Q1.a
Lowercase the text
Used nltk and converted using: `content_lower = content.lower()`
B
Then to perform tokenisation we have used : wordpunct_tokenize and then used :wordpunct_tokenize(content_lower) to save them as tokens
c
Then, removed stop words by first downloading the stop words and then removing the words from the list that are stopwords
d
Then for punctuation, we have used RegexpTokenizer(r'\w+') and tokenized the list
e
To remove spaces, Spacetokenizer was used.

For Question 2a,b and 3a,b:
For 2a,b:
i have iterated through all the folders and for each word used a dictionary ,first added the word to the dictionary and its file name and word existed just file name was added .

For 3a,b:
Similar to above just a extra index was being stored .

For pickle of both the unigram
Save

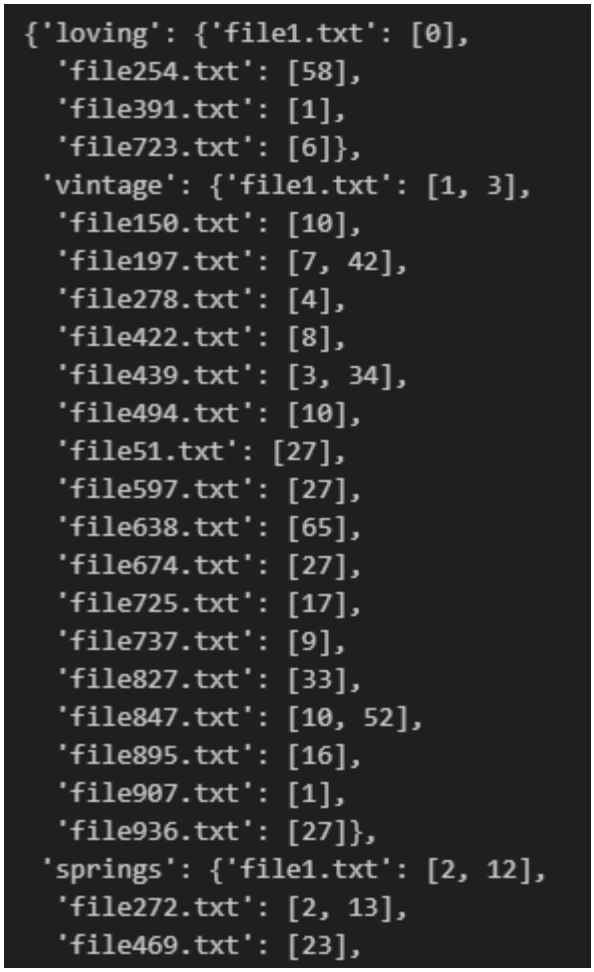
```
import pickle
with open('invd.pkl', 'wb') as file:
    pickle.dump(invd, file)
import pickle
with open('invd2.pkl', 'wb') as file:
    pickle.dump(invd2, file)
```

Load

```
with open('invd.pkl', 'rb') as file:
    invdl = pickle.load(file)
with open('invd2.pkl', 'rb') as file:
    invdl2 = pickle.load(file)
```


{'loving': {'file1.txt', 'file254.txt', 'file391.txt', 'file723.txt'},
'vintage': {'file1.txt',
'file150.txt',
'file197.txt',
'file278.txt',
'file422.txt',
'file439.txt',
'file494.txt',
'file51.txt',
'file597.txt',
'file638.txt',
'file674.txt',
'file725.txt',
'file737.txt',
'file827.txt',
'file847.txt',
'file895.txt',
'file907.txt',
'file936.txt'},
'springs': {'file1.txt',
'file272.txt',
'file469.txt',
'file806.txt',
'file937.txt'},
'strat': {'file1.txt',
...
'file746.txt',
'file880.txt',
'file883.txt',
'file953.txt'},
...}

Q2a


{'loving': {'file1.txt': [0],
'file254.txt': [58],
'file391.txt': [1],
'file723.txt': [6]},
'vintage': {'file1.txt': [1, 3],
'file150.txt': [10],
'file197.txt': [7, 42],
'file278.txt': [4],
'file422.txt': [8],
'file439.txt': [3, 34],
'file494.txt': [10],
'file51.txt': [27],
'file597.txt': [27],
'file638.txt': [65],
'file674.txt': [27],
'file725.txt': [17],
'file737.txt': [9],
'file827.txt': [33],
'file847.txt': [10, 52],
'file895.txt': [16],
'file907.txt': [1],
'file936.txt': [27]},
'springs': {'file1.txt': [2, 12],
'file272.txt': [2, 13],
'file469.txt': [23],

Q3a