

Index

A

ABBYY FineReader, 51
Accountability, 338
Accuracy, 222
Acrobat Adobe, 51
Activity bias, 341
Advanced graphs, 255–261
Advantages, 40–42, 45–47, 148, 152, 178, 196, 227
Algorithmic accountability, 339
Algorithmic confounding/biasness, 35, 338–343
Altmetrics, 34, 118, 180, 200
Amazon Textract, 51
American Library Association (ALA), 2
Annotating, 234
Anonymization of data, 337
Application programming interfaces (APIs), 1, 44, 46, 54, 56, 63, 97
Applications, 147–148, 178, 196, 214
Applications of information visualization in libraries, 270
Applications of ML, 214
Approaches for sentiment analysis, 193–194
Arc Diagram, 264
Architect, 227
Arcs, 139
Area under curve (AUC), 222, 224
Area under ROC curve (AUC), 222
arXiv, 59, 186
ASCII, 39
Association for Computing Machinery US Public Policy Council, 342
Association roles, 151
Association types, 151

Assumptions for NTA, 140

Attribute relation file format (ARFF), 301
Attribute semantics, 248
Automatic summarization, 10
AYLIEN, 66, 195, 201

B

Bag-of-words (BOW), 79, 86–87, 90, 111, 139, 148
Bar chart, 251
Beautiful Soup, 68
Benefits of data management for text mining, 326–327
Benefits of visualizing library resources, 289
Between, 236
Betweenness centrality, 144
Biasness, 226, 340–342
Bibliometrics, 6, 59, 118, 147, 157, 177, 179–180
Bibliomining, 36–37
Biblioshiny, 158, 177
Bigram, 87
Binary classification, 12, 219
Binary files, 46–47
BinderHub, 24, 130, 162, 186, 203, 291
Bipartite networks, 141
British Library, 235
BTM, 109
Bubble Chart, 258
Bubblelines, 266
Burst detection, 173–179
Burst_detection, 178
Bursts, 173, 177
Butterfly/Tornado graph, 255

C

Caret, 227
 Categorical/nominal scale, 246
 Centrality measures, 142–145
 Challenges, 230–233
 Character, 39
 Character-level representation, 81
 Chartjunk, 261
 Chart Studio, 317–318
 Chatbots, 236
 Chinese Text Project, 64
 Chronicling America, 229, 234
 Chunk, 79
 CiteSpace, 177
 Classification task, 219–220
 Classifier, 213
 Closeness centrality, 144
 Clustering, 6, 215
 Coherence, 106
 Comma-separated value (CSV), 41
 Community detection algorithm, 141
 Comparison, 145–146
 Computational analysis, 52
 Computational environment, 24
 Computational library, 1
 Computational methods, 8
 Computational thinking, 2
 Computer linguistics, 139
 Concept Drift, 341
 Concordance, 264
 Confusion matrix, 222
 Cons, 297, 300–310, 314–318
 Content analysis, 153
 ConText, 108, 147, 308–309
 Continuous bag of words (CBOW), 95
 Copyright, 335, 337
 CoreNLP, 195
 Corextopic, 109
 Corpora, 139
 Corpus, 50, 79–82, 88, 106, 142
 Correlated topic modeling (CTM), 106, 113
 Correlation explanation (CorEx), 113
 Cosine distance, 94
 Cost, benefits and barriers, 17
 Creative Commons (CC), 329
 crossrefapi, 59
 Cross-validation, 225, 226
 Crowdsourcing, 234–235
 Current awareness service (CAS), 8, 37

D

DALEX, 343
 DARIAH Topics Explorer, 109

Dashboard, 260, 290
 Data, 33–37
 Data Analysis Recommendation Assistant (DARA), 235
 Database Directive, 337
 Database rights, 335
 Data citation, 326
 DataCite, 324, 325
 Data file types, 39–47
 Data-ink ratio, 261
 Data management, 47
 Data management plan (DMP), 320
 Data maximization, 336
 Data minimization, 336
 Data mining (DM), 8, 36, 50, 295
 Data points, 27
 Data Protection Directives, 335
 Data protection rules, 335–337
 Data repositories, 324
 Data scale types, 245–246
 Data Scraper, 68
 Datasets, 33
 Data sharing, 47
 Data warehouse, 36
 Datawrapper, 314–315
 Degree centrality, 143–144
 Dendrogram, 27
 Dependencies, 90
 Dependency grammar (DG), 91
 Dependency parsing, 91
 Derivational morphology, 85
 Dictionaries, 11, 82
 Digital breadcrumbs, 34
 Digital data, 33, 54, 56
 Digital data creation, 51–54
 Digital footprint, 34
 Digital libraries, 119, 149, 270, 295
 Digital library task, 283–284
 Digital Research Tools (DiRT) Directory, 296
 Digital rights management (DRM), 330
 Digital Scholar Workbench—Constellate, 57, 229
 Digital sources, 34
 Digital text data, 51
 Digital trace data, 34, 35, 59
 Digital traces, 34
 Digitization, 52, 234, 338
 dimensionsR, 60
 Directed graphs, 145
 Disadvantages, 41, 42, 45–47, 153
 Discoverability, 235
 Dispersion Plot, 264
 Distance matrix, 26, 27
 Distances, 94

DMPonline, 320
 DMPTool, 320
 DocuBurst/SunBurst, 267
 Document classification, 12–13
 Document Frequency (DF), 89–90
 Document term matrix (DTM), 25, 88–89
 Document visualization, 269
 Dot Plot, 258
 Driftiness, 35
 Dublin Core Element Set, 48
 Dynamic topic correlation detection, 107
 Dynamic topic models (DTM), 106, 109

E

Edges, 139, 254
 EigenCentrality, 144
 Elbow method, 25
 Emotion artificial intelligence, 191
 Entropy, 106
 Ethical, 327–330
 Ethical and legal issues, 332–343
 Ethical research, 332
 Ethics Guidelines for Trustworthy AI, 339
 Euclidean distance, 94
 Euclidean distance matrix, 18
 Euclidean distance method, 26
 Evaluate bias, 339
 Evaluation metrics, 224
 Examples, 48, 142
 eXtensible Markup Language (XML), 45
 External data sources, 36
 Extraction, transformation, and loading (ETL),
 36

F

FAIR data principles, 328–330
 FairML, 342
 Fairness, 338, 339
 FAIRsharing, 324
 False positive rate (FPR), 223
 Feature engineering, 86–96
 Feature extraction, 194
 Features, 35, 213
 Features of Library 4.0, 296
 Features of ML, 215
 Flair, 195
 F-measure, 223
 FMiner, 67
 Fold, 225
 FORCE11, 325
 Formats, 39

Frequency-inverse document frequency
 (TF-IDF), 25
 f1 score, 222
 Function, 47
 Fundamental graphs, 251

G

Genealogy of text mining, 6–8
 General Data Protection Regulation (GDPR),
 335
 Generalization, 217–219
 genism, 109
 Geocode, 65
 Gephi, 147, 310–311
 Getty Provenance Index, 63
 Gibbs Sampling, 126
GloVe, 95
 Goodreads, 59
 Goodreads/Google Books, 65
 Google Cloud Vision, 51
 Google Flu Trends, 341
 Google's What-If Tool, 342
 Google Trends, 173
 Graph algorithms, 145
 Graphical decoding, 248
 Graphic variable types, 246–247
 gtrendsR, 59
 guardianapi, 59
 guidedLDA, 113
 gutenbergr, 59

H

Handwriting recognition, 234
 Harzing's Publish or Perish, 58
 HathiTrust, 234
 Heatmap, 18, 253
 Hidden Markov model (HMM), 92, 174
 Hierarchical clustering algorithm, 6, 20, 26
 Hierarchical topic modeling, 113
 h2o, 227
 Hyperparameter, 223
 HyperText Markup Language (HTML), 45, 63

I

IBM AI Fairness, 342, 360
 igraph, 147
 Imbalanced classification, 219
 iml, 342
 Import.io, 67
 In-degree, 144

- Indexing, 13
- Industry 4.0, 295
- Inflectional morphology, 85
- Infogram, 312
- Information extraction, 6
- Information layers, 262
- Information processing, 50
- Information retrieval, 6, 119, 179
- Information summarization, 6
- Information visualization, 243, 270–290
- Information visualization framework, 244–245
- Information visualization process, 244
- Information visualization skills, 289
- Informed consent, 330
- Integrated development environment (IDE), 296
- Intellectual property rights, 334, 335, 337
- Interactive storytelling, 250
- Interactive visualization, 250
- Internal data sources, 36
- Interpretability, 342
- Interval scale, 246
- Inverse document frequency (IDF), 25, 89
- ISO 13250, 149

- J**
- Jaccard similarity, 95
- JavaScript Object Notation (JSON), 9, 42, 63
- jsLDA, 109
- Jupyter Notebook, 24, 130, 162, 186, 205, 291

- K**
- Kappa, 223
- Keras, 227
- Kernlab, 227
- Keyword frequency, 96
- Keyword-in-context (KWIC), 263
- K-fold cross-validation, 225
- Kleinberg’s algorithm, 173
- Knowledge-based NER, 94
- Knowledge discovery in databases (KDD), 8
- Knowledge graphs, 153, 156
- Knowledge management, 157–158
- Knowledge organization, 119, 236

- L**
- Lancaster, 84
- LancsBox, 147, 307–308
- Language-based models, 343
- Latent Dirichlet algorithm, 106
- Latent Dirichlet Allocation (LDA), 110, 175
- Latent semantic analysis (LSA), 193
- Latent Semantic Indexing (LSI), 87, 110
- lda, 109
- LDAVis, 131
- Learning-based NER, 94
- learning-based POS tagging, 92
- Leave-one-out cross-validation, 226
- Legal, 326–329
- Lemmatization, 84–85
- Levels of granularity, 192–193
- Levels of text representation, 81
- Lexical-level representation, 81
- Lexicon, 82
- Librarian 2.0, 295
- Librarians, 289
- Libraries, 113–119, 151–158, 179–180, 197–200, 228–236, 270–290
- Libraries Ready to Code (RtC), 2
- Library 2.0, 295
- Library 4.0, 295
- Library administration, 235–236
- Library advocacy, 286–288
- Library of Congress, 65, 234
- License categories, 338
- License conditions, 337–338
- LightSide, 195
- lime, 342
- Limitations, 17, 149, 178, 196–197, 228
- Line chart, 252
- Linked data, 236
- Links, 139
- Lord of the Rings Project (LOTR Project), 96
- Louvain Community Detection method, 141, 146

- M**
- Machine learning (ML), 11, 213, 228–236, 295, 338
- Machine learning algorithms, 216–219
- Machine learning methods, 215–216
- Machine learning techniques, 105
- Machine-readable format, 90
- MALLET, 120
- MARC21, 65
- Marketing, 180, 200
- Markov model, 186
- MARKUS, 64
- Matrix, 12, 88, 253
- Mean absolute error (MAE), 222
- Mean square error (MSE), 222
- Metadata, 23, 48–50, 149, 320
- Metadata recognition and extraction, 235
- Metadata standard, 48–49

Metaknowledge, 178
 Meta-tagging, 119
 Metatags, 50
 Methods, document visualization, 269
 Methods, graphic visualization, 250
 Metrics, 34
 Microsoft Power BI, 147, 312–314
 Microsoft's Fairlearn, 342
 Mitigate biases, 342
 MIT Libraries, 1
 Model building, 226
 Model Drift, 341
 Model evaluation, 221–226
 Model selection, 226
 Modes of visualization, 250
 MonkeyLearn, 195
 Morphological normalization, 83
 Mosaic/Mekko Chart, 256
 mscstexta4r, 195
 Multi-class classification, 219
 Multi-label classification, 219
 Multiple Line Graph, 257
 Multiscale topic tomography, 107

N

Naïve Bayes, 216, 237
 Named entity recognition (NER), 10, 93–94
 Natural language processing (NLP), 6, 8, 91, 191
 Need of data management, 326–327
 Neighboring rights, 330
 Netlytic, 57
 Networks, 254
 Network text analysis (NTA), 139–149
 Network visualization, 145
 networkx, 147
 N-grams, 87–88
 NLP Architect, 195, 227
 NLTK, 195
 Nodes, 139, 249
 NodeXL, 147, 195
 Non-textual data, 34
 Normalization, 82–83

O

OCLC, 65
 Octoparse, 67
 Omitted variable bias, 341
 One-node networks, 141
 Online repositories, 56
 Ontologies, 117, 150
 OPAC, 65, 236

Open-source tools, 108, 177, 227
 Operabase, 63
 Opinion mining, 9, 191
 Optical character recognition (OCR), 51, 234
 Orange, 17, 58, 109, 147, 195, 227, 302–304
 Ordinal scale, 246
 Out-degree, 144
 Outwit, 67
 Overfitting, 217–219
 Overstemming, 84
 OverviewDoc, 309–310

P

PageRank, 145
 Palladio, 147, 316–317
 ParseHub, 67
 Parse tree, 90
 Parts-of-speech (POS), 10, 79, 84, 216
 Pattern, 68
 Pattern extraction, 17
 Perform predictive modeling, 221
 Perplexity, 106
 Personal data, 334
 Phrase-level representation, 81
 Phrase structure grammar (PSG), 91
 Pictograph, 257
 Plain text, 39–41
 Plain text files, 40, 126
 Polarity, 192
 Polinode, 147
 Porter, 84
 Portia, 67
 Positivity-Biasness, 36
 POS tagging, 10, 91–93
 Precision, 222, 223
 Predictive modeling, 213–228
 Presentation visualization, 250
 Principle of effectiveness, 249
 Principle of expressiveness, 249
 Privacy, 331, 334
 Private data, 331
 Probabilistic Latent Semantic Analysis (PLSA), 87, 110
 Probabilistic model, 18
 Process of bibliomining, 37
 Process of topic modeling, 111
 Pros, 297–301, 303, 305, 306–318
 Public domain, 336
 Purpose of visualization, 249–250
 pybursts, 178
 Pytesseract, 51
 Python, 53, 106, 109, 147, 178, 195, 227
 Python-youtube, 59

PyTorch, 227
pyvis, 147

Q

Qualitative data, 34
quanteda, 109, 195, 227

R

R, 23, 55, 130, 147, 158, 162, 177, 185, 203,
227, 290, 296–297, 342, 343
Radar/Spider/Star Chart, 260
rAltmetric, 59
RapidMiner, 58, 106, 108, 126, 147, 195, 201,
227, 236–240, 299–300
Raster graphics, 250
Ratio scale, 246
RAWGraphs, 315
rcrossref, 59
RDF, 149
Recall, 222, 223
Receiver operating characteristic (ROC) curve,
224
Recommendation service, 118
Registry of Research Data Repositories, 324
Regular expressions, 94
Relational databases, 56–63
Research data lifecycle, 320
Response bias, 341
rgoodreads, 60
rscopus, 60
R2 score, 222
RStudio Server, 24, 130, 162, 186, 203, 291
rtweet, 59
Rule-based POS tagging, 92
Rule/lexicon, 193
Rules on visual design, 261–262
rvest, 68

S

Sankey Diagram, 256
Scatter plot, 252
Scatter Text, 264
scholar, 59
Science of Science 236 (Sci2), 147, 180
Science of science (Sci2) Tool, 306–307
scihub, 59
scikit-learn, 109, 227
SciPy, 227
Scite, 200
Sci2 Tool, 177
Scopus, 158

Scraper, 68
Scrapy, 68
Selection bias, 341
Selective dissemination, 37
Selective dissemination of information (SDI),
8, 179
selectr, 68
Semantic, 17
Semantic analysis, 10
Semantic-level representation, 81
Semantic networks, 140
Semantic parsing, 86
Semantics, 63
Semantic-web technology, 149
Semi-structured, 17, 38
Semi-supervised learning, 106, 113, 215
Sensitive data, 334
Sensitivity, 223
Sentiment analysis, 9, 97, 191–197
SentimentAnalysis, 195
sentimentr, 195
Sentiment Viz, 195
SentiStrength, 195
SentiWordNet, 193
Similarity, 94–95
Skip-gram, 95
Snowball, 84
Social biases, 343
Social media data, 330–331
Social media ethics, 330–332
Social media mining (SMM), 200
Social media platform, 65
Social network analysis (SNA), 157
Societal bias, 341
Sociodemographic factors, 343
Software architecture, 151–152
Sparsing, 12
Specificity, 223
Stacked Area Chart, 257
Stacked graph, 253
Stages in text mining, 80
Standard dictionaries, 84
Stemmer, 84
Stemming, 84
Steps, 194, 221
Steps to create, 50
stm, 109
Stopwords, 85
Stratified K-fold cross-validation, 226
Streamgraph/ThemeRiver, 267, 269
Structural topic modeling (STM), 107, 113,
130
Structured data, 17, 38
Structured Query Language (SQL), 60

Subject-based classification, 149
 Subject classification, 117
 Subjectivity, 192
 Subjectivity analysis, 191
 Supervised, 215
 Supervised learning, 94, 113
 Supervised learning methods, 15–16, 94
 Supervised machine learning (ML), 194, 213, 219
 Support vector machine (SVM), 216, 237
SWOT analysis, 200
 Symbol map, 253
 Syntactical parsing, 90–91
 Syntactic-level representation, 81
 Syntactic parsing, 10, 90–91
 Syntax, 86
 System drift, 341
 syuzhet, 195, 205

T

Tableau, 147, 282
 Tableau Public, 311
 Tag Cloud/Word Cloud, 263
 Tags, 50, 86
 TDM Studio, 57, 229
 Teachable Machine, 227
 Term, 79
 Term-document matrix (TDM), 88
 Term Frequency (TF), 89
 Term frequency-inverse document frequency (TF-IDF), 89–90, 142
 Terms and conditions, 330
 Term weighting, 89
 Tesseract, 51
 Test set, 221
 tethne, 109
 Text Analysis Portal for Research (TAPoR 3), 296
 Text analytics, 9
 Text and data mining (TDM), 296, 336
 Text-based predictions, 9
 TextBlob, 195
 Text categorization, 12–13
 Text characteristics, 10–11
 Text classification, 6, 89
 Text data, 51, 327–330
 Text data management, 319–330
 textmineR, 109
 Text mining, 326–327, 332–343
 Text mining applications, 40
 Text mining tasks, 12–15
 Text mining techniques, 87
 Text mining tools, 296–307

Textnets, 146, 147, 164
 Text pre-processing, 82–86, 106, 126
 Text representation, 81
 Text retrieval, 9
 Text tools, 64
 Text transformation, 81–82
 Textual data, 9, 34, 295
 text2vec, 109
 Text visualization, 262–269
 tidytext, 109
 Tidyverse, 227
 TileBars/MicroSearch, 266
 TIMDEX, 1
 Time series, 173
 Time slicing, 177
 tinyTIM, 150
 tm, 109
 TM4J, 150
 Tokenization, 83–84
 Tokens, 79, 216
 Tools, 57
 Tools and packages, 108–109, 147, 177–178, 195, 227
 Topic Detection and Tracking (TDT), 175
 topicdoc, 109
 Topic evolution, 106–107
 Topic maps, 149–152
 Topic Maps Toolbox, 150
 Topic modeling, 8, 87, 105–110, 113–119, 142
 Topic modeling tool (TMT), 108, 119, 297–299
 topicmodels, 109
 Topic over time (TOT), 93
 Topix, 108
 Training data, 215
 Training set, 221
 Transparency, 338
 Tree diagram, 27
 Treemap, 264
 Trees, 254
 Trigram, 87
 True positive rate (TPR), 223
t-SNE, 95
 Tweepy, 59
 Twitter, 59
 Twitter Archiving Google Sheet (TAGS), 57
 Two-mode Networks, 141–142
 Types of biases, 341
 Types of classification, 219
 Types of data, 38

U

Underfitting, 217–219
 Understemming, 84

Undirected graphs, 145
 Unicode, 39, 40
 Unigram, 87
 Universal Resource Locators (URLs), 64
 Unstructured data, 9, 38
 Unsupervised, 94, 113, 215
 Unsupervised learning methods, 15–16, 94
 US Algorithmic Accountability Act, 339
 Use cases, 70, 179, 200, 282–288
 Uses, 40, 44, 46, 47, 152
 UTF-8, 40

V

VADER, 195
 Validation set, 225
 Vector graphics, 250
 Vertices, 139
 Virtual Librarian Chat, 200
 Visualizing ontology, 156
 vosonSML, 60
 VOSviewer, 147
 VoyantTools, 147, 304–305
 VyontTools, 108

W

Waikato environment for knowledge analysis
 (WEKA), 301–302
 Waterfall/Mario Chart, 259
 Weaknesses, 35
 Web 4.0, 295

Web APIs, 54, 56, 63–66, 82
 Webhose.io, 67
 Web of Science (WoS), 158
 WebOPACs, 36
 Web scraper, 68
 Web scraping/screen scraping, 54, 66–70, 82
 Web scraping tools, 67
 Weighting, 89
 WEKA, 108, 227
 Word association mining, 9
 Word Association Networks, 140
 Word-based projections, 141, 142
 Word co-occurrence, 139, 158
 Word co-occurrence network, 140, 162
 Word embedding, 95–96
 Wordij, 147, 315–316
 Word-level representation, 81
 WordNet, 193
 Word to vector, 339
 Word Tree, 263
word2vec, 95
 Workflow, 51
 wosr, 60

X

XML, 9, 17, 63

Z

Zipf's law, 89