

Index

- Absolute discounting, 130
- Abstractive text summarization, 318, 321–324
- Access modes, 73–76
- Accuracy in search engine evaluation, 168
- Ad hoc information needs, 8–9
- Ad hoc retrieval, 75–76
- Add-1 smoothing, 130, 464
- Adjacency matrices, 207–208
- Advertising, opinion mining for, 393
- Agglomerative clustering, 277, 280–282, 290
- Aggregating
 - opinions, 393
 - scores, 234
- All-vs-all (AVA) method, 313
- Ambiguity
 - full structure parsing, 43
 - LARA, 406
 - NLP, 40–41, 44
 - one-vs-all method, 313
 - text retrieval vs. database retrieval, 80
 - topics, 335, 337
- Analyzers in MeTA toolkit, 61–64, 453
- `analyzers::filters` namespace, 64
- `analyzers::tokenizers` namespace, 64
- Anaphora resolution in natural language processing, 41
- Anchor text in web searches, 201
- Architecture
 - GFS, 194–195
 - MeTA toolkit, 60–61
 - unified systems, 452–453
- Art retrieval models, 111
- Aspect opinion analysis, 325–326
- Associations, word. *See* Word association mining
- Authority pages in web searches, 202, 207
- Automatic evaluation in text clustering, 294
- AVA (all-vs-all) method, 313
- Average-link document clustering, 282
- Average precision
 - ranked lists evaluation, 175, 177–180
 - search engine evaluation, 184
- Axiomatic thinking, 88
- Background models
 - mining topics from text, 345–351
 - mixture model estimation, 351–353
 - PLSA, 370–372
- Background words
 - mixture models, 141, 351–353
 - PLSA, 368–369, 372
- Bag-of-words
 - frequency analysis, 69
 - paradigmatic relations, 256
 - text information systems, 10
 - text representation, 88–90
 - vector space model, 93, 109
 - web searches, 215
- Bar-Hillel report, 42
- Baseline accuracy in text categorization, 314
- Bayes, Thomas, 25
- Bayes' rule
 - EM algorithm, 361–363, 373–374

- Bayes' rule (*continued*)
 - formula, 25–26
 - LDA, 383
- Bayesian inference
 - EM algorithm, 361–362
 - PLSA, 379, 382
- Bayesian parameter estimation
 - formula, 458
 - overfitting problem, 28–30
 - unigram language model, 341, 359
- Bayesian smoothing, 125
- Bayesian statistics
 - binomial estimation and beta distribution, 457–459
 - Dirichlet distribution, 461–463
 - LDA, 382
 - multinomial distribution, 460–461
 - multinomial parameters, 463–464
 - Naive Bayes algorithm, 309–312
 - pseudo counts, smoothing, and setting hyperparameters, 459–460
- Berkeley study, 3
- Bernoulli distribution, 26
- Beta distribution, 457–459
- Beta-gamma threshold learning, 227–228
- Bias, clustering, 276
- Big text data, 5–6
- Bigram language model
 - abstractive summarization, 323
 - Brown clustering, 290
- Bigrams
 - frequency analysis, 68
 - sentiment classification, 394–395
 - text categorization, 305
 - words tokenizers, 149
- Binary classification
 - content-based recommendation, 223
 - text categorization, 303
- Binary hidden variables in EM algorithm, 362–364, 366, 368, 467
- Binary logistic regression, 397
- Binomial distribution, 26–27
- Binomial estimation, 457–459
- Bit vector representation, 93–97
- Bitwise compression, 159–160
- Blind feedback, 133, 135
- Block compression, 161–162
- Block world project, 42
- BM25 model
 - description, 88
 - document clustering, 279
 - document length normalization, 108–109
 - link analysis, 201
 - Okapi, 89, 108
 - popularity, 90
 - probabilistic retrieval models, 111
- BM25-F model, 109
- BM25 score
 - paradigmatic relations, 258–261
 - syntagmatic relations, 270
 - web search ranking, 210
- BM25 TF transformation
 - description, 104–105
 - paradigmatic relations, 258–259
- BM25+ model, 88, 110
- Breadth-first crawler searches, 193
- Breakeven point precision, 189
- Brown clustering, 278, 288–291
- Browsing
 - multimode interactive access, 76–78
 - pull access mode, 73–75
 - support for, 445
 - text information systems, 9
 - web searches, 214
 - word associations, 252
- Business intelligence
 - opinion mining, 393
 - text data analysis, 243
- C++ language, 16, 58
- Caching
 - DBLRU, 164–165
 - LRU, 163–164
 - META toolkit, 60
 - search engine implementation, 148, 162–165
- Categories
 - categorical distributions, 460–461

- sentiment classification, 394, 396–397
- text information systems, 11–12
- Causal topic mining, 433–437
- Centroid vectors, 136–137
- Centroids in document clustering, 282–284
- CG (cumulative gain) in NDCG, 181–182
- character_tokenizer tokenizer, 61
- Citations, 202
- Classes
 - Brown clustering, 289
 - categories, 11–12
 - sentiment, 393–396
- Classification
 - machine learning, 34–36
 - NLP, 43–44
- Classifiers in text categorization, 302–303
- classify command, 57
- Cleaning HTML files, 218–219
- Clickthroughs
 - probabilistic retrieval model, 111–113
 - web searches, 201
- Clustering bias, 276
- Clusters and clustering
 - joint analysis, 416
 - sentiment classification, 395
 - text. *See* Text clustering
- Coherence in text clustering, 294–295
- Coin flips, binomial distribution for, 26–27
- Cold start problem, 230
- Collaborative filtering, 221, 229–233
- Collapsed Gibbs sampling, 383
- Collect function, 197
- Collection language model
 - KL-divergence, 474
 - smoothing methods, 121–126
- Common form of retrieval models, 88–90
- Common sense knowledge in NLP, 40
- Common words
 - background language model, 346–347, 350–351
 - feedback, 141, 143
 - filtering, 54
 - mixture models, 352–353, 355–356
 - unigram language model, 345–346
 - vector space retrieval models, 99, 109
- Compact clusters, 281
- Compare operator, 450, 452
- Complete data for EM algorithm, 467–468
- Complete-link document clustering, 281–282
- Component models
 - background language models, 345, 347–350
 - CPLSA, 421
 - description, 143
 - EM algorithm, 359
 - mixture models, 355–356, 358–359
 - PLSA, 370–373
- Compression
 - bitwise, 159–160
 - block, 161–162
 - overview, 158–159
 - search engines, 148
 - text representation, 48–49
- Compression ratio, 160–161
- Concepts in vector space model, 92
- Conceptual framework in text information systems, 10–13
- Conditional entropy
 - information theory, 33
 - syntagmatic relations, 261–264, 270
- Conditional probabilities
 - Bayes' rule, 25–26
 - overview, 23–25
- Configuration files, 57–58
- Confusion matrices, 314–315
- Constraints in PLSA, 373
- Content analysis modules, 10–11
- Content-based filtering, 221–229
- Content in opinion mining, 390–392
- Context
 - Brown clustering, 290
 - non-text data, 249
 - opinion mining, 390–392
 - paradigmatic relations, 253–258
 - social networks as, 428–433
 - syntagmatic relations, 261–262
 - text mining, 417–419

- Context (*continued*)
 - time series, 433–439
- Context variables in topic analysis, 330
- Contextual Probabilistic Latent Semantic Analysis (CPLSA), 419–428
- Continuous distributions
 - Bayesian parameter estimation, 28
 - description, 22
- Co-occurrences in mutual information, 267–268
- Corpus input formats in MeTA toolkit, 60–61
- corpusname.dat file, 60
- corpusname.dat.gz file, 60
- corpusname.dat.labels file, 60
- corpusname.dat.labels.gz file, 60
- Correlations
 - mutual information, 270
 - syntagmatic relations, 253–254
 - text-based forecasting, 248
 - time series context, 437
- Cosine similarity
 - document clustering, 279–280
 - extractive summarization, 321
 - text summarization, 325
 - vector measurement, 222, 232
- Coverage
 - CPLSA, 420–422, 425–426
 - LDA, 380–381
 - topic analysis, 332–333
- CPLSA (Contextual Probabilistic Latent Semantic Analysis), 419–428
- Cranfield evaluation methodology, 168–170
- Crawlers
 - domains, 218
 - dynamic content, 217
 - languages for, 216–217
 - web searches, 192–194
- Cross validation in text categorization, 314
- Cumulative gain (CG) in NDCG, 181–182
- Current technology, 5
- Data-driven social science research, opinion mining for, 393
- Data mining
 - joint analysis, 413–415
 - probabilistic retrieval model algorithms, 117
 - text data analysis, 245–246
- Data types in text analysis, 449–450
- Data-User-Service Triangle, 213–214
- Database retrieval, 80–82
- DBLRU (Double Barrel Least-Recently Used)
 - caches, 164–165
- DCG (discounted cumulative gain), 182–183
- Decision boundaries for linear classifiers, 311–312
- Decision modules in content-based filtering, 225
- Decision support, opinion mining for, 393
- Deep analysis in natural language processing, 43–45
- Delta bitwise compression, 160
- Dendrograms, 280–281
- Denial of service from crawlers, 193
- Dependency parsers, 323
- Dependent random variables, 25
- Design philosophy, MeTA, 58–59
- Development sets for text categorization, 314
- Dirichlet distribution, 461–463
- Dirichlet prior smoothing
 - KL-divergence, 475
 - probabilistic retrieval models, 125–127
- Disaster response, 243–244
- Discounted cumulative gain (DCG), 182–183
- Discourse analysis in NLP, 40
- Discrete distributions
 - Bayesian parameter estimation, 29
 - description, 22
- Discriminative classifiers, 302
- Distances in clusters, 281
- Distinguishing categories, 301–302
- Divergence-from-randomness models, 87, 111
- Divisive clustering, 277
- Document-at-a-time ranking, 155

- Document clustering, 277
 - agglomerative hierarchical, 280–282
 - K*-means, 282–284
 - overview, 279–280
- Document frequency
 - bag-of-words representation, 89
 - vector space model, 99–100
- Document IDs
 - compression, 158–159
 - inverted indexes, 152
 - tokenizers, 149
- Document language model, 118–123
- Document length
 - bag-of-words representation, 89
 - vector space model, 105–108
- Documents
 - filters, 155–156
 - ranking vs. selecting, 82–84
 - tokenizing, 148–150
 - vectors, 92–96
 - views in multimode interactive access, 77
- Domains, crawling, 218
- Dot products
 - document length normalization, 109
 - linear classifiers, 311
 - paradigmatic relations, 257–258
 - vector space model, 93–95, 98
- Double Barrel Least-Recently Used (DBLRU)
 - caches, 164–165
- Dynamic coefficient interpolation in
 - smoothing methods, 125
- Dynamically generated content and
 - crawlers, 217
- E step in EM algorithm, 362–368, 373–377, 465, 469
- E-discovery (electronic discovery), 326
- Edit features in text categorization, 306
- Effectiveness in search engine evaluation, 168
- Efficiency
 - database data retrieval, 81–82
 - search engine evaluation, 168
- Electronic discovery (E-discovery), 326
- Eliza project, 42, 44–45
- EM algorithm. *See* Expectation-maximization (EM) algorithm
- Email counts, 3
- Emotion analysis, 394
- Empirically defined problems, 82
- Enron email dataset, 326
- Entity-relation re-creation, 47
- Entropy
 - information theory, 31–33
 - KL-divergence, 139, 474
 - mutual information, 264–265
 - PMI, 288
 - skewed distributions, 158
 - syntagmatic relations, 261–264, 270
- Evaluation, search engine. *See* Search engine evaluation
- Events
 - CPLSA, 426–427
 - probability, 21–23
- Exhaustivity in sentiment classification, 396
- Expectation-maximization (EM) algorithm
 - CPLSA, 422
 - general procedure, 469–471
 - incomplete vs. complete data, 467–468
 - K*-means, 282–283
 - KL-divergence, 476
 - lower bound of likelihood, 468–469
 - MAP estimate, 378–379
 - mining topics from text, 359–368
 - mixture unigram language model, 466
 - MLE, 466–467
 - network supervised topic models, 431
 - overview, 465–466
 - PLSA, 373–377
- Expected overlap of words in paradigmatic
 - relations, 257–258
- Expected value in Beta distribution, 458
- Exploration-exploitation tradeoff in
 - content-based filtering, 227
- Extractive summarization, 318–321
- F measure
 - ranked lists evaluation, 179

- F measure (*continued*)
 - set retrieval evaluation, 172–173
- F*-test for time series context, 437
- F_1 score
 - text categorization, 314
 - text summarization, 324
- Fault tolerance in Google File System, 195
- Feature generation for tokenizers, 150
- Features for text categorization, 304–307
- Feedback
 - content-based filtering, 225
 - KL-divergence, 475–476
 - language models, 138–144
 - overview, 133–135
 - search engines, 147, 157–158
 - vector space model, 135–138
 - web searches, 201
- Feedback documents in unigram language model, 466
- Feelings. *See* Sentiment analysis
- `fetch_docs` function, 154
- `file_corpus` input format, 60
- Files in Google File System, 194–195
- Filter chains for tokenization, 61–64
- Filters
 - content-based, 221–229
 - documents, 155–156
 - recommender systems. *See* Recommender systems
 - text information systems, 11
 - unigram language models, 54
- Focused crawling, 193
- `forward_index` indexes, 60–61
- Forward indexes
 - description, 153
 - k*-nearest neighbors algorithm, 308
- Frame of reference encoding, 162
- Frequency and frequency counts
 - bag-of-words representation, 89–90
 - MapReduce, 197
 - META analyses, 68–70
 - term, 97–98
 - vector space model, 99–100
- Frequency transformation in paradigmatic relations, 258–259
- Full structure parsing, 43
- G*-means algorithm, 294
- Gain in search engine evaluation, 181–183
- Gamma bitwise compression, 160
- Gamma function, 457
- Gaussian distribution, 22, 404–405
- General EM algorithm, 431
- Generation-based text summarization, 318
- Generative classifiers, 309
- Generative models
 - background language model, 346–347, 349
 - CPLSA, 419, 421
 - description, 30, 36, 50
 - LARA, 403, 405–406
 - LDA, 381
 - log-likelihood functions, 343–344, 384
 - mining topics from text, 347
 - n*-gram models, 289
 - network supervised topic models, 428–430
 - PLSA, 370–371, 380
 - topics, 338–340
 - unigram language model, 341
- Geographical networks, 428
- Geometric mean average precision (gMAP), 179
- GFS (Google File System), 194–195
- Gibbs sampling, 383
- Google File System (GFS), 194–195
- Google PageRank, 202–206
- Grammar learning, 252
- Grammatical parse trees, 305–307
- Granger test, 434, 437
- Graph mining, 49
- `gz_corpus` input format, 60
- Hidden variables
 - EM algorithm, 362–364, 366, 368, 373–376, 465, 467
 - LARA, 403

- Hierarchical clustering, 280–282
- High-level syntactic features, 305–306
- Hill-climbing algorithm, EM, 360, 366–367, 465
- HITS algorithm, 206–208
- HTML files, cleaning, 218–219
- Hub pages in web searches, 202, 207–208
- Humans
 - joint analysis, 413–415
 - NLP, 48
 - opinion mining. *See* Opinion mining
 - as subjective sensors, 244–246
 - unified systems, 445–448
- Hyperparameters
 - Beta distribution, 458–460
 - Dirichlet distribution, 461, 463
- ICU (International Components for Unicode), 61
- Icu_filter filter, 61
- Icu_tokenizer tokenizer, 61
- IDF (inverse document frequency)
 - Dirichlet prior smoothing, 126
 - paradigmatic relations, 258–260
 - query likelihood retrieval model, 122
 - vector space model, 99–101
- Illinois NLP Curator toolkit, 64
- Impact
 - CPLSA, 426–427
 - time series context, 437
- Implicit feedback, 134–135
- Incomplete data in EM algorithm, 467–468
- Incremental crawling, 193
- Independent random variables, 25
- Index sharding, 156–157
- Indexes
 - compressed, 158–162
 - forward, 153, 308
 - k*-nearest neighbors algorithm, 308
 - MapReduce, 198–199
 - META toolkit, 60–61, 453–455
 - search engine implementation, 150–153
 - search engines, 147, 150–153
 - text categorization, 314
 - web searches, 194–200
- Indirect citations in web searches, 202
- Indirect opinions, 391–392
- Indri/Lemur search engine toolkit, 64
- Inferences
 - NLP, 41
 - probabilistic, 88
 - real world properties, 248
- Inferred opinions, 391–392
- Information access in text information systems, 7
- Information extraction
 - NLP, 43
 - text information systems, 9, 12
- Information retrieval (IR) systems, 6
 - evaluation metrics, 324–325
 - implementation. *See* Search engine implementation
 - text data access, 79
- Information theory, 31–34
- Initial values in EM algorithm, 466
- Initialization modules in content-based filtering, 224–225
- Inlink counts in PageRank, 203
- Instance-based classifiers, 302
- Instructor reader category, 16–17
- Integer compression, 158–162
- Integration of information access in web searches, 213
- Integrity in text data access, 81
- Interactive access, multimode, 76–78
- Interactive task support in web searches, 216
- International Components for Unicode (ICU), 61
- Interpolation for smoothing methods, 125–126
- Interpret operator, 450–452
- Intersection operator, 449–450
- Intrusion detection, 271–273
- Inverse document frequency (IDF)
 - Dirichlet prior smoothing, 126
 - paradigmatic relations, 258–260
 - query likelihood retrieval model, 122

- Inverse document frequency (IDF)
 - (*continued*)
 - vector space model, 99–101
- Inverse user frequency (IUF), 232
- inverted_index indexes, 60
- Inverted index chunks, 156–157
- Inverted indexes
 - compression, 158
 - k -nearest neighbors algorithm, 308
 - MapReduce, 198–199
 - search engines, 150–153
- IR (information retrieval) systems, 6
 - evaluation metrics, 324–325
 - implementation. *See* Search engine implementation
 - text data access, 79
- Iterative algorithms for PageRank, 205–206
- Iterative Causal Topic Modeling, 434–435
- IUF (inverse user frequency), 232
- Jaccard similarity, 280
- Jelinek-Mercer smoothing, 123–126
- Joint analysis of text and structured data, 413
 - contextual text mining, 417–419
 - CPLSA, 419–428
 - introduction, 413–415
 - social networks as context, 428–433
 - time series context, 433–439
- Joint distributions for mutual information, 266–268
- Joint probabilities, 23–25
- K -means document clustering, 282–284
- K -nearest neighbors (k -NN) algorithm, 307–309
- Kernel trick for linear classifiers, 312
- Key-value pairs in MapReduce, 195–198
- KL-divergence
 - Dirichlet prior smoothing, 475
 - EM algorithm, 468
 - feedback, 139–140
 - mutual information, 266
 - query model, 475–476
 - retrieval, 473–474
- Knowledge acquisition in text information systems, 8–9
- Knowledge discovery in text summarization, 326
- Knowledge Graph, 215
- Knowledge provenance in unified systems, 447
- Known item searches in ranked lists
 - evaluation, 179
- Kolmogorov axioms, 22–23
- Kullback-Leibler divergence retrieval model. *See* KL-divergence
- Lagrange Multiplier approach
 - EM algorithm, 467, 470
 - unigram language model, 344
- Language models
 - feedback in, 138–144
 - in probabilistic retrieval model, 87, 111, 117
- Latent Aspect Rating Analysis (LARA), 400–409
- Latent Dirichlet Allocation (LDA), 377–383
- Latent Rating Regression, 402–405
- Lazy learners in text categorization, 302
- Learners
 - search engines, 147
 - text categorization, 302
- Learning modules in content-based filtering, 224–225
- Least-Recently Used (LRU) caches, 163–164
- length_filter filter, 61
- Length normalization
 - document length, 105–108
 - query likelihood retrieval model, 122
- Lexical analysis in NLP, 39–40
- Lexicons for inverted indexes, 150–152
- LIBLINEAR algorithm, 58
- libsvm_analyzer analyzer, 62
- libsvm_corpus file, 61
- LIBSVM package, 58, 64
- Lifelong learning in web searches, 213

- Likelihood and likelihood function
 - background language model, 349–351
 - EM algorithm, 362–363, 367–368, 376, 465–469
 - LARA, 405
 - LDA, 378, 381–382
 - marginal, 28
 - mixture model behavior, 354–357
 - MLE, 27
 - network supervised topic models, 428–431
 - PLSA, 372–374
 - unigram language model, 342–344
- line_corpus input format, 60
- Linear classifiers in text categorization, 311–313
- Linear interpolation in Jelinek-Mercer smoothing, 124
- Linearly separable data points in linear classifiers, 312
- Link analysis
 - HITS, 206–208
 - overview, 200–202
 - PageRank, 202–206
- list_filter filter, 62
- Local maxima, 360, 363, 367–368, 465
- Log-likelihood function
 - EM algorithm, 365–366, 466–467
 - feedback, 142–143
 - unigram language model, 343–344
- Logarithm transformation, 103–104
- Logarithms in probabilistic retrieval model, 118, 122
- Logic-based approach in NLP, 42
- Logical predicates in NLP, 49–50
- Logistic regression in sentiment classification, 396–400
- Long-range jumps in multimode interactive access, 77
- Long-term needs in push access mode, 75
- Low-level lexical features in text categorization, 305
- Lower bound of likelihood in EM algorithm, 468–469
- LRU (Least-Recently Used) caches, 163–164
- Lucene search engine toolkit, 64
- M step
 - EM algorithm, 361–368, 373–377, 465, 469–470
 - MAP estimate, 379
 - network supervised topic models, 431
- Machine-generated data, 6
- Machine learning
 - overview, 34–36
 - sentiment classification methods, 396
 - statistical, 10
 - text categorization, 301
 - web search algorithms, 201
 - web search ranking, 208–212
- Machine translation, 42, 44–45
- Magazine output, 3
- Manual evaluation for text clustering, 294
- map function, 195–198
- MAP (Maximum a Posteriori) estimate
 - Bayesian parameter estimation, 29
 - LARA, 404–405
 - PLSA, 378–379
 - word association mining, 271–273
- MAP (mean average precision), 178–180
- Map Reduce paradigm, 157
- MapReduce framework, 194–200
- Maps in multimode interactive access, 76–77
- Marginal probabilities
 - Bayesian parameter estimation, 29
 - mutual information, 267
- Market research, opinion mining for, 393
- Massung, Sean, biography, 490
- Matrices
 - adjacency, 207–208
 - PageRank, 204–208
 - text categorization, 314–315
 - transition, 204
- Matrix multiplication in PageRank, 205
- Maximal marginal relevance (MMR)
 - reranking
 - extractive summarization, 320–321

- Maximal marginal relevance (MMR)
 - reranking (*continued*)
 - topic analysis, 333
- Maximization algorithm for document clustering, 282
- Maximum a Posteriori (MAP) estimate
 - Bayesian parameter estimation, 29
 - LARA, 404–405
 - PLSA, 378–379
 - word association mining, 271–273
- Maximum likelihood estimation (MLE)
 - background language model, 346, 350
 - Brown clustering, 289
 - Dirichlet prior smoothing, 125–126
 - EM algorithm, 359–368, 466–467
 - feedback, 141–143
 - generative models, 339
 - Jelinek-Mercer smoothing, 124
 - KL-divergence, 475–476
 - LARA, 404
 - LDA, 382
 - mixture model behavior, 354–359
 - mixture model estimation, 352–353
 - multinomial distribution, 463
 - mutual information, 268–269
 - overview, 27–28
 - PLSA, 372–373, 378
 - query likelihood retrieval model, 118–119
 - term clustering, 286
 - unigram language models, 52–53, 341–345
 - web search ranking, 210
- Mean average precision (MAP), 178–180
- Mean reciprocal rank (MRR), 180
- Measurements in search engine evaluation, 168
- Memory-based approach in collaborative filtering, 230
- META toolkit
 - architecture, 60–61
 - classification algorithms, 307
 - design philosophy, 58–59
 - exercises, 65–70
 - overview, 57–58
 - related toolkits, 64–65
 - setting up, 59–60
 - text categorization, 314–315
 - tokenization, 61–64
 - as unified system, 453–455
- Metadata
 - classification algorithms, 307
 - contextual text mining, 417
 - networks from, 428
 - text data analysis, 249
 - topic analysis, 330
- Mining
 - contextual, 417–419
 - demand for, 4–5
 - graph, 49
 - joint analysis, 413–419
 - opinion. *See* Opinion mining; Sentiment analysis
 - probabilistic retrieval model, 117
 - tasks, 246–250
 - toolkits, 64
 - topic analysis, 330–331
 - word association. *See* Word association mining
- Mining topics from text, 340
 - background language model, 345–351
 - expectation-maximization, 359–368
 - joint analysis, 416
 - mixture model behavior, 353–359
 - mixture model estimation, 351–353
 - unigram language model, 341–345
- Mixture models
 - behavior, 353–359
 - EM algorithm, 466
 - estimation, 351–353
 - feedback, 140–142, 157
 - mining topics from text, 346–351
- MLE. *See* Maximum likelihood estimation (MLE)
- MMR (maximal marginal relevance)
 - reranking
 - extractive summarization, 320–321
 - topic analysis, 333
- Model-based clustering algorithms, 276–277
- Model files for META toolkit, 59

- Modification in NLP, 41
- Modules in content-based filtering, 224–226
- MRR (mean reciprocal rank), 180
- Multiclass classification
 - linear classifiers, 313
 - text categorization, 303
- Multi-level judgments in search engine
 - evaluation, 180–183
- Multimode interactive access, 76–78
- Multinomial distributions
 - Bayesian estimate, 463–464
 - generalized, 460–461
 - LDA, 380
- Multinomial parameters in Bayesian
 - estimate, 463–464
- Multiple-level sentiment analysis, 397–398
- Multiple occurrences in vector space model, 103–104
- Multiple queries in ranked lists evaluation, 178–180
- Multivariate Gaussian distribution, 404–405
- Mutual information
 - information theory, 33–34
 - syntagmatic relations, 264–271
 - text clustering, 278
- n*-fold cross validation, 314
- n*-gram language models
 - abstractive summarization, 322–323
 - frequency analysis, 68–69
 - sentiment classification, 394–395
 - term clustering, 288–291
 - vector space model, 109
- Naive Bayes algorithm, 309–312
- Named entity recognition, 323
- Natural language, mining knowledge about, 247
- Natural language generation in text
 - summarization, 323–324
- Natural language processing (NLP)
 - history and state of the art, 42–43
 - pipeline, 306–307
 - sentiment classification, 395
 - statistical language models, 50–54
 - tasks, 39–41
 - text information systems, 43–45
 - text representation, 46–50
- Navigating maps in multimode interactive
 - access, 77
- Navigational queries, 200
- NDCG (normalized discounted cumulative gain), 181–183
- NDCG@*k* score, 189
- Nearest-centroid classifiers, 309
- Negative feedback documents, 136–138
- Negative feelings, 390–394
- NetPLSA model, 430–433
- Network supervised topic models, 428–433
- Neural language model, 291–294
- News summaries, 317
- Newspaper output, 3
- ngram_pos_analyzer* analyzer, 62
- ngram_word_analyzer* analyzer, 62
- NLP. *See* Natural language processing (NLP)
- NLTK toolkit, 64
- no_evict_cache* caches, 60
- Nodes in word associations, 252
- Non-text data
 - context, 249
 - predictive analysis, 249
 - vs. text, 244–246
- Normalization
 - document length, 105–108
 - PageRank, 206
 - query likelihood retrieval model, 122
 - term clustering, 286
 - topic analysis, 333
- Normalized discounted cumulative gain (NDCG), 181–183
- Normalized ratings in collaborative
 - filtering, 230–231
- Normalized similarity algorithm, 279
- Objective statements vs. subjective, 389–390
- Observed world, mining knowledge about, 247–248
- Observers, mining knowledge about, 248
- Office documents, 3
- Okapi BM25 model, 89, 108
- One-vs-all (OVA) method, 313

- Operators in text analysis systems, 448–452
- Opinion analysis in text summarization, 325–326
- Opinion holders, 390–392
- Opinion mining
 - evaluation, 409–410
 - LARA, 400–409
 - overview, 389–392
 - sentiment classification. *See* Sentiment analysis
- Opinion summarization, 318
- Optimization in web searches, 191
- Ordinal regression, 394, 396–400
- Organization in text information systems, 8
- OVA (one-vs-all) method, 313
- Over-constrained queries, 84
- Overfitting problem
 - Bayesian parameter estimation, 28, 30
 - sentiment classification, 395
 - vector space model, 138
- Overlap of words in paradigmatic relations, 257–258
- p*-values in search engine evaluation, 185–186
- PageRank technique, 202–206
- Paradigmatic relations
 - Brown clustering, 290
 - discovering, 252–260
 - overview, 251–252
- Parallel crawling, 193
- Parallel indexing and searching, 192
- Parameters
 - background language model, 350–351
 - Bayesian parameter estimation, 28–30, 341, 359, 458, 463–464
 - Beta distribution, 458–460
 - Dirichlet distribution, 461–463
 - EM algorithm, 363, 465
 - feedback, 142–144
 - LARA, 404–405
 - LDA, 380–381
 - mixture model estimation, 352
 - MLE. *See* Maximum likelihood estimation (MLE)
 - network supervised topic models, 429
 - PLSA, 372–373, 379–380
 - probabilistic models, 30–31
 - ranking, 209–211
 - statistical language models, 51–52
 - topic analysis, 338–339
 - unigram language models, 52
- Parsing
 - META toolkit, 67–68
 - NLP, 43
 - web content, 216
- Part-of-speech (POS) tags
 - META toolkit, 67
 - NLP, 47
 - sentiment classification, 395
- Partitioning
 - Brown clustering, 289
 - extractive summarization, 319–320
 - text data, 417–419
- Patterns
 - contextual text mining, 417–419
 - CPLSA, 425–426
 - joint analysis, 417
 - NLP, 45
 - sentiment classification, 395
- Pdf (probability density function)
 - Beta distribution, 457
 - Dirichlet distribution, 461
 - multinomial distribution, 461
- Pearson correlation
 - collaborative filtering, 222, 231–232
 - time series context, 437
- Perceptron classifiers, 312–313
- Personalization in web searches, 212, 215
- Personalized PageRank, 206
- Perspective in text data analysis, 246–247
- Pivoted length normalization, 89, 107–108
- PL2 model, 90
- PLSA (probabilistic latent semantic analysis)
 - CPLSA, 419–428
 - extension, 377–383
 - overview, 368–377
- Pointwise Mutual Information (PMI), 278, 287–288

- Polarity analysis in sentiment classification, 394
- Policy design, opinion mining for, 393
- Pooling in search engine evaluation, 186–187
- Porter2 English Stemmer, 66–67
- porter2_stemmer filter, 62
- POS (part-of-speech) tags
 - META toolkit, 67
 - NLP, 47
 - sentiment classification, 395
- Positive feelings, 390–394
- Posterior distribution, 28
- Posterior probability in Bayesian parameter estimation, 29
- Postings files for inverted indexes, 150–152
- Power iteration for PageRank, 205
- Practitioners reader category, 17
- Pragmatic analysis in NLP, 39–40
- Precision
 - search engine evaluation, 184
 - set retrieval evaluation, 170–178
- Precision-recall curves in ranked lists
 - evaluation, 174–176
- Predictive analysis for non-text data, 249
- Predictors features in joint analysis, 413–416
- Presupposition in NLP, 41
- Prior probability in Bayesian parameter estimation, 29
- Probabilistic inference, 88
- Probabilistic latent semantic analysis (PLSA)
 - CPLSA, 419–428
 - extension, 377–383
 - overview, 368–377
- Probabilistic retrieval models
 - description, 87–88
 - overview, 110–112
 - query likelihood retrieval model, 114–118
- Probability and statistics
 - abstractive summarization, 322
 - background language model, 346–349
 - basics, 21–23
 - Bayes' rule, 25–26
 - Bayesian parameter estimation, 28–30
 - binomial distribution, 26–27
 - EM algorithm, 362–366
 - joint and conditional probabilities, 23–25
 - KL-divergence, 474
 - LARA, 403
 - maximum likelihood parameter estimation, 27–28
 - mixture model behavior, 354–358
 - mutual information, 266–270
 - Naive Bayes algorithm, 310
 - PageRank, 202–206
 - paradigmatic relations, 257–258
 - PLSA, 368–377, 380
 - probabilistic models and applications, 30–31
 - syntagmatic relations, 262–263
 - term clustering, 286–289
 - topics, 336–339
 - unigram language model, 342–344
 - web search ranking, 209–211
- Probability density function (pdf)
 - Beta distribution, 457
 - Dirichlet distribution, 461
 - multinomial distribution, 461
- Probability distributions
 - overview, 21–23
 - statistical language models, 50–54
- Probability ranking principle, 84
- Probability space, 21–23
- Producer-initiated recommendations, 75
- Product reviews in opinion mining, 391–392
- profile command, 65–66
- Properties
 - inferring knowledge about, 248
 - text categorization for, 300
- Proximity heuristics for inverted indexes, 151
- Pseudo counts
 - Bayesian statistics, 459–460
 - LDA, 381
 - multinomial distribution, 463
 - PLSA, 379, 381
 - smoothing techniques, 128, 286
- Pseudo data in LDA, 378

- Pseudo feedback, 133, 135, 142, 157–158
- Pseudo-segments for mutual information, 269–270
- Pull access mode, 8–9, 73–76
- Push access mode, 8–9, 73–76
- Python language
 - cleaning HTML files, 218
 - crawlers, 217
- Q-function, 465, 469–471
- Queries
 - multimode interactive access, 77
 - navigational, 200
 - text information systems, 9
 - text retrieval vs. database retrieval, 80
- Query expansion
 - vector space model, 135
 - word associations, 252
- Query likelihood retrieval model, 90, 113
 - document language model, 118–123
 - feedback, 139
 - KL-divergence, 475–476
 - overview, 114–118
 - smoothing methods, 123–128
- Query vectors, 92–98, 135–137
- Random access decoding in compression, 158
- Random numbers in abstractive summarization, 322
- Random observations in search engine evaluation, 186
- Random surfers in PageRank, 202–204
- Random variables
 - Bayesian parameter estimation, 28
 - dependent, 25
 - entropy of, 158, 262–263, 270
 - information theory, 31–34
 - PMI, 287
 - probabilistic retrieval models, 87, 111, 113
 - probability distributions, 22
- Ranked lists evaluation
 - multiple queries, 178–180
 - overview, 174–178
- Rankers for search engines, 147
- Ranking
 - extractive summarization, 320
 - probabilistic retrieval model. *See* Probabilistic retrieval models
 - vs. selection, 82–84
 - text analysis operator, 450–451
 - text data access, 78
 - vector space model. *See* Vector space (VS) retrieval models
 - web searches, 201, 208–212
- Ratings
 - collaborative filtering, 230–231
 - LARA, 400–409
 - sentiment classification, 396–399
- Real world properties, inferring knowledge about, 248
- Realization in abstractive summarization, 324
- Recall in set retrieval evaluation, 170–178
- Reciprocal ranks, 179–180
- Recommendations in text information systems, 11
- Recommender systems
 - collaborative filtering, 229–233
 - content-based recommendation, 222–229
 - evaluating, 233–235
 - overview, 221–222
- reduce function, 198
- Redundancy
 - MMR reranking, 333
 - text summarization, 320–321, 324
 - vector space retrieval models, 92
- Regression
 - LARA, 402–405
 - machine learning, 34–35
 - sentiment classification, 394, 396–400
 - text categorization, 303–304
 - web search ranking, 209–211
- Regularizers in network supervised topic models, 429–431
- Relevance and relevance judgments
 - Cranfield evaluation methodology, 168–169

- description, 133
- document ranking, 83
- document selection, 83
- extractive summarization, 321
- probabilistic retrieval models, 110–112
- search engine evaluation, 181–184, 186–187
- set retrieval evaluation, 171–172
- text data access, 79
- vector space model, 92
- web search ranking, 209–211
- Relevant text data, 5–6
- Relevant word counts in EM algorithm, 364–365, 376–377
- Repeated crawling, 193
- Representative documents in search engine evaluation, 183
- reset command, 57–59
- Retrieval models
 - common form, 88–90
 - overview, 87–88
 - probabilistic. *See* Probabilistic retrieval models
 - vector space. *See* Vector space (VS) retrieval models
- Reviews
 - LARA, 400–409
 - opinion mining, 391–392
 - sentiment classification, 394
 - text summarization, 318
- RMSE (root-mean squared error), 233
- robots.txt file, 193
- Rocchio feedback
 - forward indexes, 157
 - vector space model, 135–138
- Root-mean squared error (RMSE), 233
- Ruby language
 - cleaning html files, 218–219
 - crawlers, 217
- Rule-based text categorization, 301
- Scalability in web searches, 191–192
- Scanning inverted indexes, 152
- Scientific research, text data analysis for, 243
- Scikit Learn toolkit, 64
- score_term function, 154
- Scorers
 - document-at-a-time ranking, 155
 - filtering documents, 155–156
 - index sharding, 156–157
 - search engines, 147, 153–157
 - term-at-a-time ranking, 154–155
- Scoring functions
 - KL-divergence, 474
 - topic analysis, 332
- SDI (selective dissemination of information), 75
- Search engine evaluation
 - Cranfield evaluation methodology, 168–170
 - measurements, 168
 - multi-level judgments, 180–183
 - practical issues, 183–186
 - purpose, 167–168
 - ranked lists, 174–180
 - set retrieval, 170–173
- Search engine implementation
 - caching, 162–165
 - compression, 158–162
 - feedback implementation, 157–158
 - indexes, 150–153
 - overview, 147–148
 - scorers, 153–157
 - tokenizers, 148–150
- Search engine queries
 - pull access mode, 74–75
 - text data access, 78–79
- Search engine toolkits, 64
- Searches
 - text information systems, 11
 - web. *See* Web searches
- Segmentation in LARA, 405
- Select operator, 449–451, 455
- Selection
 - vs. ranking, 82–84
 - text data access, 78
- Selection-based text summarization, 318
- Selective dissemination of information (SDI), 75

- Semantic analysis in NLP, 39–40, 43, 47
- Semantically related terms in clustering, 187, 285–287
- Sensors
 - humans as, 244–246
 - joint analysis, 413–415
 - opinion mining. *See* Opinion mining
- Sentence vectors in extractive summarization, 319
- Sentiment analysis, 389
 - classification, 393–396
 - evaluation, 409–410
 - NLP, 43
 - ordinal regression, 396–400
 - text categorization, 304
- Separation in text clustering, 294–295
- Sequences of words in NLP, 46–47
- Set retrieval evaluation
 - description, 170
 - F measure, 172–173
 - precision and recall, 170–173
- Shadow analysis in NLP, 48
- Shallow analysis in NLP, 43–45
- Short-range walks in multimode interactive access, 77
- Short-term needs in pull access mode, 75
- Sign tests in search engine evaluation, 185
- Signed-rank tests in search engine evaluation, 185
- Significance tests in search engine evaluation, 183–186
- Similarity algorithm for clustering, 276
- Similarity in clustering
- Similarity functions and measures
 - extractive summarization, 319, 321
 - paradigmatic relations, 256–259
 - vector space model, 92, 109
 - description, 277
 - document clustering, 279–281
 - term clustering, 285
- Single-link document clustering, 281–282
- Skip-gram neural language model, 292–293
- sLDA (supervised LDA), 387
- Smoothing techniques
 - Add-1, 130, 464
 - Bayesian statistics, 459–460
 - KL-divergence, 474–475
 - maximum likelihood estimation, 119–128
 - multinomial distribution, 463–464
 - Naive Bayes algorithm, 310
 - unigram language models, 53
- Social media in text data analysis, 243
- Social networks as context, 428–433
- Social science research, opinion mining for, 393
- Soft rules in text categorization, 301
- Spam in web searches, 191–192
- Sparse Beta, 459
- Sparse data in Naive Bayes algorithm, 309–311
- Sparse priors in Dirichlet distribution, 461–462
- Spatiotemporal patterns in CPLSA, 425–426
- Specificity in sentiment classification, 396
- Speech acts in NLP, 47–48
- Speech recognition
 - applications, 42
 - statistical language models, 51
- Spiders for web searches, 192–194
- Split counts in EM algorithm, 374–375
- Split operator for text analysis, 449–452, 455
- Stanford NLP toolkit, 64
- State-of-the-art support vector machines (SVM) classifiers, 311–312
- Statistical language models
 - NLP, 45
 - overview, 50–54
- Statistical machine learning
 - NLP, 42–43
 - text information systems, 10
- Statistical significance tests in search engine evaluation, 183–186
- Statistics. *See* Probability and statistics
- Stemmed words in vector space model, 109
- Stemming process in META toolkit, 66–67
- Sticky phrases in Brown clustering, 291
- Stop word removal
 - feedback, 141
 - frequency analysis, 69

- META toolkit, 62, 66
 - mixture models, 352
 - vector space model, 99, 109
- Story understanding, 42
- Structured data
 - databases, 80
 - joint analysis with text. *See* Joint analysis of text and structured data
- Student reader category, 16
- Stylistic analysis in NLP, 49
- Subjective sensors
 - humans as, 244–246
 - opinion mining. *See* Opinion mining
- Subjective statements vs. objective, 389–390
- Sublinear transformation
 - term frequency, 258–259
 - vector space model, 103–104
- Summarization. *See* Text summarization
- Supervised LDA (sLDA), 387
- Supervised machine learning, 34
- SVM (state-of-the-art support vector machines) classifiers, 311–312
- Symbolic approach in NLP, 42
- Symmetric Beta, 459
- Symmetric probabilities in information theory, 32
- Symmetry in document clustering, 279–280
- Synonyms
 - vector space model, 92
 - word association, 252
- Syntactic ambiguity in NLP, 41
- Syntactic analysis in NLP, 39–40, 47
- Syntactic structures in NLP, 49
- SyntacticDiff method, 306
- Syntagmatic relations, 251–252
 - Brown clustering, 290–291
 - discovering, 253–254, 260–264
 - mutual information, 264–271
- System architecture in unified systems, 452–453
- Tags, POS
 - META toolkit, 67
 - NLP, 47
 - sentiment classification, 395
- Targets in opinion mining, 390–392
- Temporal trends in CPLSA, 424–425
- Term-at-a-time ranking, 154–155
- Term clustering, 278
 - n*-gram class language models, 288–291
 - neural language model, 291–294
 - overview, 284–285
 - Pointwise Mutual Information, 287–288
 - semantically related terms, 285–287
- Term frequency (TF)
 - bag-of-words representation, 89
 - vector space model, 97–98
- Term IDs
 - inverted indexes, 151–152
 - tokenizers, 149–150
- Term vectors, 92
- Terms, topics as, 332–335
- Terrier search engine toolkit, 64
- Test collections in Cranfield evaluation
 - methodology, 168–169
- Testing data
 - machine learning, 35
 - text categorization, 303
- Text
 - joint analysis with structured data. *See* Joint analysis of text and structured data
 - mining. *See* Mining; Mining topics from text
 - usefulness, 3–4
- Text annotation. *See* Text categorization
- Text-based prediction, 300
- Text categorization
 - classification algorithms overview, 307
 - evaluation, 313–315
 - features, 304–307
 - introduction, 299–301
 - k*-nearest neighbors algorithm, 307–309
 - linear classifiers, 311–313
 - machine learning, 35
 - methods, 300–302
 - Naive Bayes, 309–311
 - problem, 302–304
- Text clustering, 12
 - document, 279–284

- Text clustering (*continued*)
 - evaluation, 294–296
 - overview, 275–276
 - techniques, 277–279
 - term, 284–294
- Text data access, 73
 - access modes, 73–76
 - document selection vs. document ranking, 82–84
 - multimode interactive, 76–78
 - text retrieval vs. database retrieval, 80–82
 - text retrieval overview, 78–80
- Text data analysis overview, 241–242
 - applications, 242–244
 - humans as subjective sensors, 244–246
 - operators, 448–452
 - text information systems, 8
 - text mining tasks, 246–250
- Text data understanding. *See* Natural language processing (NLP)
- Text information systems (TISs)
 - conceptual framework, 10–13
 - functions, 7–10
 - NLP, 43–45
- Text management and analysis in unified systems. *See* Unified systems
- Text organization in text information systems, 8
- Text representation in NLP, 46–50
- Text retrieval (TR)
 - vs. database retrieval, 80–82
 - demand for, 4–5
 - overview, 78–80
- Text summarization, 12
 - abstractive, 321–324
 - applications, 325–326
 - evaluation, 324–325
 - extractive, 319–321
 - overview, 317–318
 - techniques, 318
- TextObject data type operators, 449, 454
- TextObjectSequence data type operators, 449, 454
- TF (term frequency)
 - bag-of-words representation, 89
 - vector space model, 97–98
- TF-IDF weighting
 - Dirichlet prior smoothing, 128
 - probabilistic retrieval model, 122–123
 - topic analysis, 333
 - vector space model, 100–103
- TF transformation, 102–105
- TF weighting, 125–126
- Themes in CPLSA, 420–422
- Therapist application, 44–45
- Thesaurus discovery in NLP, 49
- Threshold settings in content-based filtering, 222, 224–227
- Tight clusters, 281
- Time series context in topic analysis, 433–439
- TISs (text information systems)
 - conceptual framework, 10–13
 - functions, 7–10
 - NLP, 43–45
- Tokenization
 - META toolkit, 61–64, 453
 - search engines, 147–150
- Topic analysis
 - evaluation, 383–384
 - LDA, 377–383
 - mining topics from text. *See* Mining topics from text
 - model summary, 384–385
 - overview, 329–331
 - PLSA, 368–377
 - social networks as context, 428–433
 - text information systems, 12
 - time series context, 433–439
 - topics as terms, 332–335
 - topics as word distributions, 335–340
- Topic coherence in time series context, 436
- Topic coverage
 - CPLSA, 420–422, 425–426
 - LDA, 380–381
- Topic maps in multimode interactive access, 76–77
- TopicExtraction operator, 450
- TR (text retrieval)
 - vs. database retrieval, 80–82

- demand for, 4–5
 - overview, 78–80
- Training and training data
 - classification algorithms, 307–309
 - collaborative filtering, 229–230
 - content-based recommendation, 227–228
 - linear classifiers, 311–313
 - machine learning, 34–36
 - Naive Bayes, 309–310
 - NLP, 42–43, 45
 - ordinal regression, 398–399
 - text categorization, 299–303, 311–314
 - web search ranking, 209–210, 212
- Transformations
 - frequency, 258–259
 - vector space model, 103–104
- Transition matrices in PageRank, 204
- Translation, machine, 42, 44–45
- TREC filtering tasks, 228
- tree_analyzer analyzer, 62
- Trends in web searches, 215–216
- Trigrams in frequency analysis, 69
- Twitter searches, 83
- Two-component mixture model, 356
- Unary bitwise compression, 159–160
- Under-constrained queries, 84
- Unified systems
 - MeTA as, 453–455
 - overview, 445–448
 - system architecture, 452–453
 - text analysis operators, 448–452
- Uniform priors in Dirichlet distribution, 461
- Unigram language models, 51–54
 - EM algorithm, 466
 - LDA, 381
 - mining topics from text, 341–345
 - PLSA, 370
- Unigrams
 - abstractive summarization, 321–323
 - frequency analysis, 68
 - sentiment classification, 394
 - words tokenizers, 149
- Unimodel Beta, 459
- Union operator, 449–450
- University of California Berkeley study, 3
- Unseen words
 - document language model, 119–120, 122–123
 - KL-divergence, 474
 - Naive Bayes algorithm, 310–311
 - smoothing, 124, 285–287
 - statistical language models, 52
- Unstructured text access, 80
- Unsupervised clustering algorithms, 275, 278
- Unsupervised machine learning, 34, 36
- URLs and crawlers, 193
- Usability in search engine evaluation, 168
- Utility
 - content-based filtering, 224–228
 - text clustering, 294
- Valence scoring, 411
- Variable byte encoding, 161
- Variables
 - context, 330
 - contextual text mining, 419
 - CPLSA, 422
 - EM algorithm, 362–364, 366, 368, 373–376, 465, 467
 - LARA, 403
 - random. *See* Random variables
- vByte encoding, 161
- Vector space (VS) retrieval models, 87
 - bit vector representation, 94–97
 - content-based filtering, 225–226
 - document length normalization, 105–108
 - feedback, 135–138
 - improved instantiation, 97–102
 - improvement possibilities, 108–110
 - instantiation, 93–95
 - overview, 90–92
 - paradigmatic relations, 256–258
 - summary, 110
 - TF transformation, 102–105
- Vectors
 - collaborative filtering, 222

- Vectors (*continued*)
 - neural language model, 292
- Versions, META toolkit, 59
- Vertical search engines, 212
- Video data mining, 245
- Views
 - CPLSA, 420–422
 - multimode interactive access, 77
- Visualization in text information systems, 12–13
- VS retrieval models. *See* Vector space (VS) retrieval models
- Web searches
 - crawlers, 192–194
 - future of, 212–216
 - indexing, 194–200
 - link analysis, 200–208
 - overview, 191–192
 - ranking, 208–212
- Weighted k -nearest neighbors algorithm, 309
- WeightedTextObjectSequence data type, 449
- WeightedTextObjectSet data type, 449
- Weights
 - collaborative filtering, 231
 - Dirichlet prior smoothing, 127–128
 - document clustering, 279–280
 - LARA, 401–409
 - linear classifiers, 313
 - mutual information, 269–270
 - NetPLSA model, 430
 - network supervised topic models, 431
 - paradigmatic relations, 258–261
 - query likelihood retrieval model, 121–123
 - text categorization rules, 301
 - topics, 333, 335–336
 - vector space model, 92, 99–103
- Weka toolkit, 64
- whitespace_tokenizer command, 149
- Whitespace tokenizers, 149
- Wilcoxon signed-rank test, 185
- Word association mining
 - evaluation, 271–273
 - general idea of, 252–254
 - overview, 251–252
 - paradigmatic relations discovery, 254–260
 - syntagmatic relations discovery, 260–271
- Word counts
 - EM algorithm, 364–365, 376–377
 - MapReduce, 195–198
 - vector space model, 103–104
- Word distributions
 - CPLSA, 424–425
 - LARA, 405
 - topics as, 335–340
- Word embedding in term clustering, 291–294
- Word-level ambiguity in NLP, 41
- Word relations, 251–252
- Word segmentation in NLP, 46
- Word sense disambiguation in NLP, 43
- Word valence scoring, 411
- Word vectors in text clustering, 278
- word2vec skip-gram, 293
- WordNet ontology, 294
- Zhai, ChengXiang, biography, 489
- Zipf's law
 - caching, 163
 - frequency analysis, 69–70