
Index

- ABC, 49, 192
Abduction, 242
ACSys, 210, 214, 364
Adaptive filtering, 100, 105, 107–109, 116–118, 274–275, 291–292, 374, 388–391
Adversarial information retrieval, 208
Ad hoc task, 7, 24, 53–54, 184, 374
 definition, 9–10, 80, 425
 guidelines, 81, 126
 TREC-1, 83
 TREC-2, 83
Agence France Presse, 46, 47, 156, 174
Alignment of corpora, 61, 409
Alta Vista, 204, 206–207, 434, 437, 439
Anchor, 200, 202, 215, 217, 226, 437–438
Anick, Peter, 207
AQUAINT, 48, 244, 256, 413, 441
Applied Physics Laboratory at Johns Hopkins, 169–170, 178–180
Arabic
 bilingual dictionaries, 175, 178
 CLIR experiments, 174–180, 410
 documents, 47, 174
 morphology, 174
 names, 178, 271
 normalization, 175, 178, 410
 stemmers, 174–175, 177–179, 272, 410
 stop words, 175, 177
 test collections, 47, 174–176
 topics, 47, 174–176
Ashwin, Rao, 28
AskJeeves, 48, 240, 244, 436, 439
AskMSR, 242
Aspectual recall, 132–135
Assessor, 23, 31, 33, 43, 81–82, 124, 165–166, 201, 234, 243, 252, 361
Associated Press, 25, 27, 28, 34, 36, 47, 81, 102, 166, 192, 325
AT&T, 86–91, 93, 192, 194, 236
Ault, Tom, 111
Auran, Per Gunnar, 207
Australian National University, 86, 88, 91, 186–187
Autonomy Corporation, 439, 441
Average term frequency measure, 326
Averaging, evaluation effect, 56–59, 61, 67, 111
Babelfish, 386–387
Ballim, Afzal, 66
Banks, David, 65
Batch-adaptive filtering, 107–109
Batch filtering, 103, 108–109, 115
Bates, Marcia, 147
Bayes
 naive Bayes assumption, 287, 374
 theorem, 262, 287, 373, 376, 378
BBC, 48
BBN, 86, 92, 172–180, 386
Belkin, Nicholas, 147
Beaulieu, Micheline, 287–288, 429
Bellcore, 163–165
Berger, Adam, 386
Bharat, Krishna, 207
Bilingual dictionaries, 155, 163, 171–173, 175, 178, 270–271
Bliss, Bill, 207
BM25 weights, 7, 86, 289–291, 295–297, 374, 377
Bookstein, Abraham, 287, 374
Boolean, 55, 84, 128, 130, 134, 136, 168, 289, 428
Borgman, Christine, 429
Brewer, Eric, 204
Broadcast news, 46, 49, 188–196
Broder, Andrei, 147, 204, 207
Brooks, Helen, 147
Brown, Eric, 400
Browsing interfaces, 137–138, 141
Buckley, Chris, 53, 56, 79, 131, 204, 304, 400
BYBLOS Rough 'n Ready, 192

- Cahoon, Brendan, 401
Cambridge University, 194
Canadian parliamentary proceedings, 408
CEA, 163–165
Center for Intelligent Information Retrieval.
 See University of Massachusetts at Amherst
Chandrasekhar, Raman, 207
Chapman University, 139
Chaudhuri, Bidyut, 187
Chemical Abstracts, 431
Cheshire, 135, 137–138
Chinese
 bilingual dictionaries, 171–173, 338
 character-based retrieval, 158–159, 172, 270, 337, 410
 characters, 156–158, 335–336
 cross-language retrieval, 171–174, 409–410
 dictionary disambiguation, 172, 339
 documents, 46–47, 157, 171
 encoding schemes, 171
 name finding, 159
 phrase translation, 172, 339
 preprocessing documents, 337
 monolingual retrieval, 156–160, 270
 segmentation, 158–159, 179, 270, 335–336, 409–410
 stop phrase removal, 325
 test collections, 46–47, 157, 171
 thesaurus, 172
 topics, 46–47, 157, 159, 171
 translation COTS, 338
 word-based retrieval, 158–159, 172
City University London, 5, 86, 90, 93, 128–129, 132, 134, 148, 158–159, 427. *See also* Okapi
City University New York, 7, 86, 91, 93, 157–159, 171–173, 321–345. *See also* PIRCS
CLARIT, 84, 186
CLARITECH Corporation, 90–91, 128, 130, 158, 169–170, 186–187, 307
Clarity, 42
CLEF, 11, 18n2, 47, 171, 179–180, 255, 384, 386, 443
Cleverdon, Cyril, 4, 79
Closed-class question, 233–234, 247
Clustering
 documents, 22
 interfaces using, 130, 134, 137, 141–142
 topics, 90, 438
CMU (Carnegie Mellon University), 191, 206
CNN, 49, 192
CNR, Pisa, 166
Collection enrichment, 90, 327–328, 339, 341–342
Collection fusion, 401
Computer Select, 25, 27–28, 36–37, 81
Comparable corpora, 161, 168, 409
Confidence-weighted score, 245–246, 255, 413
Congressional Record, 28, 81, 331
ConQuest, 84, 128
Cooper, Edwin, 207
Coordination level, 55, 365–366
Copernic Enterprise Search, 224
Cornell University, 5, 22, 55, 79, 91, 93, 130, 154–159, 163–165, 301, 313. *See also* SMART
Correlation among evaluation measures, 60–61
Cover density ranking, 348, 365–368
Cranfield
 experimental methodology, 4, 53, 67–73, 252, 281, 288, 430
 original tests, 4, 21, 79, 421, 428
 test collection, 4, 21
Craswell, Nick, 217
Croft, W. Bruce, 389
Cross-language retrieval (CLIR), 10, 49, 160–180, 374, 384–388, 408–410
CSIRO, 47, 138–139, 141–144, 148
Cutoff level, 53, 55–5
CYC, 413–414
C-SPAN, 49

Database frontend, 234
Data fusion, 85, 91
Data heterogeneity, 434, 438–439
Definition question, 240, 247–250, 252, 412–414
Department of Defense, 130, 155–156

- Department of Energy abstracts, 27, 28, 36, 38, 81
- DIALOG, 431
- Dialogue system, 234
- Dictionary lookup algorithms, 168, 173, 270–271
- Digital library, 422
- Distributed architecture, 276, 401–403
- Documents
 - Arabic, 47, 174
 - broadcast news, 46, 188–196
 - Chinese, 46–47, 157, 171
 - CLIR, 47, 160–161, 166, 171, 174
 - copyright issues, 25
 - effects of length, 7, 36, 38, 377
 - effects of pooling, 34–35, 175
 - effects of source, 34–38
 - English ad hoc sets, 24, 81
 - English, sizes, 25–28
 - English, source, 22, 25–28, 48
 - French, 47, 160–161
 - formatting, 27–28
 - German, 47, 160–161
 - Italian, 47, 160, 166
 - non-English, 46–47
 - OCR, 46, 184–185
 - routing sets, 24
 - Schweizerische Depeschen Agentur, 47, 160–161
 - source, 22, 25–28, 48
 - Spanish, 46, 153, 156
 - time-stamped, 102
 - web vs. ad hoc, 211–214, 422–423
- Document Understanding Conference (DUC), 441
- Dolin, Ron, 207
- Dougherty, James, 384
- Dragon Systems, 192, 276
- Dublin City University, 154, 156–157, 164–165
- Duke University, 163–164
- Effectiveness
 - ad hoc trends, 80, 310–312, 432
 - ad hoc upper-bound, 45
 - ad hoc variation across topic fields, 93
 - Arabic, 175–176, 178–180
 - Chinese, 157–159, 172
 - cross-language, 164–165, 168–170, 172–180
 - effect of collection size, 211, 363–365
 - Spanish, 156–157
 - monolingual vs cross-lingual, 160–161, 164–165, 173–175, 177–180, 271
 - TREC systems vs. web search engines, 216–217
- Efficiency
 - filtering task, 117–118
 - VLC/web task, 202, 207, 210–211, 276, 361–365, 400–403
- Electronic Monk, 207
- Eleven-point average, 55, 57, 59, 71
- El Norte*, 46, 153
- EM-algorithm, 389–390
- Encarta, 48, 237
- Enterprise web, 201, 207, 225–226
- Entity recognition, INQUERY, 264–266
- Entry page, 200, 205
- Entry page search, 202, 204, 210, 216–217, 367, 378–384, 405–407, 437. *See also* Navigational search
- Environmental Research Institute of Michigan, 154
- EPFL Lausanne, 166
- Error rate, 63–64, 66–67, 70, 73
- Eurospider, 168–170
- Evaluation measure stability, 65–67, 252–255
- Evaluation
 - CLIR issues, 165–168, 171, 175
 - filtering, 109–115
 - interactive, 124–125, 131–134, 145–147
 - question answering, 252–255
 - ranked retrieval, 53–72
 - set-based retrieval, 109–115
 - strict vs. lenient scores, 238
 - web, 219–224
- Exact answer, 244–246, 254
- Excite, 48, 207, 237, 436
- Expected nonrelevant ratio, 311–312
- E measure, 112
- ETH. *See* Swiss Federal Institute of Technology

- Faceted query, 384
- Fact-finding task, 140
- FAQ Finder, 234
- Fast/AlltheWeb, 207
- Feedback. *See also* Pseudorelevance
 - feedback
 - manual feedback experiments, 132, 134–135, 137–138, 141, 293, 357, 361
 - negative feedback, 90, 137
 - relevance feedback, 99–100, 104, 113, 290, 302, 313, 388–391, 437
- Federal Register*, 27–28, 36, 38, 46, 81, 184–185, 325, 341
- Filtering, 10, 25, 53, 72, 99–119
- Financial Times*, 28, 36, 81, 103, 107
- F measure, 110, 112, 247, 250
- Foreign Broadcast Information Service, 28, 36, 81, 102, 410
- Fox, Ed, 304
- Frame sets, 223
- French
 - documents, 47, 160–161
 - near cognates experiment, 163–165, 318
 - topics, 47, 160–162, 166–167, 174
- Fudan University, 172
- Fuhr, Norbert, 313
- Fujitsu, 211, 214
- Fuzziness value, 65–66

- Gaizauskas, Rob, 234
- Garcia-Molina, Hector, 401
- General Electric Corporation, 90–91, 427
- George Mason University, 154–158, 186–187
- Georgia Institute of Technology, 130
- German
 - documents, 47, 160–161
 - topics, 47, 160–161, 166–167
- GIRT, 166, 168, 170
- Google, 204, 206, 207, 215, 220, 378, 422, 435, 437, 439
- Gopher, 206
- .GOV, 204, 207, 209–210, 225
- Graphic interfaces, 128–130
- Grefenstette, Gregory, 337
- GURU, 397–402, 410–412

- Hardness measure, 41–42
- Harman, Donna K., 128, 357
- Harter, Stephen, 287, 374, 376, 392
- Hawking, David, 47, 202, 204, 207, 211, 217
- Hearst, Marti, 123, 221
- Help-desk systems, 234
- Henzinger, Monika, 207
- Hersh, William, 103
- Hidden Markov model, 154, 158, 172, 427
- Hiemstra, Djoerd, 7, 391
- Hirschman, Lynette, 234
- Home page, 200
- Home page search. *See* Entry page search
- Huajian, 338
- Hull, David, 384
- Human-in-the-loop. *See* User-in-the-loop
- Hummingbird Technologies, 177, 180
- Hyperlink graph. *See* Link structure
- Hyphenation, 38, 304, 398

- IBM, 7, 86, 91, 134, 168–170, 172–173, 178–180, 191, 397–420
 - Arabic CLIR experiments, 410
 - Chinese CLIR experiments, 409–410
 - cross-language experiments, 408–410
 - cross-language merging, 409
 - definition questions, 412–413
 - distributed architecture, 401–402
 - effects of topic length, 400
 - efficiency in VLC track, 400–403
 - GURU, 397–402, 410–412
 - home page finding experiments, 405–407
 - interactive experiments, 407
 - JURU, 397–398
 - knowledge agents, 404–405
 - lexical affinities, 398–399
 - Okapi/INQUERY weighting, 406
 - pivot language experiments, 409
 - predictive annotation, 410–411
 - pruning experiments, 403–404
 - question-answering experiments, 410–416
 - ranked list filters, 404–405
 - spoken document retrieval, 415
 - statistical machine translation, 408–410
 - statistical question answering, 414–415
 - topic distillation, 404–405

- video retrieval, 415–416
- web indexing, 403
- Ideare SpA, 224
- Illinois Institute of Technology, 174–182
- Indexing
 - best-practices, 14
 - full-text, 4
- InFinder, 154–155, 267–268
- Informational search, 205, 208, 210, 216, 219–221, 226
- Information extraction, 12, 234, 264–265, 424, 444
- Information need. *See* Topic
- Information nugget, 248–250
- Information Technology Institute, 157–159
- InfoSeek, 204
- Inktomi, 204
- INQUERY, 40, 83, 85–89, 91–93, 127, 155, 262–264, 288, 374, 386, 427. *See also* University of Massachusetts
- InsightSoft-M, 242
- Inspec, 287, 431
- Institute of Systems Science, 159
- Intelligence analyst, 23, 34
- Interactive measures, 131–135, 138–139, 141, 145–146
- Interactive narrative reports, 133, 135
- Interactive testing protocols, 125, 133–136, 145–147
- Interactive tasks, 129–144
- Interactive track observations, 129–144
- Interfaces, 123, 128, 130, 134, 141–143
- Internet Archive, 48, 202
- Interpolated precision, 55, 56, 57
- Interpolation, 54, 56
- Italian
 - documents, 47, 160, 166
 - topics, 164–167
- IRIT, 86, 90, 165
- ITC-irst, 243
- IZ Sozialwissenschaften, Berlin, 166

- Jin, Hubert, 389
- JURU, 397–398

- Kaszkiel, Marcin, 216–217
- Katz, Slava, 326

- Kendall's tau, 60, 68, 70, 254
- Kleinberg hub and authority, 214, 404–405
- Kluck, Michael, 166
- Knowledge agents, 404–405
- Known-item search, 46, 48, 184, 190–191, 201, 205
- Kohavi, Ron, 384
- Kraaij, Wessel, 7, 389, 391
- Krause, Jürgen, 166
- Krellenstein, Marc, 204
- Kuhns, J. L., 287, 373–374, 376, 392
- Kupiec, Julien, 233
- Kwok, Kui-Lam, 427

- Lafferty, John, 386
- Lagergren, Eric, 136
- Language Computer Corporation (LCC), 241, 244, 246, 304
- Language modeling, 7, 14, 86, 324, 373–395, 427
 - background model, 375–376, 379, 390
 - document model translation, 385
 - mixture model, 376–378
 - query model translation, 386
 - rolling (adaptive), 192, 196
 - use for adaptive filtering, 388–391
 - use for ad hoc retrieval, 375–378
 - use for cross-language retrieval, 168–170, 271–272, 384–388, 408–410
 - use for entry page retrieval, 378–384
- Latent semantic indexing, 155, 163–165, 427
- Latin square design, 135
- Lavrenko, V., 389
- Leek, Tim, 389
- Le Monde*, 386
- Lewit, Alan, 400
- Lexical affinities, 398–399
- Lexis-Nexis Corporation, 28, 88, 91, 130
- LIMSI, 194, 241
- Linguistic Data Consortium, 25, 27, 47, 171, 174, 188, 244, 338
- Link structure, 200
 - use in web retrieval, 204, 214–217, 226, 378–384, 404, 435, 437
- List questions, 243–244, 247

- Local context analysis, 6, 88, 138, 267–271, 408
- Logging, 125, 130–131, 133, 135, 139, 145–146, 429
- Log odds, 379
- LOGOS Corporation, 163
- LookSmart, 218–219
- Los Angeles Times*, 28, 36, 81, 386
- Lu, Allen, 28
- Luhn, Hans Peter, 432
- Lycos, 206
- Maarek, Yoelle, 397
- MacFarlane, Andy, 207
- McGill, Michael J., 211
- McKinley, Kathryn, 401
- Machine learning, 99, 115, 116, 118, 127, 129
- Machine-readable dictionaries, 163
- Machine translation, 163, 165, 168–169, 172–173, 175–178, 384, 386, 408–410
- Management Information Technologies, 128
- Mandatory terms, 389, 434–435
- Manual query construction, 41, 45, 71, 84, 90–91, 116, 128, 154–156, 169, 353–365
- Manual query expansion, 84–85, 90–91, 127, 132, 138
- Manual query modification, 127–128, 137–138, 142, 159, 266–267, 331
- Manual query translation, 165, 167, 169
- MAP, definition of, 59, 61, 73, 311
- Maron, M. E., 287, 373–374, 376, 392
- Mauldin, Michael, 206
- Mean reciprocal rank (MRR), 61, 185, 201, 235, 238, 383
- Medline, 103, 287
- Merging CLIR results, 166, 168, 170, 409
- Merging retrieval lists, 267, 329, 339, 341
- MeSH, 48, 103–105, 119n2
- Metadata, 144–145
- Metasearcher, 201
- Microsoft, 224, 240
- Microsoft Research Cambridge, 7, 137
- Microsoft Research China, 172–173
- Miller, David, 287, 389
- MITAP, 442
- Mitra, Mandar, 131, 187
- Molino, Carmen, 28
- Monolingual retrieval, 10, 49
- Moricz, Michael, 207
- MSN Search, 48, 207, 240, 244
- MUC (Message Understanding Conference), 234
- MultiText, 7, 347–370. *See also* University of Waterloo
cover density ranking, 348, 365–368
GCL, 347, 351–353
query combination, 358–360
retrieval from structured text, 348–353
shortest substring ranking, 347, 353–365
use in question answering, 368–369
- MURAX, 233–234
- Muscat, 287
- Named page, 204–205, 217
- Navigational search, 205, 208, 219, 226, 405–407. *See also* Entry page search
Neue Zuercher Zeitung, 161
- New Mexico State University, 134, 136, 138–139, 154–157, 163–165, 169–170, 174–175
- New York Times*, 192
- New York University, 427
- NLP processing, 328, 427, 440. *See also* Phrases
- Noninterpolated average precision, 59, 61, 73, 245
- Northern Light, 204
- NTCIR, 11, 18n2, 171, 174, 180, 199, 255, 273, 443
- N-grams, 86, 154–156, 158–159, 169, 174, 177–179, 186–187, 337, 339, 373, 408
- Oddy, Robert, 147
- Okapi, 5, 7, 40, 85–87, 89–90, 92–93, 137, 139, 148, 264, 287–297, 326, 365, 374, 406, 408–409, 427. *See also* City University London
BM25, 7, 86, 289–291, 295–297, 374, 377
Chinese experiment, 294
history, 287–289
interactive experiments, 293
pseudorelevance feedback, 291

- routing/filtering, 291–292
- spoken document experiment, 294
- web track, 294
- TREC impact, 294–295
- Omsee, 207
- Open Directory Project (DMOZ), 218–219
- Open Text Corporation, 90
- Open Video Project, 48
- Operational systems, 431–432
- Optical character recognition (OCR), 11, 46, 183–184
 - effect on retrieval, 187
- Optimization
 - in filtering tasks, 108, 109, 112–114
 - query optimization, 115, 314–316
- Oracle, 128
- ORBIT, 431
- Oregon Health Sciences University, 134, 136, 138–139, 141–142
- Out-of-vocabulary words, 191, 195–196, 415
- Over, Paul, 65, 123, 136
- O'Connor, John, 429

- Page, Larry, 204, 207
- PageRank, 214, 217, 226, 406
- Panoptic, 142–143, 223–224
- Parallel corpora, 155, 168–169, 172–173, 175, 271–272, 408–409
- Participation
 - category B, 38, 322
 - list of participants, 15–17
 - number of participants, 7, 8
- Passages, 85–88, 129, 236, 264, 280
- Patents, 27, 38, 81
- Pathfinder, 134
- Pearson correlation, 42, 214
- Pedersen, Jan, 204
- Peoples Daily*, 46, 157
- Performance. *See* Effectiveness
- Peters, Carol, 166
- Phonemes, 188, 195
- Phrases, 14, 91–92, 302, 304, 325, 329, 340, 398–399, 428
- PIRCS, 7, 85–87, 89, 91, 93, 321–342. *See also* City University New York
 - average term frequency measure, 326, 330
 - Chinese bi-grams, 337
 - Chinese dictionary lookup, 339
 - Chinese query expansion, 338
 - Chinese query translation, 338–339
 - Chinese retrieval experiments, 335–339
 - Chinese word segmentation, 336–337
 - collection enrichment, 327–328, 339, 341–342
 - effects of topic length, 332
 - manual ad hoc runs, 331
 - merging retrieval lists, 329, 330, 339, 341
 - phrases, two-word, 325, 330, 340
 - pseudorelevance feedback, 327, 340
 - query zoning, 328, 330
 - recommended techniques, 340–341
 - reranking of retrieval lists, 328–329, 330
 - retrieval status formula, 322–323
 - stop phrase, sentences, 325, 330
 - stop-word removal, 324, 330
 - stemming, 324–325, 330
 - subdocument breakup, 325–326, 330, 340–341
 - suggestions for future work, 341–342
 - term weighting for short queries, 326–327, 340
 - web searching, 332–333
 - Zipf thresholds, 333, 341
- Pirkola, Ari, 386, 388
- Pirkola measure, 169, 271, 339
- Pivoted document-length normalization, 308, 326
- Ponte, Jay, 389
- Pooling, 13, 33, 56, 71, 105–106, 175, 429
- Pools
 - document sources, 34–38
 - question answering, 234–235
 - size, 34–35, 128
- Portal, 201
- Porter, Martin, 287
- Porter stemming algorithms, 154, 287
- PowerAnswer, 241–242
- Precision, definition of, 55, 56, 73, 112
- Precision-oriented measure, 110, 111, 114–115
- Predictive annotation, 410–411
- Primary measure, 109–110

- Priors
 - document, 373, 376
 - document length, 377, 380
 - inlink, 381, 383
 - url depth, 381–383
 - use for entry page retrieval, 217, 378–384
- Probabilistic argumentation, 214
- Probabilistic retrieval models, 373–375, 392
- Profile, 99–100, 107–109, 113, 117–119, 154, 274, 289, 292, 389–391
- Proportion of ties, 66–67
- Pruning experiments, 403–404
- Pseudorelevance (blind) feedback, 14, 41, 195, 307, 309, 340–342, 348, 389, 398
- ad hoc experiments, 7, 41, 88–90, 268, 291, 327
- CLIR experiments, 168–169, 179
- improvement suggestions, 342
- Spanish experiments, 154–156
- Public Radio International, 192
- Query expansion
 - ad hoc, 41–42, 88–91, 308–309, 408
 - best practices, 14
 - Arabic, 177–178
 - Chinese, 159, 270, 338
 - filtering, 116, 314, 388
 - manual, 84–85, 90–91, 127, 132, 138
 - Spanish, 154–156, 270
 - spoken document retrieval, 186, 195
- Query translation, 155, 338, 408
- Query zones, 309, 315, 328
- Question answering, 11, 48, 233, 368–369, 410–415, 436, 441
- Question answering track, 53, 233–256
 - answer correctness, 234–235, 237–238, 244–245, 248–250, 253
 - basic technique, 236
 - combined task, 247–252
 - confidence-weighted score, 245–246, 255, 413
 - definition questions, 240, 247–250, 252, 412–414
 - detecting no answer, 240–241
 - effect of question source, 237
 - evaluation, 252–255
 - exact answer, 244–246, 254
 - example questions, 235, 243
 - judging differences, 253
 - list questions, 243–244, 247
 - original task, 234–236
 - pattern-based systems, 241–243
 - question sources, 48, 237
 - question variants, 237, 239–240
 - strict vs. lenient score, 238
- qrels, 54, 68–70
- Reading comprehension, 234
- Reading time variation, 130
- Recall, definition of, 55, 56, 112
- Recall-precision graph, 58, 59, 62
- Relevance
 - definition, 34, 219
 - probability of relevance given document length, 308, 376–377, 380
- Relevance judgments
 - CLIR, 166–167, 175–176
 - completeness, 24, 34, 42–43, 47, 56, 70–72, 101, 105, 175–176, 202–204
 - consistency (agreement), 12, 29, 34, 43–45, 53, 57, 68–70, 124, 361, 412
 - creation, 23, 33–34, 44, 81–82, 104–105, 201
 - design, 33–34, 42, 45, 423
 - effect of variation, 44
 - graded, 45, 119n1, 204
 - MeSH headings as judgments, 103, 105
 - Pooling, 33, 82, 175–176
 - Reuters subject codes as judgments, 104–105
 - web presentation issues, 223–224
- Relevant documents
 - manual run contribution, 128
 - unique relevant documents, 71
- Reranking of retrieval list, 328, 404–405
- Retrieval run, 53, 55
- Retrieval status value (RSV), 55, 322–323, 326
- Reuters, 48, 104
- Risvik, Knut Magne, 207
- RMIT, 43, 87–89, 91, 93, 134, 137–144, 148, 155–159
- Robertson, Stephen, 211, 217, 287, 374, 376, 378, 382, 392

- Rocchio method, 87–89, 116, 273, 307, 313–315
- Routing task, 7, 10, 24, 100, 108–109, 116, 291, 312
- ROVER, 195
- R-precision, 58–59, 61, 221
- Rutgers University, 91, 128, 130, 132, 135, 137–139, 141–144, 186–187
- Sabir Research, 7, 92. *See also* SMART
- Sahami, Mehran, 384
- Salton, Gerard, 131, 211, 301, 428
- Sanderson, Mark, 28
- San Jose Mercury News*, 25, 27, 81, 325
- Scatter/Gather, 130
- Scholer, Falk, 438
- Schwartz, Richard, 389
- Schweizerische Depeschen Agentur, 47, 160–161, 163, 166, 409
- ScienceDirect, 431
- Searcher behavior, 130, 141
- Searcher characteristics, 125, 128, 130, 135–136, 141–142, 145, 277, 433, 436
- Semantic modeling, 154–156
- Sensitivity analysis, 63–65, 70, 255
- Service finding task, 210, 217
- Set-based evaluation, 109–115
- Shortest substring ranking, 347, 353–365
- Sibling pages, 214
- Singhal, Amit, 131, 207, 216–217, 308, 376
- SMART, 5, 6, 7, 55, 79–80, 83–87, 89, 92–93, 172, 268, 288, 301–319, 403, 421, 427. *See also* Cornell University; Sabir Research
- ad hoc task, 305–312
- confusion track, 316
- dynamic feedback optimization, 273, 314–316
- local-global similarity, 305–308
- non-English, 316–318
- pivoted document-length normalization, 308, 326
- public versions, 303–304
- query zones, 309, 315, 328
- routing task, 312–316
- SuperConcept, 309
- vector processing, 301–303
- yearly improvements, 310–312
- Smeaton, Alan, 400
- Soboroff, Ian, 224
- Spam, 208, 434
- Spanish
- analysis tools, 154–156
- bilingual dictionaries, 155
- cross-language experiments, 271
- documents, 46, 153, 156
- retrieval experiments, 153–156, 269–270
- stemmers, 154–155, 269–270
- stop words, 155, 269
- test collections, 46
- topics, 46, 153–154, 156, 161
- Sparck Jones, Karen, 3, 4, 13, 21, 53, 63, 287, 374, 378, 382, 421–448
- Sparse data problem, 375–376
- Speech recognition, 11, 46, 183, 187, 276, 373, 415
- recognition rates, 192, 415
- Spelling errors, 38, 204, 278
- SPHERE-format, 188
- SPHINX-III, 191
- Spider. *See* Web crawler
- Spitters, Martijn, 389
- Spreading activation, 7, 86–88, 214
- Stanfill, Craig, 401
- Statistical testing
- interaction among users, topics, systems, 65, 125, 136, 146
- Stemmers
- Arabic, 174–175, 178–179, 272, 410
- PIRCS, 324–325
- SMART, 302, 304
- Spanish, 154–155, 269–270
- Stop phrases, sentences, 325
- Stop words, 302, 304, 389
- PIRCS, 324
- Spanish, 155, 269
- Story boundaries, 189, 192–195
- Streaming text, 10
- Structured query, 40, 266, 351–357
- Strzalkowski, Tomek, 91
- Subdocuments, 7, 85–88, 307, 325, 340
- Success at n , 201, 219
- Sullivan, Michael, 429

- Summaries
 - for query expansion, 91
 - for interfaces, 139–140, 143
- Support vector machines, 115–116
- Swanson, Fred, 287, 374
- Swiss Federal Institute of Technology, 130,
157, 160–161, 163–165, 186–187
- System ranking, 68

- Taghva, Kazem, 187
- Target precision. *See* Precision-oriented
measure
- TDT. *See* Topic Detection and Tracking
- Technion, 397
- TechRoute, 224
- Term proximity, 91, 127–128, 262, 328
- Test collections
 - Arabic, 174–176
 - broadcast news, 46, 189
 - CACM, 21–22, 281, 325
 - Chinese, 46–47, 157, 171
 - Cranfield, 4, 21–22
 - CLIR, 47, 160–161, 166
 - design, 21–25, 34, 48–49
 - .GOV, 47, 203, 207, 209–210, 225
 - ideal test collection, 21–22, 34, 53, 288
 - NPL, 21, 281, 325
 - OCR, 46, 184
 - OHSUMED, 48, 103, 107, 119n1
 - question-answering, 48
 - reusable, 48, 124, 210, 424
 - routing/filtering, 48, 101–105
 - Schweizerische Depeschen Agentur, 47,
160–161
 - size of early collections, 4, 6, 21
 - size of TREC collections, 23–26, 206–207,
424–425
 - Spanish, 46
 - TREC English, ad hoc collections, 24, 82,
429
 - use as research tools, 4, 13, 53, 68, 73
 - VLC2, 47, 203, 206–207, 209–210
 - video, 48
 - web, 47, 203, 209, 225
 - WT2G, 47, 203, 209–210, 225
 - WT10G, 47, 203, 207, 209–210, 225
 - Textextract, 410
 - Text categorization, 99–100, 110, 112, 116
 - Thesaurii, 90, 268, 428
 - Thinking Machines, 206
 - Thistlewaite, Paul, 202
 - Three-point average, 55, 57
 - Threshold, 108–110, 113–115, 117, 118,
275, 292, 391
 - Tile Bars, 130
 - TIPSTER, 5, 22–25, 28, 34, 56
 - TNO, 86, 161, 163–165, 217
 - Tomasic, Anthony, 401
 - Topic (information need, user request)
 - Arabic, 47, 174–176
 - broadcast news, 46
 - Chinese, 46–47, 157, 159, 171
 - CLIR, 47, 161–162, 166–168, 171, 174–
176
 - CLIR example, 162
 - creation, 28–33, 36, 38–40, 43, 81, 184,
190–191, 204
 - creation, CLIR, 161, 166–168
 - design, 23, 28–33, 38–40
 - difficulty, 41–42, 128
 - drift, 101, 105, 119
 - Dutch, 161
 - effects of topic length, 41–42, 92–93, 332,
400
 - effects of topic structure, 40–41, 91
 - English ad hoc sets, 24, 29–33
 - examples of , 30, 32, 162, 184, 218
 - fields, 30–32, 38–42
 - formatting, 29–33, 38–40
 - French, 47, 160–162
 - German, 47, 160–162
 - Italian, 47, 166
 - length, 38–41
 - long-term need, 99
 - non-English, 46–47, 157, 159–162, 166–
168, 171, 174–176
 - OCR, 46
 - multilingual, 47, 161–162, 166–168, 171,
174–176
 - multimedia, 48–49
 - routing sets, 24
 - Spanish, 46, 153–154, 156, 161
 - statement of user's information need, 5,
23, 65

- translation, 167
- video, 48–49
- Topic Detection and Tracking, 117, 192, 195, 281, 429, 444
- Topic distillation, 143–144, 204–205, 210, 217–221, 404–405, 425, 430
- Tracks
 - confusion track, 183–187, 316
 - Chinese, 156–160
 - cross-language, European, 160–171
 - cross-language, Arabic, 173–180
 - cross-language, Chinese, 171–173
 - database merging track, 12, 275–276
 - filtering track, 99–119, 273–274, 428–429
 - genomics track, 12, 432, 440
 - HARD track, 10, 144–145, 267, 279–280, 429, 433
 - high-precision track, 12, 360
 - interactive track, 123–152, 223–224, 277, 293, 407, 429
 - list of tracks, 9–10
 - NLP track, 12
 - novelty track, 11, 279, 425
 - purpose of track structure, 4, 8, 424
 - query track, 12, 65, 277
 - question answering, 233–256, 277–278, 429, 440
 - robust track, 10
 - selection of tracks, 9
 - Spanish, 153–157
 - spoken document retrieval track, 187–196, 276–277, 415, 439
 - video track, 11, 48–49, 415–416, 439
 - VLC track, 200, 202–204, 361–363, 400–403
 - Web track, 142–144, 199–228, 437
- Transactional search, 208, 221, 226
- Transcripts
 - baseline, 188–189, 191, 192, 415
 - cross-recognizer, 191–192
 - one-best, word-based, 188, 191, 195
 - speech, 188–189, 192
 - reference, 188–189, 192, 195
- Translation lexica, 175
- Transliteration, 175
- Travis, Bob, 207
- TREC
 - evaluation philosophy, 426, 429
 - future challenges, 430–434, 437–444
 - goals, 5, 79, 205–210, 421, 423
 - history, 23–24, 25, 45, 47, 93–94, 99–104, 153, 202–205
 - impact of, 13–14, 49, 79–80, 93–94, 171, 180, 207–208, 224, 261, 272–273, 276, 280–282, 294–295, 301, 322, 407, 416, 421–435, 437–438, 440–444
 - trec_eval, 53–56, 59, 82, 108, 112
 - Twenty-One Consortium, 7, 86–87, 89, 93, 168–170, 373, 392
 - Two-Poisson distribution, 287, 290, 374, 392
- United Nations corpus, 155, 175, 178, 410
- University of California at Berkeley, 84, 127, 135, 137–138, 155–159, 163–165, 168–170, 177–180
- University of California at Los Angeles, 307
- University of Central Florida, 154–156
- University of Colorado, 163–165
- University of Dortmund, 84, 313
- University of Glasgow, 28, 139, 141–142
- University of Hildesheim, 168
- University of Massachusetts at Amherst, 5, 6, 93, 127, 135, 154–159, 161, 163–165, 177–180, 192. *See also* INQUERY
- Arabic experiments, 271–272
- Chinese experiments, 270–271
- entity recognition, 264–265
- HARD track, 267, 279–280
- InFinder, 154–155, 267–268
- INQUERY system description, 262–267
- INQUERY term weighting, 263–264
- interactive experiments, 277
- local context analysis, 267–271
- novelty experiments, 279
- query processing, 265–267
- query track contributions, 277–278
- question answering experiments, 278–279
- routing/filtering, 273–275
- Spanish experiments, 269–271
- spoken document experiments, 276–277
- structured query example, 266
- TREC impact, 280–282
- VLC experiments, 276

- University of Maryland, 163–165, 168–170, 174–180
- University of Michigan, 141–142
- University of Montreal, 164–165, 169
- University of Neuchatel, 177, 180, 214
- University of Nevada, 206
- University of North Carolina, 135, 137–138, 142
- University of Sheffield, 138–139, 194
- University of Southern California, 241
- University of Toronto, 91, 128, 130, 137–138, 141–142, 148
- University of Waterloo, 7, 45, 86, 88, 91, 210–211, 243. *See also* MULTITEXT
- Upstill, Trystan, 217
- URL, use in search, 202, 226, 379, 381–382
- Usenet news server, 349–351
- User-in-the-loop, 10, 18, 72, 84–85, 119, 123–152, 432–433
- User request. *See* Topic
- User studies, 18, 91, 123–152
- Utility, 106, 110–111, 113, 114–115, 117, 292, 390–391
- van Rijsbergen, Keith, 13, 21, 28, 53, 112, 287, 374–375, 400
- Variability, 13, 18, 44, 53, 61–63, 67, 193, 439
- Vector model, 302
- Verity Corporation, 84, 88, 128
- VERONICA, 206
- ViaVoice, 415
- Virginia Tech, 84
- Visualization interfaces, 130, 135–136, 138–139, 142, 277
- VLC2, 202–204, 206–207
- Voice of America, 192
- Voorhees, Ellen M., 44–45, 128, 166, 204, 357, 430
- WAIS, 206, 431
- Walker, Stephen, 287, 376
- Wall Street Journal*, 25, 27–28, 34, 36, 38, 46, 81, 306, 325, 407
- Webclopedia, 241
- Web crawl, 200, 204, 208, 216
- Web crawler, 119, 200–201
- Web indexing, 403
- Web page, 200
- Web search, 11, 13, 141, 200, 204, 215, 332, 422–423, 434–439
- Web search engines, 434–439, 442
- Web site, 200
- Web track, 199–228
- evaluation, 219–224
- history, 202–205
- limitations, 224–225
- Wilder, Dean, 28
- Williams, Hugh, 438
- Weighting
- best practices, 14, 340
- BM25 weights, 7, 86, 289–291, 295–297, 374, 377
- experiments, 84–88
- INQUERY, 263–264
- logistic regression, 84
- polynomial regression, 84
- for short queries, 326
- tf*idf weights, 303, 340, 374
- Wippern, Dorothy, 429
- Word error rate, 190, 194
- WordNet, 90, 238, 242, 342, 412, 428
- Womser-Hacker, Christa, 168
- WT2g, 47, 204, 209–210, 225
- WT10g, 47, 204, 207, 209–210, 225, 335, 383
- Xelda morphological tool, 386
- Xerox, 130, 154–157, 163–165, 309, 386
- Xinhua, 46, 157
- Xu, Jack, 207
- Yahoo!, 206, 218–220, 434
- Yang, Yiming, 111
- Zhang, Nien-Fan, 65
- Ziff-Davis Publishing Computer Select disks, 25, 27–28, 36–37, 81
- Zipf, 333, 341
- Zobel, Justin, 43, 71, 430
- ZPRISE, 133, 135, 137
- Z39.50, 206