UNIVERSITY OF WATERLOO

Faculty of Engineering

Department of Management Sciences

MSCI 541/720 –HW1

Name: Rui Zhang

Student ID: 20599896

Date: January 21st, 2018

Problems (1-3)

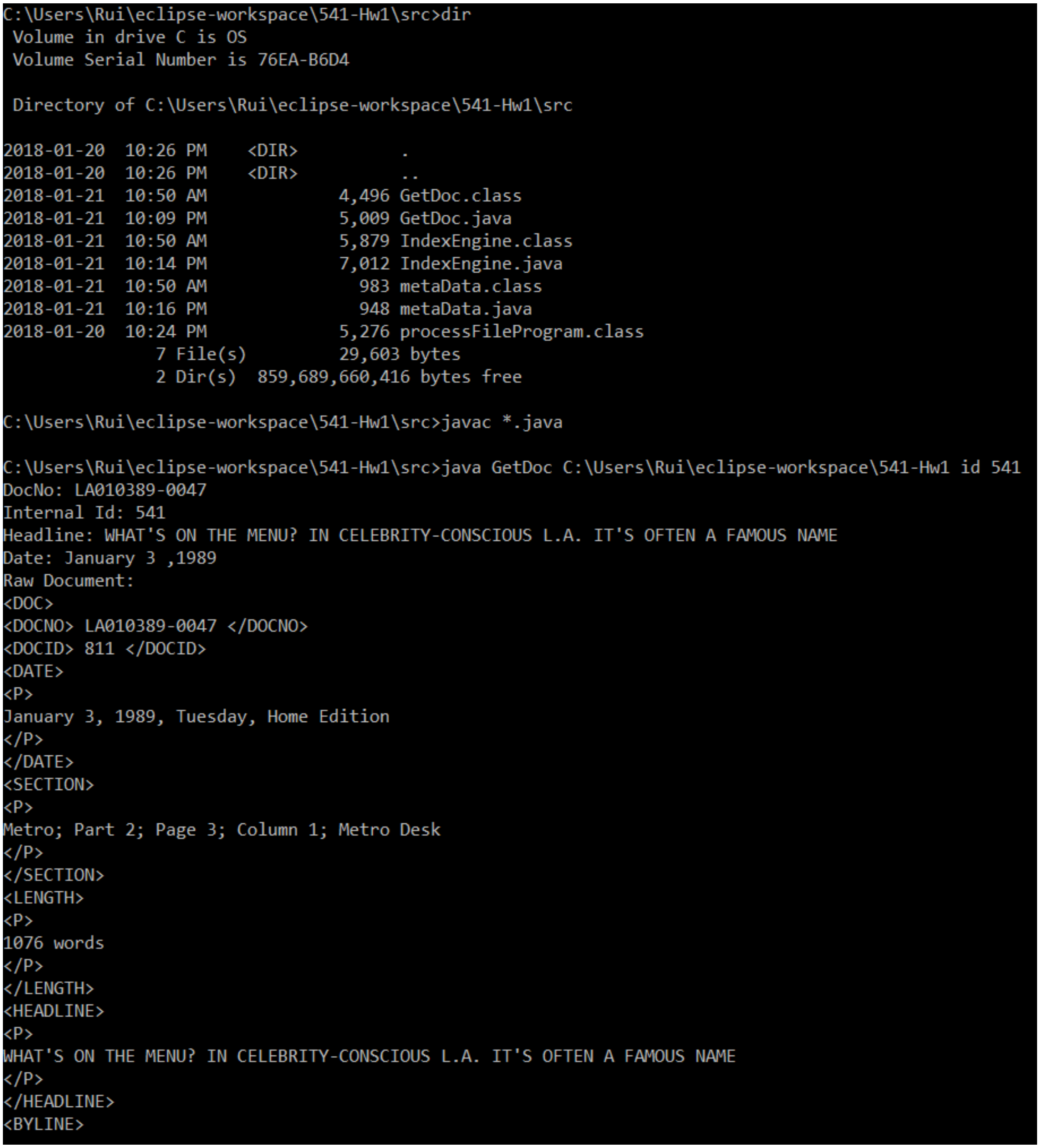
1. A) What is meant by precision enhancing and recall enhancing in context of search engine
    a. To begin, precision in terms of search engine means to retrieve the fraction of items that are relevant to the item being searched. Recall on the other hand is the fraction of items in the collection that were actually being retrieved. For example, if 2 items were retrieved and only one was relevant, then the precision would only be 0.5. If 3 relevant items were in the collection and only one was retrieved, then the recall would be 1/3. Therefore, precision enhancing means to try and maximize the number of relevant items that are already retrieved. Recall enhancing means to only retrieve the relevant items available in the collection.
    b. In the case of an alphabetical, subject card catalog, if users were to search up the catalog of Food and recipes, a precision enhancing technique would be to increase the number of items retrieved such that every item had the word "Food" or other words that go under the category of food and recipes but negates the items that use the words for other purposes. A recall enhancing technique would be to load up every single item that contained the words food and recipes.
2.
    a. Databases and search engines are similar due to their capability in their own way of organizing data. Databases for example would use tables and columns whereas search engines would index the data. However, when a user interacts with a database, the database would always output the exact answer of what the user is looking for. As a result, the user must also know exactly what they're looking for. As for search engines, the users may or may not know what they're looking and therefore the engine would output the relevant information that the users may or may not need. To dive deeper into the technical side, relational databases have tables that use foreign keys to connect between on another if the tables have a connection. In search engines, the information among them are not linked in anyway and it's the search engine's responsibility to determine whether there is a relationship or not.
3. A disadvantage of down casing all terms in a search engine is that the words that could be referred to as a company or a name is down cased such that they could just mean ordinary English names. An example would be the company Apple, if a user was to search up Apple without the down casing, functionality, he or she will get the result that was wanted about the company. However, if the down casing functionality was activated, then the user would only be prompted with the fruit apple. An advantage of down casing items would be vice versa of what was described above. For example, if users were searching up the word Lamps, the results would not return items related to "lamps" because "Lamps" and "lamps" are different from the capital "L" at the front. Therefore, down casing can result in the correct output of "lamps" if a user searching "Lamps".

Problem 4:

Test plan:

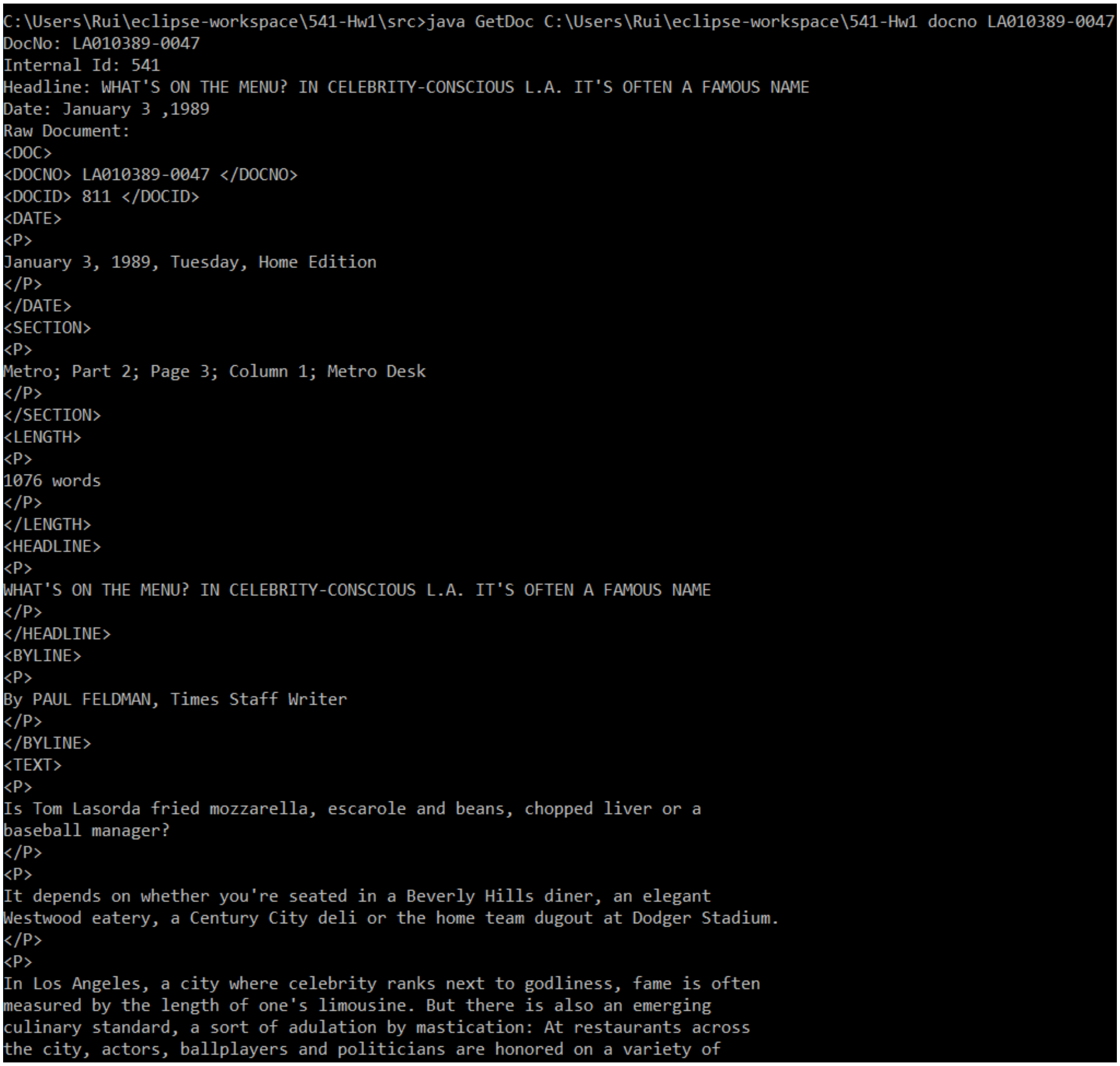
Test Case #	Description	Compliance
1	Using command line to activate the GetDoc class and search using internal id: 541	The meta data and raw document appears for that specific internal id
2	Using command line to activate the GetDoc class and search using **DOCNO: LA010389-0047**	**The raw document with the DOCID of 811 shows up** along with the corresponding raw text
3	While activating the Index engine class, the user did not input any arguments	System should stop and help the user to what should be put in
4	While activating the GetDoc engine class, the user did not input any arguments	System should stop and help the user to what should be put in
5	User inputs an invalid Docno in GetDoc	System should exit and throw a string out of bounds error
6	User inputs an invalid Docno but with the same length as a valid Docno	System should exit and throw a file not found error
7	User inputs an invalid internal id	System should exit and throw a number format error
8	**User activates the Index Engine but the directory already exists**	System should exit and tell the user that the directory already exists
9	User activating the Index Engine without the an existing final directory	Program should start reading the gzip file, process the data, and make files
10	User inputs wrong url address arguments for the paths for Index Engine	Program should stop and throw a file not found error

Test Case 1.



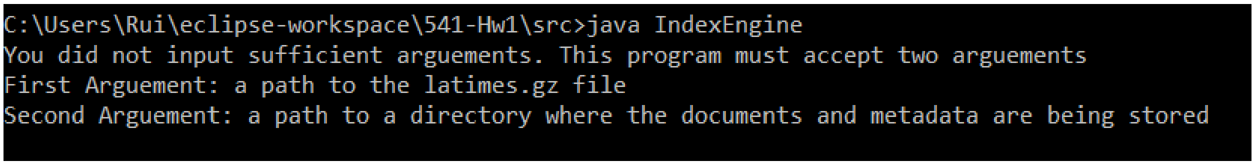
```
C:\Users\Rui\eclipse-workspace\541-Hw1\src>dir
 Volume in drive C is OS
 Volume Serial Number is 76EA-B6D4

 Directory of C:\Users\Rui\eclipse-workspace\541-Hw1\src

2018-01-20  10:26 PM    <DIR>          .
2018-01-20  10:26 PM    <DIR>          ..
2018-01-21  10:50 AM             4,496 GetDoc.class
2018-01-21  10:09 PM             5,009 GetDoc.java
2018-01-21  10:50 AM             5,879 IndexEngine.class
2018-01-21  10:14 PM             7,012 IndexEngine.java
2018-01-21  10:50 AM               983 metaData.class
2018-01-21  10:16 PM               948 metaData.java
2018-01-20  10:24 PM             5,276 processFileProgram.class
               7 File(s)         29,603 bytes
               2 Dir(s)  859,689,660,416 bytes free

C:\Users\Rui\eclipse-workspace\541-Hw1\src>javac *.java

C:\Users\Rui\eclipse-workspace\541-Hw1\src>java GetDoc C:\Users\Rui\eclipse-workspace\541-Hw1 id 541
DocNo: LA010389-0047
Internal Id: 541
Headline: WHAT'S ON THE MENU? IN CELEBRITY-CONSCIOUS L.A. IT'S OFTEN A FAMOUS NAME
Date: January 3 ,1989
Raw Document:
<DOC>
<DOCNO> LA010389-0047 </DOCNO>
<DOCID> 811 </DOCID>
<DATE>
<P>
January 3, 1989, Tuesday, Home Edition
</P>
</DATE>
<SECTION>
<P>
Metro; Part 2; Page 3; Column 1; Metro Desk
</P>
</SECTION>
<LENGTH>
<P>
1076 words
</P>
</LENGTH>
<HEADLINE>
<P>
WHAT'S ON THE MENU? IN CELEBRITY-CONSCIOUS L.A. IT'S OFTEN A FAMOUS NAME
</P>
</HEADLINE>
<BYLINE>
```

Test case 2:



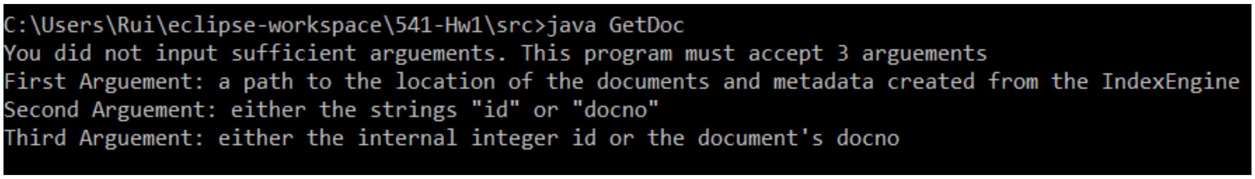
```
C:\Users\Rui\eclipse-workspace\541-Hw1\src>java GetDoc C:\Users\Rui\eclipse-workspace\541-Hw1 docno LA010389-0047
DocNo: LA010389-0047
Internal Id: 541
Headline: WHAT'S ON THE MENU? IN CELEBRITY-CONSCIOUS L.A. IT'S OFTEN A FAMOUS NAME
Date: January 3 ,1989
Raw Document:
<DOC>
<DOCNO> LA010389-0047 </DOCNO>
<DOCID> 811 </DOCID>
<DATE>
<P>
January 3, 1989, Tuesday, Home Edition
</P>
</DATE>
<SECTION>
<P>
Metro; Part 2; Page 3; Column 1; Metro Desk
</P>
</SECTION>
<LENGTH>
<P>
1076 words
</P>
</LENGTH>
<HEADLINE>
<P>
WHAT'S ON THE MENU? IN CELEBRITY-CONSCIOUS L.A. IT'S OFTEN A FAMOUS NAME
</P>
</HEADLINE>
<BYLINE>
<P>
By PAUL FELDMAN, Times Staff Writer
</P>
</BYLINE>
<TEXT>
<P>
Is Tom Lasorda fried mozzarella, escarole and beans, chopped liver or a
baseball manager?
</P>
<P>
It depends on whether you're seated in a Beverly Hills diner, an elegant
Westwood eatery, a Century City deli or the home team dugout at Dodger Stadium.
</P>
<P>
In Los Angeles, a city where celebrity ranks next to godliness, fame is often
measured by the length of one's limousine. But there is also an emerging
culinary standard, a sort of adulation by mastication: At restaurants across
the city, actors, ballplayers and politicians are honored on a variety of
```

Test case 3:



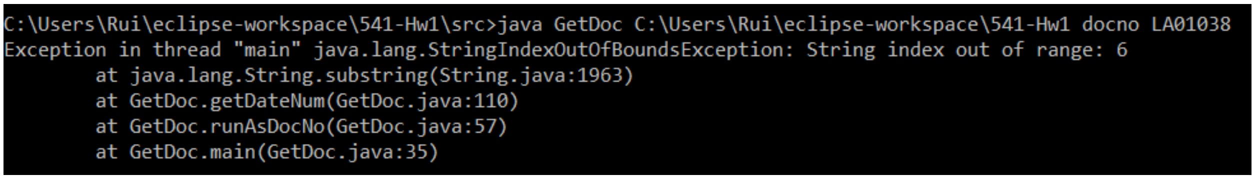
```
C:\Users\Rui\eclipse-workspace\541-Hw1\src>java IndexEngine
You did not input sufficient arguements. This program must accept two arguements
First Arguement: a path to the latimes.gz file
Second Arguement: a path to a directory where the documents and metadata are being stored
```

Test case 4:



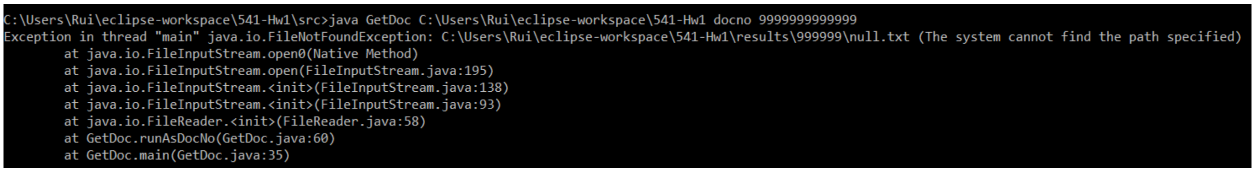
```
C:\Users\Rui\eclipse-workspace\541-Hw1\src>java GetDoc
You did not input sufficient arguements. This program must accept 3 arguements
First Arguement: a path to the location of the documents and metadata created from the IndexEngine
Second Arguement: either the strings "id" or "docno"
Third Arguement: either the internal integer id or the document's docno
```

Test case 5:



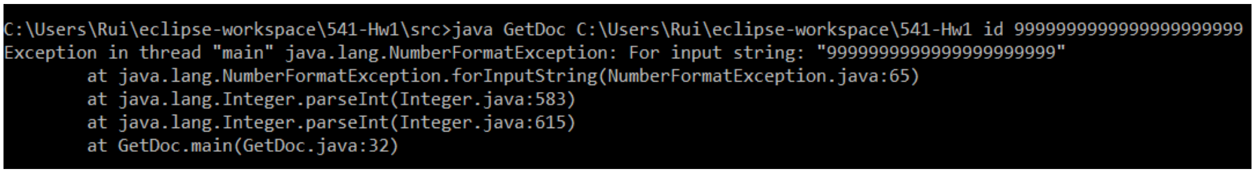
```
C:\Users\Rui\eclipse-workspace\541-Hw1\src>java GetDoc C:\Users\Rui\eclipse-workspace\541-Hw1 docno LA01038
Exception in thread "main" java.lang.StringIndexOutOfBoundsException: String index out of range: 6
        at java.lang.String.substring(String.java:1963)
        at GetDoc.getDateNum(GetDoc.java:110)
        at GetDoc.runAsDocNo(GetDoc.java:57)
        at GetDoc.main(GetDoc.java:35)
```

Test case 6:



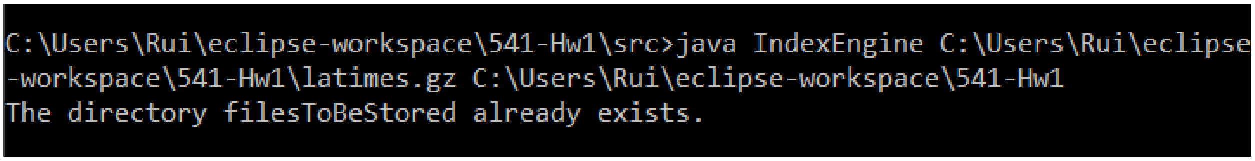
```
C:\Users\Rui\eclipse-workspace\541-Hw1\src>java GetDoc C:\Users\Rui\eclipse-workspace\541-Hw1 docno 9999999999999
Exception in thread "main" java.io.FileNotFoundException: C:\Users\Rui\eclipse-workspace\541-Hw1\results\999999\null.txt (The system cannot find the path specified)
        at java.io.FileInputStream.open0(Native Method)
        at java.io.FileInputStream.open(FileInputStream.java:195)
        at java.io.FileInputStream.<init>(FileInputStream.java:138)
        at java.io.FileInputStream.<init>(FileInputStream.java:93)
        at java.io.FileReader.<init>(FileReader.java:58)
        at GetDoc.runAsDocNo(GetDoc.java:60)
        at GetDoc.main(GetDoc.java:35)
```

Test case 7:



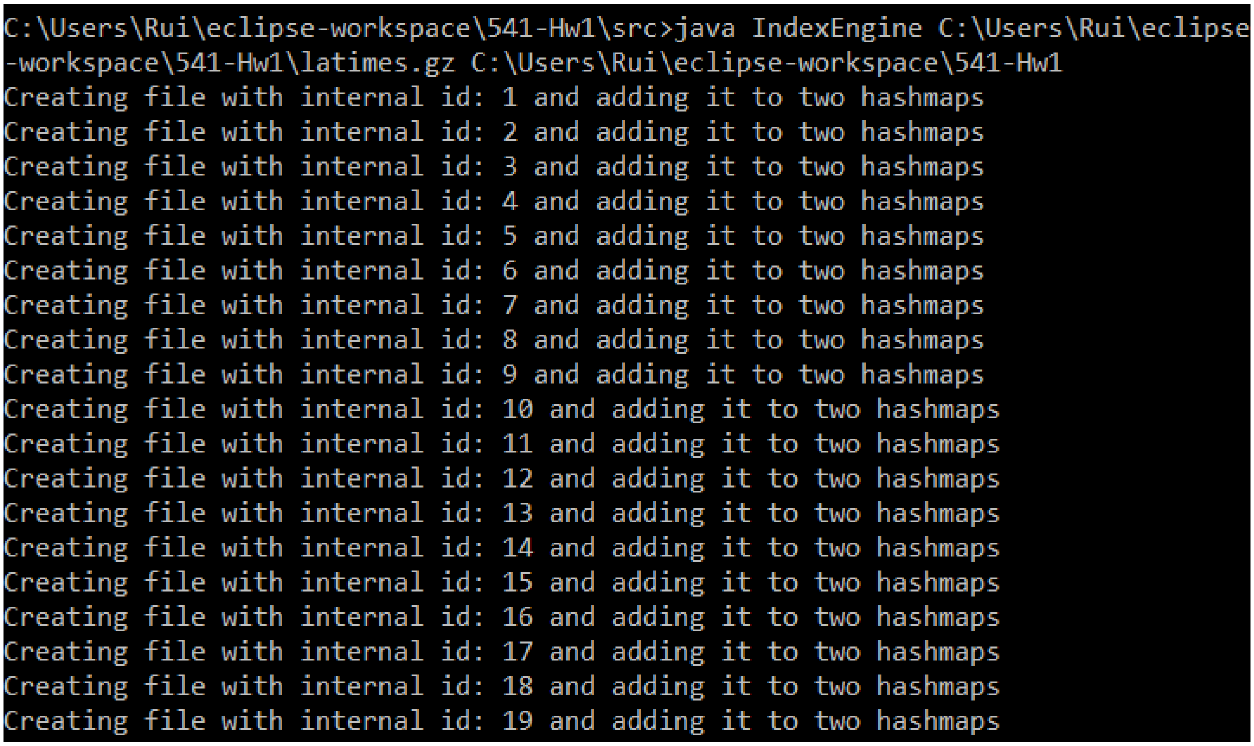
```
C:\Users\Rui\eclipse-workspace\541-Hw1\src>java GetDoc C:\Users\Rui\eclipse-workspace\541-Hw1 id 9999999999999999999999
Exception in thread "main" java.lang.NumberFormatException: For input string: "9999999999999999999999"
        at java.lang.NumberFormatException.forInputString(NumberFormatException.java:65)
        at java.lang.Integer.parseInt(Integer.java:583)
        at java.lang.Integer.parseInt(Integer.java:615)
        at GetDoc.main(GetDoc.java:32)
```

Test case 8:



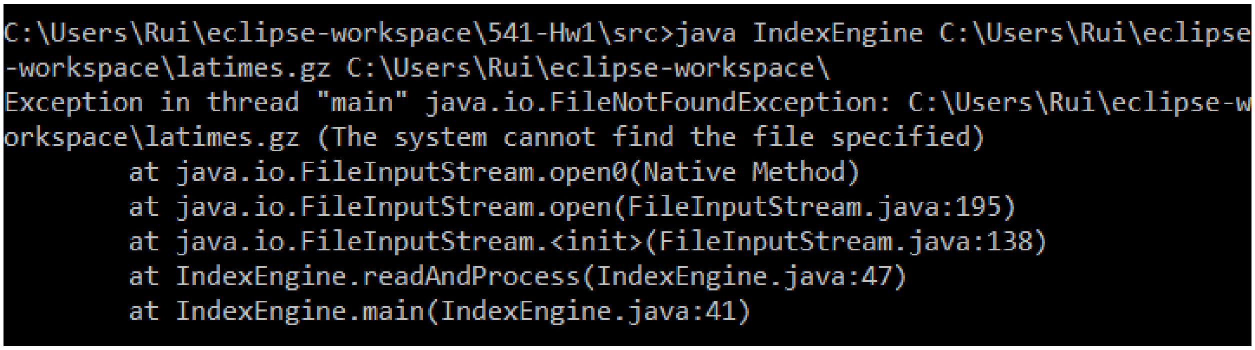
```
C:\Users\Rui\eclipse-workspace\541-Hw1\src>java IndexEngine C:\Users\Rui\eclipse
-workspace\541-Hw1\latimes.gz C:\Users\Rui\eclipse-workspace\541-Hw1
The directory filesToBeStored already exists.
```

Test case 9:



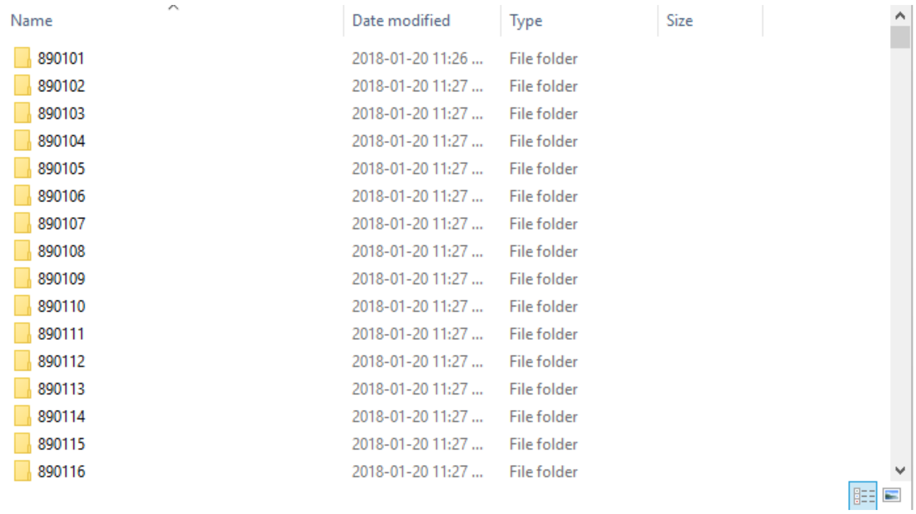
```
C:\Users\Rui\eclipse-workspace\541-Hw1\src>java IndexEngine C:\Users\Rui\eclipse
-workspace\541-Hw1\latimes.gz C:\Users\Rui\eclipse-workspace\541-Hw1
Creating file with internal id: 1 and adding it to two hashmaps
Creating file with internal id: 2 and adding it to two hashmaps
Creating file with internal id: 3 and adding it to two hashmaps
Creating file with internal id: 4 and adding it to two hashmaps
Creating file with internal id: 5 and adding it to two hashmaps
Creating file with internal id: 6 and adding it to two hashmaps
Creating file with internal id: 7 and adding it to two hashmaps
Creating file with internal id: 8 and adding it to two hashmaps
Creating file with internal id: 9 and adding it to two hashmaps
Creating file with internal id: 10 and adding it to two hashmaps
Creating file with internal id: 11 and adding it to two hashmaps
Creating file with internal id: 12 and adding it to two hashmaps
Creating file with internal id: 13 and adding it to two hashmaps
Creating file with internal id: 14 and adding it to two hashmaps
Creating file with internal id: 15 and adding it to two hashmaps
Creating file with internal id: 16 and adding it to two hashmaps
Creating file with internal id: 17 and adding it to two hashmaps
Creating file with internal id: 18 and adding it to two hashmaps
Creating file with internal id: 19 and adding it to two hashmaps
```

Test case 10:

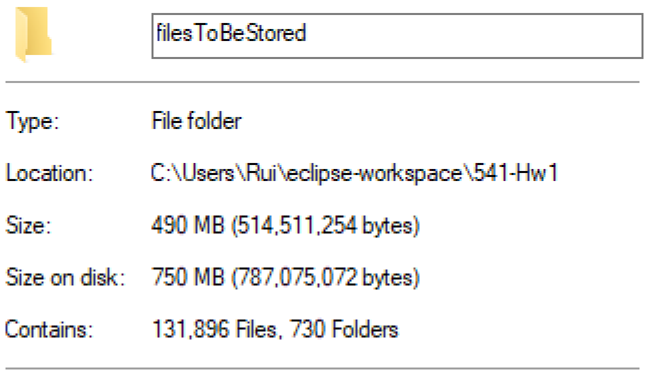


```
C:\Users\Rui\eclipse-workspace\541-Hw1\src>java IndexEngine C:\Users\Rui\eclipse
-workspace\latimes.gz C:\Users\Rui\eclipse-workspace\
Exception in thread "main" java.io.FileNotFoundException: C:\Users\Rui\eclipse-w
orkspace\latimes.gz (The system cannot find the file specified)
        at java.io.FileInputStream.open0(Native Method)
        at java.io.FileInputStream.open(FileInputStream.java:195)
        at java.io.FileInputStream.<init>(FileInputStream.java:138)
        at IndexEngine.readAndProcess(IndexEngine.java:47)
        at IndexEngine.main(IndexEngine.java:41)
```

Proof of Directories and Files



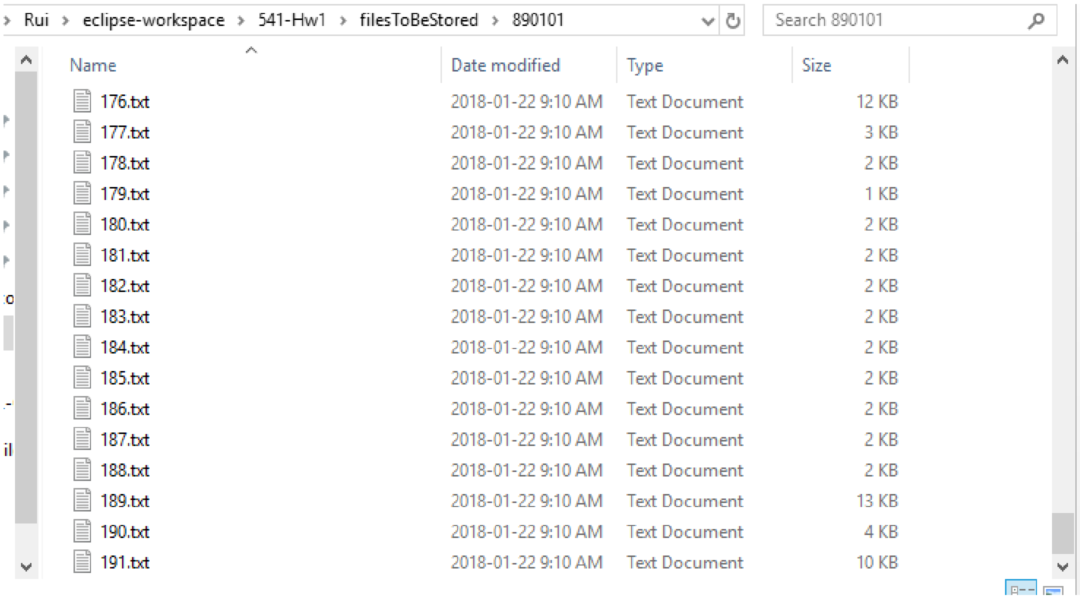
Name	Date modified	Type	Size
890101	2018-01-20 11:26 ...	File folder	
890102	2018-01-20 11:27 ...	File folder	
890103	2018-01-20 11:27 ...	File folder	
890104	2018-01-20 11:27 ...	File folder	
890105	2018-01-20 11:27 ...	File folder	
890106	2018-01-20 11:27 ...	File folder	
890107	2018-01-20 11:27 ...	File folder	
890108	2018-01-20 11:27 ...	File folder	
890109	2018-01-20 11:27 ...	File folder	
890110	2018-01-20 11:27 ...	File folder	
890111	2018-01-20 11:27 ...	File folder	
890112	2018-01-20 11:27 ...	File folder	
890113	2018-01-20 11:27 ...	File folder	
890114	2018-01-20 11:27 ...	File folder	
890115	2018-01-20 11:27 ...	File folder	
890116	2018-01-20 11:27 ...	File folder	



filesToBeStored

Type:	File folder
Location:	C:\Users\Rui\eclipse-workspace\541-Hw1
Size:	490 MB (514,511,254 bytes)
Size on disk:	750 MB (787,075,072 bytes)
Contains:	131,896 Files, 730 Folders

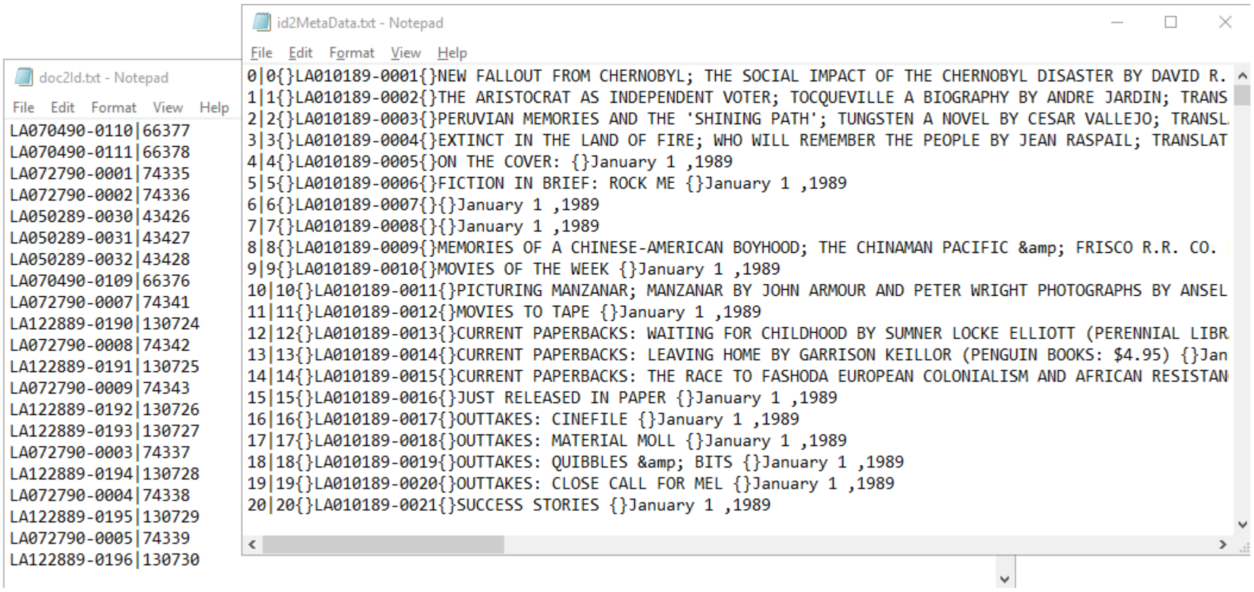
First directory dated Janauary 1st 1989 has 192 files named after their internal id starting with “0”.



› Rui › eclipse-workspace › 541-Hw1 › filesToBeStored › 890101

Name	Date modified	Type	Size
176.txt	2018-01-22 9:10 AM	Text Document	12 KB
177.txt	2018-01-22 9:10 AM	Text Document	3 KB
178.txt	2018-01-22 9:10 AM	Text Document	2 KB
179.txt	2018-01-22 9:10 AM	Text Document	1 KB
180.txt	2018-01-22 9:10 AM	Text Document	2 KB
181.txt	2018-01-22 9:10 AM	Text Document	2 KB
182.txt	2018-01-22 9:10 AM	Text Document	2 KB
183.txt	2018-01-22 9:10 AM	Text Document	2 KB
184.txt	2018-01-22 9:10 AM	Text Document	2 KB
185.txt	2018-01-22 9:10 AM	Text Document	2 KB
186.txt	2018-01-22 9:10 AM	Text Document	2 KB
187.txt	2018-01-22 9:10 AM	Text Document	2 KB
188.txt	2018-01-22 9:10 AM	Text Document	2 KB
189.txt	2018-01-22 9:10 AM	Text Document	13 KB
190.txt	2018-01-22 9:10 AM	Text Document	4 KB
191.txt	2018-01-22 9:10 AM	Text Document	10 KB

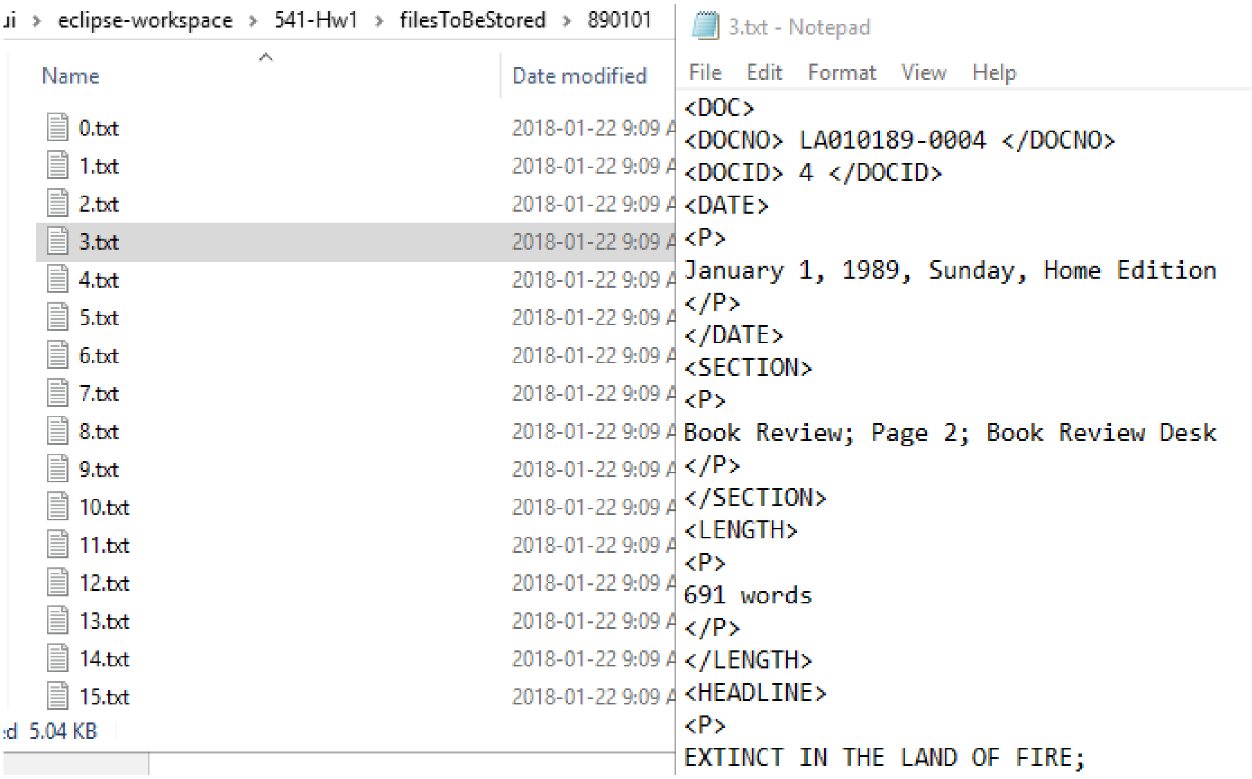
Proof that the two hashmaps are saved to file:


doc2Id.txt - Notepad
File Edit Format View Help
id2MetaData.txt - Notepad
File Edit Format View Help

```
LA070490-0110|66377
LA070490-0111|66378
LA072790-0001|74335
LA072790-0002|74336
LA050289-0030|43426
LA050289-0031|43427
LA050289-0032|43428
LA070490-0109|66376
LA072790-0007|74341
LA122889-0190|130724
LA072790-0008|74342
LA122889-0191|130725
LA072790-0009|74343
LA122889-0192|130726
LA122889-0193|130727
LA072790-0003|74337
LA122889-0194|130728
LA072790-0004|74338
LA122889-0195|130729
LA072790-0005|74339
LA122889-0196|130730
```

```
0|0{}LA010189-0001{}NEW FALLOUT FROM CHERNOBYL; THE SOCIAL IMPACT OF THE CHERNOBYL DISASTER BY DAVID R.
1|1{}LA010189-0002{}THE ARISTOCRAT AS INDEPENDENT VOTER; TOCQUEVILLE A BIOGRAPHY BY ANDRE JARDIN; TRANS
2|2{}LA010189-0003{}PERUVIAN MEMORIES AND THE 'SHINING PATH'; TUNGSTEN A NOVEL BY CESAR VALLEJO; TRANSL
3|3{}LA010189-0004{}EXTINCT IN THE LAND OF FIRE; WHO WILL REMEMBER THE PEOPLE BY JEAN RASPAIL; TRANSLAT
4|4{}LA010189-0005{}ON THE COVER: {}January 1 ,1989
5|5{}LA010189-0006{}FICTION IN BRIEF: ROCK ME {}January 1 ,1989
6|6{}LA010189-0007{}{}January 1 ,1989
7|7{}LA010189-0008{}{}January 1 ,1989
8|8{}LA010189-0009{}MEMORIES OF A CHINESE-AMERICAN BOYHOOD; THE CHINAMAN PACIFIC &amp; FRISCO R.R. CO.
9|9{}LA010189-0010{}MOVIES OF THE WEEK {}January 1 ,1989
10|10{}LA010189-0011{}PICTURING MANZANAR; MANZANAR BY JOHN ARMOUR AND PETER WRIGHT PHOTOGRAPHS BY ANSEL
11|11{}LA010189-0012{}MOVIES TO TAPE {}January 1 ,1989
12|12{}LA010189-0013{}CURRENT PAPERBACKS: WAITING FOR CHILDHOOD BY SUMNER LOCKE ELLIOTT (PERENNIAL LIBR
13|13{}LA010189-0014{}CURRENT PAPERBACKS: LEAVING HOME BY GARRISON KEILLOR (PENGUIN BOOKS: $4.95) {}Jan
14|14{}LA010189-0015{}CURRENT PAPERBACKS: THE RACE TO FASHODA EUROPEAN COLONIALISM AND AFRICAN RESISTAN
15|15{}LA010189-0016{}JUST RELEASED IN PAPER {}January 1 ,1989
16|16{}LA010189-0017{}OUTTAKES: CINEFILE {}January 1 ,1989
17|17{}LA010189-0018{}OUTTAKES: MATERIAL MOLL {}January 1 ,1989
18|18{}LA010189-0019{}OUTTAKES: QUIBBLES &amp; BITS {}January 1 ,1989
19|19{}LA010189-0020{}OUTTAKES: CLOSE CALL FOR MEL {}January 1 ,1989
20|20{}LA010189-0021{}SUCCESS STORIES {}January 1 ,1989
```

An example of the 4th document with internal id as 3


ui › eclipse-workspace › 541-Hw1 › filesToBeStored › 890101
3.txt - Notepad
File Edit Format View Help

```
Name          Date modified
0.txt         2018-01-22 9:09 A
1.txt         2018-01-22 9:09 A
2.txt         2018-01-22 9:09 A
3.txt         2018-01-22 9:09 A
4.txt         2018-01-22 9:09 A
5.txt         2018-01-22 9:09 A
6.txt         2018-01-22 9:09 A
7.txt         2018-01-22 9:09 A
8.txt         2018-01-22 9:09 A
9.txt         2018-01-22 9:09 A
10.txt        2018-01-22 9:09 A
11.txt        2018-01-22 9:09 A
12.txt        2018-01-22 9:09 A
13.txt        2018-01-22 9:09 A
14.txt        2018-01-22 9:09 A
15.txt        2018-01-22 9:09 A
d 5.04 KB
```

```
<DOC>
<DOCNO> LA010189-0004 </DOCNO>
<DOCID> 4 </DOCID>
<DATE>
<P>
January 1, 1989, Sunday, Home Edition
</P>
</DATE>
<SECTION>
<P>
Book Review; Page 2; Book Review Desk
</P>
</SECTION>
<LENGTH>
<P>
691 words
</P>
</LENGTH>
<HEADLINE>
<P>
EXTINCT IN THE LAND OF FIRE;
```