

## Numerics

---

23andme.com 345

## A

---

abstraction, content types 83  
accessing DME 202–204  
accuracy of predictive  
  model 178, 305  
activation function 297–298  
addIndexes 326  
ad-generation engine 8  
adjusted cosine 354  
adjusted cosine-based  
  similarity 45, 363  
advertisements 8, 10, 12–13,  
  208, 237–238, 276  
advertising 8  
age 351  
  attribute 32  
agglomerative 253, 262  
AJAX 11, 89, 289  
Alexa 89, 146  
Algorithm. *See* Waikato Environ-  
  ment for Knowledge Analy-  
  sis (WEKA)  
algorithms, key  
  learning 178–181  
AlgorithmSettings 195, 199,  
  269–270, 301, 303  
Amazon 10, 18, 29, 34, 37, 39,  
  350, 365, 373  
Analyzer 312  
  *See also* Lucene  
analyzers 94–95, 309

analyzeText 232, 234  
analyzing content 207–221  
AOL 150  
Apache  
  Foundation 339  
  Hadoop 339  
  Jakarta Commons 120  
  Xerces-J 120  
APIs, JDM 193–204  
application phase 274  
application server 85  
ApplyTask. *See* Java Data Mining  
  (JDM)  
apriori 181, 188, 286  
architecture 21  
  content integration 85  
  tagging 51, 62–63, 66  
arrow 285  
articles 7, 14–15, 18–19, 23,  
  51–52, 60, 80–81, 83  
artificial intelligence 11, 176  
Asian languages 210  
association algorithms 177  
association rules 275, 362  
asynchronous 21–22  
asynchronously 275  
Atom Publishing Format  
  110, 144  
Attribute 176–182, 184–191,  
  194–196  
attribute 177, 274–275, 362  
  categorical 27  
  nominal 27  
  normalized 27  
  numerical 27  
  ordinal 27

attribute selection 181  
attributes 27, 176–178  
  nominal 177  
  numerical 176  
  ordinal 176  
AttributeType 197  
attrition rates 10  
authority 148, 163  
auto-complete 338  
automated indexers 146  
average-link 253  
averages 41

## B

---

back-propagation 297–299  
BallTree 366  
banner advertisements 374  
base pairs 345  
batch process 379  
Bayes' theorem 180, 287  
  Bayes' Rule 282  
  Bayesian algorithms 178  
Bayesian belief networks  
  (BBN) 178, 275, 285, 287,  
  289, 373  
Bayesian clustering 363  
BBN. *See* Bayesian belief net-  
  works (BBN)  
Bell, Robert. *See* BellKor  
BellKor 381–385  
Bennett, Jim 381  
BFS 148  
Bialecki, Andrzej 339  
BigTable 379  
binning 177

bioinformatics 345  
 biological relationships 346  
 black box 298, 354, 363  
 blob 91  
 blog comments 88  
 blog entries 14, 51, 88, 206,  
     244–246, 260, 262–263,  
     267, 275  
     clustering 241–261  
     retrieving from  
         Technorati 244–247  
 BlogAnalysisDataItem 244  
     *See also* clustering  
 BlogDataSetCreatorImpl  
     262, 287, 313  
 Blogdigger 108, 110–111, 124,  
     139–140, 142–144  
 BlogEntry 87, 113  
 Blogflux 111  
 bloggers 107, 109–110  
 Bloglines 108, 111, 124, 135,  
     137–139, 144  
 blogosphere 9, 16, 22, 87, 89,  
     104, 107–111  
     searching 111–116  
 BlogQueryParameter 112–113,  
     117, 119–121, 131, 136  
 BlogQueryResult 112–119,  
     121–123, 125, 134, 137, 143  
 blogs 9, 52, 83–90, 103–104,  
     108–111  
     blog-tracking companies 111  
     searching 111–116  
 BlogSearcher 113  
     *See also* blogs  
 BlogSearchExample 313  
 BlogSearchResponseHandler  
     112, 115–116, 121, 123–124,  
     131, 136–137, 139  
 blog-tracking companies 111  
 Bloomberg.com 345  
 bookmark 14–15, 37, 47, 51, 64,  
     66, 80, 342  
 bookmarking 8–9, 15, 28,  
     36, 351  
 BooleanQuery 356  
     *See also* Lucene  
 boost factor 328, 357  
 boosting 311, 328  
 bots 146  
 breadth-first search 148  
 buckets 27  
 buildDataPhysicalDataSet.  
     *See* Java Data Mining (JDM)  
 building, intelligent  
     crawler 152–164

BuildSettings 195, 269  
 business intelligence 176

## C

C4.5 281  
 C5.0 281  
 caching 21  
 CachingWrapperFilter 334–335  
 Carrot2, intelligent search 342  
 CART 281  
 CARuleMiner. *See* Waikato Envi-  
     ronment for Knowledge  
     Analysis (WEKA)  
 catalog 375  
 CategorySet 197  
 cause-effect. *See* Bayesian belief  
     networks (BBN)  
     .CFS extension 320  
 Charkrabarti 149  
 CharTokenizer 210  
 chat logs 84  
 chat sessions 14  
 child intelligence 289  
 child nodes 373  
 Chinese language 354  
 chromosomes 345  
 Church, George 345  
 churn in items 379  
 CI. *See* collective intelligence  
     (CI)  
 Cinematch 381, 384  
 circle of influence 5, 18  
 classes  
     Lucene 311–312  
     search 116–127  
 classification 274–275, 362  
     *See also* regression  
 classification terms 83–84, 104  
 ClassificationModel 305  
     *See also* Java Data Mining  
     (JDM)  
 ClassificationSettings 301, 304  
 ClassificationTestTask 305  
 classifieds 84  
 clickstream 37, 241  
 click-through 241, 261, 374  
     rates. *See* decision trees  
 cloaking 149  
 cluster 257  
     *See also* Java Data Mining  
     (JDM)  
 Clusterer 242, 244, 257,  
     261–267  
 clusterer, creating 265–266  
 ClusterEvaluation 262  
     *See also* Waikato Environment  
     for Knowledge Analysis  
     (WEKA)  
 clustering 22, 28–29, 32, 240,  
     242–244, 275, 362  
     evaluating results 266–268  
     high-dimension data 261  
     sparse data 261  
     with JDM 268–272  
     with WEKA 262–268  
 clustering model 198, 262, 269  
 ClusteringModel 268–272  
 ClusteringSettings 270  
 Clusty, intelligent search 344  
 CNET Networks 339  
 Cofe 369  
 Cofi 369  
 collaborative analysis 353–354  
 collaborative approach 93  
 collaborative filtering 20–21, 29,  
     363–373  
     model-based 363  
     probabilistic 373  
 collaborative-based 20, 52  
 Collective Intelligence 30  
 collective intelligence (CI) 30,  
     54, 59, 83–84, 86, 89, 91–92,  
     94, 100–102, 169  
     benefit 9  
     classification 14–18  
     definition 4, 6  
     example 7–9  
     extracting 93–102  
     *See also* classification  
 collective power 90  
 commercial crawlers 163  
 commit lock 324  
 community 12  
 Compass 339, 341  
 Compete 146  
 complete-link 253  
 complex-event-processing 24  
 composite 362  
 CompositeContentType 103  
 CompositeContentTypes 92  
 compound files 315  
     *See also* Lucene 320  
 computational biology 345  
 computeInitialDistances 258  
 computing similarities 31  
 conditional independence. *See*  
     Naïve Bayes *and* Bayesian  
     belief networks  
 conditioning methods 287  
 connecting with other users 349

Connection 201–204, 270–272, 302–305  
 ConnectionFactory 201–204  
 ConnectionMetaData 201  
 ConnectionSpec 203  
 consists 27  
 content 82, 85, 103  
   analyzing 207–221  
   analyzing example 93  
   classification 83, 85  
   external 8  
   integration 85–86  
   personalizing 362  
   retrieving 159–160  
   types 83–93, 102–103  
 content aggregation 147  
 content visited 28  
 content-based 20, 52  
 content-based analysis 352–363  
   recommendation engine 359–362  
 ContentBasedBlogRecoEngine 360  
 content-centric applications 12, 50  
 conversion rates 374  
 core competency 12  
 corporate website 87  
 correlation coefficient 193  
 correlation matrix 363  
 cosine 34, 41–43, 45–47, 354  
 cosine-based similarity 363  
 CoverTree 366  
 co-visitation 379–380  
 crawling 146–152  
   deep 150  
   process 147–149  
 crawling the web 145  
 createInitialSingleItemClusters 258  
 creating search index 314–317  
 cross-validation 182, 266, 298  
 crowd sourcing 90  
 Cuil 347  
 Cutting, Doug 151, 164  
 cycles 286–287

## D

DAG. *See* directed acyclic graph (DAG)  
 data aggregator service 22  
 data analysis 176  
 data collection 342

data mining 4, 16–18, 175  
   core concepts 176–182  
   example 362  
   JDM 193–204  
   process 181–182  
   vendors 181  
   WEKA 182–193  
 Data Mining Group (DMG) 204  
 data mining tools vendors 181  
 data search 345–347  
   intelligent search 341  
 data, learning dataset 263–265  
 database 62–64, 66–70, 78–79, 81  
 data-based search 346  
 Datar, Mayur 378  
 DataSetApplyTask 200  
 DataSetCreator 242  
 datasets 32, 181, 183, 200, 241, 243, 279–280  
 DayPop 111  
 DBScan 268  
 deadlocks 25  
 decision trees 275–281  
   *See also* classification 178  
 decodeme.com 345  
 deep web 150  
 degree of belief 287  
 del.icio.us 15, 36  
 densely populated 241  
 derived intelligence 14, 16  
 detecting phrases 100–102, 214–218  
 diagonal matrix 370  
 dictionary of tags 8  
 Digg 15, 36, 40  
 dimensionality reduction 369–373  
 dimensions 31  
 directed acyclic graph (DAG) 171, 285  
 directed graph 171  
 Directory 311  
 diversity 363  
 DME, accessing 202–204  
   *See also* Java Data Mining (JDM)  
 DMG. *See* Data Mining Group (DMG) 204  
 DNA 345  
 DNA chips 345  
 Document 311  
   *See also* Lucene  
 document frequency 359  
 doorway pages 149

dot product 34, 42–43, 46–47, 247, 251, 253, 328, 353, 365  
 dot-com era 3  
 Dryad 169–171  
 dynamic navigation 8, 14, 51, 54, 56, 58, 69, 80

## E

eBay 83  
 Eclipse 186  
 eigen value 370, 372  
 Einstein 11  
 EM. *See* expectation maximization (EM)  
 email 36, 351  
   chain 90  
   filtering 275  
   spam. *See* classification  
 embedding intelligence 21  
 English language 354  
 Epinions.com 39  
 EqualInverseDocFreqEstimator 231  
   *See also* text analysis  
 ethics 107  
 Eurekster 344  
 Evaluation 189, 192  
 event-driven 25  
 exceptions handling 116  
 ExecutionHandle 200–201, 271, 304  
 ExecutionStatus 201, 271, 304  
 expectation maximization (EM) 261, 265  
 experimental data search 346  
 Experimenter 184  
 Explanation 317–318, 330  
 explicit information 7, 14  
 exploitation 351  
 exploration 351  
 Explorer 184  
 ExportTask 200  
 external content 8–9  
 extract phrases 94  
 extracting URLs 160–161

## F

Fair Isaac 193  
 fame factor 384  
 FAQ 90  
 FastVector 186–193, 263, 290–291  
 FeedForwardNeuralNetSettings 301

Field 314–316  
 FieldQueryParser 333  
 Fields 316  
 file-based locking 324  
 Filter 210–211  
 filter 6  
 filtering 334–335  
   *See also* Lucene  
 firewall 85, 104  
 Flickr 50, 57  
 flyweight pattern 224  
 focused crawling 147–150,  
   171, 344  
 folders 37  
 folksonomies 8, 56, 60  
 FontSizeComputationStrategy  
   72  
 FontSizeComputation-  
   StrategyImpl 73  
 forward 36  
 framework, extending 127  
 freemium 13  
 frequency count 27  
 FSDirectory 312  
   *See also* Lucene  
 fundamental concepts 21, 25  
 FuzzyQuery. *See* Lucene

## G

---

Gaussian cluster 298–299  
 Gaussian distribution 261  
 Gaussian kernel function 298  
 gender 351  
   *See also* attribute  
 General Public License  
   (GNU) 341  
 genes 345  
 genetic algorithms 180  
 genomic sequencing 345  
 geographic location 83, 351  
 German 210  
 GermanAnalyzer 212  
 GermanStemFilter 210  
 getBlogDetails 255  
 getSynonym 215, 217  
 global warming 84  
 global-lock system 323  
 Gmail 56  
 GNU. *See* General Public License  
   (GNU)  
 Google 29, 59, 111, 128,  
   144, 354  
   community-based search  
   engines 344  
   data search 345

  search results 309  
   stop word list 213  
   YouTube 5  
 Google File System 169  
 Google News 350, 377–381  
 Gospodnetic 350  
 GPL 369  
 gradient descent algorithm 180  
 gradient search 297  
 greedy recommenders 352  
 groups 83–84, 86–87, 90–93

## H

---

HAC. *See* Hierarchical Agglom-  
 erative Clustering (HAC)  
 hackability 11  
 Hadoop 146, 164, 169–171  
 Hakia 345  
 handling exceptions 116  
 handling response XML 115  
 hard-to-replicate data 10  
 Harvard Medical School 345  
 harvest from external sites 82  
 harvest rate 150  
 hashCode 256  
 Hatcher 350  
 HDFS 169  
 Heritrix 151  
 Herren, John 58  
 Hibernate 30, 341  
 Hibernate search 339, 341  
 hidden layer 297, 299  
 hidden nodes 180, 298  
 hidden unit 298  
 Hierarchical Agglomerative  
   Clustering (HAC) 253, 262  
 hierarchical clustering  
   241, 253–261  
 HierarchicalClusteringImpl  
   253, 257–261  
 HierCluster 253–260  
 HierDistance 253–256  
 high performance 21  
 high-dimension 32–33  
 Hinchcliffe, Dion 11  
 HitCollector. *See* Lucene  
 Hits 312  
   *See also* Lucene  
 Hoffman, Kevin 71  
 Hoffmann 380  
 home address 276  
 homonym 54  
 Hornick, Mark 194  
 HTMLTagCloudDecorator 236

HTTP 311  
   requests 21, 24, 85  
 HTTPS requests 21  
 Human Genome Project 345  
 hyper linking 56

## I

---

IB1 366  
 IBk 366  
 IBM 193  
 IceRocket 111  
 ID3 281  
 identity matrix 296  
 IETF 110  
 if-then rules 275, 306  
 images 354  
 immutable 223  
 implicit information 7, 14  
 ImportTask 200  
 incremental indexing 322–324  
 index  
   creating 314–317  
   files 324  
   modifying 321–322  
   optimizing  
     performance 325–327  
     searching 317–320  
 indexing 320–327  
   incremental 322–324  
   optimizing  
     performance 325–327  
 indexing service 310  
 IndexReader 311  
 IndexSearcher 312  
 IndexUpdaterService 323  
 IndexWriter 311, 315  
 info gain. *See* decision trees  
 information entropy 277  
 information retrieval 4, 17, 25,  
   27, 30  
 infrastructure 240  
 InitialContext 203  
 injecting synonyms 214–218  
 input layer 180, 297–298  
 installation guide 90  
 Instance 187–191, 262–264  
 instant gratification 379  
 instant messengers 89  
 integration 85–86  
 intelligence, extracting 93–102  
 intelligent crawling 149–150,  
   152–164  
 intelligent search 362  
 interaction history 8, 18

inverse document frequency  
 (idf) 30, 55, 149, 328, 358,  
 365, 376  
 inverse of matrix 296  
 inverse user frequency 365  
 InverseDocFreqEstimator 222  
 InverseDocFreqEstimatorImpl  
 246  
 inverted text index 320  
 invisible web 150  
 is 327  
 isValidPhrase 215, 217, 220  
 Item 102  
 item churn 377  
 item-based analysis 351–353  
 Items 37  
 items 26  
 item-to-item 61, 353  
 Amazon 376

## J

---

Jaccard coefficient 380  
 JAMA 370  
 Java 30  
 Java Community Process  
 175, 193, 205  
 Java Data Mining (JDM) 176,  
 193–204, 241, 268, 272, 275,  
 300–305  
 accessing DME 202–204  
 architecture 194  
 clustering 268  
 clustering settings 269–270  
 connections 200  
 datasets 196  
 key clustering classes 268–269  
 key objects 195  
 tasks 199  
 Java Web Start 339  
 javax.datamining 202, 268–271,  
 300–304  
 javax.datamining.clustering  
 195, 268–270  
 javax.datamining.supervised  
 195, 300–303  
 JDBC 201  
 JDM. *See* Java Data Mining  
 (JDM) 272  
 JDMConnectionExample  
 202–203  
 JDMException 201–203  
 JNDI lookup 203  
 Johnson, Dave 111  
 journal entries 206

JSON APIs 339  
 JSR 247 175, 193–194, 205  
 JSR 73 175, 193–194, 205  
 JVM 323

## K

---

k neighbors 365  
 KDTree 366  
 kernel function 298, 303  
 key learning  
 algorithms 178–181  
 Keyes, Ken 5  
 Keyword spamming 149  
 keywords 82–84, 95, 102  
 k-fold. *See* cross-validation  
 King Ping 111  
 k-means 241, 299  
 implementation 247, 249–253  
 KMeansSettings 269  
 k-nearest neighbor (k-NN)  
 363, 365, 376  
 knowledge repository 90  
 KnowledgeFlow. *See* Waikato  
 Environment for Knowl-  
 edge Analysis (WEKA)  
 Koren, Yehuda 381  
 Kosmix intelligent search 342  
 KStar 366  
 KXEN 194

## L

---

language-independent 29  
 languages 354  
 large-scale systems 365  
 latent classes 29  
 latent Dirichlet allocation 363  
 latent semantic indexing  
 (LSI) 80, 369–370  
 layer 275, 297–299  
 leaf cluster 260  
 learning dataset 181, 263–265  
 learning models 197  
 learning phase 274  
 Lemire, Daniel 369  
 LetterTokenizer. *See* Lucene  
 Lexee 345  
 Libby, Dan 109  
 life sciences 346  
 Linden, Greg 374  
 linear algebra 29  
 linear model 32  
 linear regression 180, 295–297

LinearNNSearch 366  
 Linguistic-based search 341  
 link spamming 149  
 Linkdb 166  
 LinkedIn 12  
 links, decision tree 179  
 list 7, 84  
 list of related items 206  
 load balancer 22, 311  
 Locality Sensitive Hashing  
 (LSH) 380  
 Lock 323  
 lock 324  
 log likelihood 266  
 logarithmic 71  
 LogicalAttribute 196–197  
 LogicalDataSet 195  
 look-to-book ratio 276  
 low-dimension 241  
 LowerCaseTokenizer 210, 212  
 LSH. *See* Locality Sensitive  
 Hashing (LSH)  
 LSI. *See* latent semantic indexing  
 (LSI)  
 Lucene 17, 149, 151, 164, 166,  
 169, 172, 320–327, 350  
 architecture 310–311  
 classes 311–312  
 core classes 311–312  
 download 208  
 finding similar items 355–359  
 indexing 313, 320  
 querying 330–331  
 scoring 327–330  
 term vector 324–325  
*Lucene in Action* 208, 310  
 Lucene JDBC 341  
 Lucene scoring 327–330  
 LuceneTextAnalyzer 231–235  
 Luke 339

## M

---

machine learning 176, 240, 342  
 machine-generated tags 83  
 mailing list 91  
 manufacturer 83  
 MapReduce 146, 164,  
 169–171, 380  
 margin 180  
 marketplace 12, 83  
 Markov Decision process 363  
 mass behavior 4  
 mathematical model 306  
 mathematics 176

MathWorks 370  
 Matrix 296, 370–371  
 matrix 295  
 matrix inversion 299  
 Meebo 89, 105  
 memory-based 379  
 menus 50  
 message boards 8, 26, 41, 82–86, 91–93, 206  
 messaging infrastructure 25  
 messaging server 21–22, 24–25  
 metadata 46, 51–52, 58–59, 62, 68–69, 80  
   attribute-based 26  
   content-based 27  
   example computation 41–46  
   from text 93  
   tagging 50–80  
   user actions 34–40  
   user-action-based 27  
   users and items 26–27  
 MetaDataExtractor 95  
 MetaDataVector 94, 98  
 meta-search 344  
 microarrays 345  
 Microsoft 107, 150, 170  
 mining process 181–182  
 MiningObject 195–197, 300  
 Minwise Independent  
   Permutation Hashing  
   (MinHash) 380  
 mirror sites 149, 164  
 MLP. *See* multi-layer perceptron (MLP)  
 model-based 29, 196, 379  
 ModelDetail 197  
 MOR 194  
 MoreLikeThis 358  
 movie titles 383  
 MSN 108, 111, 124, 139–140, 144  
 MultiFieldQueryParser 318, 333  
 multi-layer perceptron (MLP) 275, 297–298  
 multiple fields query 327  
 multiple indexes search 312, 327, 335  
 multiple multiplication factor model 363  
 multiple-term tokens 214  
 MultiSearcher 312, 335  
 multi-term phrases, detect 95  
 MultiTermQuery 330  
 music 17, 354  
 MyRank 59  
 MyWeb 59

---

**N**

NaïveBayes 281  
 natural language 345  
 navigation 50, 83–84  
   links 51  
   menus 8  
 navigenics.com 345  
 N-dimensional vector 353  
 nearest neighbor. *See* k-nearest neighbor (k-NN)  
 NearestNeighbourSearch 365  
 net gain 277  
 net worth 276, 351  
 Netflix 29, 350, 381–385  
 network effect 6  
 network topologies 373  
 neural network 29, 178, 274, 295, 297–298, 306  
 NeuralNetworkModelDetail 197  
 New York Times 37  
 news feeds 83  
 news items 84  
 news site 379  
 newsfeed format 110  
 NextBio 346  
 Nielsen Net Ratings, search numbers 309  
 node 179–180, 259, 276–280, 285–286, 294  
 nodes, decision tree 179  
 noisy ratings 379  
 nominal attributes 177  
 nonlinear 297–299  
 Normalize 31  
 normalizeToken 97  
 numerical attributes 176  
 Nutch 146, 149, 151, 164, 167–171, 207, 309, 339  
   running 165–167  
   searching with 168–169  
   setting up 164

---

**O**

online analytic processing (OLAP) 176  
 ontology 60, 346  
 open source crawler 164, 339  
 OPTICS 268  
 optimizing memory settings 325  
 Oracle 193–194  
 OrbiMed Advisors LLC 345  
 ordinal attributes 176  
 orthogonal matrix 370  
 overfitting 182, 298

---

**P**

PageRank 59  
 pandemic 5  
 ParallelMultiSearcher 312  
 parent nodes 285  
 path followed 28  
 pattern matcher 161, 163  
 pdf 320  
 PDFBox 320  
 Pearson-r correlation 43, 354, 363  
 Pentaho 182  
 PerFieldAnalyzerWrapper 212  
 perpetual deta 11  
 persistence model 35  
 personal health history 345  
 personal journals 83, 87  
 personalization 349, 362  
   Google News 377–381  
 personalized  
   recommendations 350, 374  
 Photo 44  
 photo 82  
 photos 26, 41–42, 48, 50–52, 80, 84  
 phrase detection 215–233  
 phrase dictionary 207  
 PhraseQuery 330–331  
 phrases 51, 53–55, 58, 80, 207, 212, 216–218, 221–222, 224, 239  
   definition 29  
   detecting 100–102, 214–218  
 PhrasesCache 215  
 PhysicalAttribute 196  
 PhysicalAttributeRole 196  
 PhysicalDataSet 195–196, 304  
 pictures 14  
 pinged 111  
 Pingoat 111  
 Pingomatic 111  
 PLSI. *See* probabilistic latent semantic indexing (PLSI)  
 PMML. *See* Predictive Model Markup Language (PMML) 204  
 podcasters 107  
 podcasts 14, 51  
 polling 25  
 polls 28, 84, 103  
 polysemy 54, 80, 369  
 Porter 99, 210–221  
 PorterStemFilter 210–214  
 PorterStemmer 327  
 PorterStemStopWordAnalyzer 212–214, 219

Postami 111  
 Powerset 345  
 precision 310  
 predictive model 274  
   intelligent search 342  
 Predictive Model Markup  
   Language (PMML) 204  
 predictive models 22, 24, 28, 32,  
   48, 102  
 PredictiveApriori 188  
 PrefixFilter 334  
 PrefixQuery 330–334  
 price 83  
 printClusterEntries 266–267  
 printer 86  
 prior history 275  
 probabilistic 373  
 probabilistic latent semantic  
   indexing (PLSI) 363, 380  
 probabilistic methods 29  
 probabilistic networks. *See* Bayes-  
   ian belief networks (BBM)  
 probability distribution 281  
 probability theory 180, 281  
 products 51, 83  
 professionally developed  
   keywords 26  
 professionally generated 52  
 profile 352  
 profile page 84  
 Profile selections 28  
 pruneDistances 258–259  
 purchasing history 375

## Q

quadratic regression 294  
 quality of the item 354  
 quality of the predictive  
   model 182  
 Quantcast 146  
 Query 312  
 query results 114  
 query terms 359  
 Query. *See* Lucene  
 QueryFilter. *See* Lucene  
 QueryParser 312  
 questions and answers 8, 26,  
   82–83, 103, 206

## R

Racofi. *See* Cofi  
 radial basis function (RBF)  
   275, 295, 298–299  
 RAM 325

RAMDirectory 312  
 random 250  
 RangeFilter 334  
 RangeQuery 330  
 RapidMiner. *See* Waikato Envi-  
   ronment for Knowledge  
   Analysis (WEKA)  
 rate 8  
 ratings 7, 14, 18, 26, 28–29,  
   35–36, 40–41, 83, 349  
   example 41–48  
   persistence model 35  
 RBF. *See* radial basis function  
   (RBF)  
 RBMs 383  
 RDF 109  
 Read A Blog 111  
 Reader 209  
 recall 310  
 RecodApplyTask 200  
 recommendation 14  
 recommendation engine 7, 9,  
   17–18, 21, 24, 37, 41, 102,  
   349–355  
   Amazon 374–377  
   collaborative-based 8  
   content-based 359–362  
   high performance 379  
   Netflix 381–385  
 recommendation engines 18  
 recommendation system 29, 238  
   hybrid 355  
   large-scale 373  
 recommendations 8, 39  
 reference weblogs 89  
 ReferenceWeblog. *See* blogs  
 registration 276  
 regression 178, 274–275, 362  
   *See also* classification  
 RegressionModel 197–198, 300  
 RegressionSettings 301  
 RelevanceTextDataItem 359  
 remixability 11  
 response XML, handling 115  
 result objects,  
   implementing 117  
 RetrievedBlogEntry 244  
 RetrievedBlogHitCollector 357  
 retrieving content 159–160  
 review 6, 9, 14  
 Reviewer 40  
 reviews 39–41, 46, 83–84  
 Revver 39  
 Rich Site Summary (RSS)  
   59, 108–109, 111, 124,  
   128–129, 139, 144

integrating providers  
   with 139–143  
 parsing 141–143  
 RSS 2.0 141–143  
 rich user experience 11  
 robots.txt 147, 156–158  
 Rolex watch. *See* classification  
 Rollyo 344  
 RSS. *See* Rich Site Summary  
   (RSS) 108  
 rule induction 178  
 running  
   Nutch 165–167  
   web crawler 162–163  
 Russian charset 210–211  
 RussianAnalyzer 212  
 RussianLetterTokenizer 210  
 RussianLowerCaseFilter 210  
 RussianStemFilter 210

## S

SaaS. *See* software-as-a-service  
   (SaaS)  
 SAP 193  
 SAS 194  
 saving 9, 28  
 SAX 116, 119–120, 123–124,  
   126, 144  
 scaling 21  
 Scuttle 63–65, 81  
 search 9, 17, 102  
   architecture 310–311  
   base classes 116–127  
   definition 310  
 search engine 8, 85, 87, 90, 92,  
   104, 207, 362  
   ranking 10, 90, 149  
 search engines 145, 147, 354  
   community-based 344  
 search history 377–378  
 search index 22, 310  
   creating 314–317  
 search parameters 113  
   implementing 117  
 search performance  
   optimization 338  
 search service 23, 311  
 search terms 84, 375  
 Searcher 323, 329, 332  
 searching  
   blogosphere 111–116  
   blogs 111–116  
   with Nutch 168–169  
 Searchme 347  
 segment 315, 320, 325

select for update 25  
 sending messages 206  
 Service-Oriented Architecture (SOA) 21  
 services, definition 21  
 setBoost 328  
 setMaxBufferedDocs 325–326  
 setMaxFieldLength 326  
 setMergeFactor 325  
 shopping basket 375  
 sigmoidal basis functions 299  
 Silicon Valley 7  
 similarity 27, 247  
 similarity computation, cosine-based 42  
 similarity matrix 364  
 similarity metric 353  
 SimpleAnalyzer 212  
 SimpleBiTermStop-WordStemmerMeta-DataExtractor 100  
 SimpleContentType 103  
 SimpleKMeans 268  
 SimpleMetaDataExtractor 95  
 SimpleStopWordMeta-Data-Extractor 98  
 SimpleStopWordStemmerMeta-DataExtractor 99  
 simulated annealing 180  
 single nucleotide polymorphisms (SNP)  
 single-link 253  
 single-signons 86  
 Singleton 323  
 singular value decomposition (SVD) 354, 369  
 sitemaps 150, 172  
 slop 331  
 SNP. *See* singular nucleotide polymorphisms (SNP)  
 SOA. *See* Service-Oriented Architecture (SOA)  
 social networking 26, 351  
 sociology 4  
 software-as-a-service (SaaS) 39, 334  
 Solr 324, 339  
 sorting 327  
 SpanQuery 330  
 sparse data 365  
 sparse matrix 353–354, 370  
 sparsely populated 32–34, 41, 47, 383  
 Sphere 111  
 spider trap 149  
 Spring 30, 59, 341  
 Spring bean 323

SPSS 193  
 spurl.net 36  
 square matrix 295, 370  
 Stack. *See* text analysis  
 standard data mining API 193  
 StandardAnalyzer 212  
 StandardTokenizer 210  
 stateless 21  
 statistical 294  
 statistics 176  
 stem 207  
 stemmer analyzers 213  
 stemming 31, 58, 94, 99–100, 102, 104  
   definition 29  
 stickier 8, 10  
 stochastic simulation 287  
 stop terms 98  
 stop words 30, 94, 98–99, 207, 358  
   removing 98–99  
 StopAnalyzer 212  
 Strategy 71  
 subcategory 50  
 supervised learning 178, 299  
 SupervisedAlgorithmSettings 300–303  
 SupervisedModel 197, 300  
 SupervisedSettings 301  
 support vector machine (SVM) 180, 295  
 Surowiecki, James 4  
 SVD. *See* singular value decomposition (SVD)  
 SVDExample 371  
 SVM. *See* support vector machines (SVM) 178  
 SVMClassificationSettings 303  
 SVMRegressionSettings 301  
 sweet spot 351  
 Synchronous services 21  
 SynonymPhraseStopWord-Analyzer 212  
 SynonymPhraseStopWordFilter 216–218  
 synonyms 31, 53–55, 58, 207  
   injecting 214–218  
 SynonymsCache 215  
 synonymy 369

## T

Tag 221  
 tag cloud 8, 15, 208  
   building 57–59, 62, 69  
   definition 57  
 TagCloud.com 59  
 TagCloudElement 71–75, 77–79  
 tagging 9, 28, 51–56, 60, 62–63, 65–69, 78–79, 342, 349  
   introduction 51  
 TagMagnitude 222, 225  
 TagMagnitudeVector 225, 244, 246, 353  
 tags 8, 14–15, 18, 207  
 tan hyperbolic functions 297  
 Task 195–196  
 Taste. *See* collaborative filtering  
 taxonomyParentId 196  
 Technorati 108, 111, 124, 128, 131, 134–135, 144, 244–247, 262, 275, 287–290  
 term frequency (TF) 30, 328, 358  
 term frequency vector 324  
 term vector 29, 32, 34, 48, 93, 207, 241, 288, 353, 362  
   infrastructure 225–231  
   representation 206  
 term vectors 30–31, 34  
 term-frequency 149  
 TermFreqVector 356–357  
 TermQuery 330  
 terms, definition 30  
 text analysis 206–221, 231, 237–239, 350  
   infrastructure 221–237  
 text analytics 359  
 text analyzers 208  
   stemmer analyzer 213  
 text clustering 242–244  
 text parsing 207  
 text processing 247, 261  
 TextAnalyzer 232  
 TextDataItem 242, 359  
*The Hundredth Monkey* 5  
*The Long Tail* 18, 41  
*The New Yorker* 4  
*The Wall Street Journal* 36  
 threshold 297  
 Time 128  
 Token 209  
 TokenFilter 214–217  
 tokenization 31, 94  
   definition 208  
 tokenize 207  
 Tokenizer 209  
 Tokenizer, European 210  
 TokenStream 209–214  
 Tomcat 339  
 toolkit, text analysis 206  
 tools 84  
 top 10 8

Top Item List 351  
 top  $n$  recommendation 352, 363  
 Top Reviewers list 39  
 TopDocCollector 336–337  
 TopDocs 336–337  
 TopFieldDocs 336–337  
 topical crawlers 149  
 top-seller lists 374  
 Toxi 63–65, 81  
 TP53 346  
 training process 298  
 transaction history 23–24, 28  
 TreeSettings 301

## U

UGC. *See* user-generated content (UGC)  
 undirected path 287  
 University of Waikato 182  
 unstructured text 25, 27, 30, 40  
 unsupervised learning 178, 240, 299  
 URLs, extracting 160–161  
 user interactions 349, 363  
 user profile 12, 351  
 user rating 20, 35, 46  
 user-based analysis 352–353  
 user-centric applications 6, 12–14  
 user-generated content (UGC) 12, 206, 275, 322  
   definition 82  
   tags 26  
 user-item dataset 363  
 user-item matrix 364, 369  
 Userland Software 109  
 users 26  
   clustered 9  
 user-user similarity matrix 363

## V

validation window 38  
 variables 177  
 vector 27, 30–34, 43, 46–48  
   space model 30, 327  
 videos 23, 33, 51–52, 80, 82, 103, 107, 354  
   content type 84, 102

dataset 33  
 example analysis 41  
 integration architecture 23  
 metadata 26  
 Rever 39  
 tagging 14  
 YouTube 5, 33  
 viral 5, 7  
 viral marketing 36  
 Vista 107  
 VisualizeTagCloudDecorator 76–77, 79  
 vocabulary 51–53, 60, 80  
 Volinsky, Chris 381  
 voting 9, 15, 28, 35, 41, 46

## W

Waikato Environment for Knowledge Analysis (WEKA) 175, 182–193, 241, 262–268, 275, 281, 287–288, 290, 292–296, 299–300, 306, 366, 370  
 APIs 186–193  
   installation 186  
   tutorial 183–185  
 watches 279  
 Web 2.0 6, 10–11, 14, 19, 39, 62  
 web 2.0 289  
 Web 3.0 9, 11, 19  
 web application 3, 20  
 web applications 3  
 web crawler 146, 149–152, 157, 163  
   building 152–164  
   running 162–163  
 web crawling 22, 82, 145–152, 342  
   deep 150  
   process 147–149  
   why 146  
 web server 24  
 web spiders 146  
 Web2.0 94, 106  
 WEKA APIs 186–193  
 WEKA. *See* Waikato Environment for Knowledge Analysis (WEKA) 275  
 weka.associations 188  
 weka.attributeselection 188

weka.classifier 187  
 weka.clusterer 187  
 weka.core 186  
 weka.filters 188  
 WEKABlogClassifier 287, 292, 300  
 WEKABlogDataSetClusterer 262  
 WEKABlogPredictor 288, 299  
 WEKAPredictiveBlogDataSet-CreatorImpl 287  
 White, Tim 164  
 WhitespaceAnalyzer 212  
 WhitespaceTokenizer 210  
 Wiki 89–90, 105  
 Wikipedia 89–90, 106, 152, 162, 164–165, 169  
 wikis 8–9, 18, 83–84, 86, 89–92, 104  
 WildCardQuery 331  
 window of terms 215  
 Winer, Dave 109  
 Wisdom of the Crowds 4  
 word frequency 7  
 works 298  
 worksheets 84  
 world wide web 145  
 write lock 324

## X

XML response, parsing 123–126

## Y

Yahoo 56, 58–59  
   blog 111  
   music 17  
 YALE. *See* Yet Another Learning Environment (YALE)  
 Yet Another Learning Environment (YALE) 183  
 YouTube 5, 33

## Z

ZoomCloud 59  
 Zopto 111