

# **Text Retrieval System**

by

**Bill Qualls**

DePaul University

CSC575 – Intelligent Information Retrieval

Dr. Bamshad Mobasher, Professor

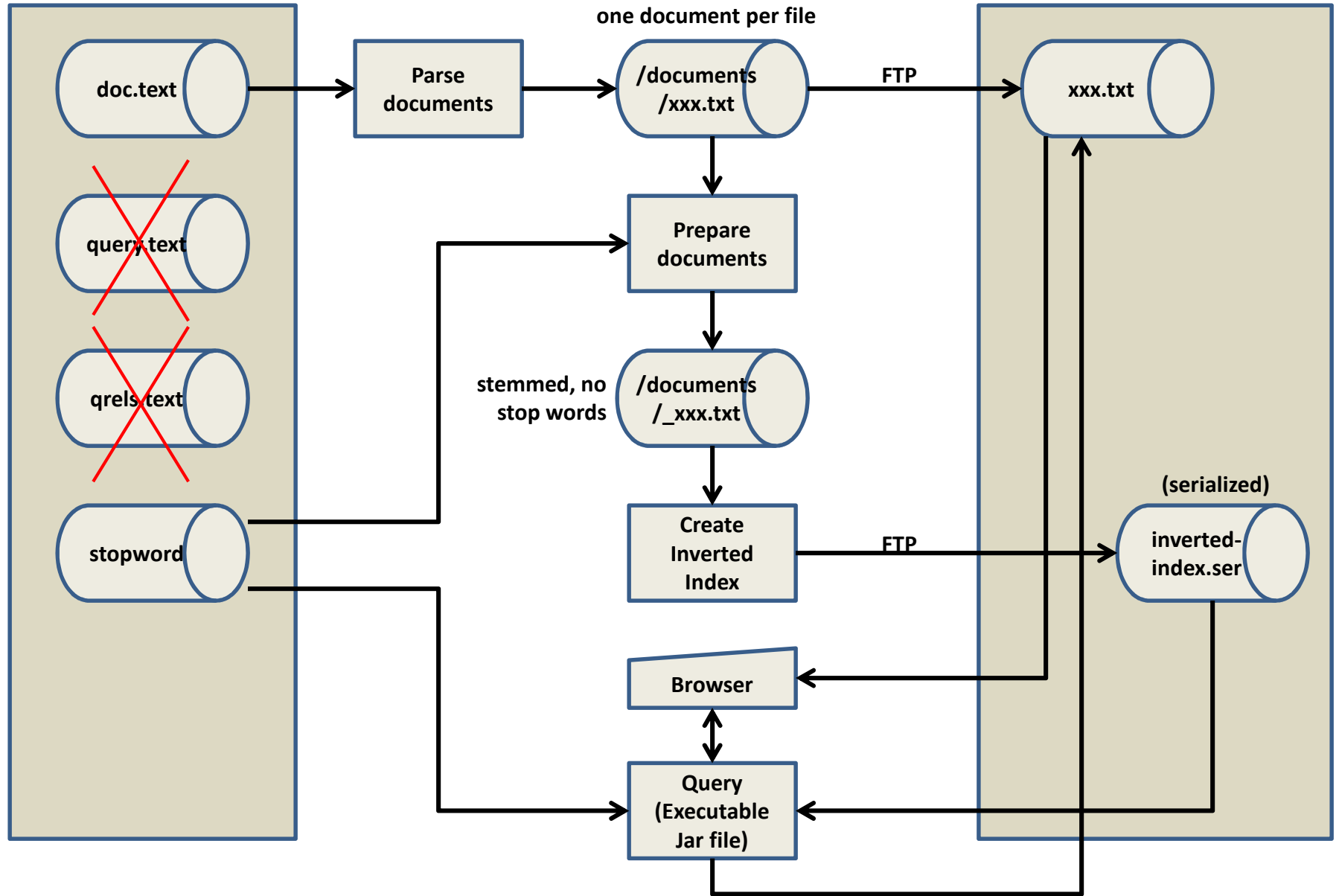
Winter Quarter 2012

# Contents

Flowchart	.	.	.	.	.	.	.	3
Source of data	.	.	.	.	.	.	.	4
Program 1: ParseDocuments	.	.	.	.	.	.	.	5
Program 2: Stemmer	.	.	.	.	.	.	.	7
Program 3: Stopword	.	.	.	.	.	.	.	9
Program 4: PrepareDocuments	.	.	.	.	.	.	.	10
Token class	.	.	.	.	.	.	.	11
TokenInfo class	.	.	.	.	.	.	.	12
TokenOccurence class	.	.	.	.	.	.	.	13
DocumentReference class	.	.	.	.	.	.	.	14
InvertedIndex class	.	.	.	.	.	.	.	15
Program 5: CreateInvertedIndex	.	.	.	.	.	.	.	16
Program 6: QueryApp (executable jar file)	.	.	.	.	.	.	.	18
Program 7: Query (servlet).	.	.	.	.	.	.	.	22
Testing	.	.	.	.	.	.	.	26

ftp.cs.cornell.edu/pub/smart/time

www.billqualls.com/corpus



# Source of Data

- This project used the Time Magazine data found at <ftp://ftp.cs.cornell.edu/pub/smart/time>.
- All documents were contained in a single file, **doc.text**. A program was written to parse this file into one file per document.
- The site contained **query.text** with 83 sample queries.
- The site also contained a **qrels.text** which was supposed to indicate relevant documents for each query, but this file was clearly flawed.

# Program 1: ParseDocuments

**Program:** ParseDocuments.java

**Purpose:** Create a separate file for each document

**Input:** doc.text (unzipped from doc.text.Z)

**Output:** project/documents/xxx.txt (where xxx = three digit number, such as 017).

First line each file is information about the source, as contained in the input file.

Output folder must exist before running program.

# Program 1: ParseDocuments

## Sample Input

\*TEXT 017 01/04/63 PAGE 020

THE ALLIES AFTER NASSAU IN DECEMBER 1960, THE U.S . FIRST  
PROPOSED TO HELP NATO DEVELOP ITS OWN NUCLEAR STRIKE FORCE . BUT  
EUROPE

.

.

GENERAL ASSEMBLY, AND IS THUS EQUAL IN VOTING POWER WITH SUCH  
NUCLEAR

GIANTS AS THE SOVIET UNION AND THE U.S .

\*STOP

## EOJ messages

23053 lines written to 425 files.

Normal end of program.

# Program 2: Stemmer

**Program:** Stemmer.java  
**Purpose:** Static method to stem a String  
**Input:** aString  
**Output:** aString, stemmed.  
**Notes:**

I download a Java implementation of Porter's Stemming Algorithm from <http://tartarus.org/martin/PorterStemmer/java.txt>.

Its main method was written to take a list of files as arguments from the command line. I modified the main() method to take the files from a static array to simplify testing while in Eclipse.  
(cont.)

# Program 2: Stemmer

I was surprised that there wasn't a method to stem a String, and I'm going to need one so I can stem the query. So I am added a static method `stem(String s)` which will return a String.

Also, as written, it does NOT remove punctuation. I would prefer to leave some punctuation in (example: I would want slashes in dates) but I am not going to spend too much time rewriting this program. Instead, I will simply remove all punctuation.

## **Sample usage**

```
String stemmedString = Stemmer.stem(rawString);
```

# Program 3: Stopword

**Program:** Stopword.java  
**Purpose:** Static method to work with stopwords  
**Input:** project/stopword  
**Output:** depends on method called  
**Notes:**

Constructor will read stopwords file found at  
<ftp://ftp.cs.cornell.edu/pub/smart/time/stopword>.

## Sample usage

```
boolean isStopword = Stopword.isStopword(word);
```

## Sample usage

```
String noStopwords = Stopword.removeStopwords(line);
```

There is a simple main() method for testing purposes.

# Program 4: PrepareDocuments

**Program:** PrepareDocuments.java  
**Purpose:** For each .txt file in the indicated directory, remove stopwords and stem  
**Input:** documents folder, populated by ParseDocuments.java (above)  
Process each file, xxx.txt.  
**Output:** xxx.txt becomes \_xxx.txt

## EOJ messages:

There are 423 input files.  
22972 lines written to 423 files.  
Normal end of program.

# Token Class

**Program:** Token.java

**Purpose:** A very simple class I created to make use of HashMap more clear. All this class contains at this time is the token (term).

# TokenInfo Class

**Program:** TokenInfo.java

**Purpose:** As presented in class. Documents where this token appears. Minor changes such as adding a toString method.

# TokenOccurrence Class

**Program:** TokenOccurrence.java

**Purpose:** As presented in class. Information about an occurrence of a token (term) within a document. Minor changes such as adding a toString method.

# DocumentReference Class

**Program:** DocumentReference.java

**Purpose:** As presented in class. Store a reference to a document. Minor changes such as adding toString() and hashCode() methods.

# InvertedIndex Class

**Program:** InvertedIndex.java

**Purpose:** A class which holds the inverted index.

All classes used within InvertedIndex (Token, TokenInfo, TokenOccurrence, and DocumentReference) implement Serializable so the index itself can be serialized.

Index is created in program CreateInvertedIndex.java.

I have been doing a lot of work with JSON recently, so I added a method to write the index as a JSON object (mostly out of curiosity, but it came in handy for checking my index!)

# Program 5: Create Inverted Index

- Program:** CreateInvertedIndex.java
- Purpose:** Create inverted index file from each `_xxx.txt` file in the indicated directory.
- Input:** documents folder, populated by PrepareDocuments.java (above)  
Process each file, `xxx.txt`.
- Output:** InvertedIndex.txt (inverted index written to a file for checking purposes)  
InvertedIndex.ser (serialized version, so it can be read into the search engine)

Developed by following the instructions contained within the "Implementation Notes" powerpoints provided by instructor.

# Program 5: Create Inverted Index

## EOJ messages

424 files indexed.

JSON version of index written to

C:/.../project/inverted\_index.json

Serialized version of index written to

C:/.../project/inverted\_index.ser

Normal end of program.

# Program 6: QueryApp

**Program:** QueryApp.java

**Purpose:** The text retrieval system implemented as an executable Jar file.

Makes use of the files and classes discussed already.

Uses the **tf x idf** algorithm.

For each retrieval, shows the document number (as a link), the relevance score, and the first two lines of the document.

The actual logic is contained in QueryGUI.java.

Access this application through the following URL:

[http://www.billqualls.com/pa/qualls\\_query.jar](http://www.billqualls.com/pa/qualls_query.jar)

→ I have also submitted a .exe version, created with Launch4j.

# 1 of 3: Opening screen

DePaul University - CSC575 - Intelligent Information Retrieval

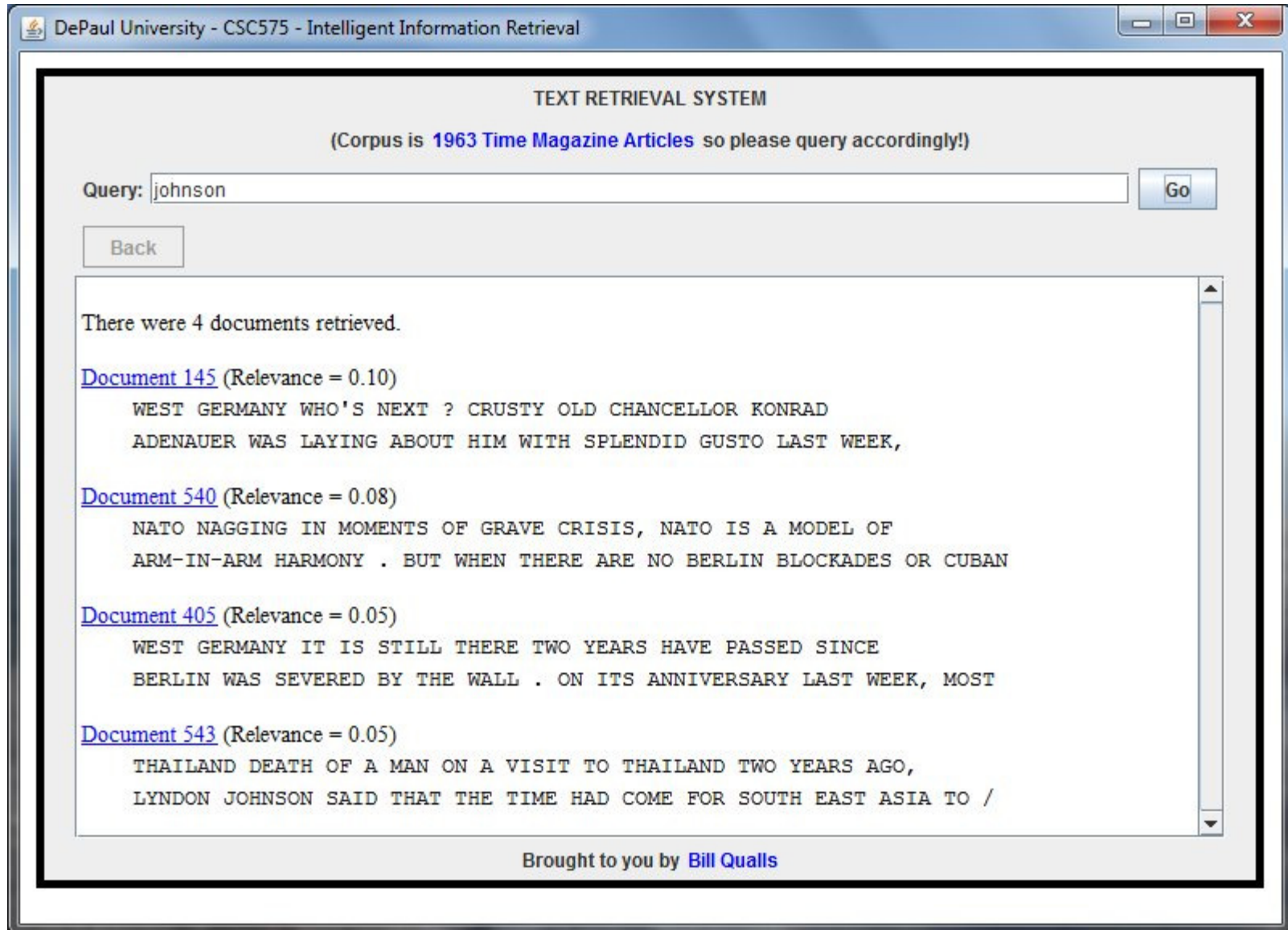
TEXT RETRIEVAL SYSTEM

(Corpus is [1963 Time Magazine Articles](#) so please query accordingly!)

Query:

Brought to you by [Bill Qualls](#)

# 2 of 3: Query Results



DePaul University - CSC575 - Intelligent Information Retrieval

TEXT RETRIEVAL SYSTEM  
(Corpus is [1963 Time Magazine Articles](#) so please query accordingly!)

Query:

There were 4 documents retrieved.

[Document 145](#) (Relevance = 0.10)  
WEST GERMANY WHO'S NEXT ? CRUSTY OLD CHANCELLOR KONRAD  
ADENAUER WAS LAYING ABOUT HIM WITH SPLENDID GUSTO LAST WEEK,

[Document 540](#) (Relevance = 0.08)  
NATO NAGGING IN MOMENTS OF GRAVE CRISIS, NATO IS A MODEL OF  
ARM-IN-ARM HARMONY . BUT WHEN THERE ARE NO BERLIN BLOCKADES OR CUBAN

[Document 405](#) (Relevance = 0.05)  
WEST GERMANY IT IS STILL THERE TWO YEARS HAVE PASSED SINCE  
BERLIN WAS SEVERED BY THE WALL . ON ITS ANNIVERSARY LAST WEEK, MOST

[Document 543](#) (Relevance = 0.05)  
THAILAND DEATH OF A MAN ON A VISIT TO THAILAND TWO YEARS AGO,  
LYNDON JOHNSON SAID THAT THE TIME HAD COME FOR SOUTH EAST ASIA TO /

Brought to you by [Bill Qualls](#)

# 3 of 3: Retrieved Document

DePaul University - CSC575 - Intelligent Information Retrieval

TEXT RETRIEVAL SYSTEM

(Corpus is [1963 Time Magazine Articles](#) so please query accordingly!)

Query:

\*TEXT 145 03/15/63 PAGE 043  
WEST GERMANY WHO'S NEXT ? CRUSTY OLD CHANCELLOR KONRAD  
ADENAUER WAS LAYING ABOUT HIM WITH SPLENDID GUSTO LAST WEEK,  
WISECRACKING ABOUT AMERICAN POLITICS ( " SAY, WHAT EVER HAPPENED TO  
LYNDON JOHNSON .Q? " ; , NEEDLING THE BRITISH (HE SAYS THEY  
DELIBERATELY SPREAD MISCONCEPTIONS), EVEN TAKING A BACKHANDED DIG AT  
HIS PAL IN PARIS, CHARLES DE GAULLE . " STUPIDITY " WAS THE CAUSE OF  
WESTERN EUROPE'S CURRENT DISUNITY, ADENAUER TOLD A DINNER MEETING OF  
THE FOREIGN PRESS IN BONN . WHOSE STUPIDITY ? " I BELIEVE THESE  
THINGS HAVE BEEN COMMITTED NOT ONLY BY BRITAIN BUT BY OTHERS AS WELL, "  
HE SIGHED . WAS ANY OF IT COMMITTED ON THE RIVER SEINE ? ASKED A  
REPORTER . " I DELIBERATELY HAVE NOT MENTIONED ANY NAMES, " RETORTED  
DER ALTE . " WHOEVER FITS THE COAT SHOULD WEAR IT . " ADENAUER'S FREE  
AND EASY REMARKS WERE MERELY SIGNS OF THE RELAXED AND REFLECTIVE MOOD  
THAT HAS COME OVER THE 87-YEAR-OLD CHANCELLOR NOW THAT HE HAS FINALLY  
MADE UP HIS MIND TO GIVE UP WEST GERMANY'S TOP JOB . LAST WEEK HE  
OFFERED NO OBJECTIONS WHEN A CAUCUS MEETING OF HIS CHRISTIAN DEMOCRATS  
AUTHORIZED C.D.U . BUNDESTAG LEADER HEINRICH VON BRENTANO TO CANVASS  
ALL THE FACTIONS AND SUGGEST A CANDIDATE TO TAKE OVER NEXT FALL AND  
LEAD THE PARTY IN THE 1965 ELECTIONS . THE HEIR-APPARENT IS STILL  
AVUNCULAR ECONOMICS MINISTER LUDWIG ERHARD, ARCHITECT OF THE WEST  
GERMAN ECONOMIC BOOM AND THE MOST POPULAR CHOICE AMONG WEST GERMAN

Brought to you by [Bill Qualls](#)

# Program 7: Query (Servlet)

**Program:** Query.java

**Purpose:** The text retrieval system implemented as a servlet.

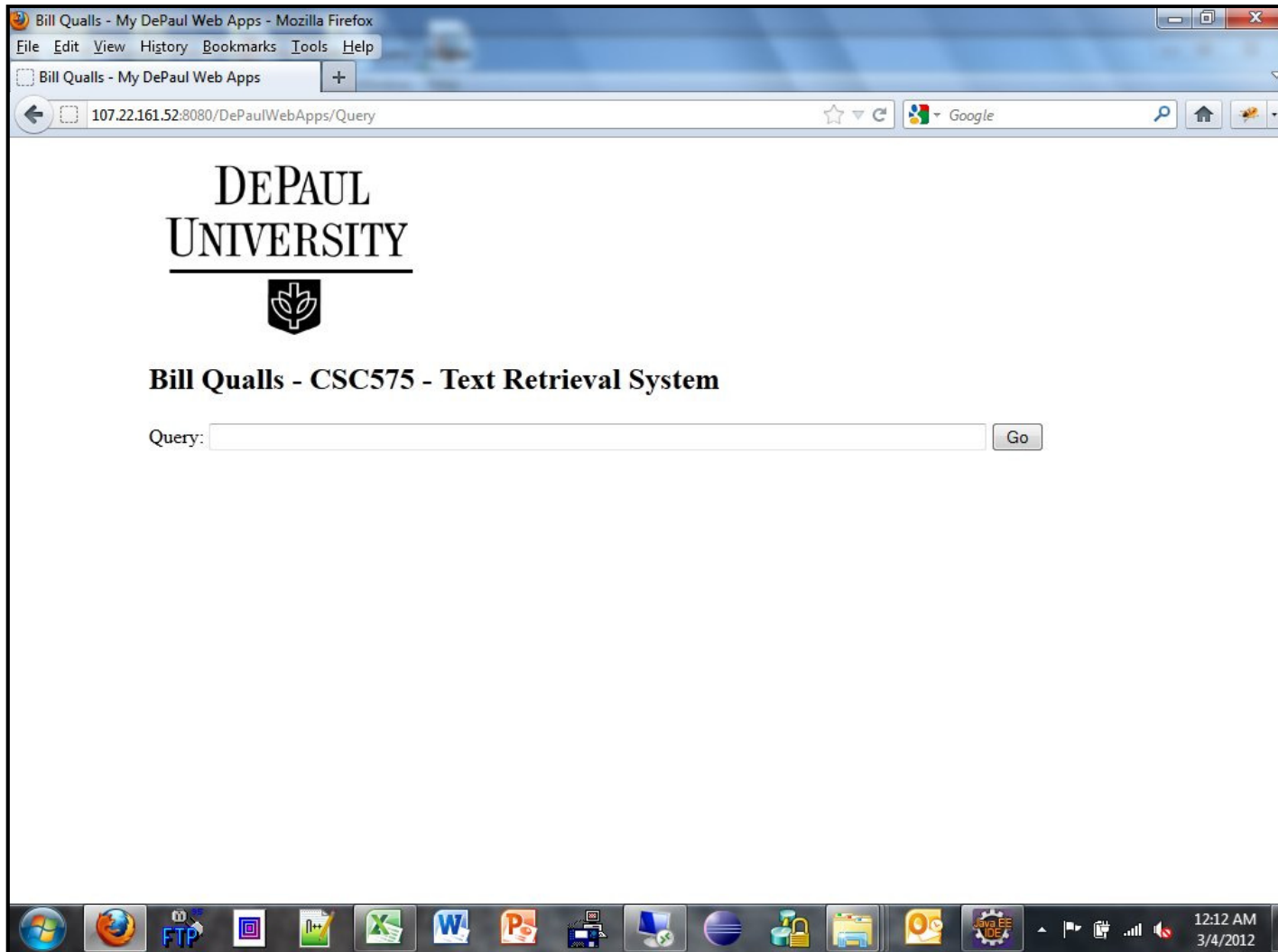
Same functionality, implemented as a servlet.

When the user clicks the document number link, the document itself appears in a separate window.

Access this servlet through the following URL (not available at night): <http://107.22.161.52:8080/DePaulWebApps/Query>

(I originally wrote this servlet, but since my current service provider does not support servlets, I cannot guarantee it will be available for viewing, so I created the Java app version as well.)

# Opening screen



# After searching for "Johnson"

The screenshot shows a Mozilla Firefox browser window with the address bar displaying `107.22.161.52:8080/DePaulWebApps/Query`. The page content includes the DePaul University logo and the title "Bill Qualls - CSC575 - Text Retrieval System". A search query of "johnson" is entered in a text box, and the results show four documents with their respective relevance scores and snippets of text.

DEPAUL  
UNIVERSITY

**Bill Qualls - CSC575 - Text Retrieval System**

Query: johnson

There were 4 documents retrieved.

[Document 145](#) (Relevance = 0.10)  
WEST GERMANY WHO'S NEXT ? CRUSTY OLD CHANCELLOR KONRAD  
ADENAUER WAS LAYING ABOUT HIM WITH SPLENDID GUSTO LAST WEEK,

[Document 540](#) (Relevance = 0.08)  
NATO NAGGING IN MOMENTS OF GRAVE CRISIS, NATO IS A MODEL OF  
ARM-IN-ARM HARMONY . BUT WHEN THERE ARE NO BERLIN BLOCKADES OR CUBAN

[Document 405](#) (Relevance = 0.05)  
WEST GERMANY IT IS STILL THERE TWO YEARS HAVE PASSED SINCE  
BERLIN WAS SEVERED BY THE WALL . ON ITS ANNIVERSARY LAST WEEK, MOST

[Document 543](#) (Relevance = 0.05)  
THAILAND DEATH OF A MAN ON A VISIT TO THAILAND TWO YEARS AGO,  
LYNDON JOHNSON SAID THAT THE TIME HAD COME FOR SOUTH EAST ASIA TO /

The taskbar at the bottom shows various application icons and the system clock indicating 12:12 AM on 3/4/2012.

# After selecting first "Johnson" document

Bill Qualls - My DePaul

Mozilla Firefox

www.billqualls.com/corpus/145.txt

\*TEXT 145 03/15/63 PAGE 043  
WEST GERMANY WHO'S NEXT ? CRUSTY OLD CHANCELLOR KONRAD  
ADENAUER WAS LAYING ABOUT HIM WITH SPLENDID GUSTO LAST WEEK,  
WISECRACKING ABOUT AMERICAN POLITICS ( " SAY, WHAT EVER HAPPENED TO  
LYNDON JOHNSON .Q? " ; , NEEDLING THE BRITISH (HE SAYS THEY  
DELIBERATELY SPREAD MISCONCEPTIONS), EVEN TAKING A BACKHANDED DIG AT  
HIS PAL IN PARIS, CHARLES DE GAULLE . " STUPIDITY " WAS THE CAUSE OF  
WESTERN EUROPE'S CURRENT DISUNITY, ADENAUER TOLD A DINNER MEETING OF  
THE FOREIGN PRESS IN BONN . WHOSE STUPIDITY ? " I BELIEVE THESE  
THINGS HAVE BEEN COMMITTED NOT ONLY BY BRITAIN BUT BY OTHERS AS WELL, "  
HE SIGHED . WAS ANY OF IT COMMITTED ON THE RIVER SEINE ? ASKED A  
REPORTER . " I DELIBERATELY HAVE NOT MENTIONED ANY NAMES, " RETORTED  
DER ALTE . " WHOEVER FITS THE COAT SHOULD WEAR IT . " ADENAUER'S FREE  
AND EASY REMARKS WERE MERELY SIGNS OF THE RELAXED AND REFLECTIVE MOOD  
THAT HAS COME OVER THE 87-YEAR-OLD CHANCELLOR NOW THAT HE HAS FINALLY  
MADE UP HIS MIND TO GIVE UP WEST GERMANY'S TOP JOB . LAST WEEK HE  
OFFERED NO OBJECTIONS WHEN A CAUCUS MEETING OF HIS CHRISTIAN DEMOCRATS  
AUTHORIZED C.D.U . BUNDESTAG LEADER HEINRICH VON BRENTANO TO CANVASS  
ALL THE FACTIONS AND SUGGEST A CANDIDATE TO TAKE OVER NEXT FALL AND  
LEAD THE PARTY IN THE 1965 ELECTIONS . THE HEIR-APPARENT IS STILL  
AVUNCULAR ECONOMICS MINISTER LUDWIG ERHARD, ARCHITECT OF THE WEST  
GERMAN ECONOMIC BOOM, AND THE MOST POPULAR CHOICE AMONG WEST GERMAN  
VOTERS . ONE CABINET MINISTER GUESSES THAT ERHARD ALSO COMMANDS THE  
LOYALTY OF 60 PER CENT OF C.D.U . POLITICIANS . BUT ERHARD STILL HAS  
ONE FORMIDABLE ENEMY DER ALTE HIMSELF WHO HAS CONDUCTED A PETULANT FEUD  
WITH PAUNCHY " UNCLE LUDWIG . " ADENAUER'S INFLUENCE IS STILL GREAT,

[Document 540](#) (Relevance = 0.08)  
NATO NAGGING IN MOMENTS OF GRAVE CRISIS, NATO IS A MODEL OF  
ARM-IN-ARM HARMONY . BUT WHEN THERE ARE NO BERLIN BLOCKADES OR CUBAN

[Document 405](#) (Relevance = 0.05)  
WEST GERMANY IT IS STILL THERE TWO YEARS HAVE PASSED SINCE  
BERLIN WAS SEVERED BY THE WALL . ON ITS ANNIVERSARY LAST WEEK, MOST

[Document 543](#) (Relevance = 0.05)  
THAILAND DEATH OF A MAN ON A VISIT TO THAILAND TWO YEARS AGO,  
LYNDON JOHNSON SAID THAT THE TIME HAD COME FOR SOUTH EAST ASIA TO /

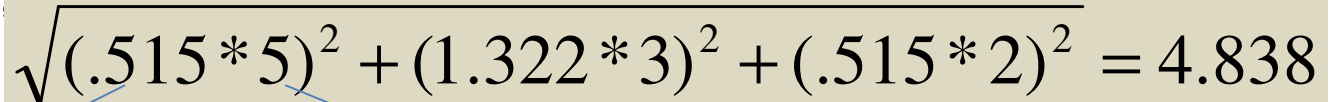
12:11 AM  
3/4/2012

# Testing

- I was unable to test with the given queries and expected results because they were flawed. (Example: queries about Syria, with supposed relevant documents about Viet Nam.)
- So to test this system I used the data contained in homework assignment #2, question 4(a).
- I created ten documents, each with some number of terms "t1" through "t8".
- I was able to validate the index by viewing the JSON output.
- Following are:
  - screenshot of Excel spreadsheet from homework assignment
  - extract of the JSON version of the inverted index
  - example of text files used as input
  - screenshot of the results from retrieving documents with term "t5"

# Inverted Index as JSON (partial)

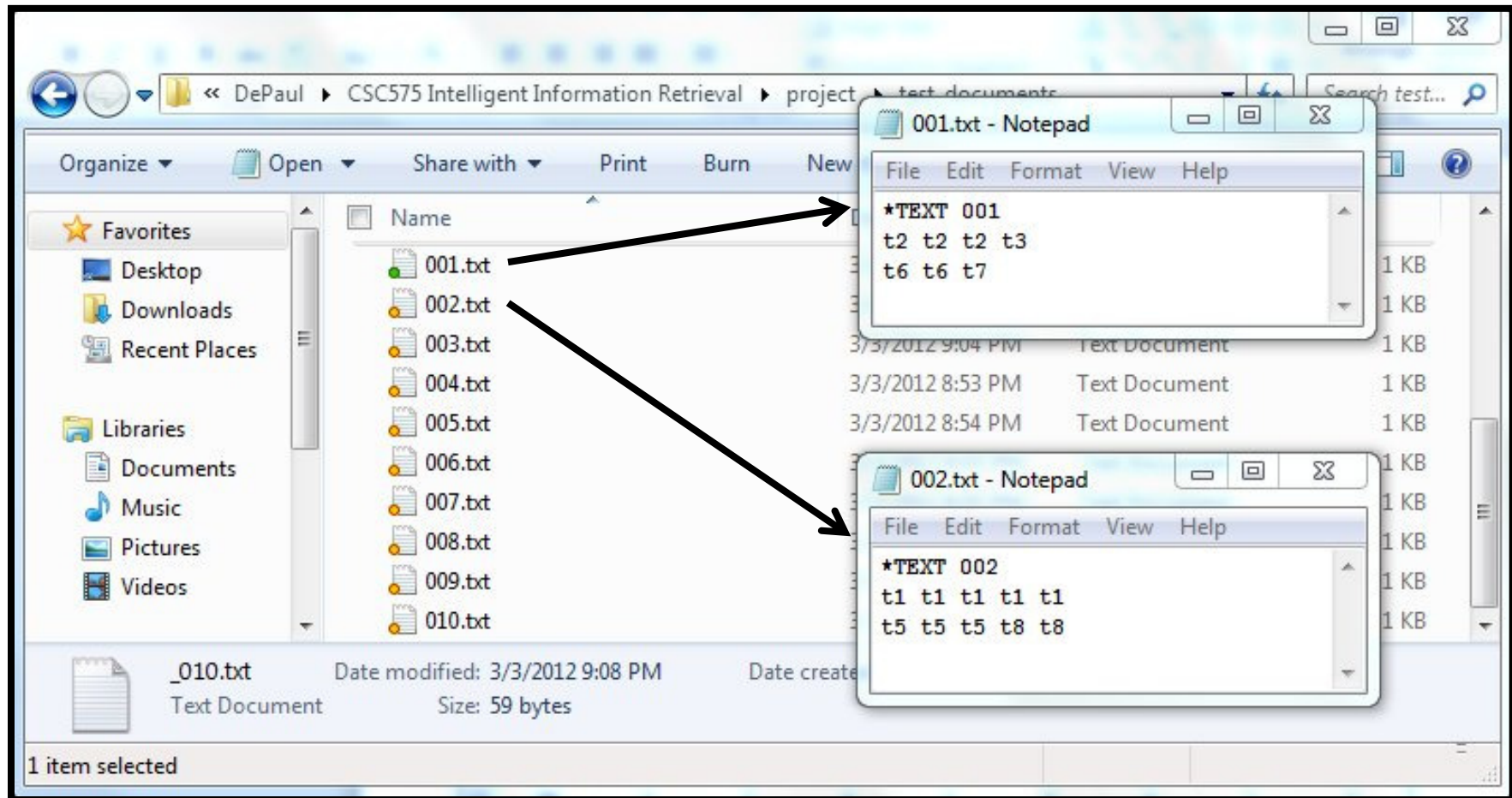
```
{
  'documentCount':10
  'index':[
    {
      'token':'t1',
      'occlist':[
        {'idf':0.515,'tokenOccurrence':
          {'docRef':{'file':'_002.txt','length':4.838},'count':5}},
        {'idf':0.515,'tokenOccurrence':
          {'docRef':{'file':'_003.txt','length':7.872},'count':3}},
        {'idf':0.515,'tokenOccurrence':
          {'docRef':{'file':'_004.txt','length':9.477},'count':1}},
        {'idf':0.515,'tokenOccurrence':
          {'docRef':{'file':'_006.txt','length':3.744},'count':2}},
        {'idf':0.515,'tokenOccurrence':
          {'docRef':{'file':'_007.txt','length':7.257},'count':2}},
        {'idf':0.515,'tokenOccurrence':
          {'docRef':{'file':'_008.txt','length':4.332},'count':3}},
        {'idf':0.515,'tokenOccurrence':
          {'docRef':{'file':'_010.txt','length':6.481},'count':1}}]
      },
      etc.
    }
  ]
}
```

$$\sqrt{(.515 * 5)^2 + (1.322 * 3)^2 + (.515 * 2)^2} = 4.838$$


# Test data taken from homework #2, part 4a

	A	B	C	D	E	F	G	H	I
1	<b>TF x IDF</b>								
2	CSC575 - Homework #2 - Question 4(a)								
3	by Bill Qualls								
4									
5	Number of documents:				10 <--ALL				
6									
7	<b>docs</b>	<b>t1</b>	<b>t2</b>	<b>t3</b>	<b>t4</b>	<b>t5</b>	<b>t6</b>	<b>t7</b>	<b>t8</b>
8	D1	0	3	1	0	0	2	1	0
9	D2	5	0	0	0	3	0	0	2
10	D3	3	0	4	3	4	0	0	5
11	D4	1	8	0	3	0	1	4	0
12	D5	0	1	0	0	0	5	4	2
13	D6	2	0	2	0	0	4	0	1
14	D7	2	5	0	3	0	1	4	2
15	D8	3	3	0	2	0	0	1	3
16	D9	0	0	3	3	3	0	0	0
17	D10	1	0	5	0	2	4	0	2
18	10	7	5	5	5	4	6	5	7

# Test data taken from homework #2, part 4a



# Test data: searching for term "t5"

Bill Qualls - My DePaul Web Apps - Mozilla Firefox

File Edit View History Bookmarks Tools Help

Bill Qualls - My DePaul Web Apps

localhost:8080/DePaulWebApps/Query

DEPAUL UNIVERSITY

Bill Qualls - CSC575 - Text Retrieval System

Query: t5 Go

There were 4 documents retrieved.

[Document 2](#) (Relevance = 0.82)  
t1 t1 t1 t1 t1  
t5 t5 t5 t8 t8

[Document 9](#) (Relevance = 0.68)  
t3 t3 t3 t4 t4 t4  
t5 t5 t5

[Document 3](#) (Relevance = 0.67)  
t1 t1 t1 t3 t3 t3 t3 t4 t4 t4  
t5 t5 t5 t5 t8 t8 t8 t8 t8

[Document 10](#) (Relevance = 0.41)  
t1 t3 t3 t3 t3 t3  
t5 t5 t6 t6 t6 t6 t8 t8

12:13 AM 3/4/2012

the end