Patrick Trinkle
CMSC676
Spring 2009
Project 4 Write-Up

## Background

After an Information Retrieval system has processed and indexed the documents in the corpus, a service must be provided for a user to query these documents via the index. Many query processing systems use phrases as the query string, however special syntax is often available to enhance the effectiveness of a query. As an example of this, the Google search engine provides special syntax keywords, such as "site:" and "intext:" to focus the search. The user interface of a search engine may hide the details of the exact syntax passed to the engine itself. Certain term weighting may be calculated for a query, as a well as indexing against the query itself to locate possible phrases.

## Implementation

This iteration of the project reads all queries from a file specified at the command line. This file can be generated by hand or via a GUI. If a GUI is used, it is required to handle processing the users query into the correct syntax for the query processing piece of the information retrieval system. The syntax requires that each term in the query is built as a term-weight pairing, "term=weight". The weight can be any value less or equal to 1. If there is more than one term in a query the term-weight pairs are space separated. The sum of the term weights should be 1, if the sum is less than one, the extra weight will be added to the longest term, if the sum is greater than 1 the query will still execute and an error message will be displayed to the user. Because the query terms are each processed via the HTML Parser used for indexing documents it is possible that a term is broken into sub-terms. Each sub-term receives a weight evenly partitioned from the weight of the parent term. At no point are stop words removed from the queries, however because the dictionary is currently void of stop words, they will not provide any benefit to a query. It might be beneficial to report this information to the user to allow them to rebalance the query. All terms are downcased, as is consistent with the document parser. The queries themselves are represented as a class, Query. Each Query object contains a table of all terms and their weights, as well as a hash table of all documents and their respective scores. Because document scores are accumulated as the query is processed, one entry is needed for each document, holding the document identification string and the current score. The score is calculated as the product of the tf-idf and the term weight, added to the previous score of the document. Future implementations can maintain which terms were found in which of the documents. A method of the Query class returns a sorted result list for the query. This full list is printed out as the top ten entries (if there are ten or more), otherwise all entries are displayed to the user. If there are no results a print out indicating this is presented to the user.

The dictionary is read into memory in one-pass and stored in a simple hash-table for easy querying. The records in the postings file are represented in memory via the RecordSet class. Each RecordSet object contains a series of document identification strings and the respective tf-idf values. The postings file is read into a cache on an as-needed basis. The cache consists of a hash table, where the key is the term, and the value is the RecordSet. Because the in-memory dictionary has an offset into the postings file for each term, if the term is not already in the cache, a simple seek is required to jump to the exact location in the postings file. The document frequency for the term provides how many records must be read for a complete RecordSet for term, and this is added to the cache, for future lookups. The name RecordSet could be confused with a database object, however namespaces can be used to disambiguate the usage.

All previous HTML parser implementations ignored numbers. To handle more queries, this was changed. However, tokens are broken on non-alphanumeric characters, therefore "123-456-7890" breaks into three terms, {"123", "456", "7890"}. This isn't the case for periods as they are always skipped over. Given a term "20.07", becomes "2007". This is a known problem in the HTML Parser, whereby if there is only a period at the end of a sentence and beginning of another the terms will combine, however given the likelihood of at least one space between sentences, this concern is documented and ignored.

Because almost all collections of information need to be queried quickly they are all implemented as hash tables. Some of the hash tables are keyed on a term, and the value is a list of information. This is slightly more complex, but provides the same lookup speed, because the list of information is all pertinent to the key.

## Conclusion

The query processing system is quick, even on a larger dataset provided that the dictionary fits in memory, and the postings file cache does not grow too large. It was unnecessary for the scope of this version to handle the case where the dictionary did not fit in memory. Managing the postings cache could be done via flag bits on the RecordSet for a term indicating when it was last used, etc. With this extra information it would be trivial to implement a cache management algorithm, such as LRU (least recently used) whereby the least recently used entry in the cache is swapped out. Processing 15 queries contributed negligible extra processing time.

An interesting feature of using strictly tf-idf vector scores is that a query with multiple terms may have results with only one of the terms higher up than a document with more than one of the query terms.

A query for the term "gift" provides an interesting results where the documents appear backwards and have very similar scores. The strangeness is due to the method of normalization used to reduce the term frequency. This indicates that the normalization method used may need to be tweaked. Another interesting result is a query for "20." The second document listed has many more occurrences, but is dramatically longer than the first document in the results list.

The program currently returns the document ID to the user, a next iterative step would be to provide a link of a full path to the source of the document matching the ID. Because our dataset is such that the IDs match the filename, it was unnecessary to provide the full path in the results. The results from the sample queries and the document snippets for the top two results in each query are in Appendix A - Query Results.

The timing results are presented in Figure 1, however they are comparable to Version 3.0.

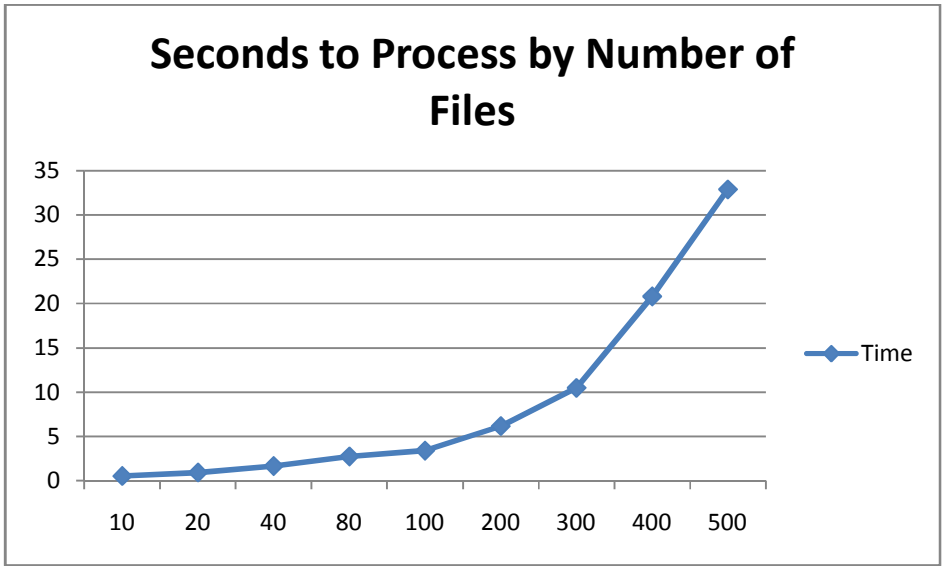


**Figure 1 - Seconds to Process Input by Number of Files**

## Appendix A - Query Results

The results presented are in decreasing order of relevance to the query.

### Query: Nicholas=1

Document Id: 192 Score: 0.004387
>>*State Department Spokesperson **Nicholas** Burns told the Committee's chair, Kati Marton, that CPJ's report Clampdown in Addis: Ethiopia's Journalists at Risk led to the policy change.*
Document Id: 279 Score: 0.000434
>>*Russia Under the Rule of Czar Alexander I and Czar **Nicholas** I*
*Rigidity, limited reform, lack of responsiveness to social changes characterizing nineteenth-century Russian monarchies. 82yr 7pgs 12fn 4s $62.65*
Document Id: 377 Score: 0.000352

### Query: elephants=1

Found no documents matching your query.

### Query: CSEE=1

Found no documents matching your query.

### Query: Maryland=1

Document Id: 399 Score: 0.004694
>>***Maryland***
*Beall 61*
*Mathias 27*
Document Id: 411 Score: 0.004618
>>***Maryland***
*Sarbanes 0*
*MATHIAS 7*
Document Id: 430 Score: 0.004618

Document Id: 400 Score: 0.004599
Document Id: 406 Score: 0.004581
Document Id: 402 Score: 0.004545
Document Id: 413 Score: 0.004527
Document Id: 404 Score: 0.004509
Document Id: 422 Score: 0.004509
Document Id: 409 Score: 0.004474

### Query: gift=1

Document Id: 149 Score: 0.013293
*>>you want to send a **gift** to a friend or relative in Canada and you are worried about sending cash.*
Document Id: 324 Score: 0.011735
*>>Tax Consequences Of A **Gift** To Joint Donees*
*A look at the **gift** tax- how many donees are involved and the amount of **gift** tax liability. What the code says. 79yr 5pgs 11fn 7s $44.75*
Document Id: 180 Score: 0.011575
Document Id: 051 Score: 0.005653
Document Id: 397 Score: 0.002214
Document Id: 074 Score: 0.001923
Document Id: 026 Score: 0.001649
Document Id: 383 Score: 0.000502
Document Id: 298 Score: 0.000273
Document Id: 043 Score: 0.000205

### Query: abcdef=1

Found no documents matching your query.

### Query: 20=1

Document Id: 035 Score: 0.073872
*>>NetShade **2.0** for Windows 95*
Document Id: 080 Score: 0.015634
*>>The range pixel spacing is 6.66 meters (assumes **20** MHz data).*
*m(1,3) = (Re(ShhShv*) + Re(ShvSvv*) ) / **2.0***
*m(1,4) = (-Im(ShhShv*) - Im(ShvSvv*) ) / **2.0***
*m(2,1) = m(1,2)*
*m(2,2) = (ShhShh* +SvvSvv* - 2ShvShv* ) / 4.0*
*m(2,3) = ( Re(ShhShv*) - Re(ShvSvv*) ) / **2.0***
*m(2,4) = (-Im(ShhShv*) + Im(ShvSvv*) )/ **2.0***
*m(3,1) = m(1,3)*
*m(3,2) = m(2,3)*
*m(3,3) = (ShvShv* + Re(ShhSvv*) / **2.0***
*m(3,4) = -Im(ShhSvv*) / **2.0***
*m(4,1) = m(1,4)*
*m(4,2) = m(2,4)*
*m(4,3) = m(3,4)*
*m(4,4) = (ShvShv* - Re(ShhSvv*) )/ **2.0***

*m(1,3) = (Re(ShhShv*) + Re(SvhSvv*) ) / **2.0***

*m(1,4) = (-Im(ShhShv*) - Im(SvhSvv*) ) / **2.0***
*m(2,1) = (ShhShh* + ShvShv* - SvhSvh* - SvvSvv*) / 4.0*
*m(2,2) = (ShhShh* + SvvSvv* - ShvShv* - SvhSvh*) / 4.0*
*m(2,3) = ( Re(ShhShv*) - Re(SvhSvv*) ) / **2.0***
*m(2,4) = (-Im(ShhShv*) + Im(SvhSvv*) )/ **2.0***
*m(3,1) = (Re(ShhSvh*) + Re(ShvSvv*) ) / **2.0***
*m(3,2) = ( Re(ShhSvh*) - Re(ShvSvv*) ) / **2.0***
*m(3,3) = (ShvShv* + Re(ShhSvv*) / **2.0***
*m(3,4) = (-Im(ShhSvv*+Im(ShvSvh*) / **2.0***
*m(4,1) = (-Im(ShhSvh*) - Im(ShvSvv*) ) / **2.0***
*m(4,2) = (-Im(ShhSvh*) + Im(ShvSvv*) )/ **2.0***
*m(4,3) = (-Im(ShhSvv* - Im(ShvSvh*) / **2.0***
*m(4,4) = (Re(ShvSvh*) - Re(ShhSvv*) )/ **2.0***
Document Id: 412 Score: 0.014021
Document Id: 231 Score: 0.010017
Document Id: 399 Score: 0.009688
Document Id: 429 Score: 0.008031
Document Id: 413 Score: 0.007008
Document Id: 432 Score: 0.00578
Document Id: 113 Score: 0.005738
Document Id: 431 Score: 0.005586

### Query: 20.07=1

Found no documents matching your query.

### Query: 123-456-7890=1

Document Id: 216 Score: 0.001845
*>>Sellers, W.D. and S. F. Kirby, 1987: Cold air drainage and urban heating in Tucson, Arizona. J. Arizona-Nevada Acad. of Sci., 22, **123**-128.*
Document Id: 122 Score: 0.00152866666666667
*>>£247 £357 £**456***
Document Id: 044 Score: 0.000766333333333333
Document Id: 213 Score: 0.000691333333333333
Document Id: 207 Score: 0.000634333333333333
Document Id: 459 Score: 0.000398
Document Id: 396 Score: 0.000397333333333333
Document Id: 222 Score: 0.000346333333333333
Document Id: 005 Score: 0.000293333333333333
Document Id: 461 Score: 0.000241

### Query: pstrink@gmail.com=1

Found no documents matching your query.

### Query: taxes=.5 federal=.5

Document Id: 388 Score: 0.0059955
*>>Leading Illinois **Federal**, State & Local Government Law Attorneys*
***Federal**, State & Local Government Law*
* ***Federal**, State & Local Government Law*

*o Public Financing*
*+ **Federal** Financing*
*+ State Financing*
*+ Local Financing*
*o Government Contracts and Technical Assistance*
*o Lobbying*
*o Dealing with Administrative Agencies*
*+ The Limited Role of Administrative Agencies*
*+ The Administrative Process*
*+ The Role of Courts*
*+ Freedom of Information Act*
*o Resources*

***Federal**, State & Local Government Law*

*Some businesses deal exclusively with one or more branches of government, while others only have to deal with the government to get permission to embark on certain projects. This chapter focuses on areas in which businesses frequently come into contact with **federal**, state, and local governments, such as public financing, government contracts, lobbying, and dealing with administrative agencies. The Environmental & Natural Resources Law Chapter examines issues relating to business interaction with agencies protecting the environment. The Intellectual Property & Computer Law Chapter discusses how a business registers a trademark with **federal** and state authorities. The Commercial Real Estate Law Chapter covers zoning and land use variances. Appendix A lists government sources of financing and management assistance.*

******There were other instances, this document is about certain parts of Federal law******

Document Id: 393 Score: 0.005459

>>******This document is similarly themed: Titled Tax Law, from the same authors******

Document Id: 408 Score: 0.00524
Document Id: 427 Score: 0.004379
Document Id: 419 Score: 0.0041995
Document Id: 426 Score: 0.0032895
Document Id: 246 Score: 0.003237
Document Id: 253 Score: 0.003201
Document Id: 324 Score: 0.002997
Document Id: 464 Score: 0.0029515

## Query: state=.5 scholarship=.5

Document Id: 199 Score: 0.002384

*>>Although the "Call to Action" conference focuses on providing women with strategies for personal and professional success, the conference also offers five $5,000 scholarships to deserving students attending California educational institutions. One **scholarship** is awarded in each of the following categories: Health/Medicine; Law Enforcement; Education; Business; and Math/Science. Winners included Theresa Marie Koppie (Health/Medicine); Angela Michelle Craig (Law Enforcement); Michele Elise Betton (Education); Wendi Nicole Smith (Business); and Gloria Maria Amador (Math/Science). Scholarships are funded by proceeds from the conference -- no state money is used to fund these scholarships.*

Document Id: 232 Score: 0.001129

*>>Divergent Opinions In Cold War **Scholarship***

*A study of American diplomacy during the Cold War years. 78yr 6pgs 8fn 4s $53.70*

Document Id: 323 Score: 0.000766
Document Id: 118 Score: 0.000603

Document Id: 249 Score: 0.0001475

### Query: college=.33 essays=.33 science=.33

Document Id: 114 Score: 0.05484506
>>***College*** *and University credit for SEA Semester*

*SEA Affiliates. These colleges and universities have entered into a formal affiliation agreement with the Sea Education Association. They list SEA Semester and SEA Summer Session courses in their academic curricula and grant credit directly to their students.*

*Boston University* *Colgate University* ***College*** *of Charleston*
*Cornell University* *Drexel University* *Eckerd College*
*Franklin & Marshall* ***College*** *Northeastern University* *The University of Pennsylvania*
*Rice University*
******Below this is a list of colleges, many with the word "college" in them******
Document Id: 155 Score: 0.01420826
>>***College*** *of Continuing Education*
Document Id: 071 Score: 0.01037816
Document Id: 259 Score: 0.00968442
Document Id: 055 Score: 0.00521908
Document Id: 063 Score: 0.00510034
Document Id: 132 Score: 0.00320824
Document Id: 226 Score: 0.00287247
Document Id: 048 Score: 0.00274758
Document Id: 257 Score: 0.00215557

### Query: zyuganov=.5 zzzzzzz=.5

Document Id: 338 Score: 0.052881
>>*EEEEEEE M M* ***ZZZZZZZ*** *EEEEEEE TTTTTTTNN*
*Yeltsin leads Communist leader Gennadii* ***Zyuganov*** *among likely voters.*
Document Id: 364 Score: 0.052881
>>*EEEEEEE M M* ***ZZZZZZZ*** *EEEEEEE TTTTTTTNN*
However, **Zyuganov** still has time to fall on his knees before me and Lebed." In the past, Zhirinovsky has refused to cooperate with the Communists and has called Lebed a "traitor." **Zyuganov** has offered to join forces with Lebed but has consistently criticized Zhirinovsky's erratic views and voting record in parliament. LDPR Duma deputies sometimes vote with the Communists but on crucial votes often back the government. -- Laura BelinYELTSIN CAMPAIGN OFFERS PRIZE FOR SOVIET-ERA COUPONS. The Yeltsin campaign headquarters in the Republic of Buryatiya is offering 1 million rubles ($200) to the person who has saved the most coupons for food and other consumer goods dating from the Brezhnev era of "developed socialism," ITAR-TASS reported on 30 May. The contest is aimed at reminding voters of the shortages and lines that were common in the Soviet period, which Yeltsin supporters warn could return if Communist leader Gennadii **Zyuganov** is elected president. -- Laura BelinPAPER: YELTSIN'S AD CAMPAIGN EFFECTIVE. President Yeltsin's paid political advertising campaign, which shows ordinary people explaining why they support the president but never shows Yeltsin himself, is very effective, according to Kommersant-Daily on 29 May. The advertising agency Video International, which developed the successful da-da-nyet-da campaign before the April 1993 referendum and worked less successfully with Yabloko before the 1995 parliamentary election, is responsible for the clips. The "man on the street" approach is connecting with ordinary people who see themselves in the advertisements,

the paper argues. The advertisements were filmed using real people speaking without any pre- written script. The whole "soap opera" will have a surprise conclusion that the authors refuse to reveal in advance. By not showing the candidate, the paper argued, the advertisements are "unobtrusive" and do not "irritate the viewer." Yeltsin himself approved this subtle approach. -- Robert Orttung in MoscowVLASOV APPEALS TO PATRIOTISM. Former world champion weight lifter and presidential candidate Yurii Vlasov on 29 May called for a policy of "people's patriotism" and accused the Communists of stealing many of his ideas, including the name of his People's Patriotic Party (**Zyuganov** calls himself the leader of the "coalition of popular-patriotic forces"). Vlasov compared his brand of nationalism with French Gaullism, claiming that it is a more effective unifying force than communist or democratic ideals. In his opinion, Yeltsin's policies have pushed 40% of the population below the poverty line and brought the government only 3% of the real value of privatized state property. He said that he expects to win 6-7% of the vote and that he will support neither Yeltsin nor **Zyuganov** in the runoff. He is running at less than 1% in the polls and the media has largely ignored his campaign. -- Robert Orttung in MoscowMOVEMENT "NYET" STICKS TO ANTI-YELTSIN COURSE. Leaders of the Movement "Nyet" on 29 May rejected accusations that their call on people to cast votes against all candidates in the second round could help **Zyuganov** beat Yeltsin. They admitted that **Zyuganov** would get 35% of the vote but argued that "against all" would gain even more votes if enough people refuse to back Yeltsin. Under the electoral law, the candidate with the most votes wins the second round as long as he gains more votes than are cast against all candidates.

*****Many other instances follow*****

Document Id: 365 Score: 0.052881
Document Id: 344 Score: 0.0067715
Document Id: 036 Score: 0.0026185
Document Id: 363 Score: 0.0022335
Document Id: 377 Score: 0.0021445
Document Id: 350 Score: 0.001706
Document Id: 366 Score: 0.0013125
Document Id: 336 Score: 0.0011345

### Query: time=.33 timeliness=.33 timely=.33

Document Id: 217 Score: 0.00553443

*>>Please plan on waiting quite a while for some of these toys to download the first **time**. After that, they should load faster the next **time** you visit.*

Document Id: 200 Score: 0.00435369

*>>The investment objective of this Fund is to provide high current income and preservation of capital primarily through investment in common stock in Canadian companies that meet the Ethical criteria. Bonds, debentures, and T-Bills may be purchased from **time** to **time**.*

Document Id: 059 Score: 0.00377817
Document Id: 145 Score: 0.00285417
Document Id: 003 Score: 0.00283932
Document Id: 135 Score: 0.00283932
Document Id: 190 Score: 0.0019437
Document Id: 116 Score: 0.00187671
Document Id: 128 Score: 0.00155496
Document Id: 122 Score: 0.00132033