Index

adaptive threshold setting, 132
centroid, 82, 83
chat rooms, 151
chi-square statistic, 173
clustering
  constrained, 81
  pairwise constrained, 81
confusion matrix, 99
constrained optimization, 83
constraints
  cannot-link, 82, 84, 87, 93, 102
  instance-level, 81
  must-link, 82, 87, 94, 102
  must–must-link, 99
correlation function, 178
cybercrime, 161
cyberbullying, 150
cyberpredators, 150
dataset
  CISI collection, 99
  Cranfield collection, 99
  Medlars collection, 99
  MPQA corpus, 15
  NIPS, 198
  Reuters, 198
distance
  Bregman, 89
  Kullback–Leibler, 82
  reverse Bregman, 81
divergence
  Bregman, 89
  Kullback–Leibler, 82
  reverse Bregman, 81
event types, 167
external prior-knowledge, 193
FeatureLens, 113
first variation, 83
function
  closed, proper, convex, 89
  distance-like, 83
FutureLens, 113
gain ratio, 66
Gaussian algorithm, 173
Gaussian distribution, 133
history flow, 110
hypothesis test, 172
information gain, 66
information retrieval (IR), 22, 95
isolating language, 23, 32

k-means, 81
  batch, 83
  constrained, 101
  incremental, 82, 83, 86, 100
  quadratic, 82, 95
  quadratic batch, 82
  spherical, 81
  spherical batch, 95, 96
  spherical constrained, 101
  spherical incremental, 97, 99
k-means, (Continued)
spherical with constraints, 82
keyword, 3, 170
  applications, 3, 4, 15
  extraction methods, 4, 5
  keyphrase, 3, 5
  metrics, 17, 18
  variants, 16

latent Dirichlet allocation, 186
  generative process, 187
  inference, 187
    Gibbs sampling, 188
  online LDA, 189
latent morpho-semantic analysis
  (LMSA), 32
    with term alignments
  (LMSATA), 33
latent semantic analysis, 185
  probabilistic, 185
latent semantic analysis (LSA), 22
  with term alignments (LSATA), 30
latent semantic indexing, 72
LDA, see latent Dirichlet allocation, 186
log-entropy term weighting, 26
LSI, see latent semantic indexing, 72
luring communication, 154

mean
  arithmetic, 90
misbehavior detection, 152
multi-parallel corpus, 21
multilingual document clustering, 21
multilingual LSA, 25
multilingual precision at five documents (MP5), 25
NMF, see nonnegative matrix factorization, 60
NMF-BCC, 74
NMF-LSI, 72

nonnegative matrix factorization, 60
  alternating least squares algorithm, 62
  classification, 70
  initialization, 65
  multiplicative update algorithm, 62
novelty mining, 130
NP-hard problem, 83

online
  communities, 150
  luring, 154
  victimization, 149

pairwise mutual information (PMI), 30
PARAFAC2, 28
partition, 83
nextFV (II), 84
quality, 83
PDDP, 99, 100
penalty, 82, 84, 87, 88, 93, 96, 99, 102
power method, 30
precision at one document (P1), 24
predatory behavior, 159
RAKE, 5
  algorithm, 6, 7
  evaluation, 9, 10, 15
  input parameters, 6
RSS, 169

SEASR, 111
semantic embedding
  data driven, 194
  external, 193
sentiment tracking, 111
sexual
  exploitation, 149
  predation, 150
singular value decomposition, 22
Sinkhorn balancing, 30
smoka, 81, 92
constrained, 101
SMT, see statistical machine translation, 30
social networking, 149
statistical machine translation, 30
stoplist, 6, 10
generation, 11
stopwords, 5
SVD, see singular value decomposition, 22
swine flu, 178
synthetic language, 23, 32
tag
cloud, 108
crowd, 108
temporal profiles, 171
tensors, 27
text stream, 169
TextArc, 111
topicality, 180
transitive closure, 88, 94, 98
TREC novelty track data, 138
Tucker1, 27
UIMA, 111
vector space model, 22, 184
vocabulary
indexing, 4
term frequency, 12
term selection, 4
VSM, see vector space model, 22
Wikipedia, see external prior knowledge, 193
Wordle, 108