Information Retrieval Assignment-1 Report

Group No: 20
Devang Prajapati (MT22050)
Heli Shah (MT22100)

Question: 1 Data Preprocessing

(i) Text Retrieval System

- In this, we had retrieved the required text from the given 1400 files using “Regular Expression (RegEx)” Module of python.
- We had stored the whole text between TITLE and TEXT tags in a string concatenated with a blank space.
- Following image shows the output of first 5 files.

Before:

```
PS D:\SEMESTER-2\IR-Ass1> python -u "d:\SEMESTER-2\IR-Ass1\test.py"
<DOC>
<DOCNO>
1
</DOCNO>
<TITLE>
experimental investigation of the aerodynamics of a
wing in a slipstream .
</TITLE>
<AUTHOR>
brenckman,m.
</AUTHOR>
<BIBLIO>
j. ae. scs. 25, 1958, 324.
</BIBLIO>
<TEXT>
  an experimental study of a wing in a propeller slipstream was
made in order to determine the spanwise distribution of the lift
increase due to slipstream at different angles of attack of the wing
and at different free stream to slipstream velocity ratios .  the
results were intended in part as an evaluation basis for different
```

```
<DOC>
<DOCNO>
5
</DOCNO>
<TITLE>
one-dimensional transient heat conduction into a double-layer
slab subjected to a linear heat input for a small time
internal .
</TITLE>
<AUTHOR>
wasserman,b.
</AUTHOR>
<BIBLIO>
j. ae. scs. 24, 1957, 924.
</BIBLIO>
<TEXT>
  analytic solutions are presented for the transient heat conduction
in composite slabs exposed at one surface to a
triangular heat rate .  this type of heating rate may occur, for
example, during aerodynamic heating .
</TEXT>
</DOC>
experimental investigation of the aerodynamics of awing in a slipstream .an ex
```

After:

```
PS D:\SEMESTER-2\IR-Ass1> python -u "d:\SEMESTER-2\IR-Ass1\Q1_i.py"
FILE: cranfield0001
experimental investigation of the aerodynamics of awing in a slipstream .an experimental study of a wing in a propeller slipstream wasmade in order to determine t
he spanwise distribution of the liftincrease due to slipstream at different angles of attack of the wingand at different free stream to slipstream velocity ratios
 .  theresults were intended in part as an evaluation basis for differentheoretical treatments of this problem .the comparative span loading curves, together wit
h supportingevidence, showed that a substantial part of the lift incrementproduced by the slipstream was due to a /destalling/ or boundary-layer-controleffect .
the integrated remaining lift increment,after subtracting this destalling lift, was found to agreewell with a potential flow theory .an empirical evaluation of th
e destalling effects was made forthe specific configuration of the experiment .

FILE: cranfield0002
simple shear flow past a flat plate in an incompressible fluid of smallviscosity .in the study of high-speed viscous flow past a two-dimensional body itis usually
 necessary to consider a curved shock wave emitting from thenose or leading edge of the body .  consequently, there exists an inviscidrotational flow region betwe
en the shock wave and the boundary layer.  such a situation arises, for instance, in the study of the hypersonicviscous flow past a flat plate .  the situation is
 somewhat differentfrom prandtl's classical boundary-layer problem . in prandtl'soriginal problem the inviscid free stream outside the boundary layer isirrotation
al while in a hypersonic boundary-layer problem the inviscidfree stream must be considered as rotational .  the possible effects ofvorticity have been recently di
scussed by ferri and libby .  in the presentpaper, the simple shear flow past a flat plate in a fluid of smallviscosity is investigated .  it can be shown that th
is problem can againbe treated by the boundary-layer approximation, the only novel featurebeing that the free stream has a constant vorticity .  the discussionher
e is restricted to two-dimensional incompressible steady flow .

FILE: cranfield0003
the boundary layer in simple shear flow past a flat plate .the boundary-layer equations are presented for steadyincompressible flow with no pressure gradient .

FILE: cranfield0004
approximate solutions of the incompressible laminarboundary layer equations for a plate in shear flow .the two-dimensional steady boundary-layerproblem for a flat
 plate in ashear flow of incompressible fluid is considered .solutions for the boundarylayerthickness, skin friction, and the velocitydistribution in the boundary
layer are obtained by the karman-pohlhausentechnique .  comparison withthe boundary layer of a uniform flow has alsobeen made to show the effect ofvorticity .

FILE: cranfield0005
one-dimensional transient heat conduction into a double-layerslab subjected to a linear heat input for a small timeinternal .analytic solutions are presented for
the transient heat conductionin composite slabs exposed at one surface to atriangular heat rate .  this type of heating rate may occur, forexample, during aerodyn
amic heating .

PS D:\SEMESTER-2\IR-Ass1>
```

(ii) Preprocessing

We had performed following 5 steps to preprocess our data in the sequence shown below:

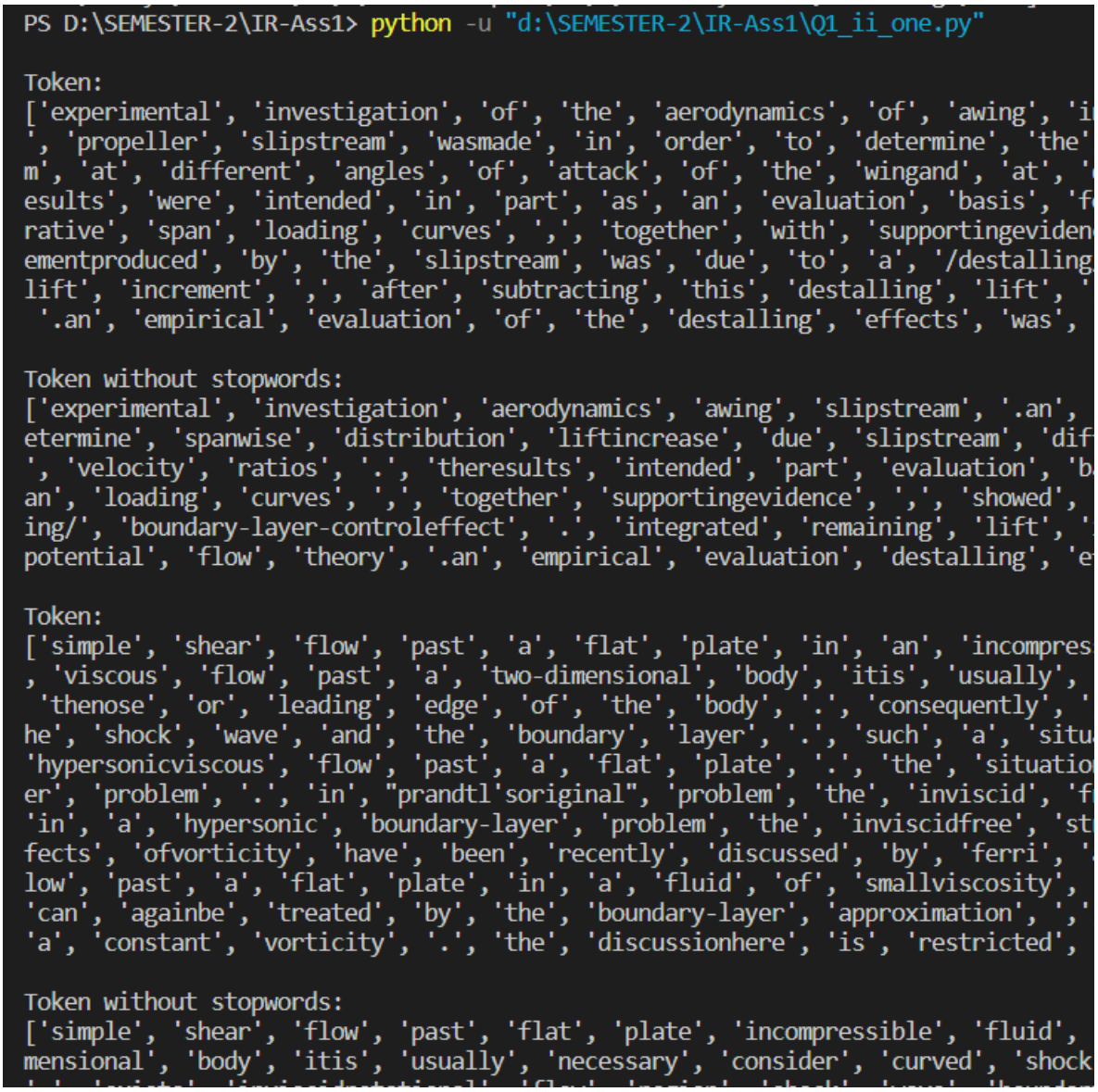
1. Lowercase the text

```
FILE: cranfield0004 in Upper Case
approximate solutions of the incompressible laminarboundary layer equations for a plate in shear flow .the two-dimensional steady boundary-layerproblem for a flat
 plate in ashear flow of incompressible fluid is considered .solutions for the boundarylayerthickness, skin friction, and the velocitydistribution in the boundary
layer are obtained by the karman-pohlhausentechnique .  comparison withthe boundary layer of a uniform flow has alsobeen made to show the effect ofvorticity .

FILE: cranfield0005 in Upper Case
one-dimensional transient heat conduction into a double-layerslab subjected to a linear heat input for a small timeinternal .analytic solutions are presented for
the transient heat conductionin composite slabs exposed at one surface to atriangular heat rate .  this type of heating rate may occur, forexample, during aerodyn
amic heating .

FILE: cranfield0001 in Lower Case
experimental investigation of the aerodynamics of awing in a slipstream .an experimental study of a wing in a propeller slipstream wasmade in order to determine t
he spanwise distribution of the liftincrease due to slipstream at different angles of attack of the wingand at different free stream to slipstream velocity ratios
 .  theresults were intended in part as an evaluation basis for differenttheoretical treatments of this problem .the comparative span loading curves, together wit
h supportingevidence, showed that a substantial part of the lift incrementproduced by the slipstream was due to a /destalling/ or boundary-layer-controleffect .
the integrated remaining lift increment,after subtracting this destalling lift, was found to agreewell with a potential flow theory .an empirical evaluation of th
e destalling effects was made forthe specific configuration of the experiment .

FILE: cranfield0002 in Lower Case
simple shear flow past a flat plate in an incompressible fluid of smallviscosity .in the study of high-speed viscous flow past a two-dimensional body itis usually
 necessary to consider a curved shock wave emitting from thenose or leading edge of the body .  consequently, there exists an inviscidrotational flow region betwe
en the shock wave and the boundary layer.  such a situation arises, for instance, in the study of the hypersonicviscous flow past a flat plate .  the situation is
 somewhat differentfrom prandtl's classical boundary-layer problem . in prandtl'soriginal problem the inviscid free stream outside the boundary layer isirrotation
al while in a hypersonic boundary-layer problem the inviscidfree stream must be considered as rotational .  the possible effects ofvorticity have been recently di
```

2. Perform tokenization

```
['experimental', 'investigation', 'of', 'the', 'aerodynamics', 'of', 'awing', 'in', 'a', 'slipstream', '.an',
', 'propeller', 'slipstream', 'wasmade', 'in', 'order', 'to', 'determine', 'the', 'spanwise', 'distribution',
m', 'at', 'different', 'angles', 'of', 'attack', 'of', 'the', 'wingand', 'at', 'different', 'free', 'stream',
esults', 'were', 'intended', 'in', 'part', 'as', 'an', 'evaluation', 'basis', 'for', 'differenttheoretical',
rative', 'span', 'loading', 'curves', ',', 'together', 'with', 'supportingevidence', ',', 'showed', 'that', 'a
ementproduced', 'by', 'the', 'slipstream', 'was', 'due', 'to', 'a', '/destalling/', 'or', 'boundary-layer-cont
lift', 'increment', ',', 'after', 'subtracting', 'this', 'destalling', 'lift', ',', 'was', 'found', 'to', 'agr
 '.an', 'empirical', 'evaluation', 'of', 'the', 'destalling', 'effects', 'was', 'made', 'forthe', 'specific',
['simple', 'shear', 'flow', 'past', 'a', 'flat', 'plate', 'in', 'an', 'incompressible', 'fluid', 'of', 'smallv
, 'viscous', 'flow', 'past', 'a', 'two-dimensional', 'body', 'itis', 'usually', 'necessary', 'to', 'consider',
 'thenose', 'or', 'leading', 'edge', 'of', 'the', 'body', '.', 'consequently', ',', 'there', 'exists', 'an',
he', 'shock', 'wave', 'and', 'the', 'boundary', 'layer', '.', 'such', 'a', 'situation', 'arises', ',', 'for',
'hypersonicviscous', 'flow', 'past', 'a', 'flat', 'plate', '.', 'the', 'situation', 'is', 'somewhat', 'differe
er', 'problem', '.', 'in', "prandtl'soriginal", 'problem', 'the', 'inviscid', 'free', 'stream', 'outside', 'th
'in', 'a', 'hypersonic', 'boundary-layer', 'problem', 'the', 'inviscidfree', 'stream', 'must', 'be', 'consider
fects', 'ofvorticity', 'have', 'been', 'recently', 'discussed', 'by', 'ferri', 'and', 'libby', '.', 'in', 'the
low', 'past', 'a', 'flat', 'plate', 'in', 'a', 'fluid', 'of', 'smallviscosity', 'is', 'investigated', '.', 'it
'can', 'againbe', 'treated', 'by', 'the', 'boundary-layer', 'approximation', ',', 'the', 'only', 'novel', 'fea
'a', 'constant', 'vorticity', '.', 'the', 'discussionhere', 'is', 'restricted', 'to', 'two-dimensional', 'inco
['the', 'boundary', 'layer', 'in', 'simple', 'shear', 'flow', 'past', 'a', 'flat', 'plate', '.the', 'boundary-
adyincompressible', 'flow', 'with', 'no', 'pressure', 'gradient', '.']
['approximate', 'solutions', 'of', 'the', 'incompressible', 'laminarboundary', 'layer', 'equations', 'for', 'a
nsional', 'steady', 'boundary-layerproblem', 'for', 'a', 'flat', 'plate', 'in', 'ashear', 'flow', 'of', 'incor
', 'for', 'the', 'boundarylayerthickness', ',', 'skin', 'friction', ',', 'and', 'the', 'velocitydistribution',
y', 'the', 'karman-pohlhausentechnique', '.', 'comparison', 'withthe', 'boundary', 'layer', 'of', 'a', 'unifor
 'the', 'effect', 'ofvorticity', '.']
['one-dimensional', 'transient', 'heat', 'conduction', 'into', 'a', 'double-layerslab', 'subjected', 'to', 'a',
imeinternal', '.analytic', 'solutions', 'are', 'presented', 'for', 'the', 'transient', 'heat', 'conductionin',
ace', 'to', 'atriangular', 'heat', 'rate', '.', 'this', 'type', 'of', 'heating', 'rate', 'may', 'occur', ',',
', '.']
PS D:\SEMESTER-2\IR-Ass1>
```

3. Remove stopwords



```
PS D:\SEMESTER-2\IR-Ass1> python -u "d:\SEMESTER-2\IR-Ass1\Q1_ii_one.py"

Token:
['experimental', 'investigation', 'of', 'the', 'aerodynamics', 'of', 'awing', 'i
', 'propeller', 'slipstream', 'wasmade', 'in', 'order', 'to', 'determine', 'the'
m', 'at', 'different', 'angles', 'of', 'attack', 'of', 'the', 'wingand', 'at', '
esults', 'were', 'intended', 'in', 'part', 'as', 'an', 'evaluation', 'basis', 'f
rative', 'span', 'loading', 'curves', ',', 'together', 'with', 'supportingeviden
ementproduced', 'by', 'the', 'slipstream', 'was', 'due', 'to', 'a', '/destalling
lift', 'increment', ',', 'after', 'subtracting', 'this', 'destalling', 'lift', '
 '.an', 'empirical', 'evaluation', 'of', 'the', 'destalling', 'effects', 'was',

Token without stopwords:
['experimental', 'investigation', 'aerodynamics', 'awing', 'slipstream', '.an',
etermine', 'spanwise', 'distribution', 'liftincrease', 'due', 'slipstream', 'dif
', 'velocity', 'ratios', '.', 'theresults', 'intended', 'part', 'evaluation', 'b
an', 'loading', 'curves', ',', 'together', 'supportingevidence', ',', 'showed',
ing/', 'boundary-layer-controleffect', '.', 'integrated', 'remaining', 'lift', '
potential', 'flow', 'theory', '.an', 'empirical', 'evaluation', 'destalling', 'e

Token:
['simple', 'shear', 'flow', 'past', 'a', 'flat', 'plate', 'in', 'an', 'incompres
, 'viscous', 'flow', 'past', 'a', 'two-dimensional', 'body', 'itis', 'usually',
 'thenose', 'or', 'leading', 'edge', 'of', 'the', 'body', '.', 'consequently', '
he', 'shock', 'wave', 'and', 'the', 'boundary', 'layer', '.', 'such', 'a', 'situ
'hypersonicviscous', 'flow', 'past', 'a', 'flat', 'plate', '.', 'the', 'situatio
er', 'problem', '.', 'in', "prandtl'soriginal", 'problem', 'the', 'inviscid', 'f
'in', 'a', 'hypersonic', 'boundary-layer', 'problem', 'the', 'inviscidfree', 'st
fects', 'ofvorticity', 'have', 'been', 'recently', 'discussed', 'by', 'ferri', '
low', 'past', 'a', 'flat', 'plate', 'in', 'a', 'fluid', 'of', 'smallviscosity',
'can', 'againbe', 'treated', 'by', 'the', 'boundary-layer', 'approximation', ','
'a', 'constant', 'vorticity', '.', 'the', 'discussionhere', 'is', 'restricted',

Token without stopwords:
['simple', 'shear', 'flow', 'past', 'flat', 'plate', 'incompressible', 'fluid',
mensional', 'body', 'itis', 'usually', 'necessary', 'consider', 'curved', 'shock
```

4. Remove punctuations

```
PS D:\SEMESTER-2\IR-Ass1> python -u "d:\SEMESTER-2\IR-Ass1\Q1_ii_one.py"
FILE: cranfield0001
experimental investigation aerodynamics awing slipstream an experimental st
e due slipstream different angles attack wingand different free stream slip
treatments problem the comparative span loading curves  together supporting
darylayercontroleffect  integrated remaining lift increment  subtracting de
 effects made forthe specific configuration experiment

FILE: cranfield0002
simple shear flow past flat plate incompressible fluid smallviscosity in st
ved shock wave emitting thenose leading edge body  consequently  exists inv
y hypersonicviscous flow past flat plate  situation somewhat differentfrom
m outside boundary layer isirrotational hypersonic boundarylayer problem in
cussed ferri libby  presentpaper  simple shear flow past flat plate fluid s
novel featurebeing free stream constant vorticity  discussionhere restricte

FILE: cranfield0003
boundary layer simple shear flow past flat plate the boundarylayer equation

FILE: cranfield0004
approximate solutions incompressible laminarboundary layer equations plate
ressible fluid considered solutions boundarylayerthickness  skin friction
he boundary layer uniform flow alsobeen made show effect ofvorticity

FILE: cranfield0005
onedimensional transient heat conduction doublelayerslab subjected linear h
omposite slabs exposed one surface atriangular heat rate  type heating rate

PS D:\SEMESTER-2\IR-Ass1>
```

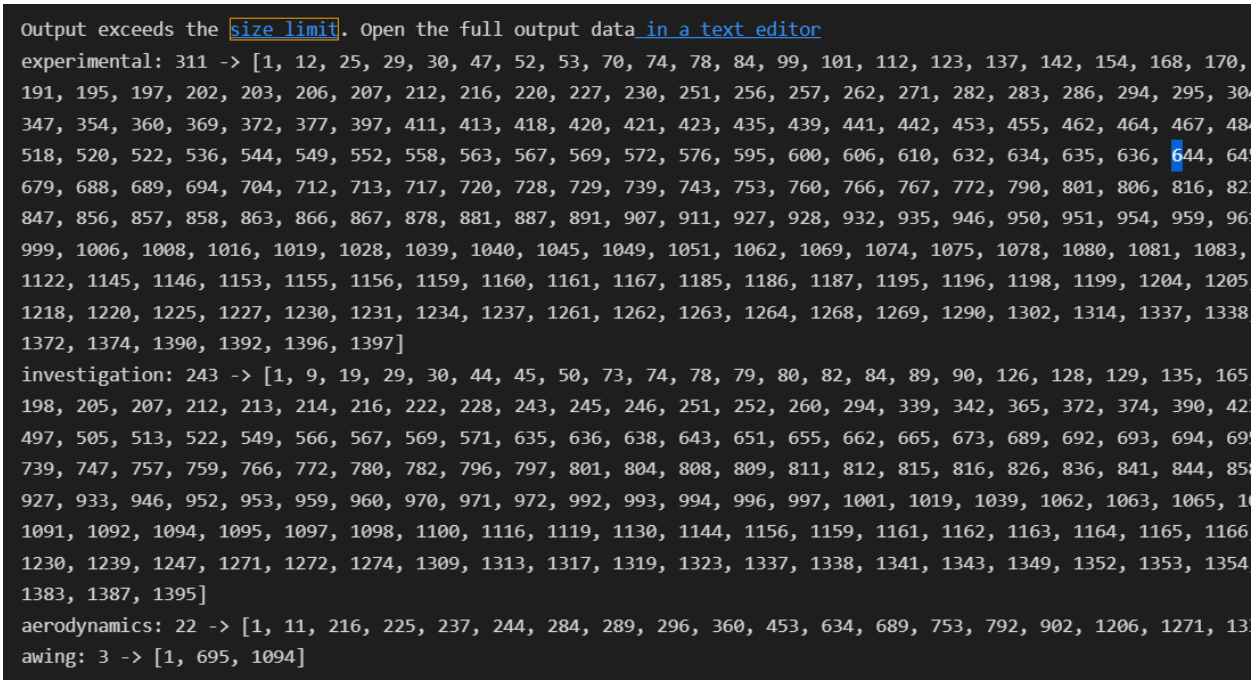
5. Remove blank space tokens

```
PS D:\SEMESTER-2\IR-Ass1> python -u "d:\SEMESTER-2\IR-Ass1\Q1_ii_one.py"
FILE: cranfield0001
experimental investigation aerodynamics awing slipstream an experimental s
e due slipstream different angles attack wingand different free stream sli
treatments problem the comparative span loading curves  together supportin
darylayercontroleffect  integrated remaining lift increment  subtracting d
 effects made forthe specific configuration experiment

FILE: cranfield0002
simple shear flow past flat plate incompressible fluid smallviscosity in s
ved shock wave emitting thenose leading edge body  consequently  exists in
y hypersonicviscous flow past flat plate  situation somewhat differentfrom
m outside boundary layer isirrotational hypersonic boundarylayer problem i
cussed ferri libby  presentpaper  simple shear flow past flat plate fluid
novel featurebeing free stream constant vorticity  discussionhere restrict

FILE: cranfield0003
boundary layer simple shear flow past flat plate the boundarylayer equatio

FILE: cranfield0004
approximate solutions incompressible laminarboundary layer equations plate
ressible fluid considered solutions boundarylayerthickness  skin friction
he boundary layer uniform flow alsobeen made show effect ofvorticity
```

Question: 2 Boolean Queries

- Initially, we had created UNIGRAM inverted Index of the dataset obtained from Question 1 for 1400 files.
- Unigram inverted Index was built from scratch using complete Brute-Force.
- Unigram Inverted Index is stored and loaded using PICKLE module in python.
- It will give the results of queries consisting of boolean expressions such as AND, OR, AND NOT and OR NOT.
- It will also give the results of documents retrieved, names of the documents and compairisions made for a particular boolean queries.
- Following are the images of all the outputs of the queries used on UNIGRAM inverted index.

Unigram Inverted Index:



Boolean Queries Performed:

```
Enter Query:
Enter 2 Operations:
Tokens : ['experimental', 'solutions', 'slipstram']
Operations:  ['AND', 'OR']
Query :  experimental AND solutions OR slipstram
No of documents retrieved:  20
Names of document retrieved:  ['cranfield0047', 'cranfield0188', 'cranfield0329', 'cranfield0435', 'cranfield0467', 'cranfield0498',
'cranfield0518', 'cranfield0576', 'cranfield0644', 'cranfield0663', 'cranfield0717', 'cranfield0753', 'cranfield0767', 'cranfield0823',
'cranfield1074', 'cranfield1078', 'cranfield1083', 'cranfield1185', 'cranfield1214', 'cranfield1352']
Minimum Number comaparisons required:  421
```

Question: 3 Phrase Queries

(i) Bigram Inverted Index

- For creating Bigram Inverted Index from scratch, firstly we had tokenized the given dataset.
- After that we combined every 2 tokens to make it bigrams.
- Store and Load Bigram Inverted Index using Pickle Module.

Bigram Inverted Index:

```
Output exceeds the size limit. Open the full output data in a text editor
experimental investigation: 43 -> [1, 29, 84, 189, 251, 372, 423, 442, 497, 505, 522, 569, 635, 636, 662, 69
858, 887, 996, 1039, 1074, 1092, 1097, 1098, 1156, 1159, 1161, 1205, 1220, 1225, 1227, 1230, 1338, 1364]
investigation aerodynamics: 1 -> [1]
aerodynamics awing: 1 -> [1]
awing slipstream: 1 -> [1]
slipstream experimental: 2 -> [1, 484]
experimental study: 17 -> [1, 74, 256, 334, 420, 464, 544, 549, 760, 772, 801, 847, 911, 1019, 1167, 1264]
study wing: 1 -> [1]
wing propeller: 1 -> [1]
propeller slipstream: 6 -> [1, 453, 1064, 1164]
slipstream wasmade: 1 -> [1]
wasmade order: 1 -> [1]
order determine: 7 -> [1, 249, 277, 354, 904, 1319, 1393]
determine spanwise: 1 -> [1]
spanwise distribution: 2 -> [1, 1289]
distribution liftincrease: 1 -> [1]
liftincrease due: 1 -> [1]
due slipstream: 1 -> [1]
slipstream different: 1 -> [1]
different angles: 1 -> [1]
angles attack: 49 -> [1, 58, 69, 197, 354, 363, 373, 420, 492, 520, 567, 673, 688, 709, 712, 713, 717, 801,
1147, 1193, 1218, 1231, 1259, 1307, 1350, 1381]
```

(ii) Positional Inverted Index

- For creating Positional Inverted Index from scratch, firstly we had tokenized the given dataset.
- And we will make a dictionary having document ID as key and its positions as value.
- After that we had stored all the common documents containing that tokens in a list.
- Now the positions of that particular token in the given document is stored in another list.
- And this position list should be inside the common documents list.
- We can also store the frequency of documents or the frequency of the positions in the document.

Positional Inverted Index:


```
{'experimental': [{1: [0, 5], 12: [25], 25: [77], 29: [121], 30: [18], 47: [103], 52: [9], 53: [6], 70: [10], 74: [0], 78: [67], 84
[121], 220: [37], 227: [69], 230: [96], 251: [42], 256: [0, 10], 257: [99], 262: [103], 271: [0, 13], 282: [127], 283: [67], 286: [
453: [13], 455: [34, 37], 462: [31], 464: [7], 467: [101], 484: [13, 18, 115], 497: [1], 498: [66], 501: [27], 504: [83], 505: [8],
[56], 666: [69, 72], 670: [10], 675: [64], 678: [26], 679: [81], 688: [75], 689: [96], 694: [19], 704: [175], 712: [162], 713: [18]
857: [0, 16], 858: [0, 60], 863: [40, 55], 866: [6, 53], 867: [36], 878: [0], 881: [38], 887: [48], 891: [21], 907: [111], 911: [0]
1045: [7], 1049: [41], 1051: [60], 1062: [0, 23], 1069: [7], 1074: [1, 46], 1075: [0], 1078: [39], 1080: [43], 1081: [63], 1083: [3
[5, 69, 134], 1212: [8], 1213: [62], 1214: [29, 44], 1216: [71], 1218: [0], 1220: [6], 1225: [71, 132], 1227: [7], 1230: [6], 1231:
1397: [39]}], 'investigation': [{1: [1], 9: [12, 165], 19: [0], 29: [122], 30: [1], 44: [81], 45: [0], 50: [0, 9], 73: [0], 74: [22
[9], 243: [0, 16], 245: [12], 246: [12], 251: [43], 252: [0], 260: [18], 294: [0], 339: [23], 342: [90], 365: [22], 372: [1, 11, 51
[0, 11, 27], 689: [0], 692: [0, 14, 54], 693: [0, 15, 59], 694: [20], 695: [14, 143], 696: [16, 108], 700: [30], 710: [6, 101], 711
859: [13], 875: [2], 887: [49], 892: [15], 905: [43], 907: [34, 72], 927: [0], 933: [67], 946: [1], 952: [27], 953: [19], 959: [7],
[1], 1100: [1], 1116: [27], 1119: [65], 1130: [61], 1144: [144], 1156: [1], 1159: [1], 1161: [8], 1162: [1, 12, 37], 1163: [1, 11,
[1], 1353: [0], 1354: [0], 1364: [1, 9, 68], 1367: [1], 1373: [60], 1377: [1], 1381: [120], 1383: [16], 1387: [12], 1395: [36]}], '
'slipstream': [{1: [4, 9, 17, 25, 47], 409: [24], 453: [53, 66, 71, 82, 97], 484: [17, 23, 32, 61], 1064: [1, 31, 35], 1094: [48],
[6], 193: [0], 198: [12], 204: [0, 40], 216: [109, 120], 221: [1], 256: [1, 11], 270: [82], 272: [20], 277: [0], 289: [1, 7], 293:
[4], 588: [18, 66], 612: [19], 626: [25], 639: [1], 640: [33, 110], 644: [0], 667: [100], 680: [11], 695: [22], 716: [0, 26], 727:
[32], 1002: [27], 1018: [33], 1019: [102], 1054: [6], 1099: [1], 1108: [0], 1109: [9], 1153: [0], 1164: [17], 1167: [1, 21], 1181:
12], 42: [6, 10, 38, 121, 147], 52: [42, 58, 100], 60: [71], 76: [32, 62, 89], 78: [77], 95: [6, 12], 146: [104], 147: [37], 189: [
289: [35, 51, 81, 92], 333: [4, 12], 379: [63], 416: [52], 420: [35, 73], 432: [7], 433: [36, 52, 106, 120, 173, 188, 193, 197, 217
[54, 128], 676: [28], 677: [87, 117], 680: [5, 14, 26], 681: [4, 9, 26], 682: [23], 683: [8, 22, 49, 63, 100], 692: [24, 85], 693:
51], 757: [21, 113, 165], 768: [6, 30], 781: [20, 80], 783: [6, 18, 53], 791: [37], 793: [4, 13, 19, 82], 794: [3, 13, 60, 109, 121
81], 970: [4, 9, 20, 24, 75], 1062: [18, 33], 1064: [17, 24], 1074: [19], 1075: [11, 21, 60], 1089: [59], 1091: [26, 35, 56], 1092:
```

(iii) Performing Phrase Queries

- With the help of Bigram Inverted Index and Positional Inverted Index we can retrieve the number of the documents and names of the documents for a given Phrase Query.

Outputs of some Queries is shown in the image: