Assumptions :

1.Used nltk inbuilt functions for tokenization, removing stop words, case folding, lemmatization and stemming.

2.If user gives a query for searching then query is also pre-processed.

3.While making unigram inverted index each file is pre-processed and there is duplicity then we remove it using set and sorting is done.

4.While removing the punctuations the letters are separated. For example u.s.a. will become u s a

unigram inverted index

a).Pre-Processing :

- Tokenization : Longer strings are splits into tokens. In this the text is tokenized into sentences and then sentences are tokenized into words. During tokenization the punctuations are also removed.
- Removing Stop-Words : The words which contribute very little to overall meaning of documents then the words are removed. This words are generally most common words in language.
- Case-Folding(Lower Case) : Converting all words to lower case to make searching easy.
- Lemmatization: It is able to capture the canonical forms of the word based on the lemma.
- Stemming : It is process of eliminating affixes from the word to get the stem word. Generally it makes word running to run.

b).Methodology :

- First step is to load the files from directory and reading the files using csopen() function.
- Once the file is read the all pre-processing is done like tokenization, removing stop words and lemmatization etc.
- After file is read then all the words are stored in dictionary whose keys are words and value is a posting list which contains the document id of that word.
- If word is new then it is added to dictionary and if word is already exist then document is added to the posting list for that word. Repeat till all files are read.
- The obtain dictionary is sorted on the basis of keys and corresponding posting list are also sorted.

c).Test-Cases :

Unigram inverted index for word “fish” and output is document containing “fish”

```
print(Dictionary_Posting['fish'])
```

```
['alt.atheism49960', 'alt.atheism51200', 'alt.atheism51309', 'alt.atheism52909', 'alt.atheism53546', 'alt.atheism53547', 'alt.
atheism54163', 'alt.atheism54258', 'alt.atheism54260', 'comp.graphics38248', 'comp.graphics38377', 'comp.graphics38398', 'com
p.graphics38851', 'comp.graphics38993', 'comp.graphics40027', 'comp.os.ms-windows.misc9632', 'comp.os.ms-windows.misc9993', 'c
omp.sys.ibm.pc.hardware58992', 'comp.sys.ibm.pc.hardware60153', 'comp.sys.ibm.pc.hardware60520', 'comp.sys.mac.hardware52305',
'comp.sys.mac.hardware52306', 'comp.windows.x67231', 'comp.windows.x67392', 'comp.windows.x67393', 'comp.windows.x67882', 'mis
c.forsale76114', 'misc.forsale76196', 'misc.forsale76227', 'misc.forsale76323', 'misc.forsale76423', 'misc.forsale76944', 're
c.motorcycles103129', 'rec.motorcycles103161', 'rec.motorcycles104433', 'rec.sport.hockey52566', 'rec.sport.hockey52630', 're
c.sport.hockey53612', 'rec.sport.hockey53824', 'rec.sport.hockey53915', 'rec.sport.hockey53963', 'rec.sport.hockey54056', 're
c.sport.hockey54074', 'rec.sport.hockey54124', 'rec.sport.hockey54168', 'rec.sport.hockey54173', 'rec.sport.hockey54181', 're
c.sport.hockey54308', 'sci.crypt14992', 'sci.crypt15329', 'sci.crypt15373', 'sci.crypt15425', 'sci.crypt15435', 'sci.crypt1545
1', 'sci.crypt15534', 'sci.crypt15603', 'sci.crypt15666', 'sci.crypt15668', 'sci.crypt15669', 'sci.crypt15712', 'sci.crypt1571
3', 'sci.crypt15722', 'sci.crypt16065', 'sci.crypt16077', 'sci.crypt16123', 'sci.crypt16142', 'sci.electronics53569', 'sci.ele
ctronics53573', 'sci.electronics53687', 'sci.med59021', 'sci.med59143', 'sci.med59189', 'sci.med59234', 'sci.med59255', 'sci.m
ed59256', 'sci.med59284', 'sci.space60242', 'sci.space61011', 'sci.space61086', 'sci.space61189', 'sci.space61373', 'sci.space
61432', 'soc.religion.christian20639', 'soc.religion.christian20703', 'talk.politics.guns54314', 'talk.politics.guns54430', 't
alk.politics.guns54487', 'talk.politics.guns55050', 'talk.politics.guns55113', 'talk.politics.mideast75958', 'talk.politics.mi
deast76021', 'talk.politics.mideast76276', 'talk.politics.misc178345', 'talk.politics.misc178420', 'talk.politics.misc178620',
'talk.politics.misc178623', 'talk.politics.misc179053', 'talk.politics.misc179108', 'talk.religion.misc82800', 'talk.religion.
misc83441', 'talk.religion.misc83444', 'talk.religion.misc83451', 'talk.religion.misc83454', 'talk.religion.misc83457', 'talk.
religion.misc84007', 'talk.religion.misc84071', 'talk.religion.misc84187', 'talk.religion.misc84188', 'talk.religion.misc8422
9']
```

Further, provide support for the following commands:

- x OR y

```
Enter the query
prometheus or califronia
['prometheus', 'califronia']
5
['talk.politics.misc178417', 'alt.atheism54163', 'alt.atheism53447', 'alt.atheism49960', 'rec.motorcycles103553']
```

- x AND y

```
Enter the query
reply and cry
['reply', 'cry']
53
['alt.atheism53519', 'alt.atheism54468', 'comp.graphics38570', 'comp.graphics38729', 'comp.graphics38817', 'comp.graphics3882
2', 'comp.graphics38852', 'comp.os.ms-windows.misc10036', 'comp.os.ms-windows.misc10144', 'comp.sys.mac.hardware52308', 'rec.a
utos103032', 'rec.sport.baseball104451', 'rec.sport.baseball104676', 'rec.sport.hockey53796', 'rec.sport.hockey54517', 'sci.cr
ypt15308', 'sci.electronics53872', 'soc.religion.christian20859', 'soc.religion.christian21647', 'soc.religion.christian2165
2', 'soc.religion.christian21662', 'talk.politics.guns54381', 'talk.politics.guns54573', 'talk.politics.mideast75878', 'talk.p
olitics.mideast75879', 'talk.politics.mideast75982', 'talk.politics.mideast75984', 'talk.politics.mideast75986', 'talk.politic
s.mideast75987', 'talk.politics.mideast76059', 'talk.politics.mideast76060', 'talk.politics.mideast76097', 'talk.politics.mide
ast76316', 'talk.politics.mideast76468', 'talk.politics.mideast76536', 'talk.politics.mideast77221', 'talk.politics.mideast772
57', 'talk.politics.mideast77315', 'talk.politics.mideast77364', 'talk.politics.mideast77365', 'talk.politics.mideast77371',
'talk.politics.mideast77388', 'talk.politics.mideast77390', 'talk.politics.misc178303', 'talk.politics.misc178305', 'talk.poli
tics.misc178757', 'talk.politics.misc179048', 'talk.religion.misc82800', 'talk.religion.misc83454', 'talk.religion.misc83457',
'talk.religion.misc83507', 'talk.religion.misc83629', 'talk.religion.misc84187']
```

- x AND NOT y

```
Enter the query
prometheus and not decompressed
3
['alt.atheism49960', 'alt.atheism53447', 'alt.atheism54163']
```

- x OR NOT y

```
Enter the query
califronia or not prometheus
19994
['sci.med58103', 'rec.sport.baseball105097', 'talk.politics.mideast76259', 'comp.os.ms-windows.misc9541', 'comp.sys.ibm.pc.h
ardware60943', 'sci.med58806', 'alt.atheism53157', 'rec.sport.hockey52565', 'talk.politics.mideast76337', 'sci.med59643', 's
ci.med58962', 'rec.sport.hockey52588', 'comp.windows.x67483', 'talk.politics.mideast77400', 'comp.graphics38828', 'talk.poli
tics.mideast77192', 'rec.sport.hockey52616', 'sci.space60869', 'comp.sys.ibm.pc.hardware60722', 'talk.politics.mideast7617
6', 'soc.religion.christian21443', 'comp.windows.x67055', 'rec.sport.hockey54035', 'soc.religion.christian21431', 'rec.spor
t.hockey54095', 'comp.sys.ibm.pc.hardware60506', 'rec.sport.baseball105080', 'talk.politics.misc178327', 'sci.space61364',
'sci.electronics52806', 'talk.politics.mideast76236', 'soc.religion.christian21805', 'talk.politics.misc178498', 'alt.atheis
m53415', 'rec.motorcycles104315', 'rec.sport.hockey54262', 'rec.sport.baseball104937', 'talk.religion.misc83764', 'misc.fors
ale76203', 'rec.sport.baseball104720', 'alt.atheism53252', 'comp.sys.ibm.pc.hardware60897', 'comp.graphics38288', 'talk.poli
tics.guns54363', 'talk.politics.misc178912', 'misc.forsale76497', 'comp.sys.mac.hardware52214', 'talk.politics.guns54738',
'soc.religion.christian21450', 'talk.politics.guns54384', 'comp.graphics38985', 'sci.electronics54480', 'comp.os.ms-windows.
misc9905', 'talk.religion.misc84347', 'talk.politics.misc178555', 'soc.religion.christian21386', 'soc.religion.christian2145
8', 'misc.forsale76607', 'sci.space61323', 'rec.autos102851', 'misc.forsale76216', 'rec.sport.hockey54116', 'sci.electronics
53939', 'talk.politics.guns55484', 'alt.atheism53619', 'rec.sport.hockey53598', 'rec.motorcycles103147', 'sci.med59635', 'co
mp.sys.mac.hardware52012', 'sci.crypt15353', 'talk.politics.guns54281', 'misc.forsale76261', 'comp.windows.x66929', 'sci.med
59389', 'talk.religion.misc84331', 'sci.electronics53621', 'alt.atheism53152', 'comp.os.ms-windows.misc10945', 'talk.politic
```

- If word is not present which user is searching

```
Enter the query
bfJHBF AND JHQWFH
['bfJHBF', 'JHQWFH']
not found

An exception has occurred, use %tb to see the full traceback.
```

Where x and y would be taken as input from the user.

Searching for phrase queries using Positional Indexes.

a). Pre-processing : Same as in question1.

b). Methodology :

- First step is to load the files from directory and reading the files using csopen() function.
- Once the file is read the all pre-processing is done like tokenization, removing stop words and lemmatization etc.
- After file is read then all the words are stored in dictionary whose keys are words and values are dictionary in which keys are document id's and values are posting list of positions of words appearing in that document.
- If word is new then it is added to dictionary and if word is already exist then document is added as key to dictionary in which keys are posting list for that word. Repeat till all files are read.
- The obtain dictionary of words is sorted and dictionary in which document id as key stored is also sorted and corresponding posting list are also sorted. c).Test-Cases :

1. For phrase query of length 2.

```
Enter the query
Santa Clara
{'comp.graphics38842': ['444'], 'comp.graphics38843': ['227'], 'comp.graphics38932': ['255'], 'rec.motorcycles101725': ['82'],
'rec.motorcycles103553': ['83'], 'rec.motorcycles104337': ['72'], 'rec.motorcycles104338': ['77'], 'rec.motorcycles104339':
['79'], 'rec.motorcycles104356': ['83'], 'rec.motorcycles104361': ['86'], 'rec.motorcycles104587': ['104'], 'rec.motorcycles10
4867': ['87'], 'rec.motorcycles105127': ['67'], 'rec.motorcycles105132': ['100']}
14
```

2. For phrase query of length 3.

```
Enter the query
useful property of
['comp.graphics38403', 'comp.graphics39078', 'comp.graphics39638']
3
```

3. For phrase query of length 5.

```
Enter the query
library of rendering routines which
['comp.graphics37916']
1
```

4. . For phrase query of length greater than 5.

```
Enter the query
to be or not to be
[]
0
```

---

5. If phrase is not present which user is searching.

```
Enter the query
hgjwhg jeb jbfjew
not found

An exception has occurred, use %tb to see the full traceback.
```