The Wumpus Information Retrieval System
- How To Use It -

Stefan Büttcher

University of
Waterloo

IR Group Seminar Talk
June 20, 2007

First Step: Downloading & Compiling

A fresh copy of Wumpus can be obtained from

```
http://www.wumpus-search.org/
```

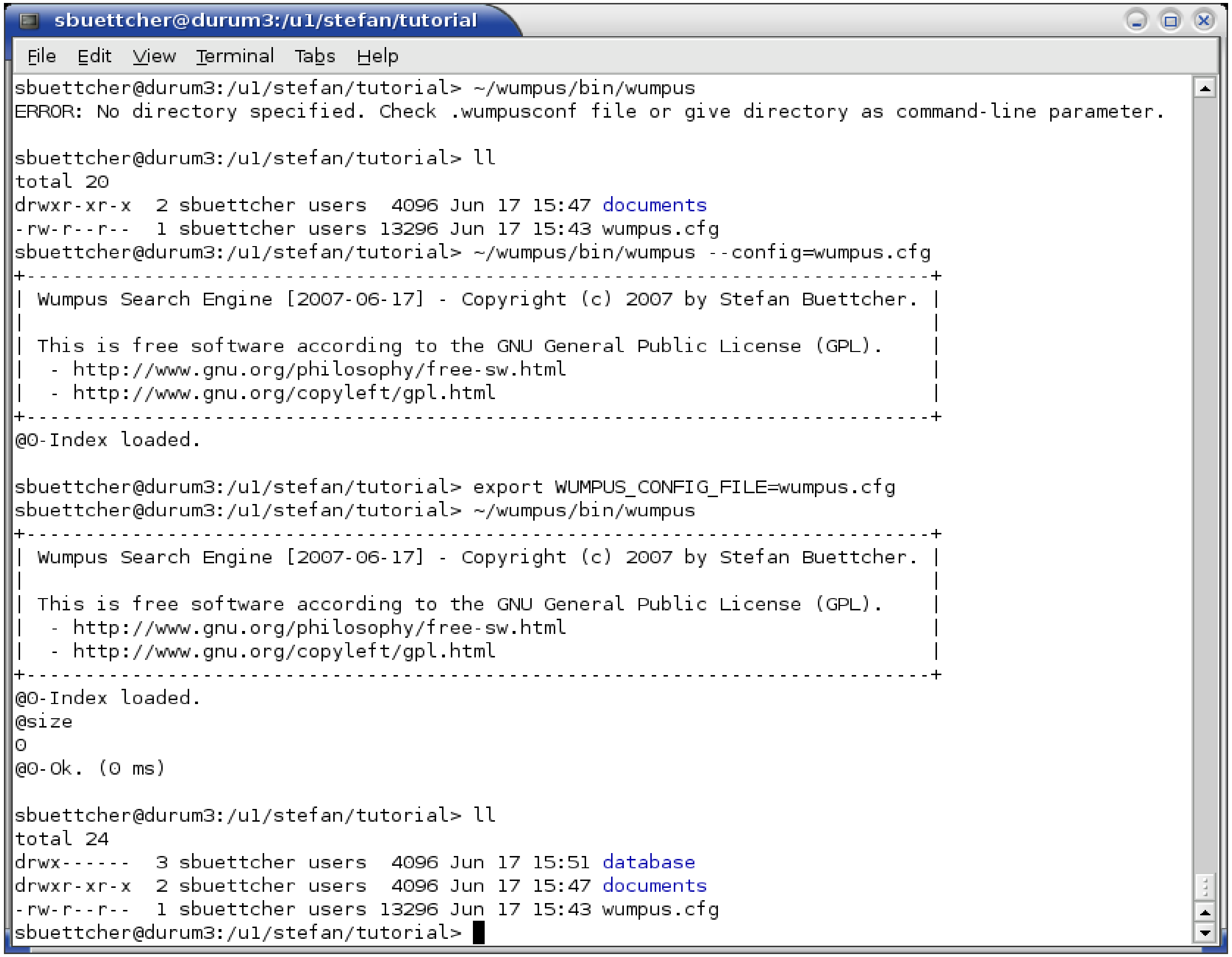
Unpacking and compiling is straightforward:

```
wget www.wumpus-search.org/download/wumpus-2007-06-18.tgz
tar xzf wumpus-2007-06-18.tgz
cd wumpus
make
```

Should work without problems under any Linux installation (32-bit or 64-bit) with gcc > 3.0.

(If not, send me an e-mail.)

Starting Wumpus



```
sbuettcher@durum3:/u1/stefan/tutorial> ~/wumpus/bin/wumpus
ERROR: No directory specified. Check .wumpusconf file or give directory as command-line parameter.

sbuettcher@durum3:/u1/stefan/tutorial> ll
total 20
drwxr-xr-x  2 sbuettcher users  4096 Jun 17 15:47 documents
-rw-r--r--  1 sbuettcher users 13296 Jun 17 15:43 wumpus.cfg
sbuettcher@durum3:/u1/stefan/tutorial> ~/wumpus/bin/wumpus --config=wumpus.cfg
+------------------------------------------------------------------------------+
| Wumpus Search Engine [2007-06-17] - Copyright (c) 2007 by Stefan Buettcher. |
|                                                                              |
| This is free software according to the GNU General Public License (GPL).    |
|   - http://www.gnu.org/philosophy/free-sw.html                              |
|   - http://www.gnu.org/copyleft/gpl.html                                    |
+------------------------------------------------------------------------------+
@0-Index loaded.

sbuettcher@durum3:/u1/stefan/tutorial> export WUMPUS_CONFIG_FILE=wumpus.cfg
sbuettcher@durum3:/u1/stefan/tutorial> ~/wumpus/bin/wumpus
+------------------------------------------------------------------------------+
| Wumpus Search Engine [2007-06-17] - Copyright (c) 2007 by Stefan Buettcher. |
|                                                                              |
| This is free software according to the GNU General Public License (GPL).    |
|   - http://www.gnu.org/philosophy/free-sw.html                              |
|   - http://www.gnu.org/copyleft/gpl.html                                    |
+------------------------------------------------------------------------------+
@0-Index loaded.
@size
0
@0-Ok. (0 ms)

sbuettcher@durum3:/u1/stefan/tutorial> ll
total 24
drwx------  3 sbuettcher users  4096 Jun 17 15:51 database
drwxr-xr-x  2 sbuettcher users  4096 Jun 17 15:47 documents
-rw-r--r--  1 sbuettcher users 13296 Jun 17 15:43 wumpus.cfg
sbuettcher@durum3:/u1/stefan/tutorial> 
```

When you start the engine, you need to give Wumpus a few pieces of information, e.g., where it should store its index data.

These setting are normally defined in the `wumpus.cfg` file (found in the Wumpus base directory). For a default config file, use $HOME/.wumpusconf or the WUMPUS_CONFIG_FILE environment variable.

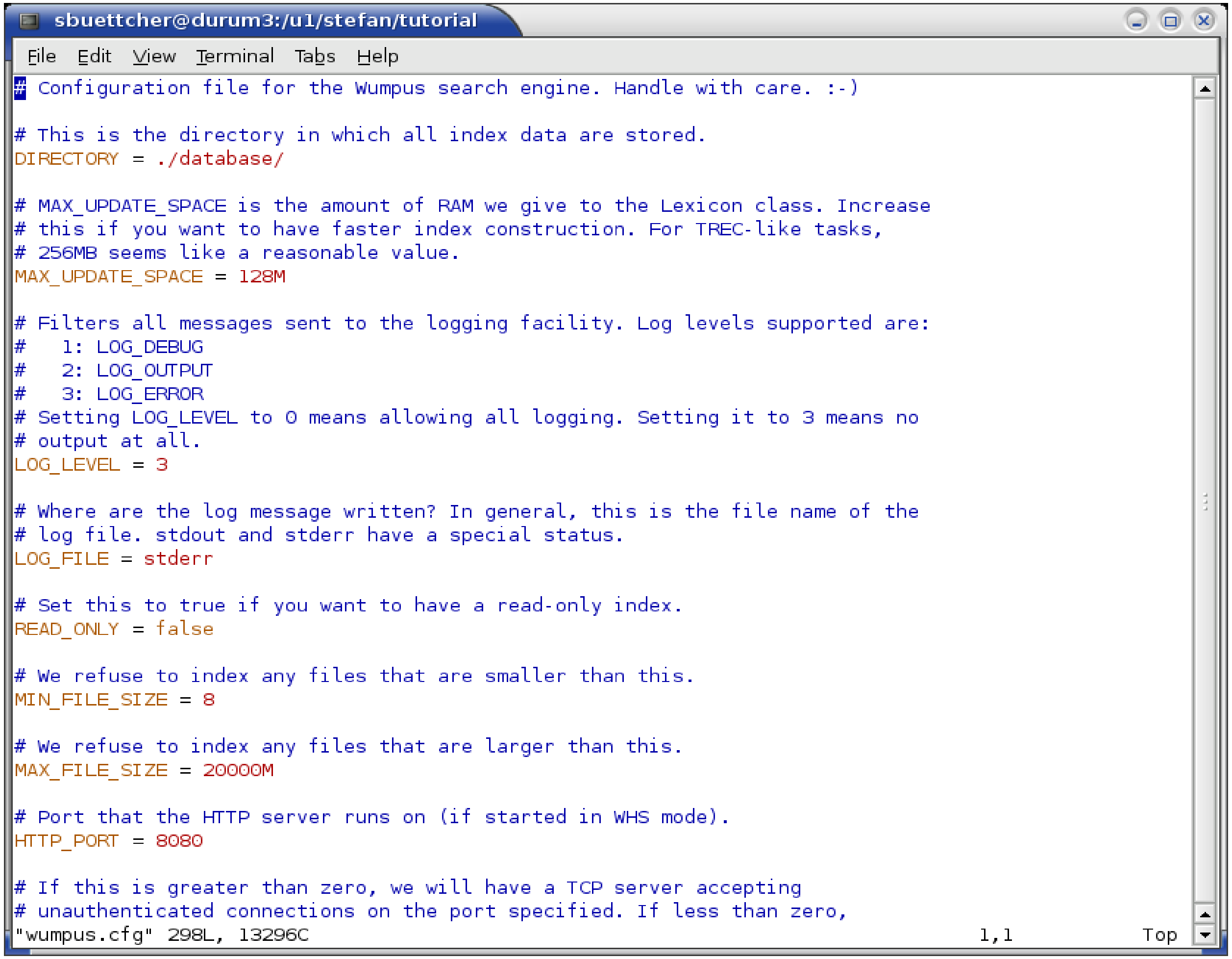
All parameters can be overridden via command-line parameters.

Example:

```
bin/wumpus --config=wumpus.cfg DIRECTORY=/wumpus/indexdir
```

Stefan Büttcher
<sbuettch@uwaterloo.ca>

The Configuration File


sbuettcher@durum3:/u1/stefan/tutorial
File Edit View Terminal Tabs Help

```
# Configuration file for the Wumpus search engine. Handle with care. :-)

# This is the directory in which all index data are stored.
DIRECTORY = ./database/

# MAX_UPDATE_SPACE is the amount of RAM we give to the Lexicon class. Increase
# this if you want to have faster index construction. For TREC-like tasks,
# 256MB seems like a reasonable value.
MAX_UPDATE_SPACE = 128M

# Filters all messages sent to the logging facility. Log levels supported are:
#    1: LOG_DEBUG
#    2: LOG_OUTPUT
#    3: LOG_ERROR
# Setting LOG_LEVEL to 0 means allowing all logging. Setting it to 3 means no
# output at all.
LOG_LEVEL = 3

# Where are the log message written? In general, this is the file name of the
# log file. stdout and stderr have a special status.
LOG_FILE = stderr

# Set this to true if you want to have a read-only index.
READ_ONLY = false

# We refuse to index any files that are smaller than this.
MIN_FILE_SIZE = 8

# We refuse to index any files that are larger than this.
MAX_FILE_SIZE = 20000M

# Port that the HTTP server runs on (if started in WHS mode).
HTTP_PORT = 8080

# If this is greater than zero, we will have a TCP server accepting
# unauthenticated connections on the port specified. If less than zero,
"wumpus.cfg" 298L, 13296C                                        1,1           Top
```

The Wumpus configuration file lets you adjust various parameters of the system, such as the location of the index files, the amount of RAM to use for index updates, and the index update strategy employed.

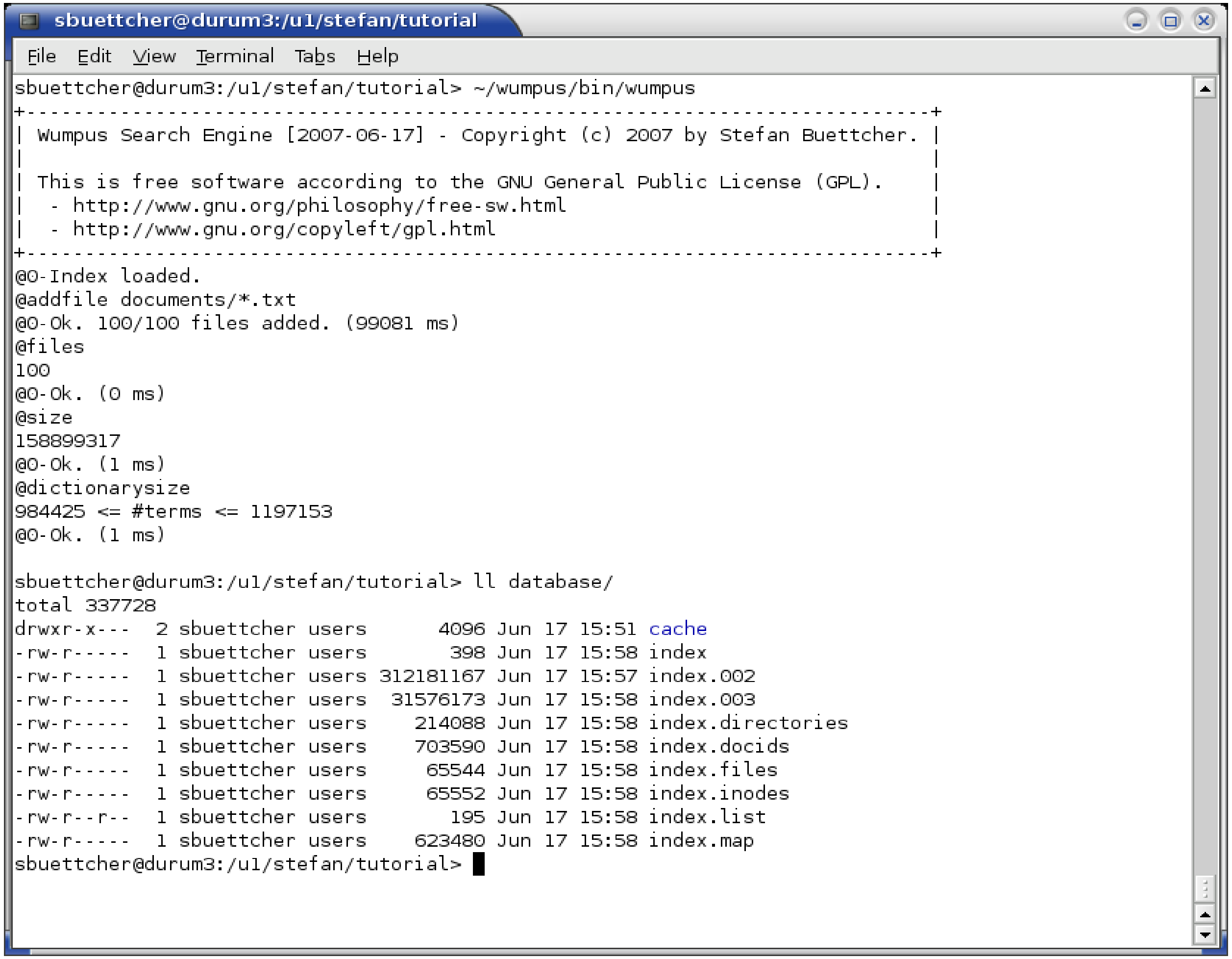
At startup, wumpus checks for the presence of the files:

1. /etc/wumpusconf
2. $HOME/.wumpusconf
3. $WUMPUS_CONFIG_FILE

and processes them in this order.

Later definitions override earlier ones (allows you to use differential config files).

Building an Index


sbuettcher@durum3:/u1/stefan/tutorial
File Edit View Terminal Tabs Help

```
sbuettcher@durum3:/u1/stefan/tutorial> ~/wumpus/bin/wumpus
+-----------------------------------------------------------------------------+
| Wumpus Search Engine [2007-06-17] - Copyright (c) 2007 by Stefan Buettcher. |
|                                                                             |
| This is free software according to the GNU General Public License (GPL).   |
|  - http://www.gnu.org/philosophy/free-sw.html                              |
|  - http://www.gnu.org/copyleft/gpl.html                                    |
+-----------------------------------------------------------------------------+
@0-Index loaded.
@addfile documents/*.txt
@0-Ok. 100/100 files added. (99081 ms)
@files
100
@0-Ok. (0 ms)
@size
158899317
@0-Ok. (1 ms)
@dictionarysize
984425 <= #terms <= 1197153
@0-Ok. (1 ms)

sbuettcher@durum3:/u1/stefan/tutorial> ll database/
total 337728
drwxr-x---  2 sbuettcher users      4096 Jun 17 15:51 cache
-rw-r-----  1 sbuettcher users       398 Jun 17 15:58 index
-rw-r-----  1 sbuettcher users 312181167 Jun 17 15:57 index.002
-rw-r-----  1 sbuettcher users  31576173 Jun 17 15:58 index.003
-rw-r-----  1 sbuettcher users    214088 Jun 17 15:58 index.directories
-rw-r-----  1 sbuettcher users    703590 Jun 17 15:58 index.docids
-rw-r-----  1 sbuettcher users     65544 Jun 17 15:58 index.files
-rw-r-----  1 sbuettcher users     65552 Jun 17 15:58 index.inodes
-rw-r--r--  1 sbuettcher users       195 Jun 17 15:58 index.list
-rw-r-----  1 sbuettcher users    623480 Jun 17 15:58 index.map
sbuettcher@durum3:/u1/stefan/tutorial> 
```

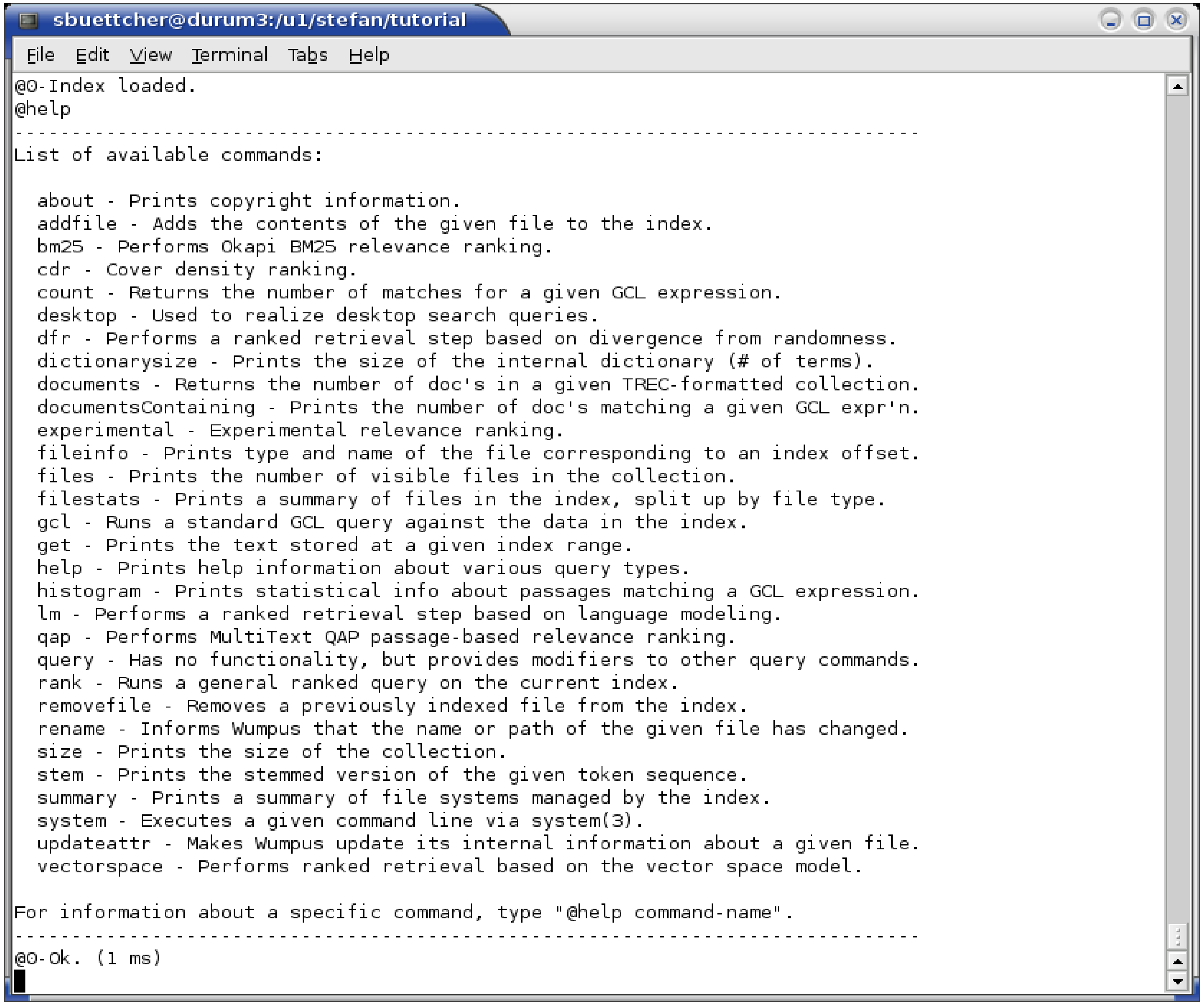
When running Wumpus for the first time, an empty index will be created. The index can easily be populated with data by issuing an @addfile command.

@addfile takes an individual file name or a file name pattern (as in the example).

Basic index statistics can be obtained through @files (number of files indexed), @size (number of tokens indexed), and @dictionarysize (number of distinct terms in index).

In the example, the exact dictionary size is unknown (lower and upper bound are given), because the index consists of multiple partitions. *For performance reasons, Wumpus does not maintain a global dictionary!*

The Helpful User Interface



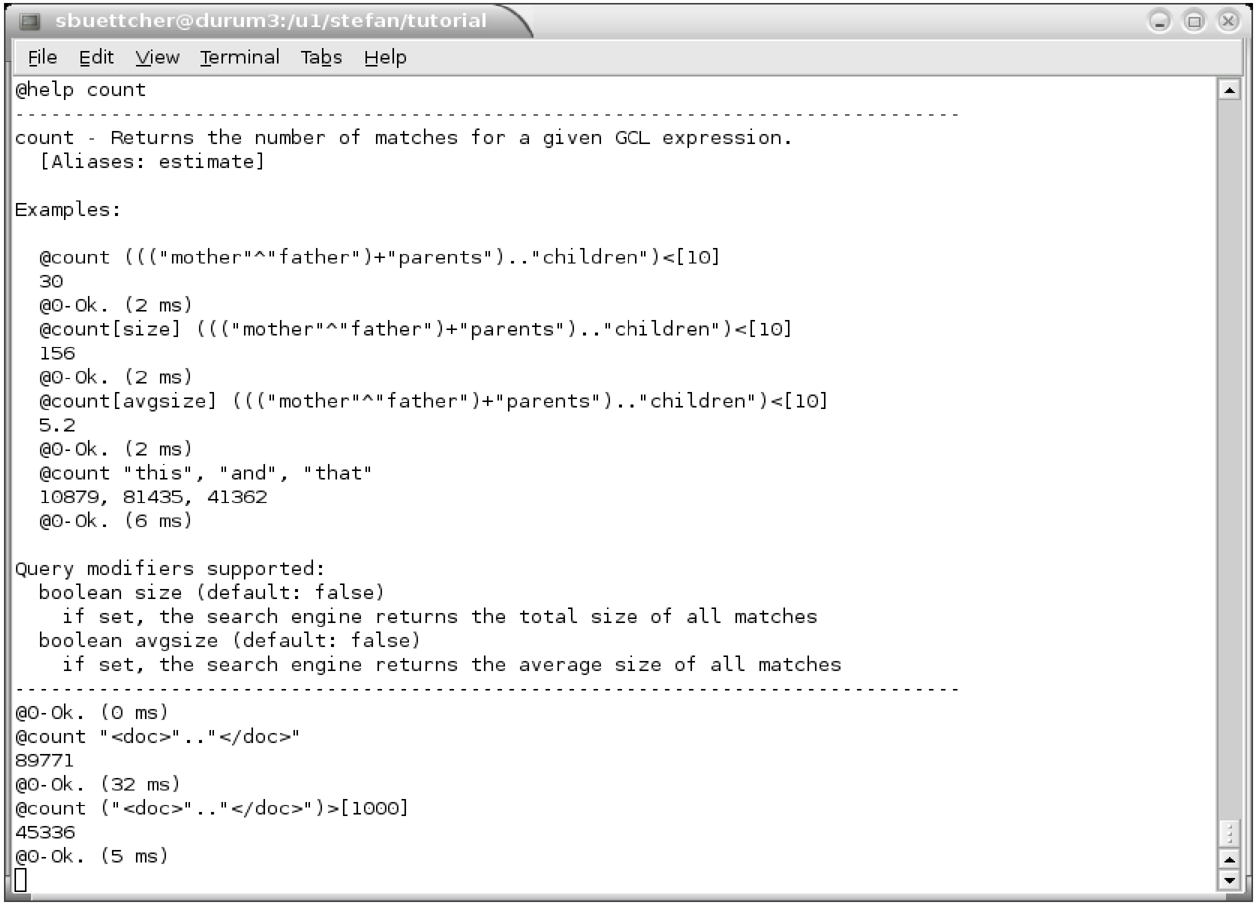
sbuettcher@durum3:/u1/stefan/tutorial
File Edit View Terminal Tabs Help

```
@0-Index loaded.
@help
--------------------------------------------------------------------------------
List of available commands:

  about - Prints copyright information.
  addfile - Adds the contents of the given file to the index.
  bm25 - Performs Okapi BM25 relevance ranking.
  cdr - Cover density ranking.
  count - Returns the number of matches for a given GCL expression.
  desktop - Used to realize desktop search queries.
  dfr - Performs a ranked retrieval step based on divergence from randomness.
  dictionarysize - Prints the size of the internal dictionary (# of terms).
  documents - Returns the number of doc's in a given TREC-formatted collection.
  documentsContaining - Prints the number of doc's matching a given GCL expr'n.
  experimental - Experimental relevance ranking.
  fileinfo - Prints type and name of the file corresponding to an index offset.
  files - Prints the number of visible files in the collection.
  filestats - Prints a summary of files in the index, split up by file type.
  gcl - Runs a standard GCL query against the data in the index.
  get - Prints the text stored at a given index range.
  help - Prints help information about various query types.
  histogram - Prints statistical info about passages matching a GCL expression.
  lm - Performs a ranked retrieval step based on language modeling.
  qap - Performs MultiText QAP passage-based relevance ranking.
  query - Has no functionality, but provides modifiers to other query commands.
  rank - Runs a general ranked query on the current index.
  removefile - Removes a previously indexed file from the index.
  rename - Informs Wumpus that the name or path of the given file has changed.
  size - Prints the size of the collection.
  stem - Prints the stemmed version of the given token sequence.
  summary - Prints a summary of file systems managed by the index.
  system - Executes a given command line via system(3).
  updateattr - Makes Wumpus update its internal information about a given file.
  vectorspace - Performs ranked retrieval based on the vector space model.

For information about a specific command, type "@help command-name".
--------------------------------------------------------------------------------
@0-Ok. (1 ms)
```

Finding your way through the user interface: the @help command.

@help prints a list of most commands. For information about a specific command, type “@help *cmd-name*”, e.g., “@help addfile”.

First Steps: Counting the Documents


sbuettcher@durum3:/u1/stefan/tutorial
File Edit View Terminal Tabs Help

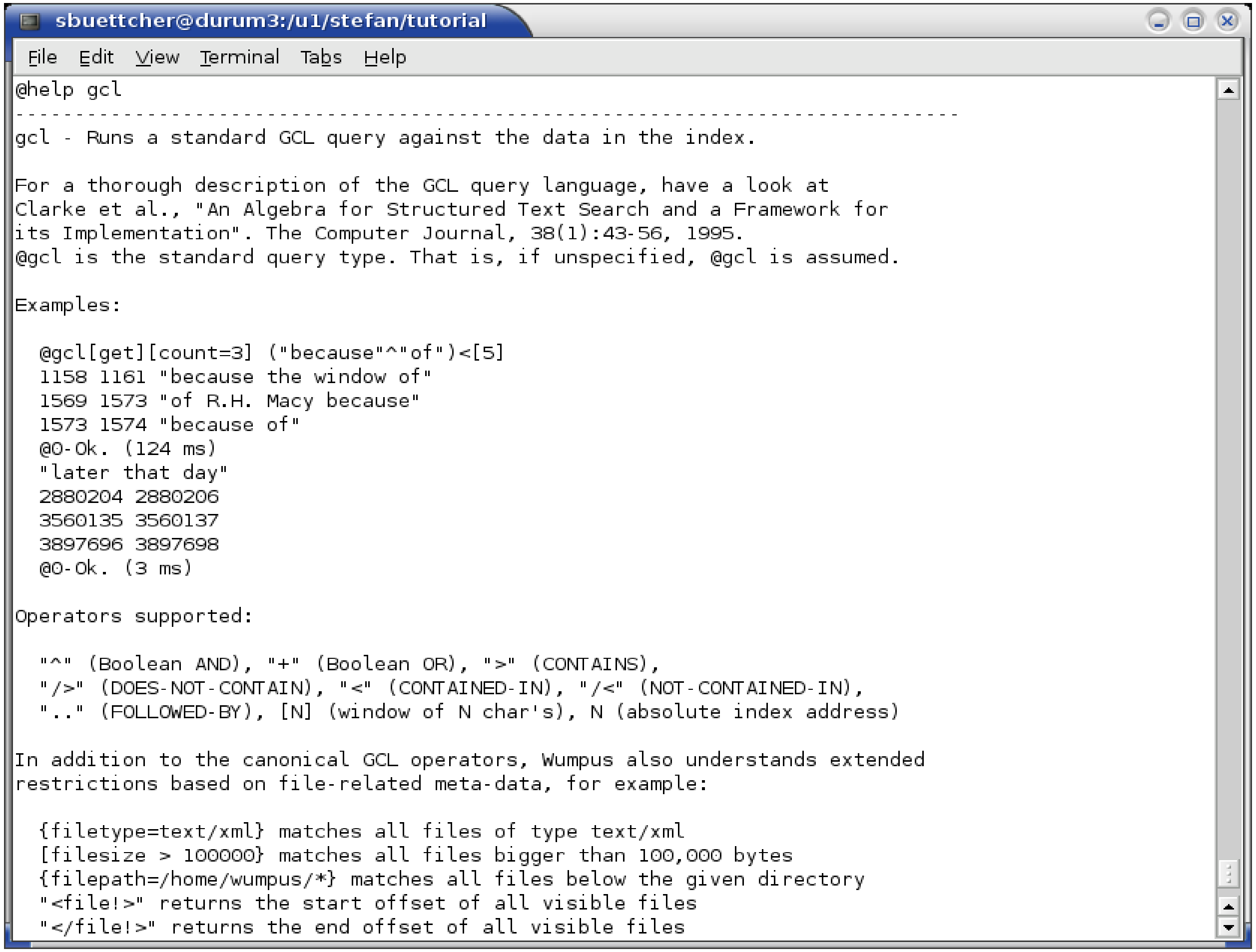
```
@help count
--------------------------------------------------------------------------------
count - Returns the number of matches for a given GCL expression.
  [Aliases: estimate]

Examples:

  @count ((("mother"^"father")+"parents").."children")<[10]
  30
  @0-Ok. (2 ms)
  @count[size] ((("mother"^"father")+"parents").."children")<[10]
  156
  @0-Ok. (2 ms)
  @count[avgsize] ((("mother"^"father")+"parents").."children")<[10]
  5.2
  @0-Ok. (2 ms)
  @count "this", "and", "that"
  10879, 81435, 41362
  @0-Ok. (6 ms)

Query modifiers supported:
  boolean size (default: false)
    if set, the search engine returns the total size of all matches
  boolean avgsize (default: false)
    if set, the search engine returns the average size of all matches
--------------------------------------------------------------------------------
@0-Ok. (0 ms)
@count "<doc>".."</doc>"
89771
@0-Ok. (32 ms)
@count ("<doc>".."</doc>")>[1000]
45336
@0-Ok. (5 ms)
```

After the collection has been indexed, queries can be run against the index.

For example: counting the number of documents in the collection, or counting the number of documents containing at least 1000 tokens.

All queries are based on the GCL retrieval framework (Clarke et al.) that lets you freely combine the contents of different postings lists in the index.

The GCL Query Language



```
@help gcl
-------------------------------------------------------------------------------
gcl - Runs a standard GCL query against the data in the index.

For a thorough description of the GCL query language, have a look at
Clarke et al., "An Algebra for Structured Text Search and a Framework for
its Implementation". The Computer Journal, 38(1):43-56, 1995.
@gcl is the standard query type. That is, if unspecified, @gcl is assumed.

Examples:

  @gcl[get][count=3] ("because"^"of")<[5]
  1158 1161 "because the window of"
  1569 1573 "of R.H. Macy because"
  1573 1574 "because of"
  @0-Ok. (124 ms)
  "later that day"
  2880204 2880206
  3560135 3560137
  3897696 3897698
  @0-Ok. (3 ms)

Operators supported:

  "^" (Boolean AND), "+" (Boolean OR), ">" (CONTAINS),
  "/>" (DOES-NOT-CONTAIN), "<" (CONTAINED-IN), "/<" (NOT-CONTAINED-IN),
  ".." (FOLLOWED-BY), [N] (window of N char's), N (absolute index address)

In addition to the canonical GCL operators, Wumpus also understands extended
restrictions based on file-related meta-data, for example:

  {filetype=text/xml} matches all files of type text/xml
  [filesize > 100000} matches all files bigger than 100,000 bytes
  {filepath=/home/wumpus/*} matches all files below the given directory
  "<file!>" returns the start offset of all visible files
  "</file!>" returns the end offset of all visible files
```

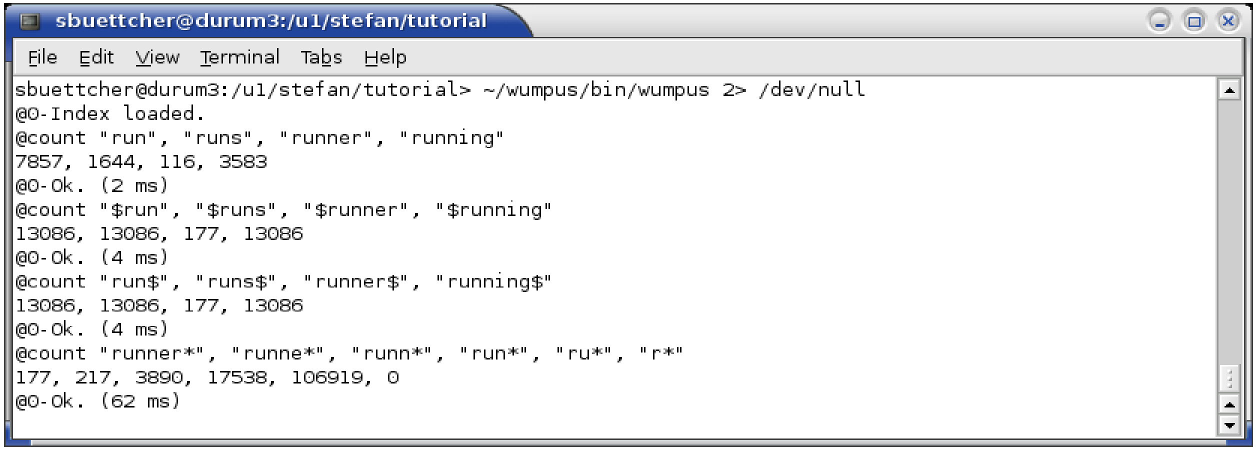
For a brief summary of the GCL query language, type “@help gcl” or have a look at Clarke's 1995 paper on the topic.

In addition to the basic GCL operators, there are some Wumpus-specific extensions that can be used to refer to the files from which the index has been built.

`“<file!>”..”</file!>”`:
A list of all files in the index.

Be careful with the curly-bracket expressions (e.g., “{filepath=...}”). They can be quite expensive to evaluate.

Stemming and Prefix Queries


sbuettcher@durum3:/u1/stefan/tutorial
File Edit View Terminal Tabs Help

```
sbuettcher@durum3:/u1/stefan/tutorial> ~/wumpus/bin/wumpus 2> /dev/null
@0-Index loaded.
@count "run", "runs", "runner", "running"
7857, 1644, 116, 3583
@0-Ok. (2 ms)
@count "$run", "$runs", "$runner", "$running"
13086, 13086, 177, 13086
@0-Ok. (4 ms)
@count "run$", "runs$", "runner$", "running$"
13086, 13086, 177, 13086
@0-Ok. (4 ms)
@count "runner*", "runne*", "runn*", "run*", "ru*", "r*"
177, 217, 3890, 17538, 106919, 0
@0-Ok. (62 ms)
```

Wumpus supports query-time stemming and prefix queries.

The “$” symbol indicates that a word should be stemmed (Porter stemmer). Can be either before or after the term.

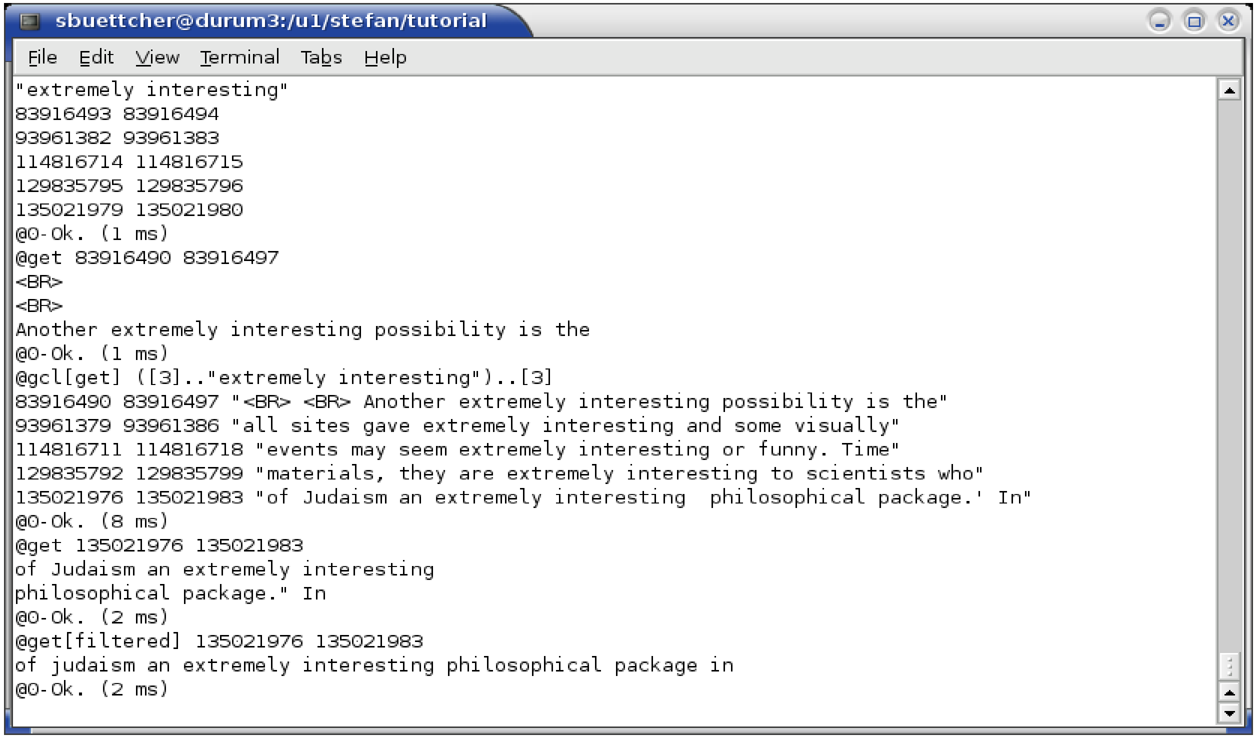
Prefix queries are only supported for a prefix of length ≥ 2. But even so, prefix queries can be quite expensive to evaluate. Use with caution.

Stemming is a time-vs.-space trade-off:

- Best query performance: Indexing each posting twice (stemmed and unstemmed). But index is twice as large.
- Smallest index: Doing all stemming at query time. But requires to collect all terms with same stem and combining their postings lists.

Controlled by the STEMMING_LEVEL configuration variable.

Obtaining Text from the Input Documents


sbuettcher@durum3:/u1/stefan/tutorial
File Edit View Terminal Tabs Help

```
"extremely interesting"
83916493 83916494
93961382 93961383
114816714 114816715
129835795 129835796
135021979 135021980
@0-Ok. (1 ms)
@get 83916490 83916497
<BR>
<BR>
Another extremely interesting possibility is the
@0-Ok. (1 ms)
@gcl[get] ([3].."extremely interesting")..[3]
83916490 83916497 "<BR> <BR> Another extremely interesting possibility is the"
93961379 93961386 "all sites gave extremely interesting and some visually"
114816711 114816718 "events may seem extremely interesting or funny. Time"
129835792 129835799 "materials, they are extremely interesting to scientists who"
135021976 135021983 "of Judaism an extremely interesting  philosophical package.' In"
@0-Ok. (8 ms)
@get 135021976 135021983
of Judaism an extremely interesting
philosophical package." In
@0-Ok. (2 ms)
@get[filtered] 135021976 135021983
of judaism an extremely interesting philosophical package in
@0-Ok. (2 ms)
```

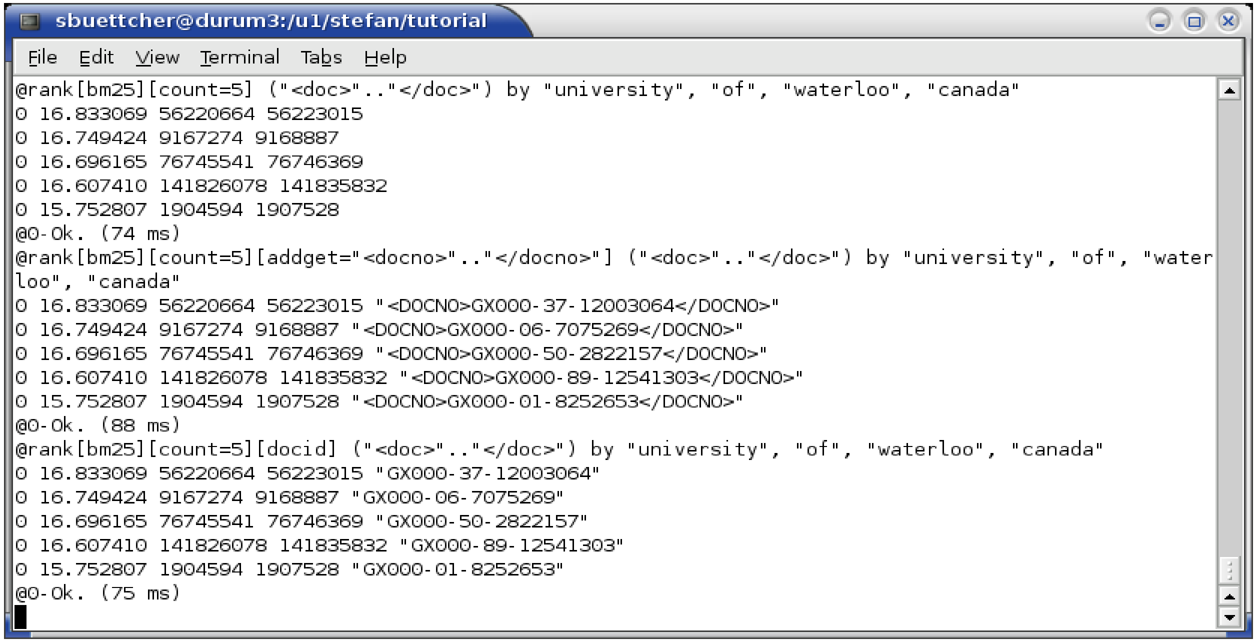
If no query type is given, it is assumed that the query is a simple @gcl query. The results to such a query are the start and end positions of all text passages matching the query.

After the matching passages have been identified, the actual text can be obtained by issuing an @get query.

For convenience, the @get query can be integrated into an @gcl query. However, the results are limited to about 80 characters (and the text returned might be slightly different from the original: note the apostrophe in the fifth result line for the @gcl[get] query).

Wumpus does not maintain a copy of the text in the input files. If you delete (or move) the file, Wumpus will not be able to obtain the text any more.

Ranked Retrieval Operations


sbuettcher@durum3:/u1/stefan/tutorial
File Edit View Terminal Tabs Help

```
@rank[bm25][count=5] ("<doc>".."</doc>") by "university", "of", "waterloo", "canada"
0 16.833069 56220664 56223015
0 16.749424 9167274 9168887
0 16.696165 76745541 76746369
0 16.607410 141826078 141835832
0 15.752807 1904594 1907528
@0-Ok. (74 ms)
@rank[bm25][count=5][addget="<docno>".."</docno>"] ("<doc>".."</doc>") by "university", "of", "water
loo", "canada"
0 16.833069 56220664 56223015 "<DOCNO>GX000-37-12003064</DOCNO>"
0 16.749424 9167274 9168887 "<DOCNO>GX000-06-7075269</DOCNO>"
0 16.696165 76745541 76746369 "<DOCNO>GX000-50-2822157</DOCNO>"
0 16.607410 141826078 141835832 "<DOCNO>GX000-89-12541303</DOCNO>"
0 15.752807 1904594 1907528 "<DOCNO>GX000-01-8252653</DOCNO>"
@0-Ok. (88 ms)
@rank[bm25][count=5][docid] ("<doc>".."</doc>") by "university", "of", "waterloo", "canada"
0 16.833069 56220664 56223015 "GX000-37-12003064"
0 16.749424 9167274 9168887 "GX000-06-7075269"
0 16.696165 76745541 76746369 "GX000-50-2822157"
0 16.607410 141826078 141835832 "GX000-89-12541303"
0 15.752807 1904594 1907528 "GX000-01-8252653"
@0-Ok. (75 ms)
```

@rank queries can be used to rank index extents (e.g., documents) according to their similarity to a given query.

Multiple implementations:
@rank[bm25], @rank[qap],
@rank[vectorspace], ...

Shortcuts exist:
@bm25, @qap, @vsm, ...

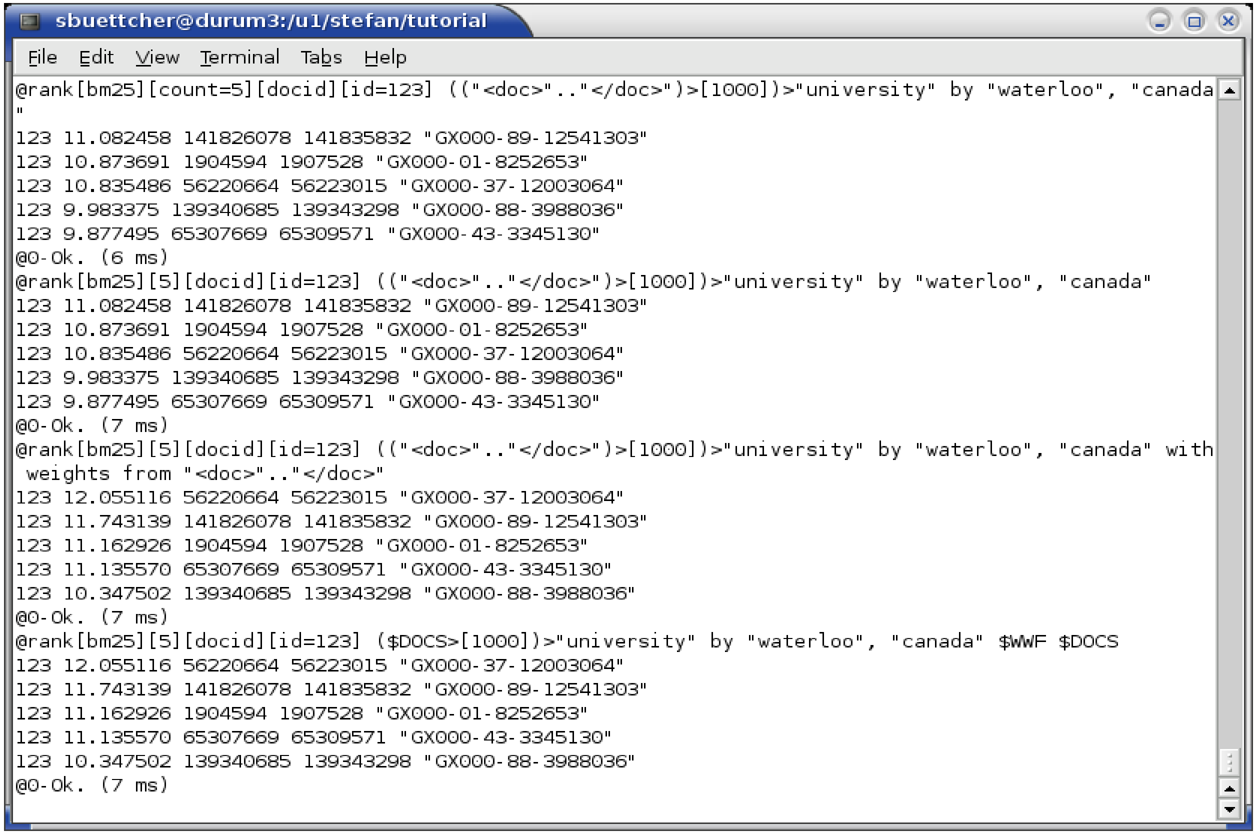
General syntax for all ranked queries:

@rank[...] *GCL* by *GCL*, *GCL*, ..., *GCL*

It is possible to obtain additional information about a matching document by passing an [addget] modifier to the query class. Here: Obtain the document identifiers in a TREC-formatted collection. Faster than [addget]: [docid] (cached in RAM).

Use the [trec] modifier to produce output that can be understood by `trec_eval`*.*

Ranked Retrieval Operations


sbuettcher@durum3:/u1/stefan/tutorial
File Edit View Terminal Tabs Help

```
@rank[bm25][count=5][docid][id=123] (("<doc>".."</doc>")>[1000])>"university" by "waterloo", "canada
"
123 11.082458 141826078 141835832 "GX000-89-12541303"
123 10.873691 1904594 1907528 "GX000-01-8252653"
123 10.835486 56220664 56223015 "GX000-37-12003064"
123 9.983375 139340685 139343298 "GX000-88-3988036"
123 9.877495 65307669 65309571 "GX000-43-3345130"
@0-Ok. (6 ms)
@rank[bm25][5][docid][id=123] (("<doc>".."</doc>")>[1000])>"university" by "waterloo", "canada"
123 11.082458 141826078 141835832 "GX000-89-12541303"
123 10.873691 1904594 1907528 "GX000-01-8252653"
123 10.835486 56220664 56223015 "GX000-37-12003064"
123 9.983375 139340685 139343298 "GX000-88-3988036"
123 9.877495 65307669 65309571 "GX000-43-3345130"
@0-Ok. (7 ms)
@rank[bm25][5][docid][id=123] (("<doc>".."</doc>")>[1000])>"university" by "waterloo", "canada" with
 weights from "<doc>".."</doc>"
123 12.055116 56220664 56223015 "GX000-37-12003064"
123 11.743139 141826078 141835832 "GX000-89-12541303"
123 11.162926 1904594 1907528 "GX000-01-8252653"
123 11.135570 65307669 65309571 "GX000-43-3345130"
123 10.347502 139340685 139343298 "GX000-88-3988036"
@0-Ok. (7 ms)
@rank[bm25][5][docid][id=123] ($DOCS>[1000])>"university" by "waterloo", "canada" $WWF $DOCS
123 12.055116 56220664 56223015 "GX000-37-12003064"
123 11.743139 141826078 141835832 "GX000-89-12541303"
123 11.162926 1904594 1907528 "GX000-01-8252653"
123 11.135570 65307669 65309571 "GX000-43-3345130"
123 10.347502 139340685 139343298 "GX000-88-3988036"
@0-Ok. (7 ms)
```

The retrieval unit, as well as the scorers, in an @rank query may be arbitrary GCL expressions.

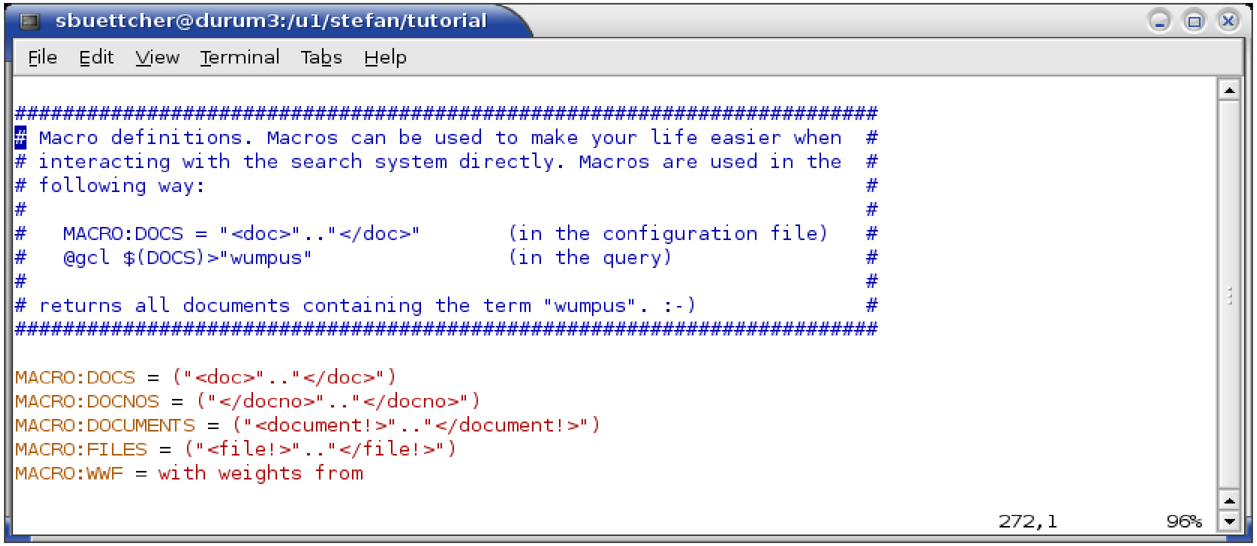
This allows you to mix ranked retrieval operations with Boolean constraints.

However, be careful: The statistics used to rank matching retrieval units are taken exclusively from the set of index extents matching the first GCL expression (=> sparse data!).

Can be circumvented by specifying the source of the corpus statistics to be used for ranking: @rank *GCL* by *GCL*, ..., *GCL* with weights from *GCL*

For lazy people: use macros (defined in wumpus.cfg).
$DOCS -> (“<doc>”..”</doc>”); $WWF -> with weights from

Macro Definitions

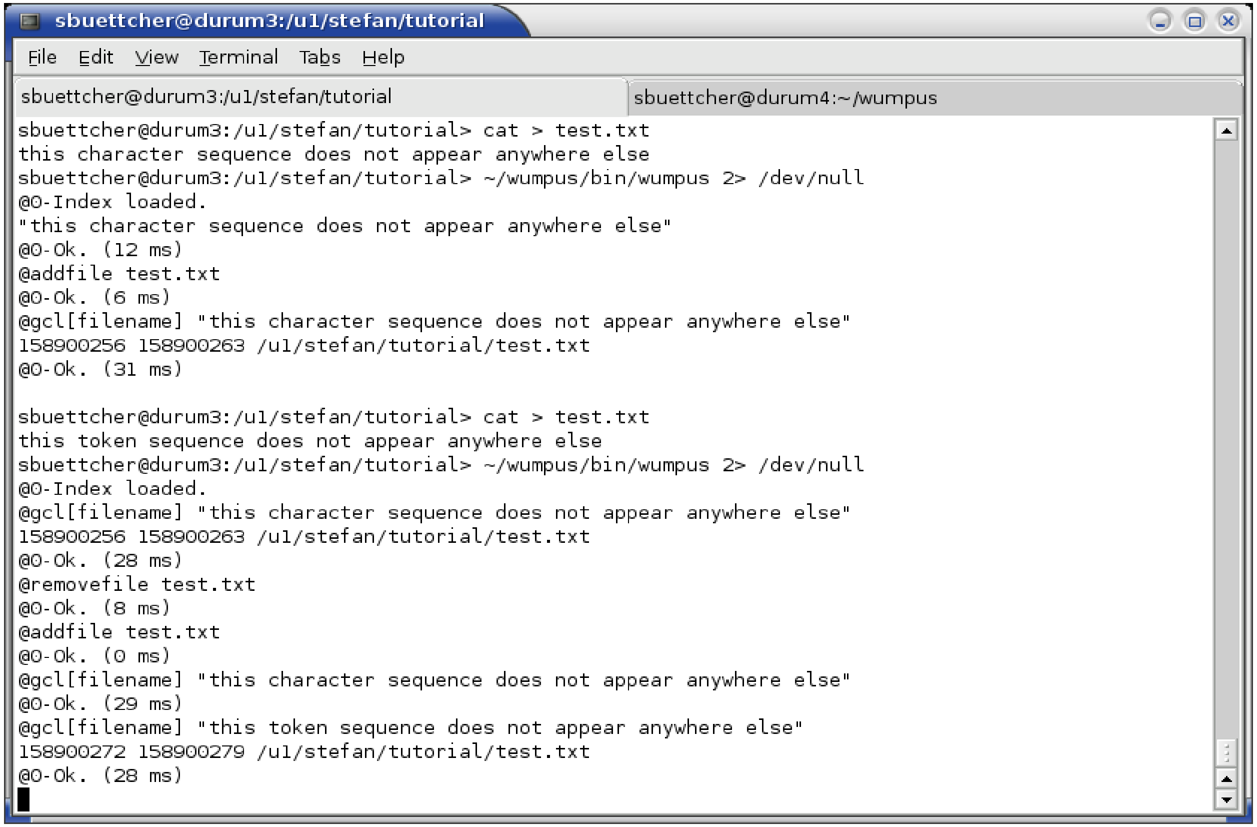

sbuettcher@durum3:/u1/stefan/tutorial
File Edit View Terminal Tabs Help

```
########################################################################
# Macro definitions. Macros can be used to make your life easier when  #
# interacting with the search system directly. Macros are used in the  #
# following way:                                                       #
#                                                                      #
#   MACRO:DOCS = "<doc>".."</doc>"        (in the configuration file)  #
#   @gcl $(DOCS)>"wumpus"                 (in the query)               #
#                                                                      #
# returns all documents containing the term "wumpus". :-)              #
########################################################################

MACRO:DOCS = ("<doc>".."</doc>")
MACRO:DOCNOS = ("</docno>".."</docno>")
MACRO:DOCUMENTS = ("<document!>".."</document!>")
MACRO:FILES = ("<file!>".."</file!>")
MACRO:WWF = with weights from
                                                                  272,1          96%
```

Ugly syntax, but pretty useful nonetheless: macro definitions in `wumpus.cfg`.

Don't forget to put parentheses around your macro definitions. Otherwise, you might be surprised by the results (same as with `#define` in C/C++).

Index Updates


sbuettcher@durum3:/u1/stefan/tutorial
File Edit View Terminal Tabs Help
sbuettcher@durum3:/u1/stefan/tutorial
sbuettcher@durum4:~/wumpus

```
sbuettcher@durum3:/u1/stefan/tutorial> cat > test.txt
this character sequence does not appear anywhere else
sbuettcher@durum3:/u1/stefan/tutorial> ~/wumpus/bin/wumpus 2> /dev/null
@0-Index loaded.
"this character sequence does not appear anywhere else"
@0-Ok. (12 ms)
@addfile test.txt
@0-Ok. (6 ms)
@gcl[filename] "this character sequence does not appear anywhere else"
158900256 158900263 /u1/stefan/tutorial/test.txt
@0-Ok. (31 ms)

sbuettcher@durum3:/u1/stefan/tutorial> cat > test.txt
this token sequence does not appear anywhere else
sbuettcher@durum3:/u1/stefan/tutorial> ~/wumpus/bin/wumpus 2> /dev/null
@0-Index loaded.
@gcl[filename] "this character sequence does not appear anywhere else"
158900256 158900263 /u1/stefan/tutorial/test.txt
@0-Ok. (28 ms)
@removefile test.txt
@0-Ok. (8 ms)
@addfile test.txt
@0-Ok. (0 ms)
@gcl[filename] "this character sequence does not appear anywhere else"
@0-Ok. (29 ms)
@gcl[filename] "this token sequence does not appear anywhere else"
158900272 158900279 /u1/stefan/tutorial/test.txt
@0-Ok. (28 ms)
```

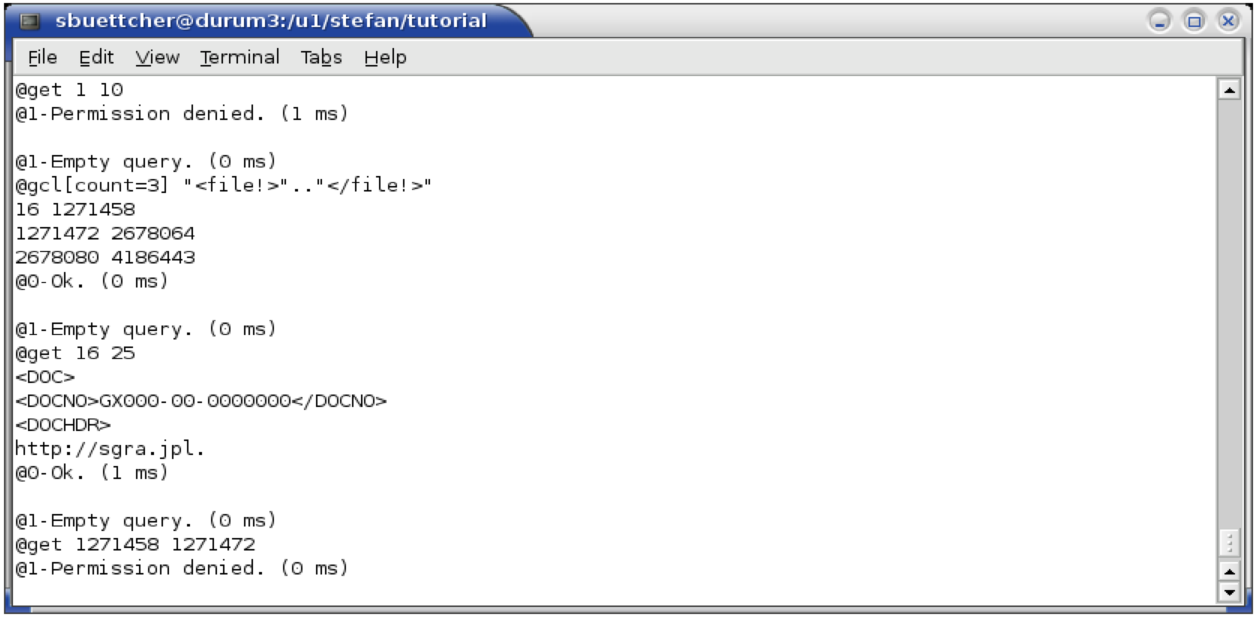
Want to add a new file to the index? Just send another @addfile command to the engine.

Want the system to re-index a previously indexed file? @removefile/@addfile is what you need.

New files are always added at the very end of the used portion of the index address space (see example).

Re-indexing a file changes its position in the address space.

Files, Offsets, and “Permission denied”

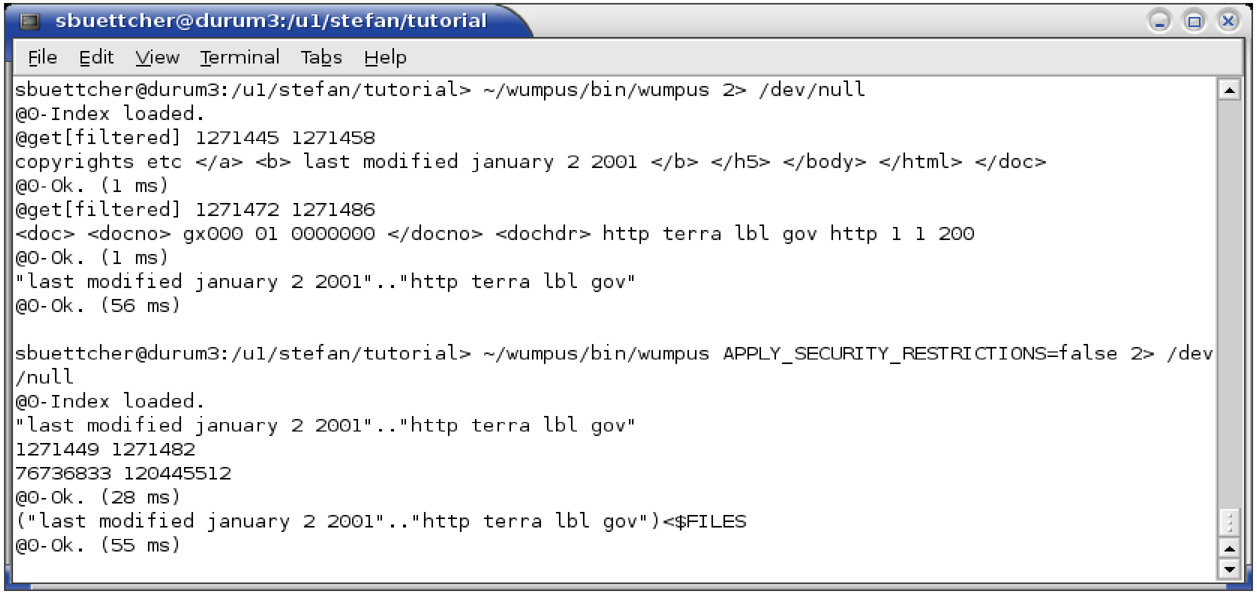

sbuettcher@durum3:/u1/stefan/tutorial
File Edit View Terminal Tabs Help

```
@get 1 10
@1-Permission denied. (1 ms)

@1-Empty query. (0 ms)
@gcl[count=3] "<file!>".."</file!>"
16 1271458
1271472 2678064
2678080 4186443
@0-Ok. (0 ms)

@1-Empty query. (0 ms)
@get 16 25
<DOC>
<DOCNO>GX000-00-0000000</DOCNO>
<DOCHDR>
http://sgra.jpl.
@0-Ok. (1 ms)

@1-Empty query. (0 ms)
@get 1271458 1271472
@1-Permission denied. (0 ms)
```

An “@get 1 10” command does not return the first 10 tokens in the collection, but a nasty error message.

Each file is associated with an index range (usually starts at a multiple of 16). An @get query cannot span beyond the boundaries of a single file.

Files, Offsets, and “Permission denied”


sbuettcher@durum3:/u1/stefan/tutorial
File Edit View Terminal Tabs Help

```
sbuettcher@durum3:/u1/stefan/tutorial> ~/wumpus/bin/wumpus 2> /dev/null
@0-Index loaded.
@get[filtered] 1271445 1271458
copyrights etc </a> <b> last modified january 2 2001 </b> </h5> </body> </html> </doc>
@0-Ok. (1 ms)
@get[filtered] 1271472 1271486
<doc> <docno> gx000 01 0000000 </docno> <dochdr> http terra lbl gov http 1 1 200
@0-Ok. (1 ms)
"last modified january 2 2001".."http terra lbl gov"
@0-Ok. (56 ms)

sbuettcher@durum3:/u1/stefan/tutorial> ~/wumpus/bin/wumpus APPLY_SECURITY_RESTRICTIONS=false 2> /dev
/null
@0-Index loaded.
"last modified january 2 2001".."http terra lbl gov"
1271449 1271482
76736833 120445512
@0-Ok. (28 ms)
("last modified january 2 2001".."http terra lbl gov")<$FILES
@0-Ok. (55 ms)
```

The same is true for GCL queries. The result to a GCL query will never span across file boundaries (this even holds for intermediate results of a GCL query).

Reason: Wumpus is a multi-user system. Security restrictions (= file permissions) are easier to enforce this way.

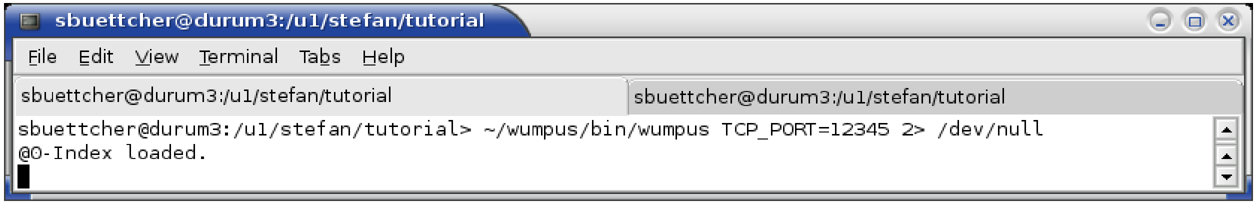
If you *really* have to use GCL expressions that match text passages spanning across multiple files, then turn off the security subsystem:

```
APPLY_SECURITY_RESTRICTIONS=false
```

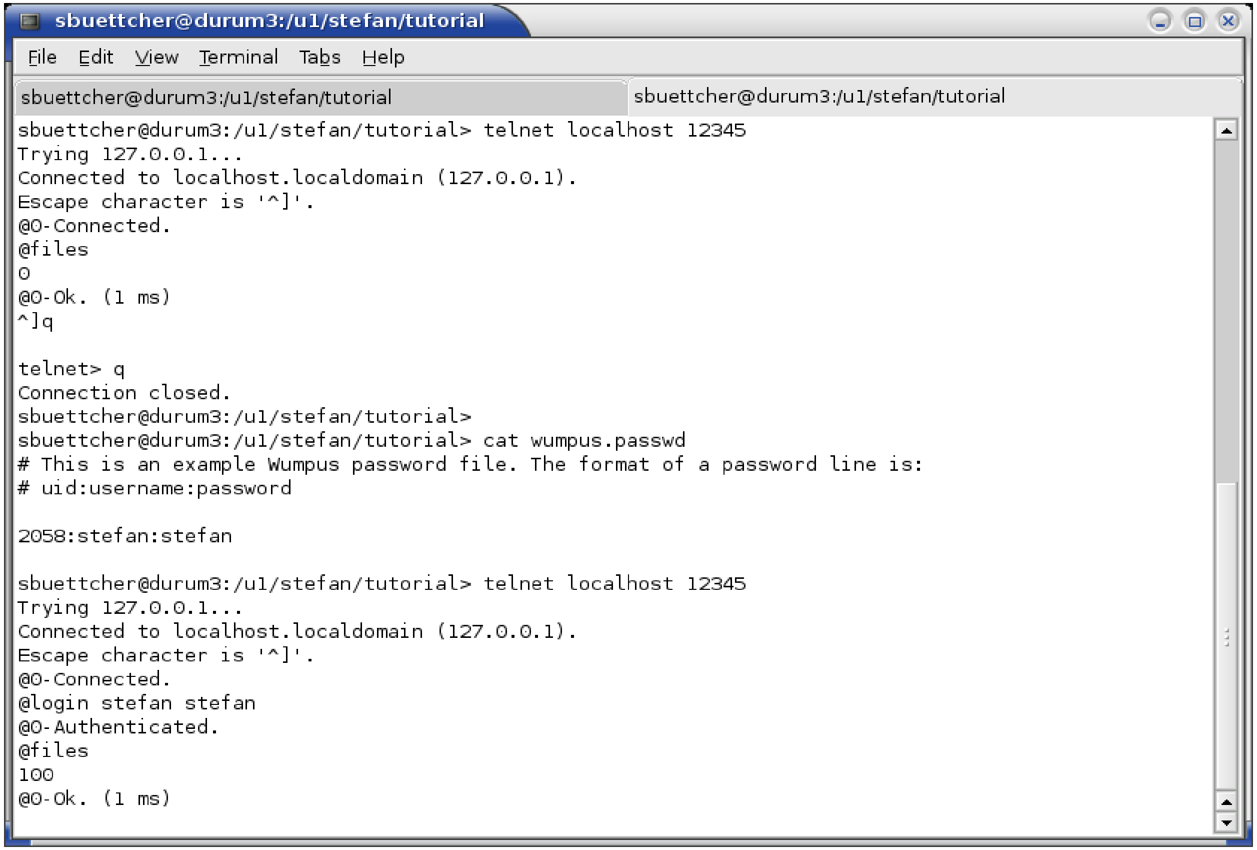
in the config file or at the command line.

(@get queries still cannot cover more than one file.)

Remote Connections


sbuettcher@durum3:/u1/stefan/tutorial
File Edit View Terminal Tabs Help
sbuettcher@durum3:/u1/stefan/tutorial
sbuettcher@durum3:/u1/stefan/tutorial

```
sbuettcher@durum3:/u1/stefan/tutorial> ~/wumpus/bin/wumpus TCP_PORT=12345 2> /dev/null
@0-Index loaded.
```


sbuettcher@durum3:/u1/stefan/tutorial
File Edit View Terminal Tabs Help
sbuettcher@durum3:/u1/stefan/tutorial
sbuettcher@durum3:/u1/stefan/tutorial

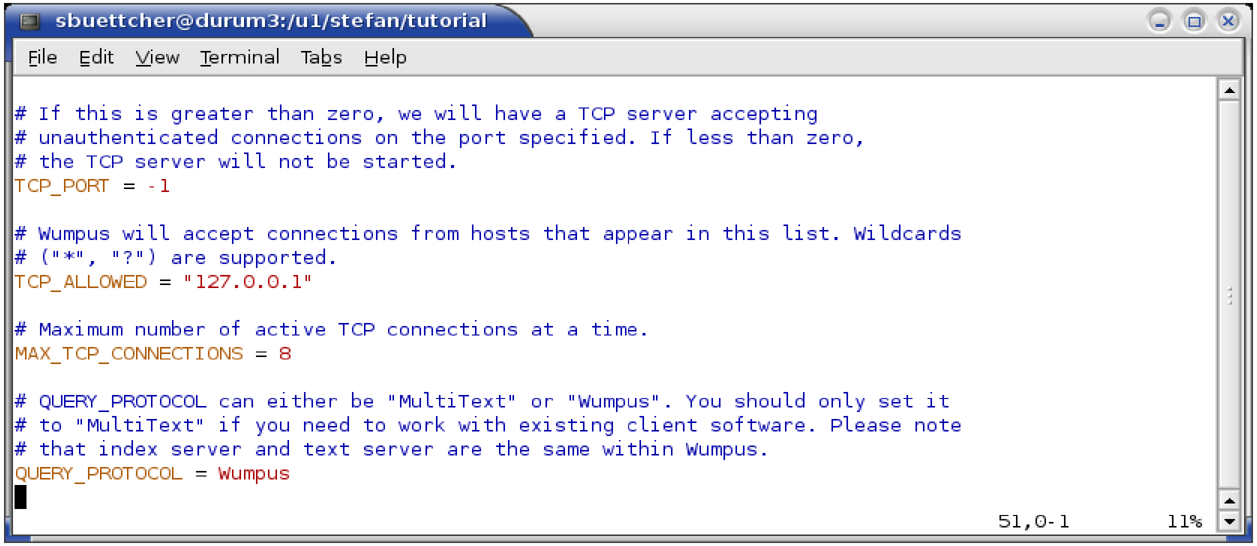
```
sbuettcher@durum3:/u1/stefan/tutorial> telnet localhost 12345
Trying 127.0.0.1...
Connected to localhost.localdomain (127.0.0.1).
Escape character is '^]'.
@0-Connected.
@files
0
@0-Ok. (1 ms)
^]q

telnet> q
Connection closed.
sbuettcher@durum3:/u1/stefan/tutorial>
sbuettcher@durum3:/u1/stefan/tutorial> cat wumpus.passwd
# This is an example Wumpus password file. The format of a password line is:
# uid:username:password

2058:stefan:stefan

sbuettcher@durum3:/u1/stefan/tutorial> telnet localhost 12345
Trying 127.0.0.1...
Connected to localhost.localdomain (127.0.0.1).
Escape character is '^]'.
@0-Connected.
@login stefan stefan
@0-Authenticated.
@files
100
@0-Ok. (1 ms)
```

Accessing Wumpus through a TCP connection is simple: Set the value of the TCP_PORT configuration variable and start the engine. Use telnet to connect.

After a connection has been established, make sure you login properly. Otherwise, you might not be able to access the entire collection.

If the server immediately closes the TCP connection, then have a look at the config file; check if TCP_ALLOWED includes your client machine.

Files, Offsets, and “Permission denied”


sbuettcher@durum3:/u1/stefan/tutorial
File Edit View Terminal Tabs Help

```
# If this is greater than zero, we will have a TCP server accepting
# unauthenticated connections on the port specified. If less than zero,
# the TCP server will not be started.
TCP_PORT = -1

# Wumpus will accept connections from hosts that appear in this list. Wildcards
# ("*", "?") are supported.
TCP_ALLOWED = "127.0.0.1"

# Maximum number of active TCP connections at a time.
MAX_TCP_CONNECTIONS = 8

# QUERY_PROTOCOL can either be "MultiText" or "Wumpus". You should only set it
# to "MultiText" if you need to work with existing client software. Please note
# that index server and text server are the same within Wumpus.
QUERY_PROTOCOL = Wumpus

                                                                  51,0-1        11%
```

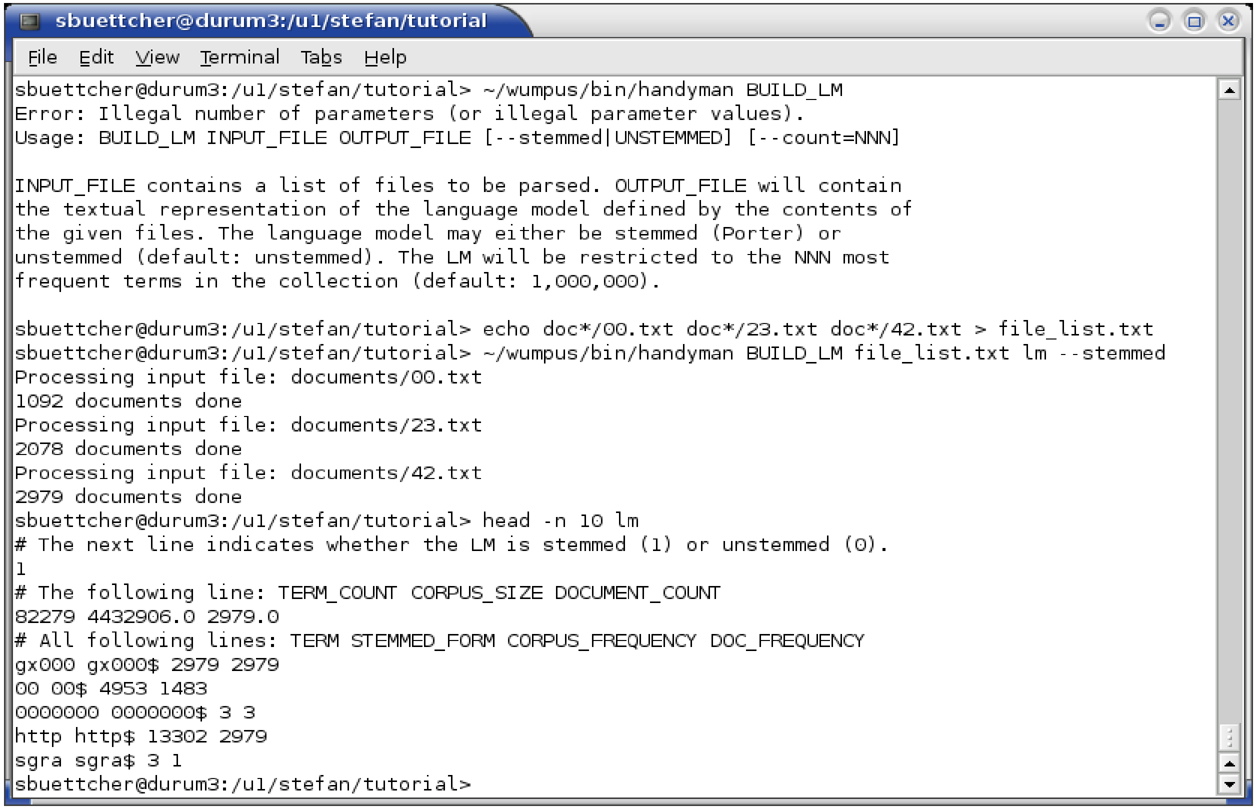
TCP_ALLOWED takes a list of IP addresses, separated by comma or whitespace. Each entry must be in quotation marks.

Also interesting:

MAX_TCP_CONNECTIONS: the maximum number of open TCP connections;

QUERY_PROTOCOL: for historic reasons; allows you to switch to the old MultiText user interface.

External Tools: The Handyman



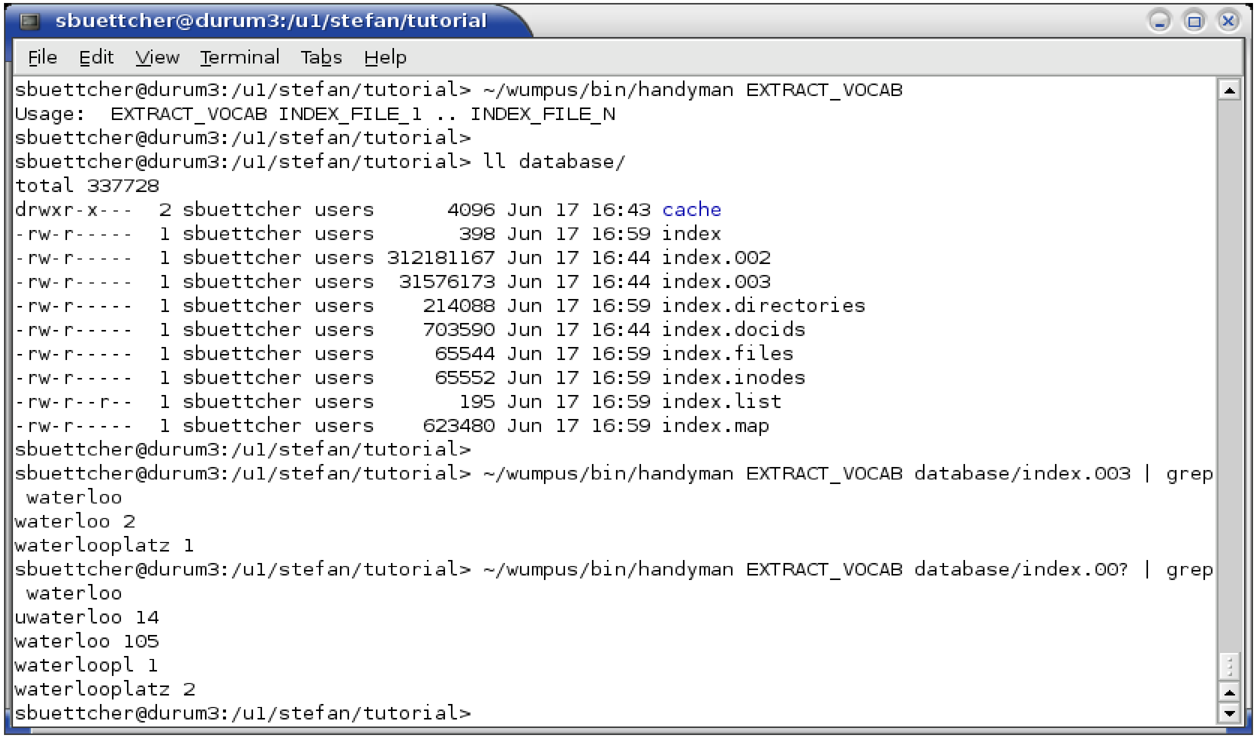
```
sbuettcher@durum3:/u1/stefan/tutorial> ~/wumpus/bin/handyman BUILD_LM
Error: Illegal number of parameters (or illegal parameter values).
Usage: BUILD_LM INPUT_FILE OUTPUT_FILE [--stemmed|UNSTEMMED] [--count=NNN]

INPUT_FILE contains a list of files to be parsed. OUTPUT_FILE will contain
the textual representation of the language model defined by the contents of
the given files. The language model may either be stemmed (Porter) or
unstemmed (default: unstemmed). The LM will be restricted to the NNN most
frequent terms in the collection (default: 1,000,000).

sbuettcher@durum3:/u1/stefan/tutorial> echo doc*/00.txt doc*/23.txt doc*/42.txt > file_list.txt
sbuettcher@durum3:/u1/stefan/tutorial> ~/wumpus/bin/handyman BUILD_LM file_list.txt lm --stemmed
Processing input file: documents/00.txt
1092 documents done
Processing input file: documents/23.txt
2078 documents done
Processing input file: documents/42.txt
2979 documents done
sbuettcher@durum3:/u1/stefan/tutorial> head -n 10 lm
# The next line indicates whether the LM is stemmed (1) or unstemmed (0).
1
# The following line: TERM_COUNT CORPUS_SIZE DOCUMENT_COUNT
82279 4432906.0 2979.0
# All following lines: TERM STEMMED_FORM CORPUS_FREQUENCY DOC_FREQUENCY
gx000 gx000$ 2979 2979
00 00$ 4953 1483
0000000 0000000$ 3 3
http http$ 13302 2979
sgra sgra$ 3 1
sbuettcher@durum3:/u1/stefan/tutorial>
```

Sometimes you want to work with a text collection without involving the search engine (or want to extract some index information that is not accessible through the text interface).

Perhaps `bin/handyman` can help you with that.

Wumpus's handyman can build a language model from a collection of text files. It can extract index statistics or postings lists. It can merge index files, recompress them using a different compression method, and many other things.

Extracting the Index's Vocabulary

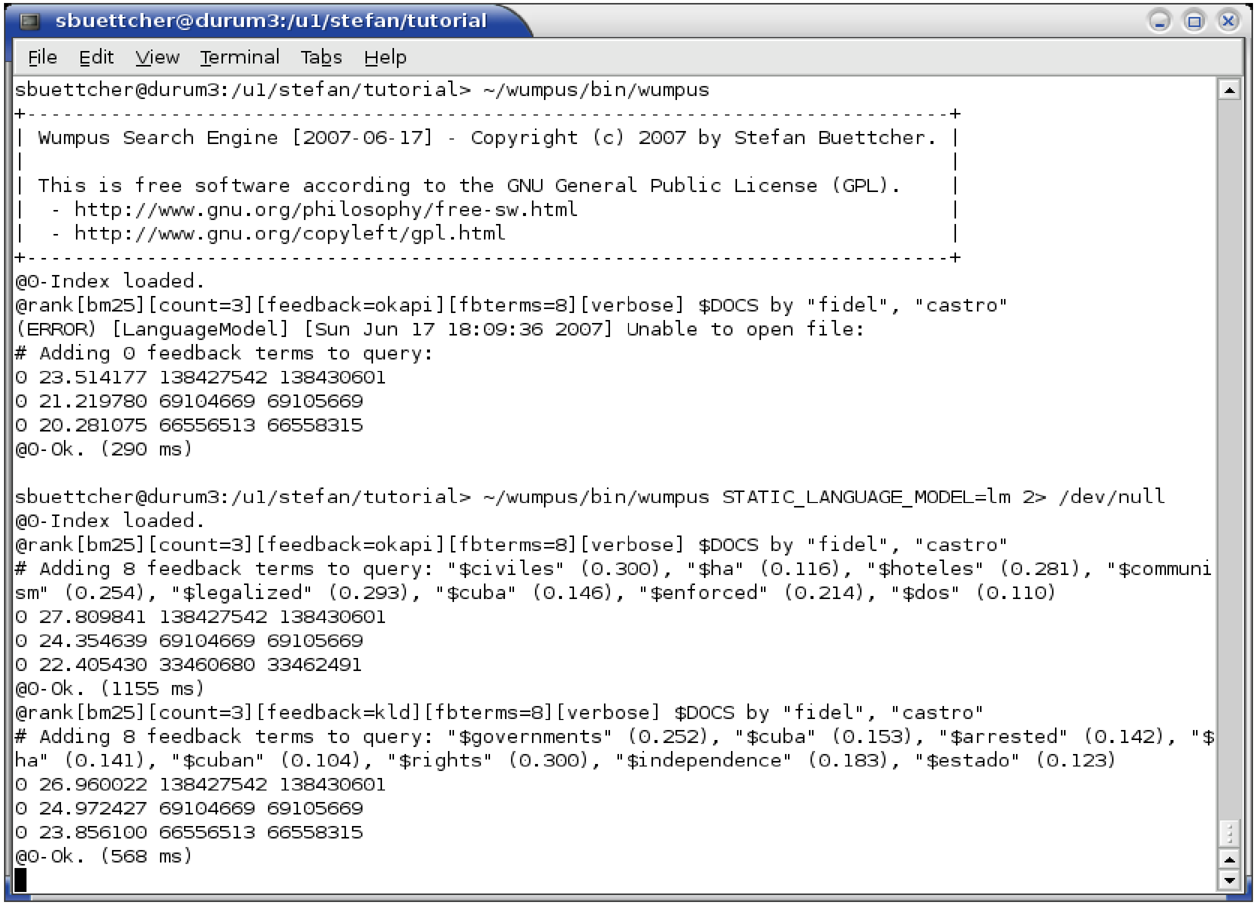

sbuettcher@durum3:/u1/stefan/tutorial
File Edit View Terminal Tabs Help

```
sbuettcher@durum3:/u1/stefan/tutorial> ~/wumpus/bin/handyman EXTRACT_VOCAB
Usage:  EXTRACT_VOCAB INDEX_FILE_1 .. INDEX_FILE_N
sbuettcher@durum3:/u1/stefan/tutorial>
sbuettcher@durum3:/u1/stefan/tutorial> ll database/
total 337728
drwxr-x---  2 sbuettcher users      4096 Jun 17 16:43 cache
-rw-r-----  1 sbuettcher users       398 Jun 17 16:59 index
-rw-r-----  1 sbuettcher users 312181167 Jun 17 16:44 index.002
-rw-r-----  1 sbuettcher users  31576173 Jun 17 16:44 index.003
-rw-r-----  1 sbuettcher users    214088 Jun 17 16:59 index.directories
-rw-r-----  1 sbuettcher users    703590 Jun 17 16:44 index.docids
-rw-r-----  1 sbuettcher users     65544 Jun 17 16:59 index.files
-rw-r-----  1 sbuettcher users     65552 Jun 17 16:59 index.inodes
-rw-r--r--  1 sbuettcher users       195 Jun 17 16:59 index.list
-rw-r-----  1 sbuettcher users    623480 Jun 17 16:59 index.map
sbuettcher@durum3:/u1/stefan/tutorial>
sbuettcher@durum3:/u1/stefan/tutorial> ~/wumpus/bin/handyman EXTRACT_VOCAB database/index.003 | grep
 waterloo
waterloo 2
waterlooplatz 1
sbuettcher@durum3:/u1/stefan/tutorial> ~/wumpus/bin/handyman EXTRACT_VOCAB database/index.00? | grep
 waterloo
uwaterloo 14
waterloo 105
waterloopl 1
waterlooplatz 2
sbuettcher@durum3:/u1/stefan/tutorial>
```

Using the handyman to extract the index's vocabulary from Wumpus's data files.

Note that the Wumpus data directory contains two index partitions. Passing only one to the handyman will give you wrong results.

Pseudo-Relevance Feedback


sbuettcher@durum3:/u1/stefan/tutorial
File Edit View Terminal Tabs Help

```
sbuettcher@durum3:/u1/stefan/tutorial> ~/wumpus/bin/wumpus
+-----------------------------------------------------------------------------+
| Wumpus Search Engine [2007-06-17] - Copyright (c) 2007 by Stefan Buettcher. |
|                                                                             |
| This is free software according to the GNU General Public License (GPL).   |
|  - http://www.gnu.org/philosophy/free-sw.html                               |
|  - http://www.gnu.org/copyleft/gpl.html                                     |
+-----------------------------------------------------------------------------+
@0-Index loaded.
@rank[bm25][count=3][feedback=okapi][fbterms=8][verbose] $DOCS by "fidel", "castro"
(ERROR) [LanguageModel] [Sun Jun 17 18:09:36 2007] Unable to open file:
# Adding 0 feedback terms to query:
0 23.514177 138427542 138430601
0 21.219780 69104669 69105669
0 20.281075 66556513 66558315
@0-Ok. (290 ms)

sbuettcher@durum3:/u1/stefan/tutorial> ~/wumpus/bin/wumpus STATIC_LANGUAGE_MODEL=lm 2> /dev/null
@0-Index loaded.
@rank[bm25][count=3][feedback=okapi][fbterms=8][verbose] $DOCS by "fidel", "castro"
# Adding 8 feedback terms to query: "$civiles" (0.300), "$ha" (0.116), "$hoteles" (0.281), "$communi
sm" (0.254), "$legalized" (0.293), "$cuba" (0.146), "$enforced" (0.214), "$dos" (0.110)
0 27.809841 138427542 138430601
0 24.354639 69104669 69105669
0 22.405430 33460680 33462491
@0-Ok. (1155 ms)
@rank[bm25][count=3][feedback=kld][fbterms=8][verbose] $DOCS by "fidel", "castro"
# Adding 8 feedback terms to query: "$governments" (0.252), "$cuba" (0.153), "$arrested" (0.142), "$
ha" (0.141), "$cuban" (0.104), "$rights" (0.300), "$independence" (0.183), "$estado" (0.123)
0 26.960022 138427542 138430601
0 24.972427 69104669 69105669
0 23.856100 66556513 66558315
@0-Ok. (568 ms)
```

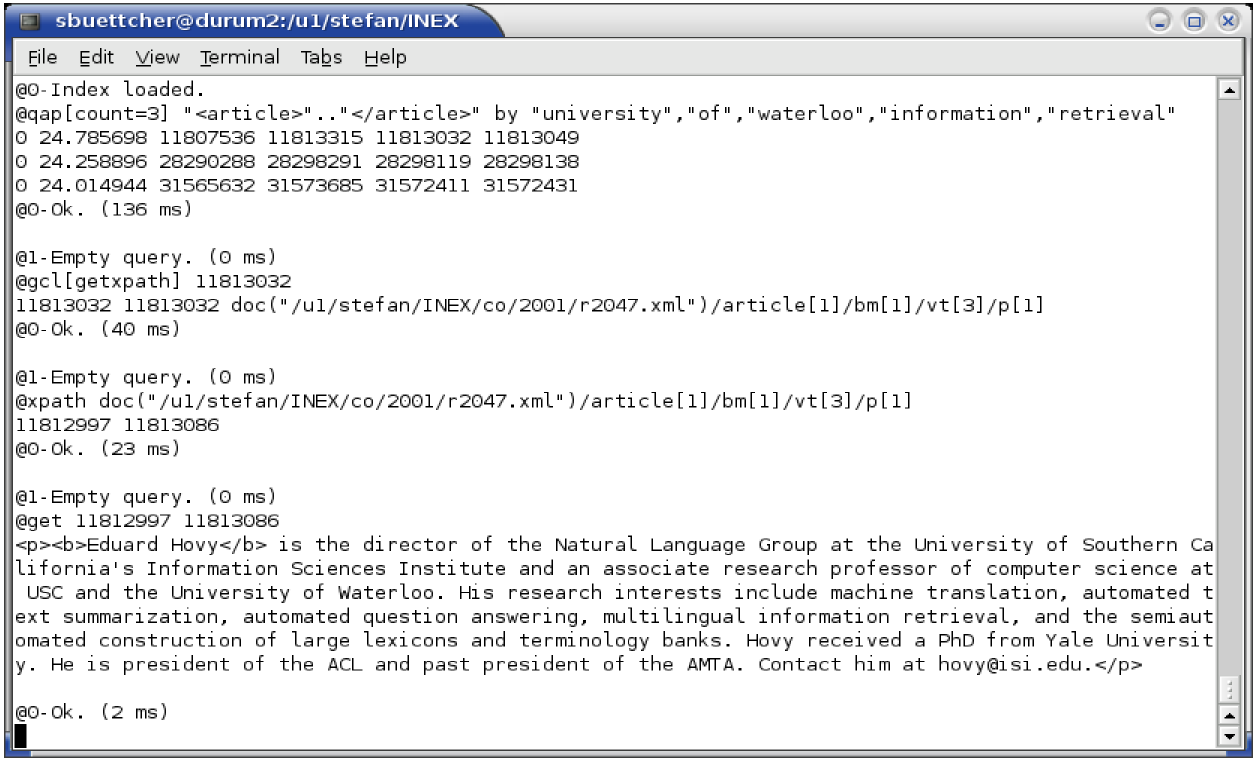
Wumpus has built-in support for pseudo-relevance feedback: Okapi-type feedback (Billerbeck and Zobel, 2004) and KLD-based feedback (Carpineto et al., 2001).

Pretty slow, though – has to read and parse the top *k* documents (default: 15; can be changed via [fbdocs]).

Number of expansion terms can be specified via the [fbterms] query modifier.

*If you want to use PRF, you **must** create a static language model file first (use the handyman). This is for your protection: feedback without a precomputed language model is incredibly slow!*

XML/XPath Support


sbuettcher@durum2:/u1/stefan/INEX
File Edit View Terminal Tabs Help

```
@0-Index loaded.
@qap[count=3] "<article>".."</article>" by "university","of","waterloo","information","retrieval"
0 24.785698 11807536 11813315 11813032 11813049
0 24.258896 28290288 28298291 28298119 28298138
0 24.014944 31565632 31573685 31572411 31572431
@0-Ok. (136 ms)

@1-Empty query. (0 ms)
@gcl[getxpath] 11813032
11813032 11813032 doc("/u1/stefan/INEX/co/2001/r2047.xml")/article[1]/bm[1]/vt[3]/p[1]
@0-Ok. (40 ms)

@1-Empty query. (0 ms)
@xpath doc("/u1/stefan/INEX/co/2001/r2047.xml")/article[1]/bm[1]/vt[3]/p[1]
11812997 11813086
@0-Ok. (23 ms)

@1-Empty query. (0 ms)
@get 11812997 11813086
<p><b>Eduard Hovy</b> is the director of the Natural Language Group at the University of Southern Ca
lifornia's Information Sciences Institute and an associate research professor of computer science at
 USC and the University of Waterloo. His research interests include machine translation, automated t
ext summarization, automated question answering, multilingual information retrieval, and the semiaut
omated construction of large lexicons and terminology banks. Hovy received a PhD from Yale Universit
y. He is president of the ACL and past president of the AMTA. Contact him at hovy@isi.edu.</p>

@0-Ok. (2 ms)
```

Wumpus has basic support for XPath queries.

If you want to use this, set the ENABLE_XPATH config variable to “true” and build a new index.

This will add special postings lists of the form `<level!N>` and `</level!N>` to the index, indicating the start and end of XML elements at nesting level *N*. Can be used to XPath stuff.

However:

1. *Wumpus only supports a very limited form of XPath.*
2. *I am not sure how bug-free the whole thing is. Haven't tested it very much.*

So, You Want It To Be Fast...

Most likely, you don't need dynamic indexing, security restrictions, optional query-time stemming, etc.

You probably want to index a collection once and then query the index a billion times. And you want it to be fast. In order to get the best performance out of Wumpus, use the following settings:

```
UPDATE_STRATEGY = NO_MERGE
MERGE_AT_EXIT = true
APPLY_SECURITY_RESTRICTIONS = false
CACHED_EXPRESSIONS = GCL expressions that you use a lot
COMPRESSED_INDEX_CACHE = false
STEMMING_LEVEL = (do you always stem your query terms ? 3 : 2)
ALL_INDICES_IN_MEMORY = (index small enough for this ? true : false)
```

If you run lots of phrase queries, you might also want to set BIGRAM_INDEXING=true before you build the index. This will increase the index quite a bit, but make most phrase queries much faster.

Getting Your Hands Dirty

If you want to start changing the code, then

```
query/languagemodel_query.[h|cpp]
```

is probably the best place to start. They contain a very simple implementation of multinomial LM with Dirichlet priors.

Within a sub-class of `RankedQuery` (e.g., `LanguageModelQuery`), index access is fairly straightforward. Use `getListForGCLExpression` to get a result list for a given GCL query. Use the classes in `extentlist/` to combine multiple lists via GCL operators.

Example:

```
ExtentList *list1, *list2, *list3;
list1 = getListForGCLExpression(“\”<article>\”..\”</article>\””);
list2 = getListForGCLExpression(“\”university\” ^ \”waterloo\””);
list3 = new ExtentList_Containment(list1, list2, true, false);
offset start = -1, end;
while(list3->getFirstStartBiggerEq(start + 1, &start, &end)) {
  printf(“Article from %lld to %lld: %lld occurrences.\n”,
    start, end, list2->getCount(start, end));
}
```