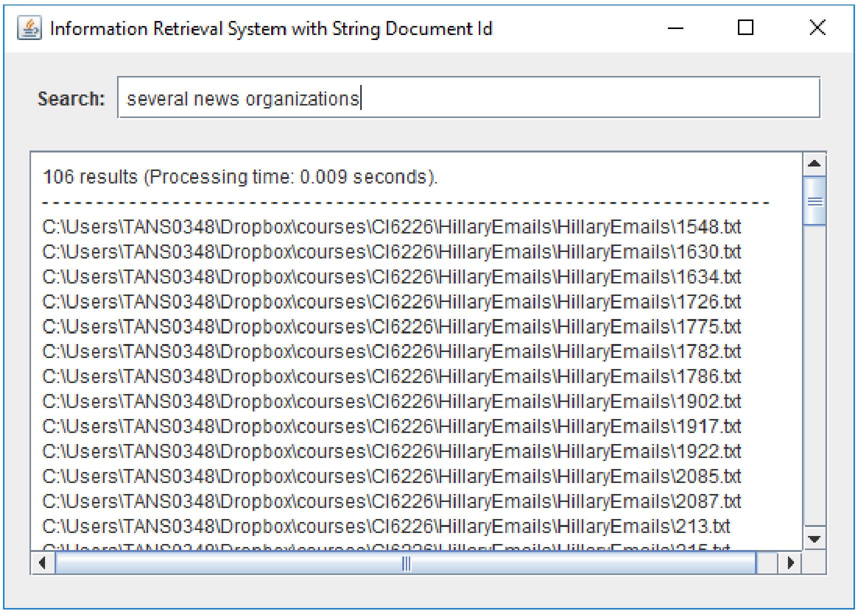
CI-6226 Information Retrieval & Analysis

Group 5: Tan Sang Sang (G1500559E), Sirin Haddad (G1603795G), and Divya Gupta (G1602461E)

1 Introduction

In this project, we built an information retrieval system using Java for retrieving documents from an email corpus of 7,945 emails. Our information retrieval system was created with the simple interface shown in Figure 1. The search function is triggered when a user presses <*Enter*> in the *Search* text field. The system then responds to the search by returning the followings:

- The paths to the files that fulfil the user's search query, in the sense that the files contain all the terms that constitute the query. In other words, the system considers all terms in the user's query as joined by Boolean AND operators.
- Single-line summary that provides quick information about the number of files that fulfil the user's query and the processing time of the query.


Information Retrieval System with String Document Id

Search: several news organizations

```
106 results (Processing time: 0.009 seconds).
----------------------------------------------------------------
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\1548.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\1630.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\1634.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\1726.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\1775.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\1782.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\1786.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\1902.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\1917.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\1922.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\2085.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\2087.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\213.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\215.txt
```

Fig. 1. The information retrieval system

To achieve efficient retrieval, the system first builds an inverted index, and the major steps involved in the construction of the index are described in section 2. Since this system handles multi-word queries as Boolean AND queries, a postings merging algorithm is an essential component in the query processing stage. Section 3 presents the basic postings merging algorithm used in the system. The evaluation of the system—in terms of time and memory requirements for indexing, as well as its query processing speed—is discussed in section 4. Finally, section 5 describes two optimization attempts and provides a comparison of the time and memory for indexing and query processing before and after the optimization steps.

2 Creating an Inverted Index

The four major steps in creating the inverted index are described in detail in subsection 2.1-2.4.

2.1 Tokenization

This tokenization component takes as input the textual contents of the 7,945 emails from the corpus and splits the contents by whitespace characters. The output from this component is a list of <token, document id> pairs, where document id is a String containing the absolute file path to the document. In our implementation, the <token, document id> pairs are stored as objects of a user-defined class called Pair, of which the class structure is shown in Figure 2. That said, the output list is actually a list of Pair objects. For optimal processing, we also ensure that the list only stores non-empty tokens.

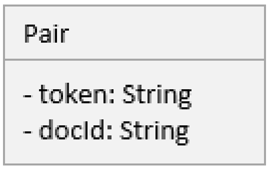


Fig. 2. The class structure of Pair class

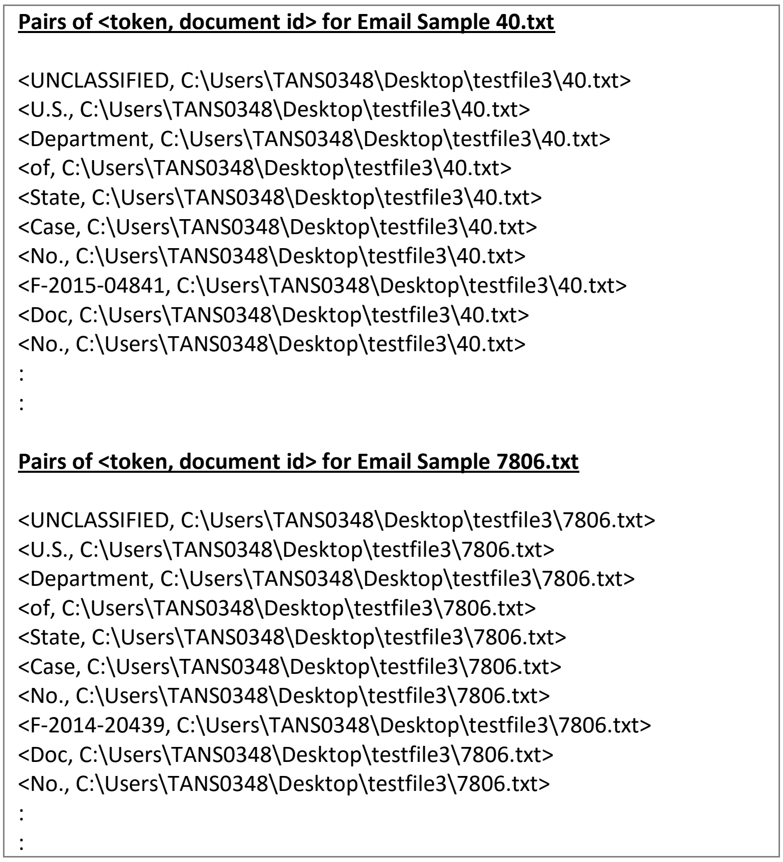
To demonstrate the tokenizer's functionality, we chose two email samples (see Figure 3 and 4) randomly from the corpus. The list of <token, document id> pairs generated from these emails is shown in Figure 5. Note that the same email samples will also be used for demonstration purpose in the following sections that describe other indexing and query processing steps of the system.

```
UNCLASSIFIED
U.S. Department of State
Case No. F-2015-04841
Doc No. C05739607
Date: 05/13/2015
STATE DEPT. - PRODUCED TO HOUSE SELECT BENGHAZI COMM.
SUBJECT TO AGREEMENT ON SENSITIVE INFORMATION & REDACTIONS. NO FOIA WAIVER.
RELEASE IN
FULL
From Mills, Cheryl D <MillsCD@state.gov>
Sent: Friday, September 14, 2012 4:58 PM
To: Marshall, Capricia P; Kennedy, Patrick F
Cc
Subject: You do great work - thanks for making our heros have the homecoming
they deserved.
……
```

Fig. 3. Portion of content for email sample 40.txt

```
UNCLASSIFIED U.S. Department of State Case No. F-2014-20439 Doc No. C05777095
Date: 08/31/2015
RELEASE IN PART
B6
From: Jiloty, Lauren C <JilotyLC@state.gov>
Sent: Saturday, December 4, 2010 8:55 AM
To: H; monica.hanle
Cc: Huma Abedin
Subject: Re: Followup
Ok
……
```

Fig. 4. Portion of content for email sample 7806.txt



```
Pairs of <token, document id> for Email Sample 40.txt

<UNCLASSIFIED, C:\Users\TANS0348\Desktop\testfile3\40.txt>
<U.S., C:\Users\TANS0348\Desktop\testfile3\40.txt>
<Department, C:\Users\TANS0348\Desktop\testfile3\40.txt>
<of, C:\Users\TANS0348\Desktop\testfile3\40.txt>
<State, C:\Users\TANS0348\Desktop\testfile3\40.txt>
<Case, C:\Users\TANS0348\Desktop\testfile3\40.txt>
<No., C:\Users\TANS0348\Desktop\testfile3\40.txt>
<F-2015-04841, C:\Users\TANS0348\Desktop\testfile3\40.txt>
<Doc, C:\Users\TANS0348\Desktop\testfile3\40.txt>
<No., C:\Users\TANS0348\Desktop\testfile3\40.txt>
:
:

Pairs of <token, document id> for Email Sample 7806.txt

<UNCLASSIFIED, C:\Users\TANS0348\Desktop\testfile3\7806.txt>
<U.S., C:\Users\TANS0348\Desktop\testfile3\7806.txt>
<Department, C:\Users\TANS0348\Desktop\testfile3\7806.txt>
<of, C:\Users\TANS0348\Desktop\testfile3\7806.txt>
<State, C:\Users\TANS0348\Desktop\testfile3\7806.txt>
<Case, C:\Users\TANS0348\Desktop\testfile3\7806.txt>
<No., C:\Users\TANS0348\Desktop\testfile3\7806.txt>
<F-2014-20439, C:\Users\TANS0348\Desktop\testfile3\7806.txt>
<Doc, C:\Users\TANS0348\Desktop\testfile3\7806.txt>
<No., C:\Users\TANS0348\Desktop\testfile3\7806.txt>
:
:
```

Fig. 5. List of <token, document id> pairs generated for email samples 40.txt and 7806.txt

2.2 Linguistic Module

The output produced by the tokenization component described above was then processed by a linguistic module component that modifies each token in the <token, document id> pair to generate a list of <modified token, document id> pairs. This linguistic module component puts the tokens through the following steps:

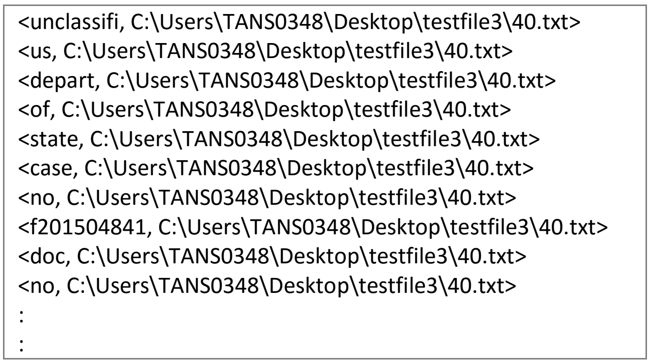
- Removes all non-alphanumeric characters, keeping only numbers and letters, i.e., A-Z, a-z, and 0-9.
- Converts the tokens to lower case.
- Uses Porter-stemmer[1] to stem the tokens.

Figure 6 shows examples of modified tokens generated by the linguistic module component after applying the three processing steps to the tokens obtained from the email samples.

2.3 Sorting the Tokens

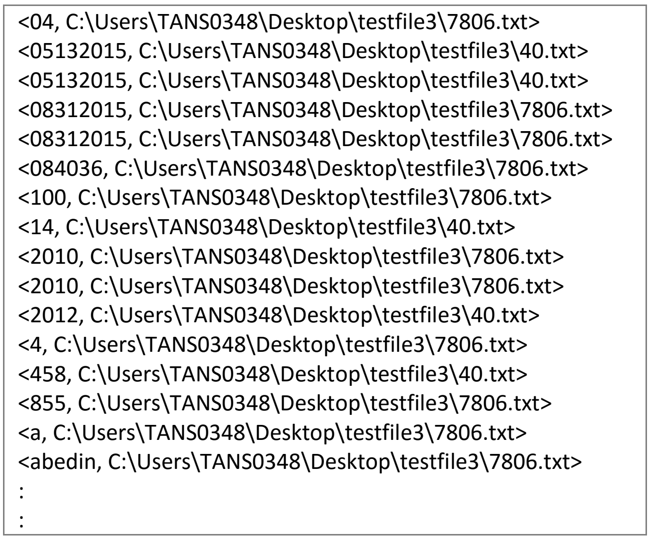
The list of <modified token, document id> pairs is sorted by two orders: first, by alphabetical order of the tokens; second, by alphabetical order of the document ids (i.e., the documents' absolute file paths). This sorting step is achieved in Java by implementing the *Comparable* interface and overriding the *compareTo* method in the Pair class we defined. Part of the sorted list for the email samples is shown in Figure 7.

[1] https://github.com/caarmen/porter-stemmer



```
<unclassifi, C:\Users\TANS0348\Desktop\testfile3\40.txt>
<us, C:\Users\TANS0348\Desktop\testfile3\40.txt>
<depart, C:\Users\TANS0348\Desktop\testfile3\40.txt>
<of, C:\Users\TANS0348\Desktop\testfile3\40.txt>
<state, C:\Users\TANS0348\Desktop\testfile3\40.txt>
<case, C:\Users\TANS0348\Desktop\testfile3\40.txt>
<no, C:\Users\TANS0348\Desktop\testfile3\40.txt>
<f201504841, C:\Users\TANS0348\Desktop\testfile3\40.txt>
<doc, C:\Users\TANS0348\Desktop\testfile3\40.txt>
<no, C:\Users\TANS0348\Desktop\testfile3\40.txt>
:
:
```

Fig. 6. Examples of <modified token, document id> pairs for email sample 40.txt

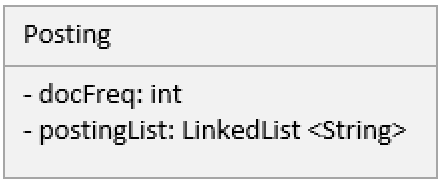


```
<04, C:\Users\TANS0348\Desktop\testfile3\7806.txt>
<05132015, C:\Users\TANS0348\Desktop\testfile3\40.txt>
<05132015, C:\Users\TANS0348\Desktop\testfile3\40.txt>
<08312015, C:\Users\TANS0348\Desktop\testfile3\7806.txt>
<08312015, C:\Users\TANS0348\Desktop\testfile3\7806.txt>
<084036, C:\Users\TANS0348\Desktop\testfile3\7806.txt>
<100, C:\Users\TANS0348\Desktop\testfile3\7806.txt>
<14, C:\Users\TANS0348\Desktop\testfile3\40.txt>
<2010, C:\Users\TANS0348\Desktop\testfile3\7806.txt>
<2010, C:\Users\TANS0348\Desktop\testfile3\7806.txt>
<2012, C:\Users\TANS0348\Desktop\testfile3\40.txt>
<4, C:\Users\TANS0348\Desktop\testfile3\7806.txt>
<458, C:\Users\TANS0348\Desktop\testfile3\40.txt>
<855, C:\Users\TANS0348\Desktop\testfile3\7806.txt>
<a, C:\Users\TANS0348\Desktop\testfile3\7806.txt>
<abedin, C:\Users\TANS0348\Desktop\testfile3\7806.txt>
:
:
```

Fig. 7. Sorted list of pairs for email samples 40.txt and 7806.txt

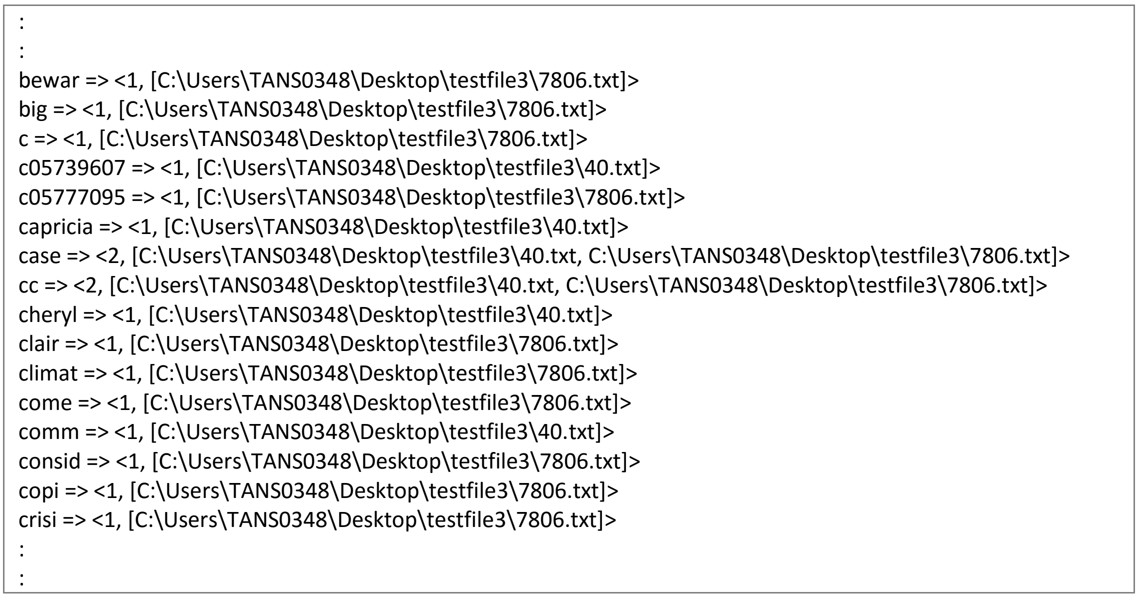
2.4 Transformation into Postings

In this step, we iterate through the sorted list, merge the duplicated document ids for every token, and coalesce the list entries by their tokens. The result is a hash table where every unique term is the key of an entry in the hash table, and each term is mapped to an object of a user-defined class called Posting in our implementation. The class structure of the Posting class and examples of entries in the hash table are shown in Figure 8 and 9 respectively. Each object of the Posting class is composed of the document frequency of the term and the postings list (i.e., list of document ids) of the term.



```
Posting
- docFreq: int
- postingList: LinkedList <String>
```

Fig. 8. The class structure of Posting class



```
:
:
bewar => <1, [C:\Users\TANS0348\Desktop\testfile3\7806.txt]>
big => <1, [C:\Users\TANS0348\Desktop\testfile3\7806.txt]>
c => <1, [C:\Users\TANS0348\Desktop\testfile3\7806.txt]>
c05739607 => <1, [C:\Users\TANS0348\Desktop\testfile3\40.txt]>
c05777095 => <1, [C:\Users\TANS0348\Desktop\testfile3\7806.txt]>
capricia => <1, [C:\Users\TANS0348\Desktop\testfile3\40.txt]>
case => <2, [C:\Users\TANS0348\Desktop\testfile3\40.txt, C:\Users\TANS0348\Desktop\testfile3\7806.txt]>
cc => <2, [C:\Users\TANS0348\Desktop\testfile3\40.txt, C:\Users\TANS0348\Desktop\testfile3\7806.txt]>
cheryl => <1, [C:\Users\TANS0348\Desktop\testfile3\40.txt]>
clair => <1, [C:\Users\TANS0348\Desktop\testfile3\7806.txt]>
climat => <1, [C:\Users\TANS0348\Desktop\testfile3\7806.txt]>
come => <1, [C:\Users\TANS0348\Desktop\testfile3\7806.txt]>
comm => <1, [C:\Users\TANS0348\Desktop\testfile3\40.txt]>
consid => <1, [C:\Users\TANS0348\Desktop\testfile3\7806.txt]>
copi => <1, [C:\Users\TANS0348\Desktop\testfile3\7806.txt]>
crisi => <1, [C:\Users\TANS0348\Desktop\testfile3\7806.txt]>
:
:
```

Fig. 9. Examples of entries in the hash table (implemented as LinkedHashMap in Java)

The Java data structure that we chose to store the hash table is LinkedHashMap, for its key-value structure allows efficient retrieval from the hash table using the term as the key. The time complexity of *Get* operation (i.e., retrieving a value using a key) for LinkedHashMap is O(1). In addition, unlike some other data structures for implementing hash table in Java, LinkedHashMap keeps track of the order in which items are added (or accessed). This characteristic of LinkedHashMap allows the iteration order of the hash table to follow the insertion order of items. Some other hash table data structures like HashMap, on the other hand, make no guarantee about the iteration order. Using LinkedHashMap thus gives us more controls and might help to avoid possibly unforeseen issues in the implementation.

3 Postings Lists Merging for Processing Boolean AND Queries

To process Boolean AND queries, a basic merging algorithm was implemented for intersection of postings lists. The code snippet showing the implementation of the algorithm in Java is attached in Appendix A. We acknowledge that there are other more advanced algorithms for merging such as merging with a skip list to allow intersection to be processed in sublinear time instead of in time linear in the sizes of the postings lists. However, for this project, we did not invest additional time and effort in any of these algorithms because the size of the corpus is relatively small, so even with a basic merging algorithm, the system was able to process most if not all queries within reasonable time spans.

This basic merging algorithm takes two postings lists as its inputs, walks through the two postings lists in parallel, and returns the intersection of the lists. The complexity of the merging algorithm is O(m+n), where m and n are the number of entries in the two postings lists that are being merged.

The processing of queries follows the procedures below:

- Pass the query through the same tokenizer and linguistic module used for creating the index.
- For all modified tokens returned by the linguistic module, look them up in the index to get their postings lists. Note that if one of the words in the query does not exist in the index, our

system would stop the look-up process immediately and return empty (or null) result so that no computation is wasted on completing the look-up and intersection steps.

- Get the intersection of these postings lists using the basic merging algorithm.
- The entries (file paths) of the intersection are displayed as search results.

4 System Evaluation

4.1 Time and Memory Requirements for Indexing

The time it took the system to create an index and the memory necessary for holding the index were measured. Having considered that the measured values might fluctuate slightly from one execution to another, we repeated the measurement over five executions (see Table 1). The average time for indexing is 10.704 seconds whereas the average memory requirement is 40.206 megabytes.

Execution	**Time (seconds)**	**Memory (megabytes)**
1	10.903	40.206
2	10.571	40.206
3	10.862	40.206
4	10.760	40.205
5	10.423	40.207
Average	**10.704**	**40.206**

Table 1. Time for indexing and memory for holding the index

4.2 Query Processing Time

In the interest of learning how the system would behave in various scenarios, we put the system through a series of tests with queries that are chosen to test out different cases. Table 2 shows the queries used in the tests with the descriptions of test scenarios for which they are intended. In general, case 1 through 5 examine the relation between query processing time and query length, and they also serve to find out whether the document frequencies of terms in a query might have any effects on the processing time. Case 6 and 7 are chosen to verify whether the system is able to handle, in a decent manner, cases of non-existent or misspelled words and non-alphanumeric characters.

Table 3 summarizes the number of search results and processing time for the seven cases described above. Screenshots of the query processing results are given in Appendix B. From the observations of the test outcomes, it seems the processing time largely depends on the length of queries. This is somehow expected because this system considers all queries as Boolean AND queries, the time spent on intersecting postings lists is thus positively correlated with the length of queries. In other words, the longer the queries, the more time would be spent on merging postings lists of the query words. Therefore, the processing time for single-word queries (case 1 and 2) is apparently smaller than the processing time for multi-word queries (case 3, 4, and 5). Whether the queries consist of common or uncommon words did not seem to affect the processing time in these cases. However, the query processing time for certain multi-word queries can be improved with a simple query optimization step that we will describe in the next section.

Case 6 has demonstrated that the system is able to handle non-existent or misspelled word efficiently. The processing time for case 6 is negligible (close to 0 second) because the system aborted the index look-up process and returned 0 result when the word ‘Egyptt’ was encountered in the query.

In addition, case 7 has verified that queries containing non-alphanumeric characters are also properly handled by the system.

Case No.	**Query**	**Test Case Description**
1	Date	Single word query of a common word. The document frequency of the word ‘date’ is 7944.
2	Gimmick	Single word query of an uncommon word. The document frequency of the word ‘gimmick’ is 1.
3	sent from to	Multi-word query containing only common words. The document frequencies of ‘sent’, ‘from’, and ‘to’ are 7670, 7823, and 7857 respectively.
4	sent from to mushroom	Multi-word query containing common words and uncommon words. The document frequencies of ‘sent’, ‘from’, and ‘to’ are 7670, 7823, and 7857 respectively, whereas the document frequency of ‘mushroom’ is 1.
5	Sources with direct access to the Libyan National Transitional Council, as well as the highest levels of European Governments, and Western Intelligence and security services.	Exceptionally long query.
6	has sparked violence in **Egyptt** and Libya and led to the deaths	Query containing non-existent or misspelled words (i.e., Egyptt).
7	Case No. F-2015-04841	Query containing non-alphanumeric characters.

Table 2. Queries and their intended test scenarios

Case No.	**Query**	**Number of Results**	**Processing Time (seconds)**
1	Date	7944	0
2	Gimmick	1	0
3	sent from to	7665	.119
4	sent from to mushroom	1	.115
5	Sources with direct access to the Libyan National Transitional Council, as well as the highest levels of European Governments, and Western Intelligence and security services.	53	.139
6	has sparked violence in **Egyptt** and Libya and led to the deaths	0	0
7	Case No. F-2015-04841	296	.093

Table 3. Number of search results and processing time case 1-7

5 Optimization

We experimented with two optimization steps:

- Query optimization by processing Boolean AND queries in ascending order of document frequencies (which are also stored with postings lists in the index).
- Optimization of the index by using numeric document ids instead of String document ids.

5.1 Query Optimization with Document Frequencies

To improve the efficiency of query processing, we added a sorting step to the query processing procedure described in section 3 so that the merging of postings lists is carried out in ascending order of document frequencies. For instance, given the multi-word query ‘sent from to mushroom’ in case 4, the intersection of the query terms would be carried in the following order: (((mushroom AND sent) AND from) AND to).

Case 1 through 5 were tested again after this optimization step was incorporated into the system. Table 4 shows the new search results and processing time for these five cases. The screenshots of the query results are attached in Appendix C. This optimization step only affected the processing time for multi-word queries, while the processing time for single-word queries remained unchanged. The improvement was only prominent for case 4, in which the multi-word query contains an uncommon word ‘mushroom’ which has a document frequency of 1. In case 4, the processing time was reduced from 0.115 seconds to 0.002 seconds. However, in cases where all the words in the queries have similar document frequencies, the sorting step did not help the performance. In fact, in case 5, the query processing time increased slightly from 0.139 seconds to 0.148 seconds after the sorting step was introduced.

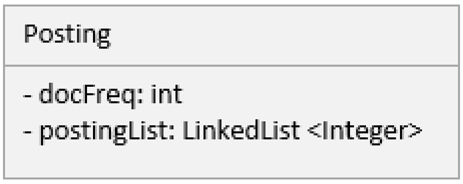
Case No.	Query	Time Before Optimization (seconds)	Time After Optimization (seconds)
1	Date	0	0
2	Gimmick	0	0
3	sent from to	.119	.111
4	sent from to mushroom	.115	.002
5	Sources with direct access to the Libyan National Transitional Council, as well as the highest levels of European Governments, and Western Intelligence and security services.	.139	.148

Table 4. Comparison of processing time before and after query optimization

5.2 Index Optimization with Numeric Document Id

To improve the time and memory requirements for indexing, we used numeric document ids instead of String ids in the postings lists. The class structure of our modified user-defined class (Posting) is shown in Figure 10. Other data structures were also changed accordingly.

Surprisingly, the change from String ids to numeric ids did not yield much improvement in indexing time, and the average memory requirements even increased from 40.206 megabytes to 57.678 megabytes (see Table 5). A plausible explanation is that the *List* data structure in Java cannot store primitive data types. Therefore, the numeric document ids in the postings lists are stored as *Integer* objects, which are wrappers for the primitive *int* data type, and these wrapper objects occupy more memories than the primitive *int* data type.



```
Posting
- docFreq: int
- postingList: LinkedList <Integer>
```

Fig. 10. The modified class structure of Posting class

Execution	Time (seconds)	Memory (megabytes)
1	10.002	57.679
2	9.970	57.678
3	9.790	57.678
4	9.792	57.678
5	9.967	57.678
Average	**9.904**	**57.678**

Table 5. Time for indexing and memory for holding the index after changing to numeric document ids

Case 1 through 7 were tested again after this index optimization step. Table 6 shows the new search results and processing time. The screenshots of the query results are attached in Appendix D. In general, using numeric document ids has increased the query processing time. This degradation in performance is probably due to the additional look-up required to retrieve the file paths that correspond to the numeric document ids.

Case No.	Query	Time Before Optimization (seconds)	Time After Optimization (seconds)
1	date	0	.001
2	gimmick	0	0
3	sent from to	.119	.133
4	sent from to mushroom	.115	.123
5	Sources with direct access to the Libyan National Transitional Council, as well as the highest levels of European Governments, and Western Intelligence and security services.	.139	.138
6	has sparked violence in **Egyptt** and Libya and led to the deaths	0	0
7	Case No. F-2015-04841	.093	.099

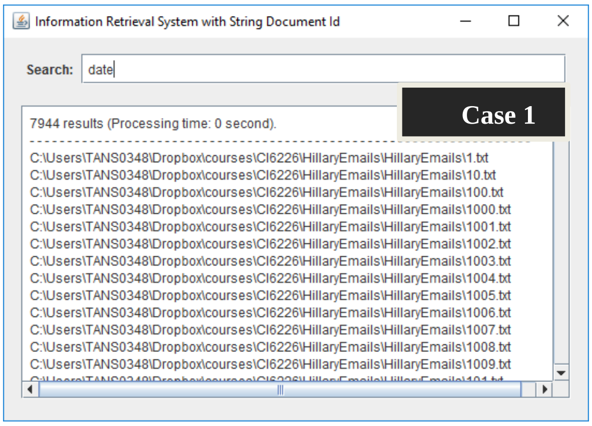
Table 6. Comparison of processing time before and after index optimization

Appendix

Appendix A

```
public static LinkedList<String> intersect(LinkedList<String> p1, LinkedList<String> p2) {
     LinkedList<String> answer = new LinkedList<String>();
     int step1 = 0;
     int step2 = 0;
     int n1 = p1.size();
     int n2 = p2.size();
     while(step1 < n1 && step2 < n2) {
        String docId1 = p1.get(step1);
        String docId2 = p2.get(step2);
        int compareValue = docId1.compareTo(docId2);
        if(compareValue == 0) { //docId1 = docId2
        answer.add(docId1);
        step1++;
        step2++;
        } else if(compareValue < 0) { //docId1 < docId2
        step1++;
        } else { //docId1 > docId2
        step2++;
        }
     }
     return answer;
}
```

Appendix B

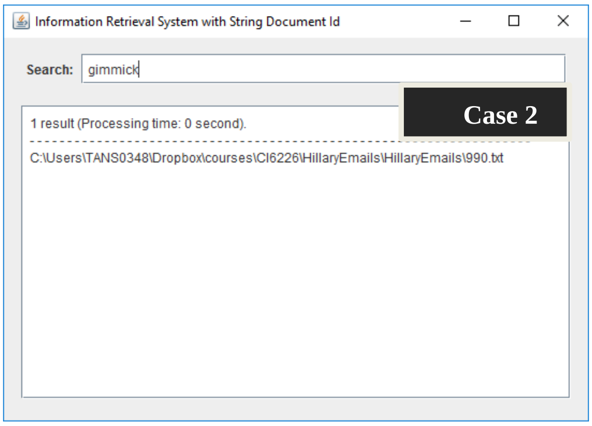


Information Retrieval System with String Document Id

Search: date

Case 1

```
7944 results (Processing time: 0 second).
------------------------------------------------------------------
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\1.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\10.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\100.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\1000.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\1001.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\1002.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\1003.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\1004.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\1005.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\1006.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\1007.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\1008.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\1009.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\101.txt
```

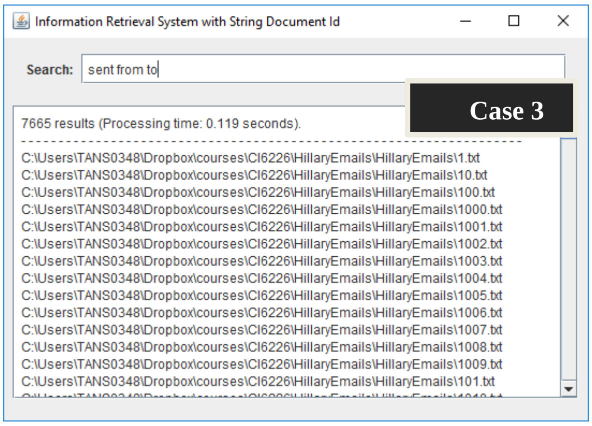


Information Retrieval System with String Document Id

Search: gimmick

Case 2

```
1 result (Processing time: 0 second).
------------------------------------------------------------------
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\990.txt
```

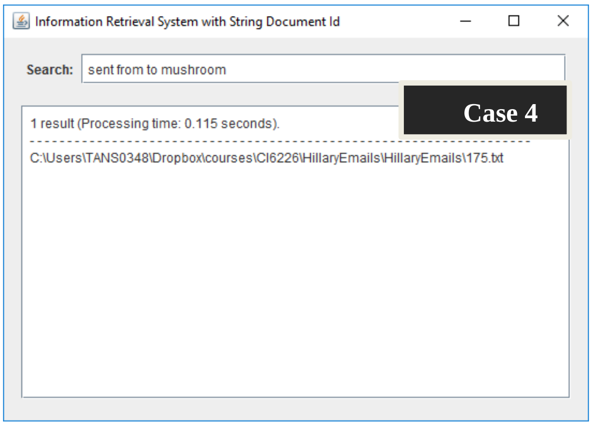


Information Retrieval System with String Document Id

Search: sent from to

Case 3

```
7665 results (Processing time: 0.119 seconds).
------------------------------------------------------------------
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\1.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\10.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\100.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\1000.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\1001.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\1002.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\1003.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\1004.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\1005.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\1006.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\1007.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\1008.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\1009.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\101.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\1010.txt
```

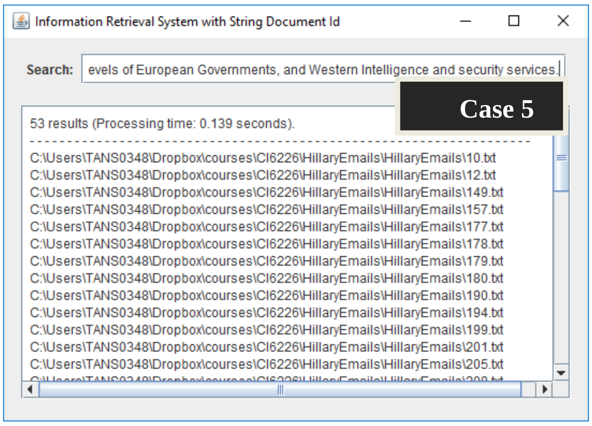


Information Retrieval System with String Document Id

Search: sent from to mushroom

Case 4

```
1 result (Processing time: 0.115 seconds).
------------------------------------------------------------------
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\175.txt
```

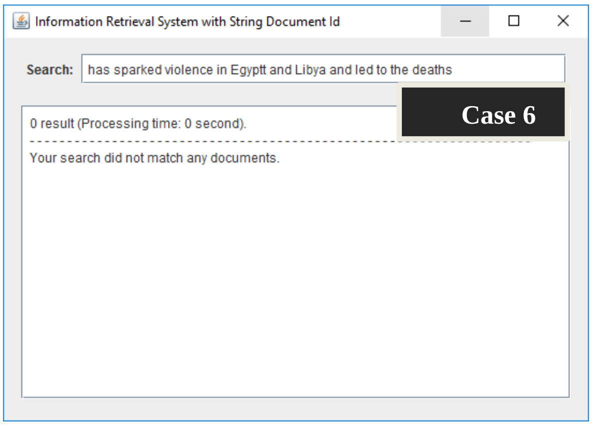


Information Retrieval System with String Document Id

Search: evels of European Governments, and Western Intelligence and security services.

Case 5

```
53 results (Processing time: 0.139 seconds).
------------------------------------------------------------------
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\10.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\12.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\149.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\157.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\177.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\178.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\179.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\180.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\190.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\194.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\199.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\201.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\205.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\209.txt
```

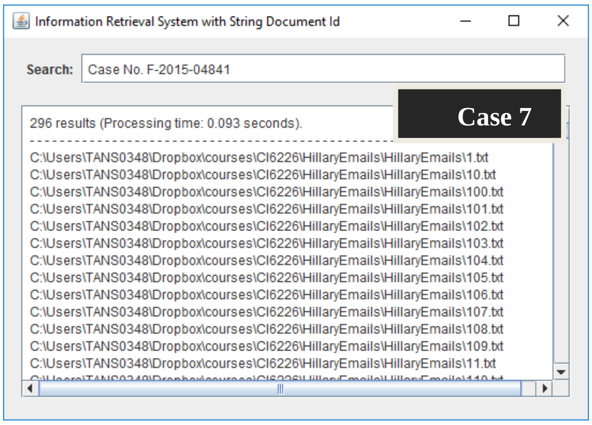


Information Retrieval System with String Document Id

Search: has sparked violence in Egyptt and Libya and led to the deaths

Case 6

```
0 result (Processing time: 0 second).
------------------------------------------------------------------
Your search did not match any documents.
```



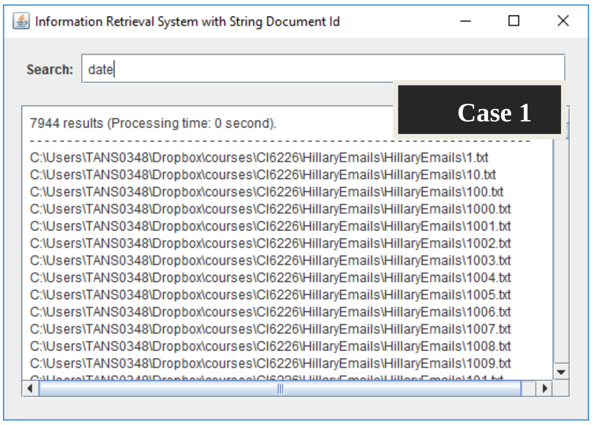
Information Retrieval System with String Document Id

Search: Case No. F-2015-04841

Case 7

```
296 results (Processing time: 0.093 seconds).
------------------------------------------------------------------
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\1.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\10.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\100.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\101.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\102.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\103.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\104.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\105.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\106.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\107.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\108.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\109.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\11.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\110.txt
```

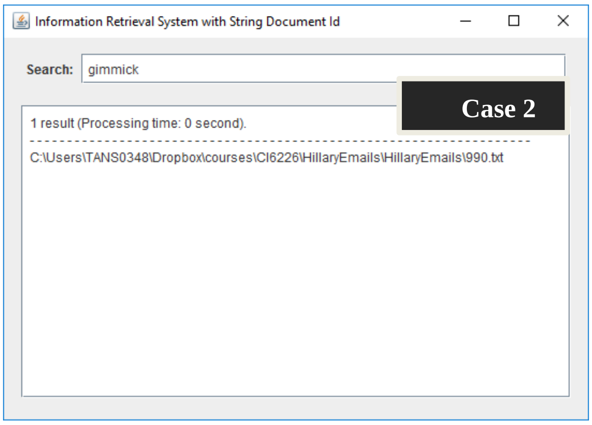
Appendix C



Case 1

Search: date

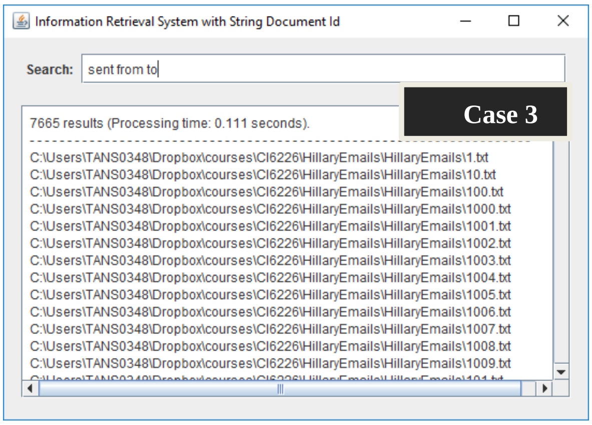
```
7944 results (Processing time: 0 second).
--------------------------------------------------------------------------------
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\1.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\10.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\100.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\1000.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\1001.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\1002.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\1003.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\1004.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\1005.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\1006.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\1007.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\1008.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\1009.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\101.txt
```



Case 2

Search: gimmick

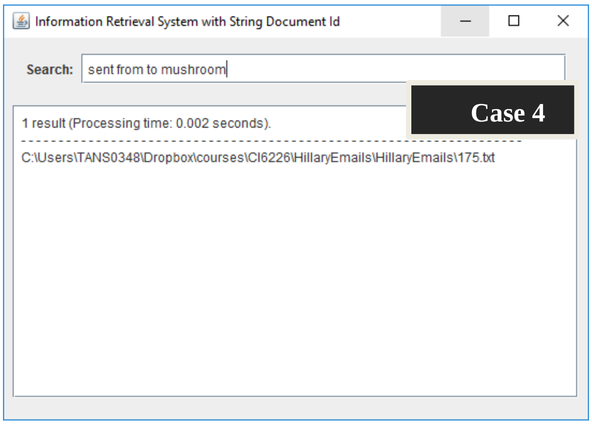
```
1 result (Processing time: 0 second).
--------------------------------------------------------------------------------
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\990.txt
```



Case 3

Search: sent from to

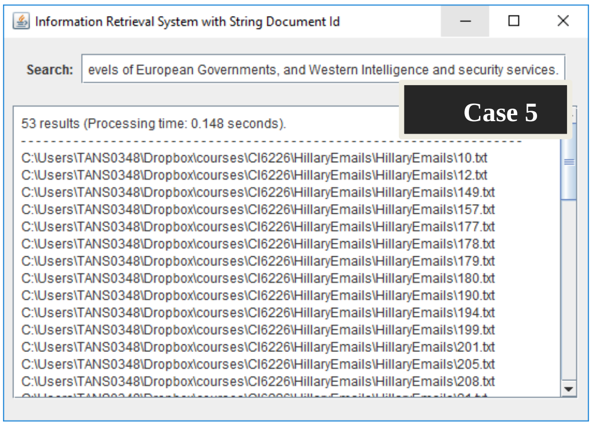
```
7665 results (Processing time: 0.111 seconds).
--------------------------------------------------------------------------------
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\1.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\10.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\100.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\1000.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\1001.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\1002.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\1003.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\1004.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\1005.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\1006.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\1007.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\1008.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\1009.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\101.txt
```



Case 4

Search: sent from to mushroom

```
1 result (Processing time: 0.002 seconds).
--------------------------------------------------------------------------------
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\175.txt
```

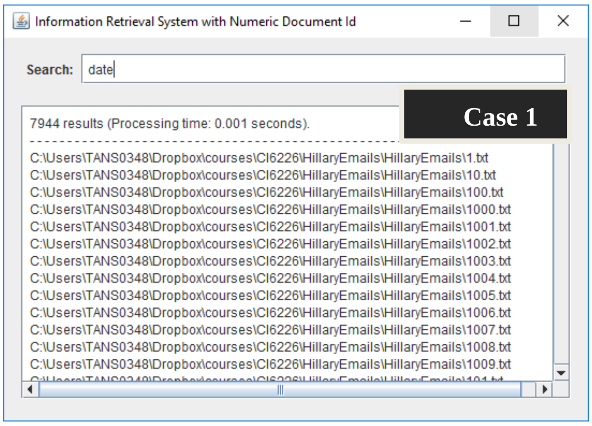


Case 5

Search: evels of European Governments, and Western Intelligence and security services.

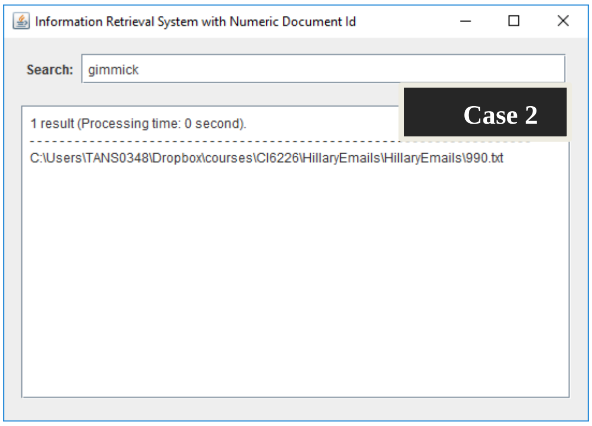
```
53 results (Processing time: 0.148 seconds).
--------------------------------------------------------------------------------
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\10.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\12.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\149.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\157.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\177.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\178.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\179.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\180.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\190.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\194.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\199.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\201.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\205.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\208.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\21.txt
```

Appendix D



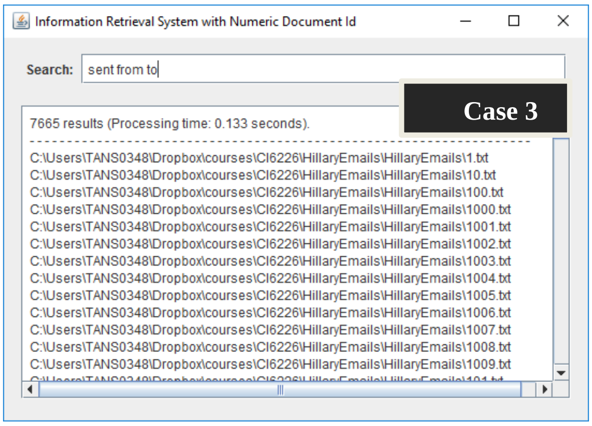
Search: date

```
7944 results (Processing time: 0.001 seconds).
-----------------------------------------------------------
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\1.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\10.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\100.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\1000.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\1001.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\1002.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\1003.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\1004.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\1005.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\1006.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\1007.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\1008.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\1009.txt
```



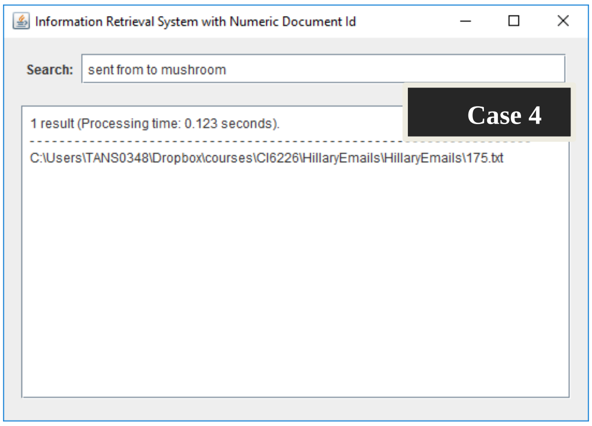
Search: gimmick

```
1 result (Processing time: 0 second).
-----------------------------------------------------------
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\990.txt
```



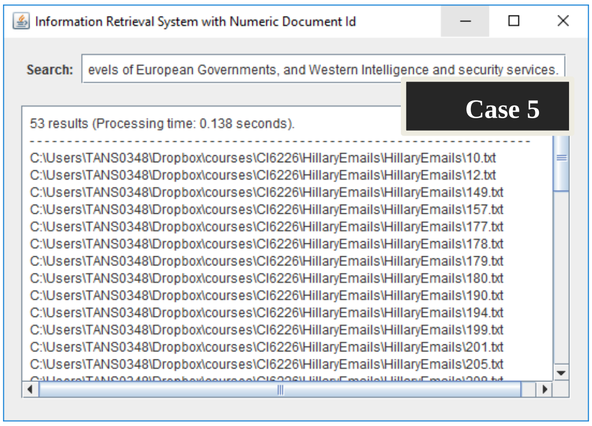
Search: sent from to

```
7665 results (Processing time: 0.133 seconds).
-----------------------------------------------------------
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\1.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\10.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\100.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\1000.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\1001.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\1002.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\1003.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\1004.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\1005.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\1006.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\1007.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\1008.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\1009.txt
```



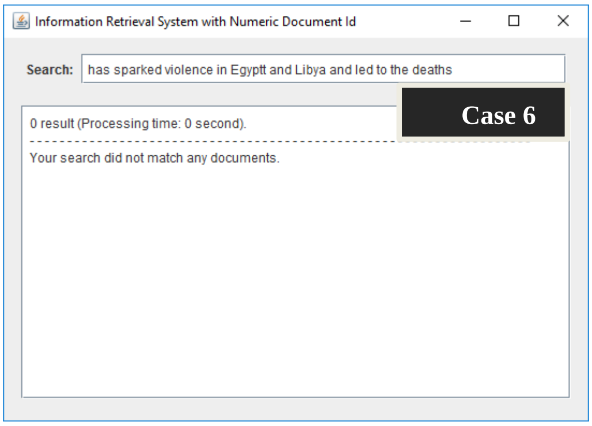
Search: sent from to mushroom

```
1 result (Processing time: 0.123 seconds).
-----------------------------------------------------------
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\175.txt
```



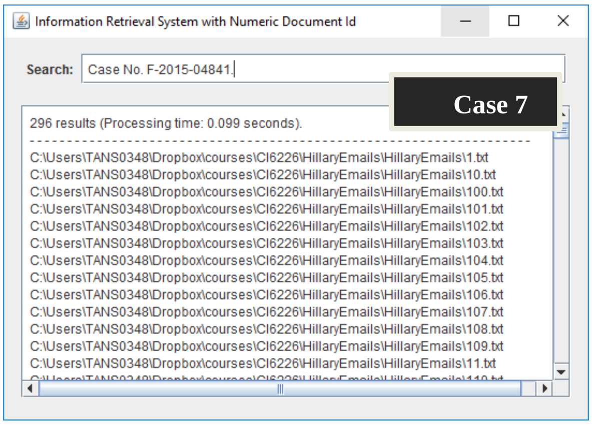
Search: evels of European Governments, and Western Intelligence and security services.

```
53 results (Processing time: 0.138 seconds).
-----------------------------------------------------------
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\10.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\12.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\149.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\157.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\177.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\178.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\179.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\180.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\190.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\194.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\199.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\201.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\205.txt
```



Search: has sparked violence in Egyptt and Libya and led to the deaths

```
0 result (Processing time: 0 second).
-----------------------------------------------------------
Your search did not match any documents.
```



Search: Case No. F-2015-04841.

```
296 results (Processing time: 0.099 seconds).
-----------------------------------------------------------
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\1.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\10.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\100.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\101.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\102.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\103.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\104.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\105.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\106.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\107.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\108.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\109.txt
C:\Users\TANS0348\Dropbox\courses\CI6226\HillaryEmails\HillaryEmails\11.txt
```