

Demo for Project 2 of COMP 479 by Silas Kalinowski ID: 40256077

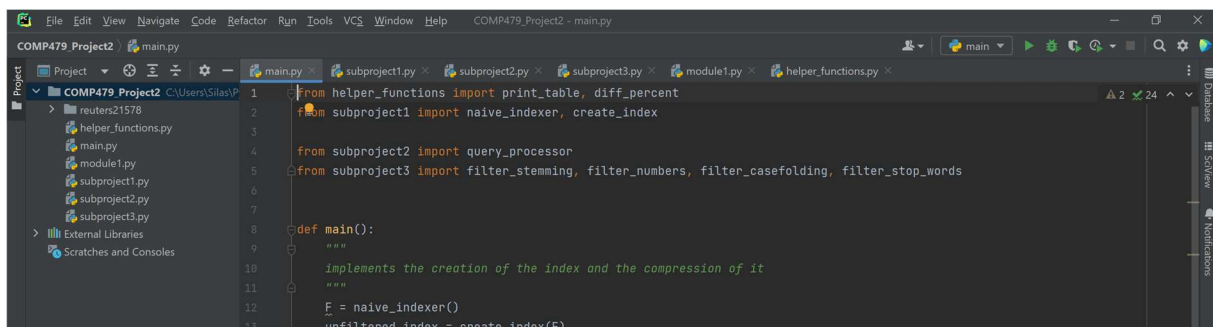
The program can be started by running the main.py file in a Python environment. For the program to work all modules and all input files need to be in the local directory. Also, the program uses NLTK, which needs to be downloaded. For the stemmer `nltk.download('stopwords')` might have to be used. The output files are saved in the local directory and are named in the following manner:

Index_TYPEOFCOMPRESSION.txt for the indices

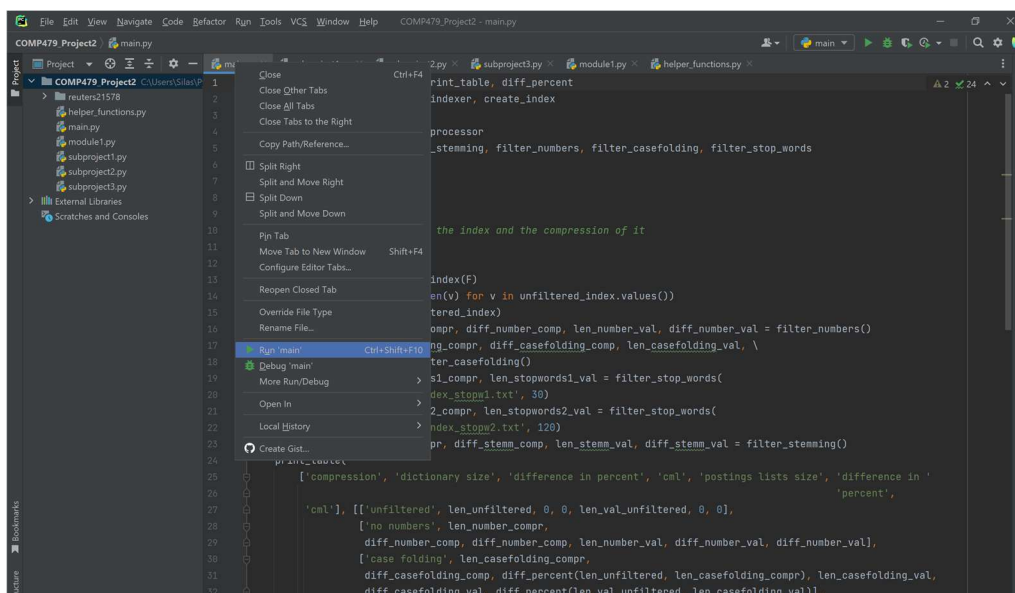
File F is saved as F_filtered_sorted.txt

Running the main file starts the creation of the index, the compression of the index and starts the query in the console. The user can then type terms in the console which he wants to search for or type t to run all tests and sample queries. The query starts again after every output and can be stopped by typing stop.

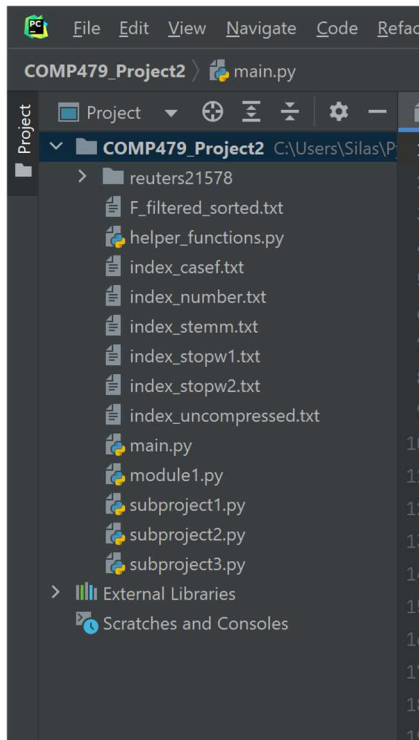
The following screenshots display the whole process of reading in the articles, creating the index and compressing it. The last three screenshots explain how to use the query processor and how to start the test.



This screenshot shows the directory and main module of project before running the main.py file.



By running main.py the files from reuters21578 are read in and the index is created and compressed. Also, the query is started.



After running the main function the output files can be found in the local directory.

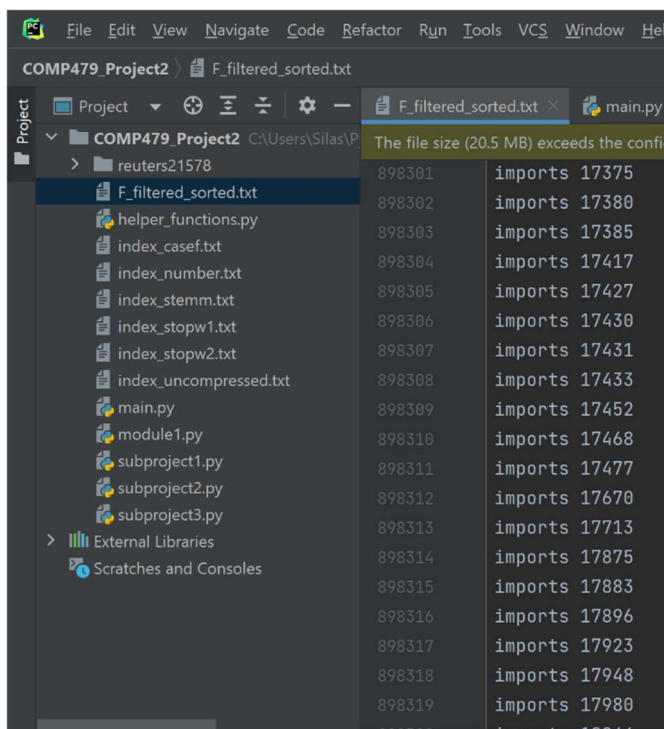
```

compression  dictionary size  difference in percent  cml  postings lists size  difference in percent  cml
unfiltered   76793              0                    0    1582940              0                    0
no numbers   53329              30.56               30.56 1467499              7.29               7.29
case folding 46544              12.72               39.39 1422115              3.09               10.16
30 stopw's   46514              0.06                39.43 1142682              19.65              27.81
150 stopw's  46394              0.32                39.59 870265               38.8               45.02
stemming     36017              22.37               53.1  831166               4.49               47.49

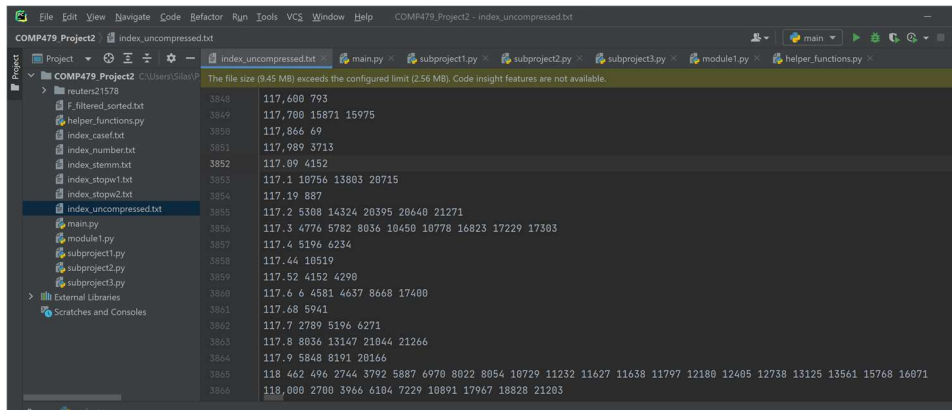
Process finished with exit code 0

```

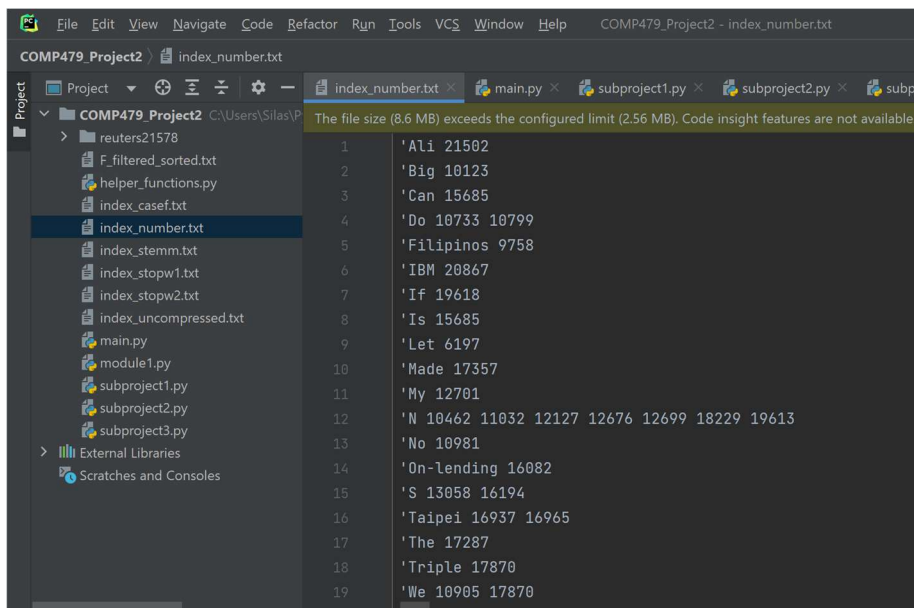
Also, the compression table is printed in the console.



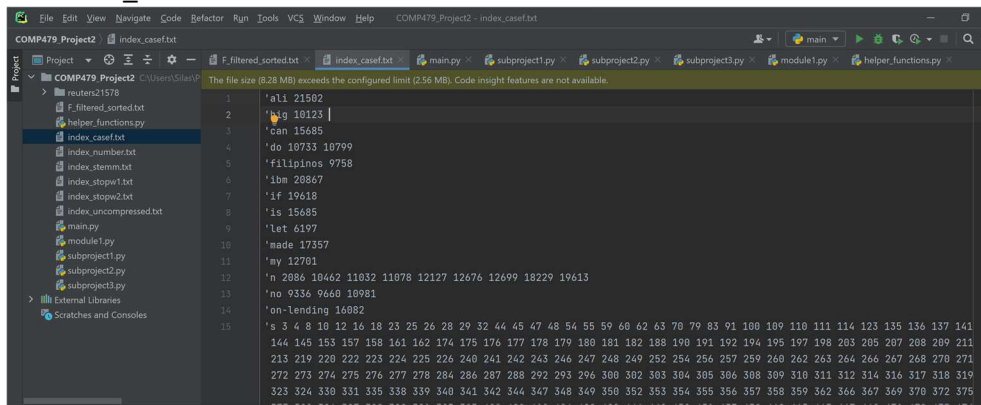
The file F contains all the token-documentID pairs in the format seen in the screenshot.



Index_uncompressed.txt contains the uncompressed index.

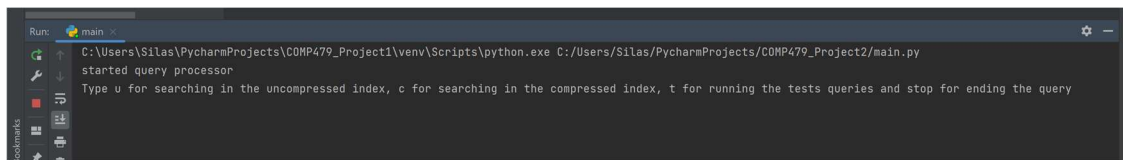


The index_number.txt contains the index where all numbers were removed.



Index_casefolding contains the index where numbers were removed and casefolding was performed on.

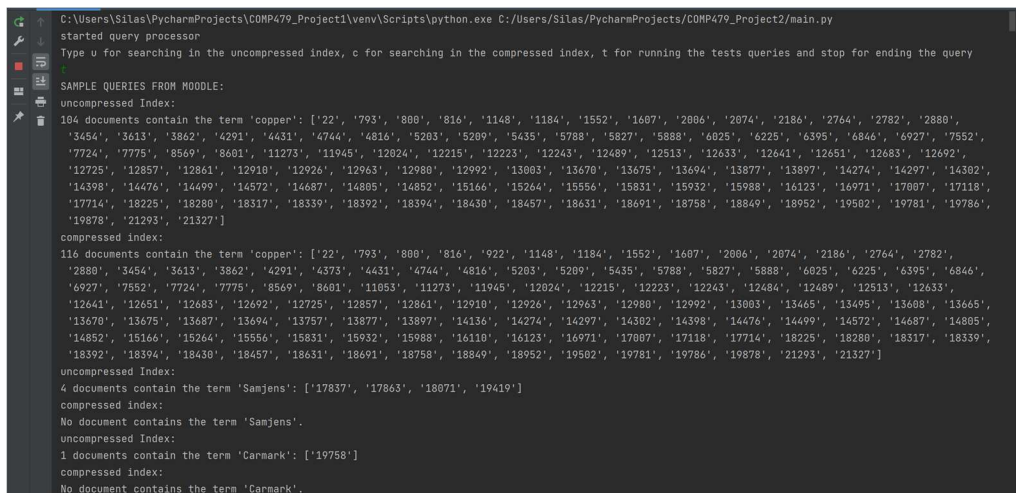
Using the query



```
Run: main
C:\Users\Silas\PycharmProjects\COMP479_Project1\venv\Scripts\python.exe C:\Users\Silas\PycharmProjects\COMP479_Project2\main.py
started query processor
Type u for searching in the uncompressed index, c for searching in the compressed index, t for running the tests queries and stop for ending the query
```

The user just must follow the print in the console, the query starts automatically after a search again and stops when typing stop.

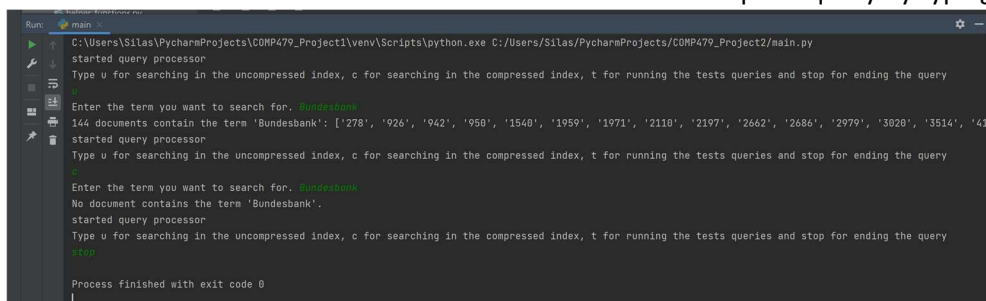
The user can start the sample and test queries by typing t.



```
C:\Users\Silas\PycharmProjects\COMP479_Project1\venv\Scripts\python.exe C:\Users\Silas\PycharmProjects\COMP479_Project2\main.py
started query processor
Type u for searching in the uncompressed index, c for searching in the compressed index, t for running the tests queries and stop for ending the query

SAMPLE QUERIES FROM MOODLE:
uncompressed Index:
184 documents contain the term 'copper': ['22', '793', '880', '816', '1148', '1184', '1552', '1607', '2886', '2874', '2186', '2764', '2782', '2880',
'3454', '3613', '3862', '4291', '4431', '4744', '4816', '5203', '5209', '5435', '5788', '5827', '5888', '6025', '6225', '6395', '6846', '6927', '7552',
'7724', '7775', '8569', '8601', '11273', '11945', '12024', '12215', '12223', '12243', '12489', '12513', '12633', '12641', '12651', '12683', '12692',
'12725', '12857', '12861', '12910', '12926', '12963', '12980', '12992', '13003', '13670', '13675', '13694', '13877', '13897', '14274', '14297', '14302',
'14398', '14476', '14499', '14572', '14687', '14885', '14852', '15166', '15264', '15556', '15831', '15932', '15988', '16123', '16971', '17087', '17118',
'17714', '18225', '18280', '18317', '18339', '18392', '18394', '18430', '18457', '18631', '18691', '18758', '18849', '18952', '19502', '19781', '19786',
'19878', '21293', '21327']
compressed index:
116 documents contain the term 'copper': ['22', '793', '880', '816', '922', '1148', '1184', '1552', '1607', '2006', '2074', '2186', '2764', '2782',
'2880', '3454', '3613', '3862', '4291', '4373', '4431', '4744', '4816', '5203', '5209', '5435', '5788', '5827', '5888', '6025', '6225', '6395', '6846',
'6927', '7552', '7724', '7775', '8569', '8601', '11053', '11273', '11945', '12024', '12215', '12223', '12243', '12484', '12489', '12513', '12633',
'12641', '12651', '12683', '12692', '12725', '12857', '12861', '12910', '12926', '12963', '12980', '12992', '13003', '13665', '13675', '13670', '13675', '13694', '13757', '13877', '13897', '14136', '14274', '14297', '14302', '14398', '14476', '14499', '14572', '14687', '14805',
'14852', '15166', '15264', '15556', '15831', '15932', '15988', '16110', '16123', '16971', '17007', '17118', '17714', '18225', '18280', '18317', '18339',
'18392', '18394', '18430', '18457', '18631', '18691', '18758', '18849', '18952', '19502', '19781', '19786', '19878', '21293', '21327']
uncompressed Index:
4 documents contain the term 'Samjens': ['17837', '17863', '18071', '19419']
compressed index:
No document contains the term 'Samjens'.
uncompressed Index:
1 documents contain the term 'Carmark': ['19758']
compressed index:
No document contains the term 'Carmark'.
```

The user can search for term in one of the indices and can stop the query by typing stop.



```
Run: main
C:\Users\Silas\PycharmProjects\COMP479_Project1\venv\Scripts\python.exe C:\Users\Silas\PycharmProjects\COMP479_Project2\main.py
started query processor
Type u for searching in the uncompressed index, c for searching in the compressed index, t for running the tests queries and stop for ending the query

Enter the term you want to search for. Bundesbank
146 documents contain the term 'Bundesbank': ['278', '926', '942', '950', '1540', '1959', '1971', '2110', '2197', '2662', '2686', '2979', '3020', '3514', '41
started query processor
Type u for searching in the uncompressed index, c for searching in the compressed index, t for running the tests queries and stop for ending the query

Enter the term you want to search for. Bundesbank
No document contains the term 'Bundesbank'.
started query processor
Type u for searching in the uncompressed index, c for searching in the compressed index, t for running the tests queries and stop for ending the query

stop
Process finished with exit code 0
```

I certify that this submission is my original work and meets the Faculty's Expectations of Originality.



Signature

ID: 40256077

Date 2022-10-08