Question 1

PART 1

a. Lowercase the text
Using string.lower()

b. Perform tokenization
Using nltk.tokenize.word_tokenize()

c. Remove stopwords
Using nltk.corpus.stopwords()

d. Remove punctuations
Removed using string.punctuation

e. Remove blank space tokens
Using .strip()

PART 2

Sample output is:

```
Original Content of ./text_files/file258.txt:
Poor design doesn't line their own pedals up when connected using these. The offset isn't enough and results in each pedal along the
 line mounted a bit lower than the one next to it. As you can see in the picture, the third pedal from the chain is already almost o
ff the board. How can they overlook this?


Lowercased Content:
poor design doesn't line their own pedals up when connected using these. the offset isn't enough and results in each pedal along the
 line mounted a bit lower than the one next to it. as you can see in the picture, the third pedal from the chain is already almost o
ff the board. how can they overlook this?

Tokenized Content:
['poor', 'design', 'does', "n't", 'line', 'their', 'own', 'pedals', 'up', 'when', 'connected', 'using', 'these', '.', 'the', 'offset
', 'is', "n't", 'enough', 'and', 'results', 'in', 'each', 'pedal', 'along', 'the', 'line', 'mounted', 'a', 'bit', 'lower', 'than', '
the', 'one', 'next', 'to', 'it', '.', 'as', 'you', 'can', 'see', 'in', 'the', 'picture', ',', 'the', 'third', 'pedal', 'from', 'the'
, 'chain', 'is', 'already', 'almost', 'off', 'the', 'board', '.', 'how', 'can', 'they', 'overlook', 'this', '?']

Content after Removing Stopwords:
['poor', 'design', "n't", 'line', 'pedals', 'connected', 'using', '.', 'offset', "n't", 'enough', 'results', 'pedal', 'along', 'line
', 'mounted', 'bit', 'lower', 'one', 'next', '.', 'see', 'picture', ',', 'third', 'pedal', 'chain', 'already', 'almost', 'board', '.
', 'overlook', '?']

tokens after Removing Blank Spaces:
['poor', 'design', "n't", 'line', 'pedals', 'connected', 'using', '.', 'offset', "n't", 'enough', 'results', 'pedal', 'along', 'line
', 'mounted', 'bit', 'lower', 'one', 'next', '.', 'see', 'picture', ',', 'third', 'pedal', 'chain', 'already', 'almost', 'board', '.
', 'overlook', '?']

Content after Removing Punctuations and Blank Spaces:
poor design nt line pedals connected using  offset nt enough results pedal along line mounted bit lower one next  see picture  third
 pedal chain already almost board  overlook

Processed Content of ./text_files/file258.txt:
poor design nt line pedals connected using  offset nt enough results pedal along line mounted bit lower one next  see picture  third
 pedal chain already almost board  overlook
```

This output is printed for the first 5 files

Also saved each file by overwriting the original files.

Question 2:

Part 1 & 2

Create a unigram inverted index:

- I made this by maintaining a dictionary
- Then parsed through each word in all files
    - if that word wasn't in the dictionary initialize a list (with that file) in the dictionary with that word as key
    - else append that file in pre-existing list
- Dump this dictionary in a pickle file

Part 3 & 4

Support for the following operations:

a. T1 AND T2

This can be done by taking the intersection of the list attached to T1 term and T2 term

b. T1 OR T2

This can be done by taking the union of the list attached to T1 term and T2 term

c. T1 AND NOT T2

This can be done by taking the difference of the list attached to T1 term and T2 term

d. T1 OR NOT T2

This can be done by:

- ➢ Subtracting list attached to T2 from all 999 files
- ➢ Finding the union of list attached to T1 and the above list

Some sample Inputs and Outputs

Sample I/0 1 : (As provided in Assignment)

```
2
Car bag in a canister
OR, AND NOT
Coffee brewing techniques in cookbook
AND, OR NOT, OR
Query 1:car OR bag AND NOT canister
Number of documents retrieved for query 1: 31
Names of the documents retrieved for query 1: file797.txt, file956.txt, file404.txt, file174.txt, file981.txt, file3.txt, file313.tx
t, file860.txt, file886.txt, file686.txt, file780.txt, file264.txt, file459.txt, file738.txt, file665.txt, file166.txt, file573.txt,
 file363.txt, file682.txt, file73.txt, file118.txt, file864.txt, file892.txt, file930.txt, file699.txt, file542.txt, file863.txt, fi
le746.txt, file466.txt, file942.txt, file698.txt
Query 2:coffee AND brewing OR NOT techniques OR cookbook
Number of documents retrieved for query 2: 999
Names of the documents retrieved for query 2: file258.txt, file288.txt, file154.txt, file368.txt, file429.txt, file470.txt, file54.t
xt, file364.txt, file582.txt, file327.txt, file115.txt, file139.txt, file972.txt, file390.txt, file845.txt, file859.txt, file281.txt
, file169.txt, file548.txt, file443.txt, file486.txt, file696.txt, file752.txt, file633.txt, file771.txt, file598.txt, file852.txt,
file1.txt, file998.txt, file228.txt, file205.txt, file371.txt, file224.txt, file541.txt, file95.txt, file658.txt, file923.txt, file6
97.txt, file313.txt, file946.txt, file980.txt, file284.txt, file823.txt, file799.txt, file483.txt, file614.txt, file710.txt, file939
.txt, file372.txt, file294.txt, file890.txt, file907.txt, file711.txt, file801.txt, file584.txt, file642.txt, file274.txt, file950.t
xt, file839.txt, file944.txt, file615.txt, file67.txt, file471.txt, file51.txt, file706.txt, file82.txt, file354.txt, file361.txt, f
ile613.txt, file365.txt, file163.txt, file538.txt, file786.txt, file138.txt, file695.txt, file325.txt, file37.txt, file549.txt, file
750.txt, file992.txt, file25.txt, file732.txt, file881.txt, file578.txt, file195.txt, file384.txt, file670.txt, file637.txt, file227
.txt, file273.txt, file434.txt, file813.txt, file234.txt, file684.txt, file497.txt, file927.txt, file591.txt, file104.txt, file527.t
xt, file749.txt, file103.txt, file860.txt, file71.txt, file277.txt, file301.txt, file997.txt, file49.txt, file69.txt, file553.txt, f
ile9.txt, file32.txt, file929.txt, file975.txt, file218.txt, file985.txt, file430.txt, file773.txt, file291.txt, file824.txt, file92
8.txt, file359.txt, file616.txt, file575.txt, file576.txt, file479.txt, file622.txt, file91.txt, file426.txt, file809.txt, file92.tx
t, file973.txt, file306.txt, file508.txt, file751.txt, file780.txt, file468.txt, file189.txt, file191.txt, file503.txt, file777.txt,
 file899.txt, file671.txt, file926.txt, file971.txt, file715.txt, file295.txt, file350.txt, file864.txt, file620.txt, file800.txt, f
ile53.txt, file836.txt, file251.txt, file48.txt, file88.txt, file302.txt, file902.txt, file942.txt, file47.txt, file135.txt, file460
.txt, file457.txt, file334.txt, file501.txt, file43.txt, file709.txt, file500.txt, file705.txt, file345.txt, file232.txt, file7.txt,
 file40.txt, file446.txt, file152.txt, file499.txt, file321.txt, file87.txt, file118.txt, file319.txt, file762.txt, file898.txt, fil
e315.txt, file389.txt, file912.txt, file233.txt, file236.txt, file760.txt, file979.txt, file121.txt, file349.txt, file440.txt, file7
34.txt, file478.txt, file378.txt, file272.txt, file312.txt, file451.txt, file570.txt, file666.txt, file988.txt, file703.txt, file784
.txt, file208.txt, file776.txt, file757.txt, file977.txt, file16.txt, file28.txt, file106.txt, file330.txt, file335.txt, file495.txt
, file693.txt, file10.txt, file477.txt, file250.txt, file746.txt, file181.txt, file532.txt, file603.txt, file909.txt, file187.txt, f
ile150.txt, file667.txt, file38.txt, file444.txt, file484.txt, file123.txt, file794.txt, file941.txt, file560.txt, file664.txt, file
19.txt, file455.txt, file160.txt, file748.txt, file916.txt, file401.txt, file536.txt, file634.txt, file193.txt, file989.txt, file237
.txt, file397.txt, file726.txt, file145.txt, file681.txt, file332.txt, file14.txt, file415.txt, file125.txt, file105.txt, file385.tx
t, file566.txt, file631.txt, file52.txt, file114.txt, file464.txt, file174.txt, file966.txt, file906.txt, file383.txt, file215.txt,
file2.txt, file197.txt, file531.txt, file262.txt, file203.txt, file246.txt, file609.txt, file873.txt, file58.txt, file595.txt, file4
```

Sample I/0 2 : (All operations performed)

```
4
loving vintage
AND
loving vintage
OR
loving vintage
AND NOT
loving vintage
OR NOT
Query 1:loving AND vintage
Number of documents retrieved for query 1: 1
Names of the documents retrieved for query 1: file1.txt
Query 2:loving OR vintage
Number of documents retrieved for query 2: 21
Names of the documents retrieved for query 2: file827.txt, file1.txt, file725.txt, file597.txt, file254.txt, file723.txt, file674.tx
t, file638.txt, file907.txt, file278.txt, file936.txt, file847.txt, file197.txt, file439.txt, file422.txt, file737.txt, file51.txt,
file895.txt, file391.txt, file494.txt, file150.txt
Query 3:loving AND NOT vintage
Number of documents retrieved for query 3: 3
Names of the documents retrieved for query 3: file723.txt, file254.txt, file391.txt
Query 4:loving OR NOT vintage
Number of documents retrieved for query 4: 982
Names of the documents retrieved for query 4: file437.txt, file102.txt, file331.txt, file110.txt, file97.txt, file489.txt, file801.t
xt, file28.txt, file561.txt, file281.txt, file109.txt, file772.txt, file865.txt, file93.txt, file554.txt, file276.txt, file162.txt,
file373.txt, file987.txt, file682.txt, file138.txt, file384.txt, file672.txt, file776.txt, file864.txt, file980.txt, file477.txt, fi
le602.txt, file758.txt, file3.txt, file128.txt, file516.txt, file239.txt, file896.txt, file533.txt, file551.txt, file599.txt, file47
.txt, file729.txt, file13.txt, file55.txt, file375.txt, file920.txt, file22.txt, file823.txt, file455.txt, file653.txt, file90.txt,
file927.txt, file861.txt, file969.txt, file799.txt, file641.txt, file953.txt, file50.txt, file14.txt, file58.txt, file904.txt, file9
78.txt, file553.txt, file966.txt, file779.txt, file913.txt, file406.txt, file417.txt, file971.txt, file280.txt, file527.txt, file302
.txt, file256.txt, file507.txt, file5.txt, file474.txt, file916.txt, file114.txt, file875.txt, file36.txt, file243.txt, file719.txt,
 file745.txt, file868.txt, file186.txt, file313.txt, file614.txt, file845.txt, file854.txt, file522.txt, file676.txt, file536.txt, f
ile706.txt, file940.txt, file763.txt, file607.txt, file518.txt, file582.txt, file696.txt, file540.txt, file311.txt, file339.txt, fil
e400.txt, file576.txt, file616.txt, file251.txt, file242.txt, file648.txt, file43.txt, file4.txt, file73.txt, file669.txt, file141.t
xt, file589.txt, file81.txt, file968.txt, file1.txt, file303.txt, file80.txt, file174.txt, file649.txt, file720.txt, file416.txt, fi
le584.txt, file164.txt, file948.txt, file541.txt, file343.txt, file335.txt, file323.txt, file826.txt, file803.txt, file418.txt, file
309.txt, file890.txt, file704.txt, file64.txt, file508.txt, file86.txt, file119.txt, file207.txt, file493.txt, file332.txt, file359.
txt, file386.txt, file678.txt, file959.txt, file668.txt, file59.txt, file337.txt, file930.txt, file628.txt, file664.txt, file23.txt,
```

Sample I/0 3 : (Showcasing the pre processing)

```
4
Loving, vintage
AND
loving and VINtage
OR
loving is are ? vintage
AND NOT
loving vintage
OR NOT
Query 1:loving AND vintage
Number of documents retrieved for query 1: 1
Names of the documents retrieved for query 1: file1.txt
Query 2:loving OR vintage
Number of documents retrieved for query 2: 21
Names of the documents retrieved for query 2: file391.txt, file674.txt, file597.txt, file422.txt, file51.txt, file737.txt, file494.t
xt, file827.txt, file278.txt, file895.txt, file150.txt, file254.txt, file907.txt, file847.txt, file638.txt, file197.txt, file1.txt,
file723.txt, file439.txt, file725.txt, file936.txt
Query 3:loving AND NOT vintage
Number of documents retrieved for query 3: 3
Names of the documents retrieved for query 3: file723.txt, file391.txt, file254.txt
Query 4:loving OR NOT vintage
Number of documents retrieved for query 4: 982
Names of the documents retrieved for query 4: file543.txt, file304.txt, file711.txt, file215.txt, file569.txt, file441.txt, file444.
txt, file47.txt, file860.txt, file969.txt, file624.txt, file578.txt, file683.txt, file395.txt, file825.txt, file363.txt, file752.txt
, file879.txt, file400.txt, file671.txt, file921.txt, file59.txt, file342.txt, file983.txt, file686.txt, file854.txt, file147.txt, f
ile435.txt, file475.txt, file285.txt, file524.txt, file148.txt, file310.txt, file385.txt, file390.txt, file598.txt, file708.txt, fil
e805.txt, file512.txt, file173.txt, file274.txt, file377.txt, file718.txt, file824.txt, file869.txt, file163.txt, file678.txt, file9
45.txt, file378.txt, file119.txt, file647.txt, file90.txt, file12.txt, file755.txt, file823.txt, file768.txt, file70.txt, file506.tx
t, file495.txt, file277.txt, file288.txt, file436.txt, file772.txt, file821.txt, file688.txt, file279.txt, file876.txt, file311.txt,
 file44.txt, file527.txt, file185.txt, file662.txt, file817.txt, file991.txt, file754.txt, file812.txt, file797.txt, file634.txt, fi
le894.txt, file331.txt, file83.txt, file404.txt, file484.txt, file581.txt, file998.txt, file261.txt, file350.txt, file68.txt, file24
3.txt, file32.txt, file260.txt, file571.txt, file474.txt, file130.txt, file438.txt, file302.txt, file409.txt, file56.txt, file319.tx
t, file862.txt, file301.txt, file131.txt, file906.txt, file653.txt, file910.txt, file903.txt, file101.txt, file333.txt, file842.txt,
 file693.txt, file846.txt, file448.txt, file964.txt, file270.txt, file140.txt, file868.txt, file610.txt, file216.txt, file188.txt, f
ile515.txt, file673.txt, file849.txt, file596.txt, file34.txt, file424.txt, file464.txt, file53.txt, file442.txt, file684.txt, file2
38.txt, file658.txt, file654.txt, file413.txt, file410.txt, file451.txt, file69.txt, file829.txt, file873.txt, file687.txt, file830.
txt, file625.txt, file211.txt, file486.txt, file952.txt, file129.txt, file77.txt, file323.txt, file425.txt, file866.txt, file764.txt
, file963.txt, file759.txt, file593.txt, file397.txt, file756.txt, file953.txt, file339.txt, file715.txt, file815.txt, file630.txt,
```

Question 3:

Part 1 & 2

Create a positional index:

- I made this by maintaining a dictionary
- Then parsed through each word in all files
    - if that word wasn't in the dictionary initialize a list (with that file and the words position in the file as a nested list) in the dictionary with that word as key
        - If the word occurs in the file again, add the position to the nested list again.
    - else append that file in pre-existing list (and the position of the word in the file)
- Dump this dictionary in a pickle file

Part 3 & 4

For phrase queries we use the 2 pointer approach:

- First I initialize a set(say result) of all the files containing the first word in our query

- Now using a for loop I iterate through the next words in the phrase and find the intersection of files in result and this word
- For these given files I check that if there's at least one instance of this word occurring at i+1th position (if previous word was at ith position)
    - If the above condition is satisfied then we keep that file
    - Else we remove it from the result
- We do this for all terms

Sample I/0 1 : (As provided in Assignment)

```
2
Car bag in a canister
Coffee brewing techniques in cookbook
Number of documents retrieved for query 1 using positional index: 0
Names of documents retrieved for query 1 using positional index:
Number of documents retrieved for query 2 using positional index: 0
Names of documents retrieved for query 2 using positional index:
```

Sample I/0 2 : (few operations performed)

```
4
car bag
even light gauged strings
bef
today pleasantly
Number of documents retrieved for query 1 using positional index: 0
Names of documents retrieved for query 1 using positional index:
Number of documents retrieved for query 2 using positional index: 1
Names of documents retrieved for query 2 using positional index: file987.txt
Number of documents retrieved for query 3 using positional index: 0
Names of documents retrieved for query 3 using positional index:
Number of documents retrieved for query 4 using positional index: 1
Names of documents retrieved for query 4 using positional index: file990.txt
```

Sample I/0 3 : (Showcasing the pre processing)

```
4
Loving Vintage
loving is vintage
Loving ? are vintage
loving vintage
Number of documents retrieved for query 1 using positional index: 1
Names of documents retrieved for query 1 using positional index: file1.txt
Number of documents retrieved for query 2 using positional index: 1
Names of documents retrieved for query 2 using positional index: file1.txt
Number of documents retrieved for query 3 using positional index: 1
Names of documents retrieved for query 3 using positional index: file1.txt
Number of documents retrieved for query 4 using positional index: 1
Names of documents retrieved for query 4 using positional index: file1.txt
```