge_lists function with 'OR NOT' operation. The result (fileNames) will be the documents that contain either of the words but not both.

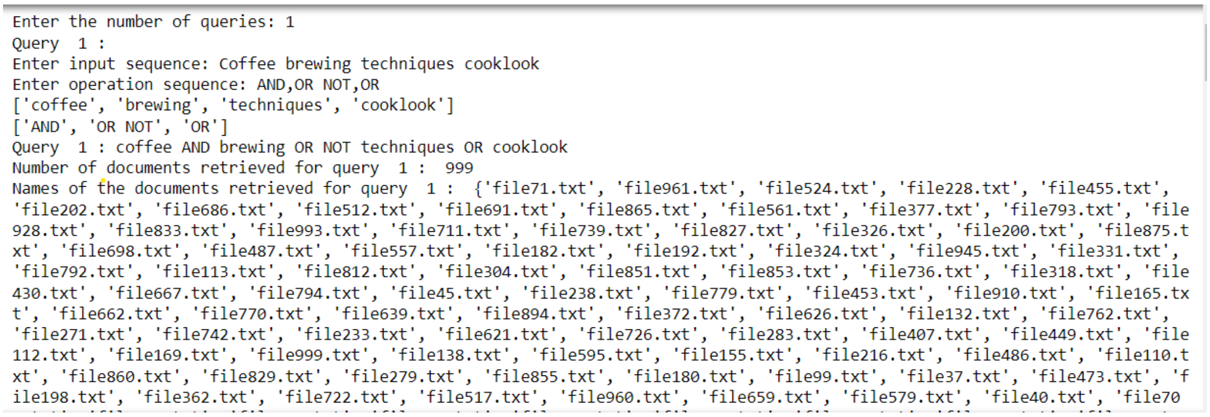
Displaying Results:

The function then prints the processed query and displays the number of retrieved documents and their names based on the calculated fileNames.

User Input Loop:

The code prompts the user to input the number of queries and the queries with corresponding operators.

It processes each query using the queryProcess function and displays the results.



```
Enter the number of queries: 1
Query  1 :
Enter input sequence: Coffee brewing techniques cooklook
Enter operation sequence: AND,OR NOT,OR
['coffee', 'brewing', 'techniques', 'cooklook']
['AND', 'OR NOT', 'OR']
Query  1 : coffee AND brewing OR NOT techniques OR cooklook
Number of documents retrieved for query  1 :  999
Names of the documents retrieved for query  1 :  {'file71.txt', 'file961.txt', 'file524.txt', 'file228.txt', 'file455.txt',
'file202.txt', 'file686.txt', 'file512.txt', 'file691.txt', 'file865.txt', 'file561.txt', 'file377.txt', 'file793.txt', 'file
928.txt', 'file833.txt', 'file993.txt', 'file711.txt', 'file739.txt', 'file827.txt', 'file326.txt', 'file200.txt', 'file875.t
xt', 'file698.txt', 'file487.txt', 'file557.txt', 'file182.txt', 'file192.txt', 'file324.txt', 'file945.txt', 'file331.txt',
'file792.txt', 'file113.txt', 'file812.txt', 'file304.txt', 'file851.txt', 'file853.txt', 'file736.txt', 'file318.txt', 'file
430.txt', 'file667.txt', 'file794.txt', 'file45.txt', 'file238.txt', 'file779.txt', 'file453.txt', 'file910.txt', 'file165.tx
t', 'file662.txt', 'file770.txt', 'file639.txt', 'file894.txt', 'file372.txt', 'file626.txt', 'file132.txt', 'file762.txt',
'file271.txt', 'file742.txt', 'file233.txt', 'file621.txt', 'file726.txt', 'file283.txt', 'file407.txt', 'file449.txt', 'file
112.txt', 'file169.txt', 'file999.txt', 'file138.txt', 'file595.txt', 'file155.txt', 'file216.txt', 'file486.txt', 'file110.t
xt', 'file860.txt', 'file829.txt', 'file279.txt', 'file855.txt', 'file180.txt', 'file99.txt', 'file37.txt', 'file473.txt', 'f
ile198.txt', 'file362.txt', 'file722.txt', 'file517.txt', 'file960.txt', 'file659.txt', 'file579.txt', 'file40.txt', 'file70
```

Fig- Unigram Inverted Index and Boolean Queries

QUESTION 3

Document Structure and Sample Data:

The code defines a data structure (data_structure) for a sample postings list, which stores information about the frequency and positions of words in documents.

Building Positional Index:

1. It initializes an empty positional_index dictionary.

2. Iterates through each document in a given list (file_list).

3. Reads the document, identifies unique words, and calculates the number of occurrences and positions of each unique word.

4. Populates the positional_index with information about word occurrences and positions in each document.

Preprocessing Input Queries:

1. Defines a function (preprocess_input_query) to preprocess user input queries.

2. Removes stop words, converts to lowercase, and tokenizes the input query.

Executing Phrase Queries:

1.Asks the user to input the number of phrase queries (n) to execute.

2.Enters a loop where the user inputs phrase queries, and the code processes them.

3.For each input query, it retrieves relevant documents using the built positional index.

4.Displays the number and names of documents retrieved for each query.

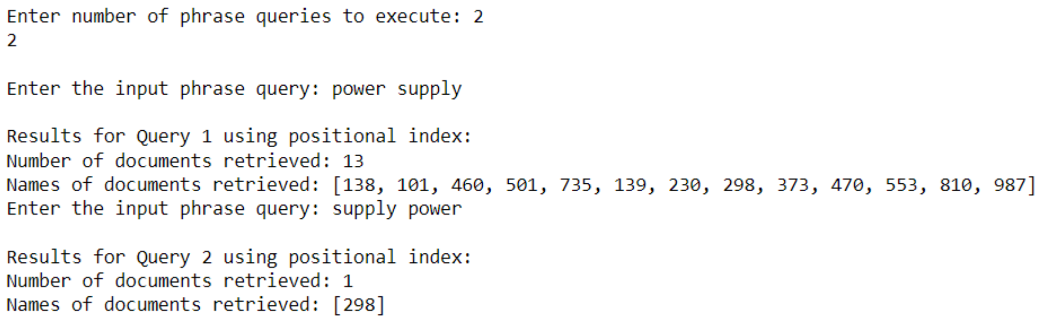
Common Documents and Resultant List:

1. The code finds common documents where all query tokens appear.

2. Iterates through the common documents and calculates a resultant list of positions based on the positions of query tokens in each document.

Displaying Results:

1.Displays the results for each query, including the number of retrieved documents and their names using the positional index.



```
Enter number of phrase queries to execute: 2
2

Enter the input phrase query: power supply

Results for Query 1 using positional index:
Number of documents retrieved: 13
Names of documents retrieved: [138, 101, 460, 501, 735, 139, 230, 298, 373, 470, 553, 810, 987]
Enter the input phrase query: supply power

Results for Query 2 using positional index:
Number of documents retrieved: 1
Names of documents retrieved: [298]
```

Fig-Positional index of phase query