

A

- add_document method (IndexWriter), 16
- add_field method (Reader), 80
- add_indexes method (IndexWriter), 30
- analysis, 67–78
 - Analyzer class and built-in analyzers, 75
 - custom analyzers, 77
 - Token class, 67
 - TokenFilters, 72
 - Tokenizer class, 68
 - TokenStream class, 68
- Analyzer class, 75
- analyzers
 - Analyzer class, 75
 - choosing for QueryParser, 51
 - PerFieldAnalyzer, 76
 - StandardAnalyzer, 76
 - stemming and removing stop words in queries, 38
 - StemmingAnalyzer, 65
- AND keyword, 54
- antiword utility, 88
- Apache Lucene, fields, 10
- Array class, 9
 - sort method, 61
- arrays, 10
 - indexes as arrays of documents, 17
 - representing fields with arrays of strings, 9
- AsciiLetterTokenizer, 70
- AsciiLowerCaseFilter class, 72
- AsciiWhiteSpaceTokenizer, 69

- :auto_flush parameter (Index), 33

B

- BitVector class, 58
- boolean queries, FQL, 54
- BooleanQuery class, 37, 38
 - sloppy phrase queries, 40
- boosts, 3
 - boost attribute, 9
 - BoostMixin, 11
 - FieldInfo#boost property, 12
 - queries, 49
 - queries in FQL, 56
 - Query objects, 37

C

- C compiler, 1
- caching
 - filter, 59
 - sorts, 62
- :category parameter, PrefixQuery, 43
- character encodings
 - multibyte, 42
 - support by stemming algorithms, 74
- :chunk_size parameter, 26
- clean_string method (QueryParser), 52
- close method (Index), 3
- command-line indexing program, 2
 - running, 4
- commit locks, 32
- commit method, IndexWriter class, 17
- :compressed option (FieldInfo), 12
- concurrency issues, index locking and, 32

ConstantScoreQuery class, 46
:create parameter, 3

D

datatypes, indexing non-string types, 19
dates and times
 adding time to a document, 19
 date format, 19
 recording time for adding a file to the index, 88
 sorting by date, 21, 64
DBM class, 88
delete method
 Index class, 18
 IndexReader and IndexWriter, 18, 32
 IndexReader and multithreaded environment, 33
DESC sort modifier, 61
directories to add to an index, 89
Directory class, 7
:doc_skip_interval parameter, 28
Document class, 3, 8
document IDs, 4
documents, 3, 8
 adding to an index, 10, 16
 deleting with IndexReader or IndexWriter, 17
 retrieving from the index, 17
DOC_ID sort modifier, 62
downcasing
 using LowerCaseFilter with StandardTokenizer, 72
 using WhiteSpaceTokenizer, 69

E

end and start offsets (Token), 67
escaping special characters in
 QueryParser, 53
excerpting query results, 64
EXIF tags, 85
exifr, 85
EXTENSIONS constant (subreaders), 80

F

Ferret Query Language (see FQL)
ferretfind utility, 90
Field class, 9

:field_infos parameter
 IndexWriter class, 82
:field parameter
 QueryParser class, 52
 RangeFilter class, 57
FieldInfo class, 12
 :index parameter, 13
 :store parameter, 12
 :term_vector parameter, 14
FieldInfos class, 15, 81
 load method, 16
fields, 3, 8
 added to indexes, 11
 date fields, 19
 Ferret versus Apache Lucene, 10
 indexed, 13
 number fields, 19
 representing by strings or arrays of strings, 9
 sort fields, 20
 SortField class, 62
:field_infos parameter, 16
file descriptors, open, 28
file type, detecting, 81
filesystem
 indexing on personal computers, 2–5
 storing an index, 8
:filter_proc parameter, Searcher#search methods, 60
FilteredQuery class, 47
filters (analysis)
 TokenFilters, 72
 HyphenFilter, 75
 LowerCaseFilter, 72
 StemFilter, 74
 StopFilter, 72
filters (search)
 ConstantScoreQuery, 46
 for search results, 57
 QueryFilter class, 58
 RangeFilter, 57
 writing your own, 58
FQL (Ferret Query Language), 35, 37, 52
 boolean queries, 54
 boosting queries, 56
 fuzzy queries, 56
 phrase queries, 54
 range queries, 55

- term queries, 53
- wildcard queries, 55

FSDirectory class, 7

fuzzy queries, FQL, 56

FuzzyQuery class, 45

G

gcc, 1

getoptlong utility, 90

get_links method (HtmlReader), 84

grouping results with :filter_proc, 61

H

hashes, 10

- Hash class, 8
- Hash object, 5
- storing query results, 37

highlight method, 64

highlighting search results, 11, 64, 81

Hpricot library, 82

HtmlReader class, 82

HyphenFilter class, 75

I

:id field, 13, 18

ID3 tags, 86

id3lib-ruby, 86

in-memory indexing, 25

Index class, 5, 8

- :auto_flush parameter, 33
- highlight method, 64
- :key property, 18

:index parameter (FieldInfo), 13

:index_skip_interval parameter, 28

indexes

- locking, 32
- optimizing for searching, 3

indexing, 7–21

- adding boost attribute to any class
 - using BoostMixin, 11
- adding documents to an index, 10, 16
- boosts, 9
- command-line index program, 2
- deleting documents, 17
- documents, 8
- FieldInfos class, 15
- fields, 8

- filesystem on your computer, 2–5
- flow chart depicting the process, 24
- getting indexed documents, 17
- non-string datatypes, 19
- other improvements, 88
- performance tuning, 25
- running indexer from command line,
 - 4
 - setting up the index, 11
 - storing indexes, 7
 - updating the index, 18
- indexing parameters, 25
 - :doc_skip_interval, 28
 - :index_skip_interval, 28
 - :max_buffered_docs, 27
 - :max_buffer_memory
 - and :chunk_size, 26
 - :max_field_length, 27
 - :max_merged_docs, 27
 - :merge_factor, 26
 - testing, 29
 - :use_compound_file, 28
- IndexReader class, 17
 - deleting documents, 17
 - locking indexes, 32
 - multithreaded environment, 33
- IndexSearcher class, 35
- IndexWriter class, 16, 23, 81
 - add_indexes method, 30
 - deleting documents, 17
 - locking indexes, 32
 - multithreaded environment, 33
 - optimize method, 31
- inherited method, 80
- initialize method (Reader), 80
- irb session, 1

J

JpegReader class, 85

K

:key field, 18

L

LetterTokenizer, 70

- regular expression for a German language tokenizer, 71

- Levenshtein distance, 45
- Linux, Ruby on, 1
- load_readers method (Reader), 80
- locale, setting, 70
- locking, 3
 - index locking and concurrency issues, 32
- LowerCaseFilter, 69, 72
- Lucene fields, Ferret versus, 10

M

- Macintosh
 - Ruby on, 1
 - Spotlight on OS X, 2
- MatchAllQuery class, 46
- :max_buffer_memory parameter, 25, 26
- :max_buffered_docs parameter, 25, 26
- :max_field_length parameter, 27
- :max_merged_docs parameter, 27
- :max_terms property
 - MultiTermQuery, 43
 - setting for FuzzyQuery, 45
 - setting for PrefixQuery, 44
 - setting for WildcardQuery, 44
- memory
 - use by filters, 60
 - use by sort indexes, 62
- :merge_factor parameter, 24, 26
- Microsoft Word documents, 88
- :min_prefix_length, setting for FuzzyQuery, 46
- :min_similarity parameter, setting for FuzzyQuery, 46
- Mp3Reader class, 86
- multiprocess environment, 33
- MultiSearcher class, 36
- MultiTermQuery class, 42
- multithreaded environment, 33
- :must parameter (BooleanQuery), 39

N

- number fields, 19
 - padding to a fixed width, 19
 - padding with custom analyzer, 77
 - sorting, 21

O

- :omit_norms property, 27
- OOoReader (OpenOffice.org Reader), 84
- operating systems
 - id3lib, 86
 - limits on open file descriptors, 27
- optimize method
 - Index class, 3
 - IndexWriter class, 17, 31
- optimizing the index, 31
- OR keyword, 54
- ostrucut utility, 90

P

- padding numbers to a fixed width, 19
- parallel indexing, 30
- :path parameter, 3
- path to the index, 89
- PDFBox library, 87
- pdftotext utility, 87
- PerFieldAnalyzer class, 76
- performance tuning, 25
 - in-memory indexing, 25
 - indexing parameters, 25
 - testing indexing parameters, 29
- WildcardQueries, 45
- phrase queries, FQL, 54
- PhraseQuery class, 39
- Porter, Michael, 74
- position increment attribute (Token), 68
- PrefixQuery class, 43
- Proc object, 60
- processes, multiprocess environment, 33
- Project Gutenberg, 5
- punctuation, WhiteSpaceTokenizer and, 69

Q

- queries, 36
 - BooleanQuery class, 38
 - boosting, 49
 - building, 37
 - ConstantScoreQuery class, 46
 - FilteredQuery class, 47
 - FuzzyQuery class, 45
 - MatchAllQuery class, 46
 - MultiTermQuery class, 42

- PhraseQuery class, 39
- PrefixQuery class, 43
- RangeQuery class, 41
- span queries, 47
- SpanFirstQuery, 48
- SpanNearQuery, 49
- SpanNotQuery, 48
- SpanOrQuery, 48
- SpanTermQuery, 47
- TermQuery class, 38
- WildcardQuery class, 44
- QueryFilter class, 58
- QueryParser class, 36, 50
 - parameters, 50

R

- RAMDirectory class, 8, 25
- range queries on number fields, 19
- range queries, FQL, 55
- RangeFilter class, 57
 - custom analyzer that pads numbers, 77
- RangeQuery class, 41
 - custom analyzer that pads numbers, 77
- read method
 - HtmlReader class, 84
 - Reader class, 80
- Reader class, 80
 - HtmlReader subclass, 82
 - JpegReader subclass, 85
 - Mp3Reader subclass, 86
 - OoOReader subclass, 84
 - TextReader subclass, 82
- RegExpTokenizer, 71
- Reuters-21578 indexer testing collection, 30
- RubyGems, 1
- rubyzip, 84

S

- SCORE sort modifier, 62
- Searcher class, :filter_proc parameter, 60
- searches, 1
 - building queries, 37
 - date formats, 20
 - filtering results, 57

- filters, 37
- improvements in, 89
- queries, 36
 - QueryParser, 36
 - search classes, 35
 - sorting results, 37, 61
 - writing search code, 4
- search_each() method, 35
- SegmentReader class, 31
- :should parameter (BooleanQuery), 39
- skip lists, 28
- slop, calculating, 40
- sloppy phrase queries, 55
 - boosts and, 57
 - SpanNearQuery versus, 49
- sloppy phrases, 39
- Snowball stemmer, 74
- Sondergaard, Thomas, 84
- Sort class, 37, 62, 63
- SortField class, 62
 - constant SortField objects, 63
- sorting
 - dates, 64
 - results of range queries on string fields, 42
 - search results, 37, 61
 - searches, improvements in, 89
 - sort fields, 20
- span queries, 47
- SpanFirstQuery class, 48
- SpanNearQuery class, 49
- SpanNotQuery class, 48
- SpanOrQuery class, 48
- SpanTermQuery class, 47
- special characters, escaping in
 - QueryParser, 53
- Spotlight on OS X, 2
- StandardAnalyzer class, 71, 76
- StandardTokenizer, 70
- start and end offsets (Token), 67
- StemFilter class, 74
- stemmers, 38
- stemming, 74
- StemmingAnalyzer, 65
- Stocker, Robin, 86
- StopFilter class, 72
 - StandardAnalyzer and, 76
- :store parameter (FieldInfo), 12

- strings, 19
 - converting time to, 19
 - parsing of query strings, 53
 - range queries, 42
 - representing fields, 9
 - sort strings, 61

T

- tagging formats
 - EXIF, 85
 - ID3, 86
- Technorama Ltd., 84
- term queries, FQL, 53
- :term_vector parameter
 - FieldInfo class, 14
 - FieldInfos class, 81
- TermDocEnum, 59
- TermQuery class, 38
- text attribute (Token), 67
- TextReader class, 82
- threads, multithreaded environment, 33
- Token class, 67
- TokenFilter class, 72
 - HyphenFilter, 75
 - LowerCaseFilter and AsciiLowerCaseFilter, 72
 - StemFilter, 74
 - StopFilter, 72
- Tokenizer class, 68
 - LetterTokenizer, 70
 - RegExpTokenizer, 71
 - StandardTokenizer, 70
 - WhiteSpaceTokenizer, 69
- tokenizing
 - fields, 13
 - number fields and, 19
 - query strings, 53
 - query terms, 37
 - sort fields and, 20
- TokenStream class, 68

U

- :untokenized option (FieldInfo), 14
- :untokenized_omit_norms option (FieldInfo), 14
- :use_compound_file parameter, 27
- UTF-8 character encoding, 75

- UTF-8 locales, 70

V

- van 't Veer, Remco, 85

W

- weightings, 3
- WhiteSpaceTokenizer, 69
- wildcard characters, 44
- wildcard queries, FQL, 55
- WildcardQuery class, 44
- Windows
 - built-in search, 2
 - Ruby on, 1
- :with_offsets option (FieldInfo), 14
- :with_positions option (FieldInfo), 14
- :with_positions_offsets option (FieldInfo), 14
- Word documents, 88
- write locks, 32

Y

- YAML files, 15
- :yes_omit_norms option (FieldInfo), 14