i1Patrick Trinkle
CMSC676
Spring 2009
Project 1 Write-Up

Background

Information Retrieval systems need to index a corpus to provide users the ability to query for documents based on terms. These documents are not necessarily basic text files, but can also be PDF documents, emails, etc. Therefore an intelligent IR system needs to have methods for identifying and handling these varying document formats. This indexing does not need to be extremely fast; as long as it is accurate and precise. An important point in this is that the documents returned should reflect the goal of the query itself. I designed a document parser which strictly tokenizes Hyper Text Markup Language formatted documents, or basic web pages.

Implementation

Ruby is an entirely Object-Oriented programming language. Having no real experience using it; I felt it a good sandbox for working on the document parser. As a modern scripting language it automatically handles certain objects in a convenient way; such as strings and file streams. Although it is Object-Oriented in nature, it does have support for procedural methods.

For this implementation of the tokenizer all tokens are stored in entirely lowercase. Ruby provides casing methods for string objects; therefore downcasing is applied to each line in the process, versus each token. The HTML parser does not tokenize all character groupings or words. Special characters in ampersand-semicolon notation are disregarded as meta character information. Originally the script handled these special cases, but it was determined too costly for the limited benefit.

Tokens are built strictly from letters; ignoring numbers and periods; even if they appear in the string itself. Also any term that is hyphenated is broken into two or more pieces. URLs are not cleanly handled by the parser, and are broken into pieces. Instead of skipping over certain characters, anything not a letter or period is considered an end token character. Also when an HTML tag starts, if we were processing a token it is saved off. Some strangeness can appear in the term list because some terms can be parts of PGP keys, md5 hashes, et cetera. Also there is no error checking logic to throw out terms which are likely useless. By breaking on certain characters in tokens the following case is common: "It's" => "It" + "s". Because we do accept periods into a token, they are deleted once a termination character is reached. Therefore the following conversion occurs: "u.s.a." => "usa". Similarly because of the exclusion of periods as termination characters, the parser incorrectly converts "good.neat" => "goodneat". Therefore any sentences separated by a period without a space are merged.

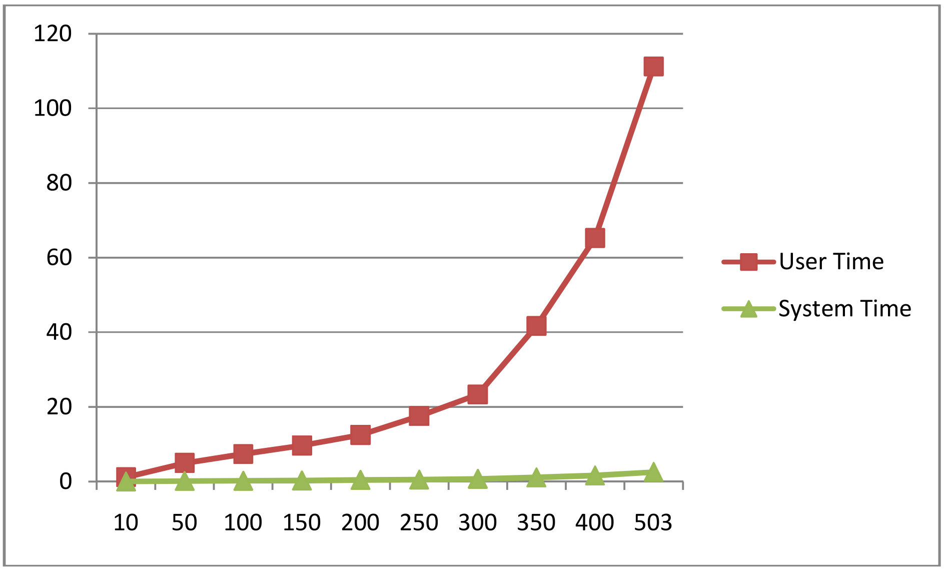
Originally the parser used an object oriented approach whereby each file required instantiating a parser object. This object read in and tokenized the file. This implementation proved rather slow (5 minutes or more) and therefore I switched to a procedural approach. The procedural implementation was still slow. I found that handling special cases as token endings there was a significant slowdown in processing. Also, each line calls block code to handle each byte individually. For all comparisons I had to convert the byte to character. This conversion is not optimized by the interpreter; therefore I had to assign the byte as a character to a character variable and then did the comparisons against the new variable. This detail sliced approximately 90 seconds off the processing time for the 503 input files in the CMSC676 corpus.

Another student approached the problem in a very similar fashion, but with the Python scripting language. In this implementation a stateM object is created to handle parsing of each input file. This object reads in one character at a time. Although a new object is required for each input file, Python is faster. The Python code includes number and characters only and breaks a token on either a whitespace character, new line character or the start of a tag. This can provide better terms if there are meaningful numbers; although since hyphens and periods are not included in terms some values such as IP addresses can lose value as a document term. Also because his tokens just step over non-alphanumeric characters various interesting cases can arise, such as: "It's" => "Its". Many interesting cases provide no detriment to the quality of the term list.

An even faster approach to the HTML tokenizer is to write the processing library in C and then importing into python script. Originally, I wrote the parser in C# and it was very quick as it is entirely in managed system code. This approach provided speed, but was less useful for rapid editing as each incarnation needed compilation.

Conclusion

The other implementation using strictly Python parsed the 503 document corpus in approximately 60 seconds where my Ruby script took 2 minutes. Term frequency was calculated as a simple incrementing of a value in a hash table keyed via the term. Due to the nature of processing in a simple way the time the process takes is directly proportional to the size of the input files; versus the number. This is due to the processing time required per file taking longer than the system IO required for each file.



Time in Seconds by Number of Input Files

Appendix A - Ruby Script

```
#!/usr/bin/env ruby

# Course: CMSC676 - Information Retrieval
# Semester: Spring 2009
# Professor: Dr Nicholas
# Student: Patrick Trinkle <tri1@umbc.ed>
# Date: 20090204
# Title: HTMLParser Object Definition
# Note: This was far trickier than doing it in C# or Perl or C, but has provided a good learning experience
#       Also this method is rather slow.  Doing it as a procedural script might be faster as it won't have to
#       instantiate objects, etc.  But since speed wasn't a primary concern; I felt this approach was
#       appropriate.

#       Okay all special characters within a token (either in the middle or the end) are
#       disregarded: be&#146s => bes
#       Numbers are also ignored.
#       Tokens are terminated by anything that isn't a letter or a '.'; unless it's in a special character code
#       Also note: "good.neat" => "goodneat"

require 'fileutils'

# Ruby doesn't have quite the notion of an enumeration
class HTMLParserState
 InsideTag = 1
 InsideToken = 2
 InsideSpecial = 3
end

def HTMLParserTokenize( filename )
 state = 0
 templine = ""
 termList = Hash.new( 0 )

 file = File.open( filename )

 file.each {
  |line|

  line.downcase.each_byte {
   |c|
   b = c.chr

   case b
    when '<':
     if state == HTMLParserState::InsideToken then
      # end current token
      if templine.size > 0 then
       templine.delete!( '.' )
       termList[templine] += 1
      end

      templine = ""
      state = HTMLParserState::InsideTag
     end
     if state != HTMLParserState::InsideTag then
      state = HTMLParserState::InsideTag
     end
    when '>':
     if state == HTMLParserState::InsideTag then
      state = HTMLParserState::InsideToken
     end # end if insidetag
    when '&':
     # we only go into specialstate if we are inside a token and by that i
     # mean in the middle/end of a token
     if state == HTMLParserState::InsideToken && templine.size > 0 then
      state = HTMLParserState::InsideSpecial
     end # end if state == insidetoken
    when ';':
     if state == HTMLParserState::InsideSpecial then
      # we go back to regular token because we had to have been in this state before special
      state = HTMLParserState::InsideToken
     end # end if state == insidespecial
    else
     if state == HTMLParserState::InsideToken then
      if (b > 'z' || b < 'a') && b != '.' then
       if templine.size > 0 then
        templine.delete!( '.' )
        termList[templine] += 1
       end
       templine = ""
      elsif (b >= 'a' && b <= 'z') then # we currently ignore urls
       templine.concat( b )
      end # end if b.chr == ...
     elsif state == HTMLParserState::InsideSpecial
      if (b == '\n' || b == ' ' || b == '\t' || b == '\r') then
       templine.delete!( '.' )
       termList[templine] += 1
      end
     end #end if state == HTMLParserState::InsideToken
   end #end case
   } #end each byte
  } #end each line

 file.close
 # we now have all the terms for that file

 return termList

end #end tokenize

# Main Execution
# Parse input parameters
directory = ""
index = ""

if ARGV.size != 2 then
  puts "usage: <input file directory> <index directory>"
  exit
else
  directory = ARGV[0]
  index = ARGV[1]
end

# Is there an index directory?
if File.exists? index then
  puts "Index Directory Already Exists"
else
  FileUtils.mkdir( index )
end

files = Array.new

# Glob up the HTML Files in the Directory
dir = Dir.open( directory )
files = dir.to_a
dir.close

# Tokenize each file
i, j = 0, 0
totalTokens = Hash.new( 0 )
temporaryList = Array.new

# We only want .html files
while i < files.size do
  if files[i] !~ /html$/ then
    files.delete_at( i )
    i -= 1
    else
      shortname = String.new( files[i] )
      shortname =~ /(.*?)\.html/
      temporaryList = HTMLParserTokenize( (files[i].insert 0, directory) ).to_a.sort
      outputFile = File.new( index + $1 + ".txt", "w" )
      j = 0
      while j < temporaryList.size do
        totalTokens[temporaryList[j][0]] += temporaryList[j][1]
        outputFile.puts temporaryList[j][0] + " : " + temporaryList[j][1].to_s
        j += 1
      end
      outputFile.close
      puts "Finished: " + files[i] + " Found: " + temporaryList.size.to_s + " Tokens"
    end
  i += 1
end

puts "Total Tokens: " + totalTokens.size.to_s

# Print out Total Terms List sorted by term
puts "Writing out Complete Term List by Term"
outputFile = File.new( index + "TermListSortedByTerm.txt", "w" )
temporaryList = totalTokens.to_a.sort
i =0
while i < temporaryList.size do
  outputFile.puts temporaryList[i][0] + " : " + temporaryList[i][1].to_s
  i += 1
end
outputFile.close

# Print out Total Terms List sorted by frequency
puts "Writing out Complete Term List by Frequency"
outputFile = File.new( index + "TermListSortedByFrequency.txt", "w" )
# reverse sort code: .sort {|a,b| -1*(a[1]<=>b[1]) }
temporaryList = totalTokens.sort {|a,b| a[1] <=> b[1]}.to_a
i = 0
while i < temporaryList.size do
  outputFile.puts temporaryList[i][0] + " : " + temporaryList[i][1].to_s
  #puts temporaryList[i].to_s
  i += 1
end
outputFile.close
```

Appendix B - Top 50 Tokens - Bottom 50 Tokens by Alphabet

1. a : 24640
2. aa : 10
3. aaa : 24
4. aaaaaamajgaaaaaacwapaaaaaaadadyaaaaaaeaaoqaae : 1
5. aaaeaaiwaqaaaauaaabttvrqaaaaab : 1
6. aaaeab : 1
7. aaaeealktqti : 13
8. aaarundel : 1
9. aacheninformatik : 4
10. aachenoph : 1
11. aachenrad : 1
12. aacutebrndtani : 1
13. aacutegoston : 1
14. aacutelczott : 1
15. aacuteldozatos : 2
16. aacuteldozattal : 1
17. aacuteldozunk : 1
18. aacuteldva : 1
19. aacutelellenzkiek : 1
20. aacutell : 3
21. aacutellam : 6
22. aacutellamalaptsa : 1
23. aacutellamhatrait : 1
24. aacutellami : 3
25. aacutellamisg : 1
26. aacutellamnyelv : 1
27. aacutellamok : 1
28. aacutellamokbl : 1
29. aacutellamotismeri : 1
30. aacutellampolgrainak : 1
31. aacutellampolgri : 1
32. aacutellampolgrok : 4
33. aacutellamprti : 2
34. aacutellamrend : 3
35. aacutellamrendet : 1
36. aacutellamtitkr : 1
37. aacutellamtitkra : 2
38. aacutellamtitkrokat : 1
39. aacutellamuk : 1
40. aacutellamukat : 1
41. aacutelland : 4
42. aacutellandan : 1
43. aacutellapotban : 3
44. aacutellaptanunk : 1
45. aacutelliberlis : 1
46. aacutellitani : 1

47. aacutellitja : 1
48. aacutellitsunk : 1
49. aacutelljunk : 1
50. aacutellnak : 4

51. zugpiacokon : 1
52. zugtak : 2
53. zuhan : 1
54. zuhanast : 1
55. zuhanni : 1
56. zuhanok : 1
57. zuhant : 4
58. zuhantak : 2
59. zuhanttavaly : 1
60. zullese : 1
61. zullest : 1
62. zulleszteni : 1
63. zullhet : 1
64. zullott : 1
65. zullottseg : 1
66. zum : 1
67. zundel : 9
68. zundelrol : 1
69. zundelsite : 1
70. zundelwww : 1
71. zunkel : 1
72. zurcherzeitungban : 1
73. zurck : 1
74. zurich : 3
75. zuros : 1
76. zurzavarba : 1
77. zurzavarban : 1
78. zurzavaros : 1
79. zurzavart : 1
80. zuzzuk : 2
81. zva : 1
82. zvet : 1
83. zvezda : 2
84. zvonimircicak : 1
85. zvornik : 2
86. zvyahilsky : 2
87. zw : 1
88. zwach : 3
89. zwack : 5
90. zwalm : 1
91. zweck : 1
92. zwei : 1
93. zwiers : 1

94. zwxsawdl : 1
95. zxqgpg : 13
96. zyanon : 1
97. zycie : 1
98. zylyftar : 1
99. zyuganov : 63
100. zzzuk : 1
101. zzzzzzz : 78

Appendix C - Top 50 Tokens - Bottom 50 Tokens by Frequency

1. hitelezo : 1
2. cadalanatomy : 1
3. earlham : 1
4. engedelyre : 1
5. tomegdemokraciakantihumanizmusarol : 1
6. edupurdueansc : 1
7. edupurduepmu : 1
8. eduarizonaenrlsvcs : 1
9. helytartok : 1
10. felszolalasok : 1
11. netsunbeltscclmat : 1
12. operal : 1
13. gossip : 1
14. rappaccine : 1
15. govnoaakc : 1
16. kozossegkent : 1
17. jpachokudaimathgalois : 1
18. sherwin : 1
19. hibapontot : 1
20. onegism : 1
21. asegedeszkozoke : 1
22. szobrok : 1
23. ervenyesulnekha : 1
24. bekerulhet : 1
25. kampanyokat : 1
26. specifies : 1
27. jelenunk : 1
28. nehanapjan : 1
29. toucan : 1
30. seikea : 1
31. ludnut : 1
32. osszefonodassal : 1
33. almodozas : 1
34. tanarat : 1
35. tandijrendelet : 1
36. irhato : 1
37. arrogance : 1

38. koraven : 1
39. battlecrhtm : 1
40. nzcosouthnet : 1
41. merlegerol : 1
42. operar : 1
43. edundstudent : 1
44. bik : 1
45. stratosphere : 1
46. jpactiu : 1
47. adougyeinek : 1
48. retorika : 1
49. ertelmisegnek : 1
50. tnccsoportok : 1

51. m : 917
52. its : 936
53. e : 948
54. new : 951
55. la : 960
56. us : 964
57. meg : 965
58. will : 993
59. has : 998
60. have : 1042
61. en : 1070
62. their : 1070
63. analysis : 1072
64. examines : 1076
65. el : 1093
66. was : 1116
67. magyar : 1128
68. his : 1148
69. not : 1156
70. which : 1208
71. this : 1773
72. from : 1789
73. are : 1893
74. be : 2060
75. nem : 2086
76. it : 2148
77. wb : 2319
78. or : 2554
79. hogy : 2742
80. with : 2820
81. es : 2872
82. an : 2889
83. by : 3228
84. de : 3277

85. that : 3356
86. at : 3361
87. as : 3441
88. on : 3905
89. for : 4523
90. is : 5188
91. az : 5594
92. fn : 5898
93. yr : 6430
94. pgs : 6431
95. in : 9857
96. to : 11063
97. s : 11379
98. and : 16994
99. of : 20676
100. a : 24640
101. the : 32589