

## Numerics

---

- 10-fold cross-validation 221
- 2 distribution 255–256
- 2 statistic 251, 274–275

## A

---

- abstract syntax tree. *See* AST
- Accuracy 221
- accuracy differences 251
- activation rule 203
- AdaBoost 234, 265, 267, 276
- adaptive resampling 265, 276
- adjacency
  - list 62
  - matrix 134
- adjustable mortgages 234
- adjustment
  - primitivity 36
  - stochasticity 36
- Aggarwal, C.C. 160
- agglomerative hierarchical algorithms 129
- aggregated content 5
- aggregating classifiers 263
- AI 16
  - utility problem 190
- Aitken extrapolation 63
- algorithms 5
  - agglomerative hierarchical 129
  - applicability limitations 18
  - application context 278
  - arc-x4 269
  - average-length 137
  - average-link 138
  - BIRCH 131
  - Borvka\_fs 141
  - classification 164
  - clustering 302
  - computational time estimation 19
  - constrained clustering 130
  - density-based 151
  - divisive hierarchical 129
  - Expectation-Maximization 161
  - gradient-descent learning 218
  - graph theoretic 139
  - KISS 19
  - k-means 142
  - Kruskal\_fs 141
  - link-based 132
  - NaiveBayes 172
  - nearest neighbor 129
  - parallelization 18
  - partitional 129, 142
  - ranking 286
  - regression 172
  - Rete 192
  - ROCK 147, 300
  - scalability 18
  - single-link 135
  - SQLEM 161
  - structural 171
- alpha 62
- Amazon.com 108–109
  - among first with recommendations 3
  - item to item approach 92
- analysis
  - link 22
  - user click 22, 33
- analyzer
  - lexical 31
- analyzing 30, 282
- and 227
- Android 287
- Anscombe’s quartet 112
- ANTLR 193
- Apache Axis 15
- Apache CXF
  - supports numerous standards 15
- Apache POI project 30, 283
- Apache Shindig 9
- application intelligence 279
- arc-fs 276
- arcing 265
- arc-x4 234, 265, 267, 276
- arc-x4 algorithms 269
  - crux 271
  - fourth power 271
- Armstrong, Lance 23, 175
- arrays
  - sorting 125
- Arthur, David 145
- artificial data 241
  - generation 239
- artificial intelligence. *See* AI
- assignTopicToCluster 297–298
- assignTopicToStory 297, 299
- assumption of independence 254
- AST 193

- Atom 13
    - not RFD-based 14
    - syndication format 14
  - Attribute 175
  - attribute selection 273
    - news portal 313
  - automatic categorization 174
  - average distance 137
  - average-link algorithm 137–138
  - AverageLinkAlgorithm 134
- B**
- 
- back propagation algorithms
    - termination conditions 218
  - backward chaining 189
  - BadUserType 239
  - bagging 257, 263
    - algorithm 258
    - data diversity 258
    - independent classifiers 265
    - premise of 258
    - tweaks 260
    - tweaks and tips 276
  - BaggingCreditClassifier
    - 258–260
    - accuracy and execution time 259
  - bankruptcy
    - significance 235
  - base recommenders
    - enhancement 107
  - BaseConcept 297
  - BaseInstance 178, 297
  - BaseLayer 216
  - BaseNN 208, 213–214
  - basic search
    - load, index, search 24
  - BasicWebCrawler 27, 282
  - Bayes theorem 49, 172, 182, 227
    - attribute independence
      - assumption 183
    - conditional probability 182
    - evidence 182
    - likelihood 182
    - naïve assumption 51
    - posterior probability 182
    - prior probability 182
  - Bayes theorem formula
    - output 184
  - Bayes theorem input probability
    - evidence 48
    - likelihood 48
    - prior probability 48
  - Bayes theorem output
    - probability
    - posterior probability 49
  - Bayesian
    - belief networks 182
    - networks 172
    - neural networks 330
  - BeanShell 317
  - Bernoulli process 221
  - Beyer, Kevin 160
  - bias vs. generalization 19
  - binary classification 174, 178
  - binary information
    - difficulty in processing 9
  - BIRCH
    - algorithm 131
    - clustering algorithm 158
  - black box trap 205
  - Black-Sholes
    - option pricing model 331
  - Boltzmann machines 330
  - Boorah 13
  - BoostCreditClassifier 267
  - boosting 234, 265
    - computational
      - performance 267
      - main idea 265
      - strategy 67
  - BoostingARCX4Classifier
    - 268–269
  - BoostingCreditClassifier 268
  - bootstrap 222, 257
    - aggregating 234, 257
    - process 258
  - BootstrapTrainingSetBuilder
    - 261, 276
  - Borvka\_fs algorithm 141
  - Borvka, Otaker 161
  - Bradley, Paul S. 131
  - Breiman, Leo 265
  - Brin, Sergey 33
  - BSD
    - license 30
  - building intelligence 12
  - business news 284
- C**
- 
- C5.0 171
  - calculation of similarity 84
  - car ownership 236
  - categorical data 130
  - categories 164
    - internet newsgroups 166
    - newspaper articles 166
    - restaurant menu 166
  - categorization
    - automatic 174
    - email 187
  - Celtix by IONA. *See* Apache CXF
  - centroid 142
    - initial selection 145
    - role 143
  - Cereghini, Paul 161
  - CF 80
    - item based 89
    - requirements 80
  - CFOI 166
  - chain effect 142
  - Chebyshev polynomial 63
  - Chi2 251
  - chi-square 251, 274
  - chronological age 235
  - Cinematch 3, 107
  - city block metric 118
  - CJK
    - tokenizing 31
  - classification 48, 289
    - algorithms 164–165
    - binary 169, 174, 178
    - content noise reduction 293
    - continuous values 169
    - correct 243
    - cost 221
    - cross-referencing 304
    - data noise effects 206
    - discrete values 169
    - distance-based
      - algorithms 169
    - email 228
    - erroneous 243
    - flat class structure 169
    - forecasting 169
    - generalization vs.
      - specialization 178
    - groups of instances 294
    - hierarchical class
      - structure 169
    - majority vote rule 295
    - multiclass 169, 174, 178
    - neural networks 165, 169
    - news categories 294
    - news groups 294
    - order effect 288
    - overview 169
    - performance
      - characteristics 250
    - region of influence 233
    - regression algorithms 169

- classification (*continued*)
    - representative news story 299
    - rule-based 188
    - rule-based algorithms 169
    - runtime 250
    - runtime performance 224
    - specialization vs.
      - generalization 178
    - statistical algorithms 169, 172
    - strategy 288
    - structural algorithms 169–170
    - training time 250
    - utility problem 225
    - wrong decision impact 222
  - classification accuracy
    - statistically insignificant 258
  - classification attributes
    - large number effect 166
  - classification system
    - CFOI 166
    - ICD-10 166
    - Library of Congress 166
    - Linnaean 166
    - OIIC 166
    - Schatzker 166
    - SOII 166
  - classification training
    - attribute value coverage 223
    - representative data 223
    - scaling characteristics 224
    - statistical assessment 224
  - ClassificationStrategy 297, 314
  - ClassificationStrategyImpl
    - 294–295, 297, 299, 314
  - classifier ensemble
    - incremental growth 265
  - classifier training
    - user clicks 48
  - ClassifierEnsemble 260–261, 263, 268, 271
  - ClassifierResults 251
  - classifiers 48
    - aggregating 263
    - combination 232
    - comparison 233
    - decision tree 266
    - ensembles 263
    - fusion 232
    - lifecycle stages 173
    - metaclasser scheme 173
    - pair-wise comparisons 250
    - selection 232, 256, 275
    - sensitivity 207
    - stable 245
    - training stage 173
    - unstable 245, 258
    - validation stage 173
  - classify-cluster approach 293
  - Classify-Cluster-DS 292
  - classifyClusters 299
  - classifyStories 299
  - cleansing news stories 285
  - Clearspring 9
  - Clementine 171
  - CLIPS 170
  - cluster
    - discovery 125
    - formations 140
    - invariance 306
    - structure 128
  - cluster centroid
    - center of mass analogy 143
  - cluster formation
    - goodness 315
  - cluster-classify approach 293
  - Cluster-Classify-DS 292
  - clustering 289
    - by age 127
    - agglomerative hierarchical
      - algorithms 129
    - algorithm 302
    - arbitrary objects 128
    - array sorting 125
    - average distance 137
    - average-link algorithm 137
    - BIRCH algorithm 131
    - book example 122
    - by cluster structure 129
    - categorical data 130
    - categorization 128
    - centroid 142
    - computational
      - complexity 157
    - conceptual modeling 129
    - constrained algorithms 130
    - curse of dimensionality 159
    - data normalization 127
      - by data size 131
    - data squashing 158
      - by data structure 130
    - by data type 129–130
    - DBSCAN 151
    - dendrograms 132
    - density-based algorithms 151
    - divisive hierarchical
      - algorithms 129
    - epsilon neighborhood 154
    - Euclidian distance 127
    - fine tuning 316
    - goodness measure 150
  - hierarchical 316
  - hierarchical algorithms 129
  - high dimensionality 158
  - human expert 127
  - in high dimensions 157
  - iterative optimization 129
  - k-means algorithm 129, 142
  - lack of normalization 134
  - large databases 131
  - link-based algorithms 134
  - many dimensions 128
  - mean value 142
  - MST 139
  - news articles 129
  - objective 150
    - and ordering 124
  - overview 128
  - partitional algorithms 129
  - performance
    - characteristics 157
  - point density 151
  - proximity threshold 135
  - ROCK 147
  - R-trees 158
  - single-link algorithm 135
  - singletons 140
  - Sourceforge-like case
    - study 123
  - spectral methods 130
  - SQL limitations 125
  - SQLEM algorithm 125
  - threshold parameter 129
  - very large datasets 157
  - visual identification 124
  - VLDB 131
  - wavelet methods 130
    - with SQL 124
- clustering algorithm
  - BIRCH 158
  - combinations 162
- clustering applications
  - creation of social
    - networks 122
  - like-minded individuals 122
  - targeted advertisement 122
- clustering attribute
  - age 123
  - education level 123
  - income range 123
  - paid participation 123
  - professional skills 123
  - social relationship rating 123
- clustering distribution
  - optimality 315

- Cochran's Q test 233, 250, 255–256
  - Codehaus XFire. *See* Apache CXF
  - collaboration
    - as opposed to intelligence 4
  - collaborative filtering. *See* CF
  - collaborative platforms 5
  - collective intelligence 4
  - collective knowledge
    - capture 4
  - combination of classifiers 232
    - computational
      - robustness 233
    - representational
      - advantage 233
    - risk reduction 233
  - combining classifiers
    - bagging 234
    - boosting 234
  - comparator 126
  - complexity
    - multiclass classification 224
  - computational
    - cluster 63
    - complexity 157
    - cost 316
    - linguistics 327
    - nodes 63
  - Concept 175, 182
  - CONCEPT\_LABEL\_FRAUD 211
  - CONCEPT\_LABEL\_VALID 211
  - ConceptMajorityVoter 264
  - conceptPriors 50
    - map 183
  - conditional probabilities
    - 50, 183
    - user clicks 48
  - confidence interval 221
  - conflict resolution 196–198
  - confusion matrix 220, 243, 259, 274
  - constrained clustering
    - algorithms 130
  - content 13
    - aggregator 6, 9
    - annotation 9
    - cleansing 283
    - field 27, 31
    - impurities 283
    - reconciliation 7
  - content aggregation
    - digg.com 99
  - content similarity
    - case study 93
    - normalization 103
    - text analysis sensitivity 96
  - content-based
    - accumulation and analysis 80
    - recommendation 70
  - Corcho, Oscar 165
  - correlation
    - complete negative 111
    - complete positive 111
  - cosine similarity 95, 324
  - CosineDistance 152
  - CosineSimilarity 149, 187
  - CosineSimilarityMeasure 95
  - cost
    - function 223, 230
    - matrix 230
  - craigslist 13
  - crawler 13
    - collecting data 22
    - custom 281
    - fetching documents 24
    - known URLs 24
    - page links 24
    - processed documents 24
  - crawling 23, 30, 281–282
    - Apache Tika 321
    - custom web crawler 320
    - depth of 13
    - Heritrix 321
    - Nutch 321
    - retrieved content
      - structure 282
    - stages of 320
  - CrawlResultsNewsDataset 284
  - createClusters 300–301
  - createClustersWithinTopics 300, 305
  - credibility of classification 219
  - credit
    - risk 233
    - score 236
  - credit card activity 236
  - credit worthiness
    - attributes 235
    - case study 233
    - overview 234
  - CreditErrorEstimator 244, 266, 274
  - criminal record 236
  - cross product calculation 111
  - cross-referencing 304
  - curse of dimensionality 159, 166
  - CustomAnalyzer 95
  - Cutting, Doug 22
- 
- ## D
- 
- DAG 172, 202
  - damping factor 36
  - DangerousUserType 239
  - dangling node 62
    - heuristic 67
  - data
    - diversity 265
    - incongruent 17
    - missing values 17
    - noisy 259
    - normalization 17, 156, 204
    - preprocessing 204
    - reliability 17
    - renormalization 115
    - representation
      - inaccuracies 17
      - size issues 18
      - squashing 158
      - understanding 207
      - understanding
        - importance 279
      - variability 17
    - data normalization 110
      - PearsonCorrelation 113
  - DataGenerator 240
  - DataPoint 146, 154
  - Datapoint 134
  - dataset
    - dimensionality 156
  - DatasetAdapter 311
  - DBSCAN 151
    - algorithm 162
    - border point 154
    - core point 154
    - directly density reachable 154
    - eps variable 154
    - ink drops analogy 151
    - minPoints variable 154
  - DBSCANAlgorithm 152–154
  - decision tree 170, 245, 258, 273
    - accuracy 246
    - algorithms 171
    - classifier 234, 266–267
  - decision tree classification
    - instability 247
    - interpretation 247
  - decisionTree
    - printTree 247
  - declarative programming 188
  - default analyzer 31

- degree
    - of belief 81
    - of credibility 223
    - of freedom 251, 255–256
  - Delphi 310–311
  - Dataset interface 81
  - inner workings 86
  - recommend 87
  - recommendation engine 80–81
  - similarity between users 82
  - DelphiUC 103
  - delphiUC 103
  - DelphiUR 103
  - Dendrogram 132
  - dendrogram
    - data structure 132–133
    - initialized 138
    - two linked hash maps 132
    - visual representation of 132
  - density-based
    - algorithms 151
    - spatial clustering of applications with noise. *See* DBSCAN
  - dez 165
  - Dhillon, Inderjit S. 145
  - diagnosis
    - of diseases 166
    - of injuries 166
  - Diff2PropTest 253
  - difference of proportions
    - test 233, 250, 253
  - Digg
    - API 99, 146
    - RESTful services 14
  - Digg stories
    - blood donors 146
    - CSV file 146
    - Facebook 146
  - DiggCategory 100
  - DiggDelphi
    - findSimilarUsers 103
    - getTopNFriends 103
    - inner workings 102
    - recommend 103–104
  - dimensionality
    - curse of 157
  - directed acyclic graphs. *See* DAG
  - directed graph 34
  - discourse 288, 328
  - Distance 154
  - distance
    - city block 324
    - Euclidean 324
    - L2 324
    - properties 73
    - symmetry 74
    - taxi cab 324
    - triangle inequality 74
  - distributed computing
    - fallacies 17–19
  - distribution of clusters 305
  - divisive hierarchical
    - algorithms 129
  - docid field 27
  - DocRank 55–56, 280, 286
    - inner workings 57
    - matrix builder 57
    - relational tables 61
    - values reused 61
  - doctype field 27
  - document
    - distance 92
    - heuristic importance 59
    - terms 286, 288
  - document collection
    - business news 23
    - Lance Armstrong 23
    - U.S. politics 23
    - world news 23
  - domain of discourse 5
  - dot (inner) product 96
  - Drools 165, 170, 189, 193
    - ReteOO 190
  - Drools attribute
    - no-loop 197
    - ruleflow-group 197
    - salience 197
  - DTCreditClassifier 246, 259, 266
  - Dunham, Margaret 158
- E**
- 
- ECLiPSe 189
  - Eisner, Jason 161
  - elements of intelligence
    - synergy 100
  - EM algorithm
    - E-step 161
    - M-step 161
  - email categorization
    - 174, 178, 187
  - email classification
    - blacklists 175
    - header tests 175
    - idiosyncracies 175
    - real-time blackhole lists 175
    - whitelists 175
  - email concept
    - NOT SPAM 178
    - SPAM 178
  - email content
    - congressional elections 175
    - global warming 175
    - Lance Armstrong 175
    - marathon 175
    - newspaper advertisement 175
    - Nicaragua elections 175
    - NVIDIA stock 175
    - Ortega 175
    - spam 175
    - U.S. politics 175
    - world news 175
  - email messages
    - sorting 174
  - EmailClassifier 175–176, 178, 184
  - EmailData 176
  - EmailDataset 176
    - getTrainingSet 178
    - setBinary(false) 187
  - EmailInstance 178
  - EmailRuleClassifier 192
  - embedding intelligence 11
  - Engage 9
  - ensembles
    - accuracy 260
    - of classifiers 263
  - Epictetus 232
  - epsilon neighborhood 154
    - selecting value 156
  - error
    - type I 220
    - type II 220
  - ESPN 314
  - Ester, Martin 151
  - estimateUserBasedRating 88
  - Euclidean distance 77, 127, 130, 145, 160, 324
  - EuclideanDistance 127
  - evaluation
    - 10-fold cross-validation 221
  - evaluation for
    - recommendations 116
  - ExcellentUserType 239
  - Expectation-Maximization
    - algorithm 161
- F**
- 
- F distribution 256
  - F statistic 256
  - F test 233, 250, 255

Facebook 2, 6  
 RESTful API 14  
 fact checking 2  
 fallacies  
 intelligent applications 17  
 Fan, James 145  
 FASTCLUS 145  
 Fawcett, Tom 222  
 Fayyad, Usama M. 131  
 feed formats  
 Atom 13  
 RSS 13  
 FetchAndProcessCrawler  
 23, 282  
 addUrl 27  
 purpose of class 24  
 fetched 282  
 field content  
 indexed 27  
 stored 28  
 unstored 27  
 Fielding, Roy T. 14  
 FileListNewsDataset  
 285, 292, 297  
 filesystem analogy  
 Attributes 48  
 Concepts 48  
 Instances 48  
 financial  
 assets 238  
 turbulence 234  
 findSimilarUsers 81  
 fine tuning clustering 316  
 Fisher, Ronald A. 255  
 Fisher-Snedecor  
 distribution 256  
 flat reference structures 167  
 floating-point arithmetic 119  
 FN rate 230  
 folksonomy 5  
 FoodieBytes 13  
 forecasting  
 example 169  
 foreclosures 234  
 Forgy, Charles 190  
 Forgy, E.W. 144  
 forward chaining 170, 189  
 FP rate 221, 230  
 fraud  
 benefit application forms 199  
 detection 229  
 internet auction 199  
 purchasing transactions 199  
 telecommunications 199

TenUsersSample 200  
 transactional data 200  
 fraud detection 199  
 biases 214  
 hidden layer 214  
 use case 199  
 FraudErrorEstimator 203, 205  
 fraudulent transactions  
 identify 204  
 frequency  
 of occurrence 59  
 of terms 95  
 Friedman, Jerome 172  
 Friendster 9  
 F-score 221  
 FTest 256  
 functional analysis 325

---

**G**


---

Gabow, Harold 161  
 Galil, Zvi 161  
 games  
 online 10  
 garbage collection 61  
 GATE 328  
 gating network 275  
 Gaussian  
 distribution 200, 262  
 processes 182, 330  
 generalization 227  
 generated-test-txns.txt 201  
 geometries  
 flat vs. curved 79  
 getNoisyType 241  
 goodness measure 150  
 GoodUserType 239  
 Google 33  
 began it all 2  
 Finance 3  
 maps 13  
 matrix 35  
 News 3, 8, 279–280, 286  
 Google PageRank. *See* PageRank  
 Gospodnetić, Otis 22  
 gradient-descent learning  
 algorithm 218  
 grammar-based tokenizer 31  
 graph  
 directed 34  
 theoretic algorithms 139  
 group classification  
 representative news story 294  
 grouping  
 discrepancy 292

GroupLens 107  
 groups of news groups 304  
 Guan, Yuqiang 145  
 Guha, Ramanathan V. 161

---

**H**


---

Hadoop 63  
 Hadoop distributed filesystem.  
*See* HDFS  
 Hastie, Trevor 172  
 Hatcher, Erik 22  
 HDFS 63  
 health news 284  
 Hebbian learning 229  
 hexagon 143  
 hi5 9  
 hierarchical  
 agglomerative algorithm 130  
 clustering 129  
 reference structures 167  
 hierarchical clustering 129  
 news stories 316  
 high dimensionality  
 specifics 158  
 high-dimensional clustering 157  
 HITS 34  
 Hits object 29  
 hm 141  
 home equity  
 lines of credit 236  
 HousingMaps 12  
 HTML 30, 283  
 parser 30  
 hyperbolic tangent function  
 59, 78  
 Hypertext Induced Topic  
 Search. *See* HITS  
 Hyves 9

---

**I**


---

IBM DeveloperWorks 7  
 IETF 14  
 if-then clauses 170  
 imeem 9  
 income 237  
 indexDocument 27  
 indexing 22, 30, 282  
 searching beyond 32  
 stage 31  
 IndexSearcher 29  
 IndexWriter 27

- inference 2
    - response time limits 18
  - information content 227
  - information retrieval. *See* IR
  - inner product 324
    - applet 324
  - Instance 175, 182
  - intelligence
    - building 12
    - collective 4
    - embedding 11
    - milestone 33
    - as opposed to collaboration 4
    - triangle of 5
  - intelligent
    - combination potential 304
    - crawling 312
    - document system 4
    - sanity check note 81
    - searching 288
  - intelligent application
    - elements 5
    - fallacies 17
    - prerequisites 11
  - internet
    - behavioral characteristic 33
    - structural characteristic 33
  - Internet Engineering Task Force. *See* IETF
  - intrapoint distance statistics 156
  - IR 21, 64
    - traditional steps 22
  - Item 71
  - item similarity 89
    - efficiency 97
    - large size data 92
  - ItemBasedSimilarity
    - calculate 90
  - Items 308
  - ithm 161
- J**
- 
- Jaccard
    - coefficient 130, 146, 186
    - metric 84, 91
    - similarity 79, 117, 149
    - similarity measure 130
  - JaccardCoefficient 149, 187, 213, 229
  - jaccardThreshold 187
  - JANINO 194
  - Janino
    - embedded Java compiler 194
  - JAR 81
  - Java
    - embedded compiler
      - Janino 194
  - Java Archive. *See* JAR
  - JavaScript 14
  - JavaScript Object Notation. *See* JSON
  - JAX-WS 2.0 15
  - JAX-WSA 15
  - JBoss
    - Drools 165
    - Rules 170, 189
  - Jess 170, 189
  - JFlex 31
  - job classes 237
  - JSON 1, 14
  - JSR-181 15
- K**
- 
- K nearest neighbors. *See* kNN
  - Karger, David R. 161
  - Kendall's tau 112, 119
  - Klein, Philip N. 161
  - Kleinberg, Jon 34
  - k-means algorithm 129, 142, 162, 304
    - centroids 144
    - core algorithm 143
    - pickInitialMeanValues 144
  - k-means clustering 129
  - k-means++ 145
  - kNN 171
  - knowledge representation 165
  - knownurls 282
  - Kruskal\_fs algorithm 141
  - Kullback-Leibler
    - divergences 145
- L**
- 
- L2 norm 324
  - labels
    - top-level 6
  - land property 238
  - language
    - morphology 327
    - syndication-specific 14
    - syntax 327
  - language detection 284
  - Langville, Amy 36
  - large databases
    - algorithm properties 131
  - large-scale cleansing
    - effectiveness 286
  - large-scale crawling
    - efficiency 286
  - large-scale searching
    - computing constraints 61
    - data structures 62
    - PageRank accuracy 62
  - learning rule 203
  - learning vector quantization. *See* LVQ
  - leave-one-out 222
  - legitimate transactions 204
  - level of significance
    - statistical test 251, 253, 255
  - lexical analyzer 31
  - lexicographic ordering 130
  - LGPL 55
  - library
    - call number 166
  - linear
    - correlation coefficient 110
    - regression 171
  - link algorithms
    - comparison 139
    - visualization 135
  - link analysis 3, 34
    - documents 55
  - link-based algorithms 132, 134
  - LinkedIn 9
  - LinkMatrix 149
  - links 201
  - linkThreshold 303
  - Linnaean classification 166
  - Lloyd, S.P. 144
  - loan defaults 234
  - logistic
    - function 172
    - regression 172
  - lookup table
    - problems with approach 16
  - Lp norm 324
  - Lucene 22, 283, 286
    - boosting 31
    - Document class 27
    - Document object 30
    - document score 286
    - Field class 27
    - index files 25
    - Query 32
    - query expression 31
    - QueryParser class 29
    - searching 28
    - searching with 22–32
    - StandardAnalyzer class 31

Lucene analyzers 30  
 non-English languages 31  
 stop words 31  
 synonyms 31  
 text 30  
 Lucene and PageRank  
 combining scores 43–45  
 Lucene Documents  
 removal and update 31  
 variety 31  
 Lucene text  
 analyzers 31  
 Lucene, PageRank, and naïve  
 Bayes  
 combining scores 46, 51  
 LuceneIndexBuilder 25, 66  
 LucenePDFDocument 30, 283  
 LVQ 171

## M

---

machine accuracy 119  
 MapReduce 63  
 Markov chain theory 35  
 mashups  
 aggregated content 7  
 defined 7  
 mathematical formulas 323  
 matrices 323  
 sparse 35  
 matrix  
 adjacency 134  
 confusion 220, 243, 274  
 cost 230  
 similarity 134  
 transition probability 35  
 matrix H  
 basic link contribution 40  
 dangling node  
 contribution 40  
 substochastic version 40  
 symmetric reordering 62  
 teleportation contribution 40  
 Word documents 57  
 maxBatchSize 27  
 McNemar 253  
 test 233, 250, 274  
 McNemar test  
 bagging vs. boosting 267  
 McNemarTest 251  
 mean value. *See* centroid  
 media-sharing sites  
 binary format 9  
 MegaUpload 9  
 mergeClusters 139  
 MergeGoodnessMeasure 149  
 meta-algorithm 279  
 metaclassifier scheme 174  
 metadata  
 web page 24  
 metric 76  
 spaces 325  
 Meyer, Carl 36  
 Microsoft 15  
 Microsoft OLE 2 Compound  
 Document 30  
 Microsoft Word 30  
 97, 2000, XP, and 2003 283  
 documents 55  
 documents parsing 55  
 parser 30  
 Microsoft Word 2000 30  
 Microsoft Word 2003 30  
 Microsoft Word 97 30  
 Microsoft Word XP 30  
 minimum spanning tree. *See*  
 MST  
 mining opportunity 280  
 MinorThird 328  
 misclassification  
 cost 243  
 misclassified  
 news stories 293  
 missing attribute value 184  
 mixture of experts 256, 275  
 module  
 pattern-matching 189  
 mortgage 236  
 application 234  
 down payment 237  
 financing 234  
 mortgage mess  
 United States 234  
 mortgage rates  
 teaser 234  
 motorcycle ownership 237  
 MovieLens 107  
 MovieLens dataset 108  
 large 116  
 RMSE 116  
 small 110  
 MovieLensData  
 createDataset 108  
 MovieLensDelphi 108, 113, 116  
 MovieLensItemSimilarity 113  
 MST 129, 139  
 algorithms 161  
 Borvka\_fs algorithm 141  
 chain effect 142  
 findMinimumEdge 142

Kruskal\_fs algorithm 141  
 randomized algorithm 161  
 MST class 140  
 Edge 141  
 MST link algorithm  
 time complexity 142  
 MSTSingleLinkAlgorithm 134  
 MSWordDocumentParser 55  
 multiclass classification 174, 178  
 complexity 224  
 multidimensional data  
 ordering 126  
 multilingual text 284  
 MusicData 81  
 MusicUser 71, 73  
 getSimilarity 72, 76  
 plot 74  
 MyDiggSpace.com 99  
 case study 99  
 data statistics 100  
 Find friends 100  
 MyDiggSpaceDataset 146  
 MySearcher 28, 31, 43, 51  
 MySpace 2, 6, 9  
 MySQL 124

## N

---

naïve Bayes algorithms 284, 288  
 robustness 51  
 naïve Bayes classifier  
 243, 267, 294  
 naïve similarity 78  
 beta 78  
 NaiveBayes 48, 50, 175–176,  
 179, 227, 244, 294, 297  
 classification 172  
 classifier 46, 48  
 TrainingSet 182  
 natural language elements  
 high-level 288  
 natural language processing. *See*  
 NLP  
 NBCreditClassifier 244, 259,  
 266  
 NBLanguageDetector 284  
 NBStoryClassifier 294–295,  
 297, 299  
 Neapolitan, Richard E. 173  
 nearest neighbor  
 algorithms 129  
 NekoHTML 30, 283  
 NetFlix  
 Cinematch 3

- Netflix 3, 108–109
    - movies selection 107
  - Netflix prize
    - competition 118
    - RMSE 116
  - Netflix.com 92
  - Netscape
    - Rich Site Summary 14
  - network
    - topology 213
  - neural network 169, 171, 234, 258
    - architecture 203
    - BaseNode 217
    - calculateWeightAdjustments 219
    - connectFully 217
    - credit classifier 247
    - design complexity 249
    - disadvantages 172
    - essential elements 202
    - feedback 202
    - feedforward 202
    - fireNeuron 217
    - fireNeuronDerivative 217
    - fully connected 202
    - layers 202
    - learning rate 217
    - learningRate 217
    - LinearNode 216
    - links 217
    - overview 201
    - SigmoidNode 216
    - structure 216
    - training phase 202
    - updateWeights 219
  - neural network classifiers
    - accuracy 248
  - neural networks 330
    - complex valued 331
  - neurons 201
  - news
    - content 280
    - portal 279–280
    - topic 297
  - News Alerts 281
  - news browser
    - create and display 291
    - window 286
  - news categories 297, 305
    - assignment 288
  - news clustering
    - analysis 302
  - news group
    - clustering robustness 293
  - groups of 304
    - juxtaposition 292
  - news stories 279
    - arrangement 290
    - misclassified 293
    - searching 279
  - NewsCategory 297
  - NewsClusterBuilder 302–303
  - NewsCrawler 281–282, 312
  - NewsDataset 284, 292, 301
  - NewsProcessor 286, 295, 299
    - training phase 294
  - NewsStory 297
  - Niemeyer, Pat 317
  - Ning 9
  - NLP 97, 281, 283
  - NNCreditClassifier 248, 259, 266
  - NNFraudClassifier 203–204, 208, 210
  - nodes
    - dangling 36, 62
    - hidden 201
  - noise
    - elements 153
    - levels 240, 242–243
  - noisy data 259
  - nonparametric
    - correlation 112, 119
    - method 257
  - normal distribution 242
  - null hypothesis 250
  - numerical representation 130
- O**
- 
- OASIS 328
  - Object Management Group. *See* OMG
  - OMG 168
  - Octave 331
  - OHC 166
  - OLE 2 283
  - OMG 168
  - online games 10
  - ontology 165, 167
    - analogy with OOD 167
    - attributes 167
    - concepts 167
    - engineering 165
    - example 167
    - instances 167
    - management 165
    - semantic 167
  - OpenSocial
    - premise 8
  - Oracle 9, 15
  - order of operations 289
  - ordering
    - and clustering 124
    - food 2
  - Ordóñez, Carlos 161
  - Orkut 79
  - orkut 9
  - outlinks 24, 35
  - overfitting 178, 227
- P**
- 
- Package 194
  - Page, Larry 34
  - pagelinks 282
  - PageRank 33–45, 280, 286
    - acceleration techniques 63
    - Aitken extrapolation 63
    - alpha 36
    - alpha coefficient 38
    - alpha effect on
      - convergence 38
    - alpha selection 38
    - approximate aggregation
      - technique 63
    - calculation 36
    - convergeness and
      - uniqueness 35
    - damping factor 36
    - dangling nodes 36, 67
    - direct methods (solvers) 38
    - epsilon effects 42
    - hyperlink matrix 34
    - key idea 34
    - power method 34–35, 38
    - primitivity adjustment 36
    - quadratic extrapolation
      - technique 63
    - random surfer 36
    - scaling 67
    - score scaling 45
    - stochasticity adjustment 36
    - teleportation effect 36, 38, 67
    - vector 34
  - PageRankMatrixH 38
  - pair-wise classifier
    - comparisons 250
  - parsing 30, 282
  - partitional algorithm 142
  - pattern recognition 9
  - PDF 30, 283
    - documents 55
    - indexing 30
  - PDFBox 30, 283

- Pearson's r 110
    - counterexample 112
  - Pearson's r calculation
    - singular case 111
  - PearsonCorrelation 113
    - roundoff error 119
  - pecuniary aid 234
  - perceptions 164
  - personalization 46, 280
    - temporal effects 47
    - vector 67
  - phonetics 327
  - phonology 327
  - PhraseQuery 32
    - slope 32
  - Plaxo 9
  - point density 151
  - pointwise deviations 111
  - politics news 284
  - portal feature
    - In the News 280
  - portals
    - aggregated content
      - dispersed 8
  - PorterStemFilter 178
  - posterior probability 184
    - heuristics 50
  - power laws 45
  - power method
    - acceleration 38
    - number of iterations 62
  - pragmatics 288, 328
  - Precision 221
  - precision 64–65
  - PredictedNewsStoryRating 311
  - prediction
    - degree of belief 81
  - PredictWallStreet 3
  - preprocessing stage 23
  - prerequisites
    - for intelligent applications 11
  - prior probability
    - heuristics 51
  - probabilities
    - conditional 50, 182–183
    - posterior 50
    - prior 50, 182–183
  - probability 182
    - distribution 251
    - of linkage 316
  - processed 282
  - processing
    - natural-language 22
  - production rules 170
  - product-moment correlation
    - coefficient 110
  - ProgrammableWeb 7
  - programming
    - declarative 188
    - imperative 188
  - Prolog 189
  - proximity
    - relative 134
    - threshold 134–135
- Q**
- 
- quality assurance 293
  - query
    - context 286
    - “google ads” 47
    - terms 288
  - QueryParser 32
- R**
- 
- Random 242
  - random
    - samples 257
    - surfer 35
  - Rank 40
    - error evaluation 42
  - rank correlation 119
  - ranking algorithms 286
  - RapidShare 9
  - Rating 71
  - rating
    - value range note 81
  - rating storage
    - advantages 86
  - RatingCountMatrix 84, 91
  - Ratings 308
  - ratings
    - artificial bias 104
    - range 71
  - RDF 13
    - Site Summary 13
  - Recall 221
  - recall 64–65
  - recommendation engine 70–71
    - based on CF 82
    - basic concepts 308
    - code optimizations 89
    - combinations 99–100
    - content based item-item 99
    - data normalization 108
    - ensemble methods 118
    - neighbor selection 108
    - score normalization 103
    - similarity 71
    - types 79
    - user based 80
  - recommendation example
    - music song online store 70
    - online music store 80
  - recommendation heuristics 103
  - recommendation ratings
    - value range 118
  - recommendation system
    - news stories 308
    - size issues 115
  - recommendations 3
    - accuracy evaluation 116
    - ancient proverb 80
    - based on content 92
    - cost evaluation 105
    - evaluation for 116
    - large systems 108
    - live update 115
    - quality 115
    - real-time 115
    - response time 110
    - roundoff-error
      - minimization 118
  - RecommendationType
    - ITEM\_PENALTY\_BASED 311
  - recommender combination
    - based on averaging 105
    - based on voting 105
  - reference encoding 62
  - reference structures 5, 165
    - dictionaries 5
    - knowledge bases 5
    - ontologies 5
  - regression algorithms 172
  - Reina, Cory A. 131
  - relative
    - proximity 134
    - ranking 102
  - relevance
    - subjectivity 46
  - relevance score 29, 286
    - combination 68
    - generalization 92
  - repository of knowledge 9
  - representational motivation 233
  - Representational State Transfer.
    - See* REST
  - representations of
    - knowledge 165
  - Resource Description Frame-  
work. *See* RDF
  - REST 14

- Rete 170
    - algorithm 189, 192
  - ReteOO 190
  - retirement 238
  - RFC 4287 14
  - Richardson, Leonard 14
  - RMSE 116
  - RMSEEstimator 116
  - Robust Clustering Using Links.
    - See* ROCK
  - ROC
    - curves 222
    - graphs 222
  - ROCK 130, 146
    - algorithm 161–162, 300, 304
    - algorithm details 148
    - algorithms tweaks 315
    - formula explanation 150
    - goodness measure 149
    - initialization stage 149
    - key idea 147
    - link structure 149
    - termination criteria 149
  - ROCKAlgorithm 146–147, 303, 315
  - ROCKClusters class 149
  - root mean square error. *See* RMSE
  - roundoff error 111
    - magnitude of 119
    - minimize 118
  - RSS 13
  - RSS 2.0 14
  - R-trees 158
  - Ruby, Sam 14
  - rule engine 189
    - authoring 193
    - runtime 193
  - rule-based classification 188
  - RuleEngine 228
  - RuleQuest 171
  - rules
    - AccumulateFunction 194
    - attribute salience 197
    - ChainedProperties 194
    - ClassificationResult 191
    - Dialect 194
    - Email 191
    - engine 165
    - global statement 191
    - isSpamEmail 191
    - Package instance 194
    - PackageBuilder 194
    - PackageBuilderConfiguration 194
  - RuleBase 194
  - RuleEngine 193
  - StatefulSession 194
  - WorkingMemory 194
- S**
- 
- SAAJ 15
  - Salesforce 9
  - saliency 198
  - sampling
    - distribution estimation 257
    - with replacement 257
  - SAP 15
  - SAS 145
  - Scalable Vector Graphics. *See* SVG
  - ScanScout 9
  - Schatzker
    - classification system 166
  - score
    - relevance 29
  - scoring
    - index page 45
  - screen scraping 13
  - scripting 317
  - search engines
    - tuning 64
    - turning point in web history 3
  - search personalization
    - user clicks 45
  - search quality 64
    - metrics 64
  - search validation
    - precision 64
    - precision/recall plot 65
    - recall 64
  - searching 21, 30, 282
    - analysis stage 31
    - basic stages 29
    - beyond indexing 32
    - credibility 65
    - efficient data structures 61
    - intention 33
    - large-scale issues 61
    - link analysis 32
    - Lucene and PageRank
      - scores 43
    - PageRank 34
    - reference encoding 62
    - relevance 46
    - relevance score 29
  - SearchResult 29
  - second order effects 101
  - See5 171
  - selectBestMatchingTopic 299
  - selection strategies 264
  - selectLongestStory 299, 315
  - selectRepresentativeStory 315
  - self-organizing maps 229
  - semantic ontology 167
  - semantics 288, 328
  - semiotics 198
  - serialized PHP 14
  - setTopTerms 297
  - SFDataSet 134
  - Sheikholeslami,
    - Gholamhosein 130
  - shortest path metric 130
  - similarity 71
    - ad hoc threshold 104
    - best empirical results 79
    - between friends 81
    - calculation of 76, 84
    - code optimization 82
    - common misconception 79
    - compares proximity 71
    - cosine 95
    - evaluation façade 87
    - formula selection 79
    - formulas 77
    - heuristic 78
    - hybrid models 99
    - item 89
    - Jaccard metric 84
    - large scale comparisons of 79
    - linear correlation
      - coefficient 110
    - measures 73
    - metrics 117
    - music songs 89
    - naïve 78
    - normalization 89, 104
    - of content 92
    - plotting 74
    - ratio of the common items 78
    - related to cognition 74
    - shortcomings 77
    - symmetric matrix 84
    - symmetrical property 84
    - visual representation 74
  - similarity matrix 134
  - sparsity 84
  - upper triangular form 84
  - Simple Access Object Protocol.
    - See* SOAP
  - single-link algorithms 135
    - chain effect 142
    - computational
      - complexity 136

- single-link algorithms (*continued*)
    - MST 139
    - proximity threshold 135
  - SingleLinkAlgorithm 134
  - singletons 140, 146, 294, 301, 305
    - large number 303
  - Six Apart 9
  - SOAP 15
  - SOAP 1.1 15
  - SOAP 1.2 15
  - Soar 170
  - social networking sites
    - two most visited 6
  - SOII 166
  - SortedArrayClustering 126–127
  - SourceForge.net 123
  - SourceForge-like data 134
  - spam
    - documents 56
    - filtering 174
    - pages 32
  - Spam Assassin 175
  - spamRules.drl 191
  - spamRulesWithConflict.drl 198
  - spanning tree 139
  - sparse matrices 35
  - Spearman rank-order
    - correlation 112
    - coefficient 119
  - specialization 227
  - Specificity 221
  - spectral
    - clustering 130
    - methods 63
  - speech synthesis 327
  - Spencer, Thomas H. 161
  - spiders 13
  - sports news 284
  - SQL
    - clustering with 124
    - statements 124
  - SQL query
    - ORDER BY 124
  - SQLEM algorithm 161
  - Staab, Steffan 165
  - standard
    - deviation 111
    - normal distribution 254
  - StandardAnalyzer 30–31, 95, 178, 283
  - StandardTokenizer 31
  - statistic choice 251
  - statistic  $q$  255
  - statistically significant 250
  - stochasticity adjustment 36
  - stock
    - forecasting 3
  - StoryRecommender 310–311
  - structural algorithms 171
    - functional 171
    - numerical approximation 171
  - structures
    - flat reference 167
    - hierarchical reference 167
  - Studer, Rudi 165
  - supercluster 302, 305
    - avoidance 308
    - breakup 307
  - supervised learning 220
    - semiempirical approach 224
  - support vector machines. *See* SVM
  - surfer
    - random 35
  - SVG 1
  - SVM 16
  - Swing 292
    - client 301–302
  - synapse 201–202
    - weight 202
  - syndicated content 13
  - syndication-specific language 14
- T**
- 
- Tanimoto metric 117
  - Tarjan, Robert E. 161
  - Taxicab Geometry 118
  - teaser rates 234
  - technology articles
    - email spam 290
  - technology news 284
  - teleportation mechanism 34
  - tells 200
  - term vectors 59
  - termination criteria 129, 149
  - test
    - difference of
      - proportions 233, 250
  - test-users.txt 243
  - text
    - multilingual 284
    - tokenizing 30
    - understanding freely typed 4
  - text analysis 59
    - noise reduction 95
    - stemming 95
  - stop words 95
    - word disambiguation 97
  - TextMining 30, 55, 283
    - See also* tm-extractor
  - textual analysis 178
  - textual information
    - representation 95
  - threshold parameter 129
  - threshold value
    - statistical test 250
  - Tianji 9
  - Tibshirani, Robert 172
  - timestamp 47
  - title field 27
  - tm-extractor 55
  - token
    - frequency of occurrence 59
  - tokenizer
    - grammar-based 31
  - tokenizing
    - acronyms 31
    - alphanumerics 31
    - Chinese, Japanese, and Korean 31
    - computer host names 31
    - email addresses 31
    - text 30
  - top N frequent terms 178
  - Top stories 280
  - TopicalNewsClusterBuilder 307, 316
  - top-level labels 6
  - training
    - bootstrap 222
    - leave-one-out 222
  - TrainingSet 48, 50, 176, 178, 297
  - training-txns.txt 204
  - training-users.txt 243
  - trainOnAttribute 183
  - TransactionDataset 204
  - TransactionInstanceBuilder 210, 213
  - TransactionNN 204, 208, 210–211, 213–214, 229
  - transactions
    - identify fraudulent 204
    - legitimate 204
  - transition probability
    - matrices 35
  - transparent learning
    - algorithms 16
  - trap
    - black box 205
  - triangle of intelligence 5

triangulation effect 101  
 type I error 220  
 type II error 220

## U

---

UIMA 328  
 uncorrelated items 111  
 underfitting 227  
 unsolicited bulk email. *See* spam  
 unstable classifiers 258  
 unsupervised learning 128  
 url field 27  
 UseCaseData 258  
   createUserTypes 241  
 User 71  
 user click analysis 33  
 user clicks 287  
   news portal 312  
 user similarity  
   reliability 97  
 UserBasedSimilarity 82  
 UserClick 48  
 user-clicks.csv 46, 313  
 user-content similarities 103  
 UserContentBasedSimilarity 93  
 UserCreditNN 248–249, 274  
 UserInstanceBuilder 245, 261  
 user-item similarities  
   content-based 97  
 user-item-content similarity 103  
 Users 308  
 UserStatistics 204  
 UserType  
   addNoiseLevel 242  
 utility problem 225

## V

---

Vassilvitskii, Sergei 145  
 vector  
   personalization 67  
 vectors 323  
 very large database. *See* VLDB  
 VeryGoodUserType 239  
 Viadeo 9  
 visual pattern recognition 151  
 visualization 243  
 VLDB 131  
 voice recognition 9

## W

---

WaveCluster 130  
 wavelet clustering 130  
 web  
   semantic interpretation 14  
 web crawler  
   categories 320  
   components 319  
   custom 66  
 web crawling 13, 319  
 web services  
   providers 15  
 websites  
   inference 2  
   learning capacity 2  
   syndication 13  
 WhitespaceAnalyzer 31  
 WHO 166  
 wiki  
   defined 9  
 Wikipedia 9  
 world news 284

WS-Addressing 15  
 WS-I Basic Profile 15  
 WS-Policy 15  
 WS-RM 15  
 WS-Security 15

## X

---

Xerox  
   Palo Alto Research Center 80  
 XForms 1  
 Xignite  
   financial web services 15  
 XING 9  
 XML 14, 30, 283  
 XML Path Language. *See* XPath  
 XML User Interface Language.  
   *See* XUL  
 XOR gate 213  
 XORNetwork 213  
 XPath 1  
 XSL Transformations. *See* XSLT  
 XSLT 1  
 XUL 1

## Y

---

Yahoo!  
   RSS feeds 14  
 YouTube  
   media sharing 9

## Z

---

z statistic 254–255