Index

adjacency matrix, 142
American politics
  2016 election, 15, 142
  2018 election, 69
Congress, 25, 30
APIs, 12
  Academic API, 13, 14, 24
Botometer, 143
gecoded data, 91
  keys, 23
  limitations, 16
  sample stream, 36
search API, 13
SourceForge, 152
streaming API, 14, 36, 153
Arab Spring, 1

Belmont Report, 146
big data, 2, 9, 36, 155
  dealing with, 30
Black Lives Matter, 15
Botometer, 143
bots, 142
  2016 election, 142
cyborgs, 160
detection, 142
Brazil, 116, 117

Cambridge Analytica, 159
Center for Social Media and Politics, 6
centrality, 139
China, 42
censorship, 10
collective action, 4
Color Revolutions, 1
Common Rule, 146, 148
  communications technologies, 4
  compressing data, 38
  concept drift, 75
Congress, 25, 75, 77, 123, 124, 130, 135, 141, 153
district shapefiles, 96
  measuring partisanship, 77
  polarity of tweets, 69
content analysis, 63
  random sampling, 77
CoreNLP library, 69
CSMap, 6
CSV files, 29
Cyrillic, 42
data missingness, 15
database
  alternate database software, 31
  setting up for tweets, 30
SQLite, 34
Dataverse, 151
denial-of-service attacks, 148
Digital Society Project, 7
directed graphs, 134
downloading tweets, 25
  by keyword, 44
  by location, 91
  by timeline, 124
  by user, 124
tweets, 13
  entire language stream, 45
  friend networks, 133
  worldwide sample, 36

emotional analysis, 67
entities object, 20
Euromaidan, 14, 20, 157
extended mode tweets, 27
Facebook, 10, 150, 159
firehose, 15
followers, 133
France, 117
friends, 133

GenSim library, 70
gecoded data, 86, 91
  caching techniques, 98
  coastline problem, 105
  coordinates, 89
data inconsistencies, 90
downloading all, 92
grid cells, 113
history, 89
identifying country of origin, 97
identifying geocode from user activity, 117
place matching, 112
statistics, 116
subnational matching, 110
variation by country, 118
Gephi, 141
GitHub
emotion recognition tool, 69
repository for book’s code, 24
Global Administrative Areas (GADM)
project, 96
grid cells, 113
Gutenberg, 2
Harvard Dataverse, 25
hashtags, 11, 46
analyzing, 49
cohashtags network, 137
extracting, 47
JSON object, 20
network analysis, 122, 137, 138
removing, 66
human subjects research, 146
images
analysis, 59
downloading, 55
how social media changed, 1
JSON object, 20
pretrained classifiers, 60
Instagram, 121
International Telecommunications
Union, 3
Iran, 76, 117
IRB, 146, 155
Ireland, 25
Islamic State, 15
JSON
Twitter object, 16
language code inconsistencies, 20
languages, 41
by name
Farsi, 76
Persian, 76
Russian, 20
Swedish, 45
Ukrainian, 20
field in JSON object, 19
frequently used on Twitter, 41
language codes, 20
machine detection, 19
limitations, 42
text tools for, 65
Latent Dirichlet allocation, 70
limitations, 75
latitude and longitude, 86
Lazer Lab, 7
lemmatization, 64, 65
literacy, 2, 3
location identification, 115
location problem, 115
mentions, 11, 49
content analysis, 83
JSON object, 20
removing, 66
naive text processing, 70
Natural Language Toolkit, 64
network analysis, 122
adjacency matrices, 142
centrality, 139
cofriend networks, 141
visualization, 134
neural networks, 76
emotion identification, 69
evaluating, 82
evaluating performance, 78
implementing, 78
performance, 82
training with other tweets, 83
Nigeria, 117
NIST library, 64, 70
Norway, 117
object-oriented programming, 32
observational data, 149
partisanship, 77
place definitions, 87
places
JSON object, 86
points of interest, 89
types, 87
point-in-polygon, 96
points of interest, 89
polarity of text, 69
Polyglot library, 68
print media, 2, 5
Putin
Twitter accounts, 132
pyLDAvis library, 70
quote tweets, 19
radio, 2, 5
randomly sampling tweets, 36
Reddit, 10, 121
removing short words, 67
replication, 22, 25, 146, 151
Index

retweets, 11, 116
quote tweets, 19
replies, 19
text field behavior, 19
Russia, 19, 42, 117, 150
disinformation operations, 15
Putin on Twitter, 132

sentiment analysis, 67
shapefiles, 95
creating, 109
SMS messages, 12
social media
academic literature, 5
effect on society, 2
social movements, 4
SoMe Lab, 6

source code of book
GitHub repository, 24
organization, 25
SourceForge, 152
Stanford CoreNLP library, 69
stemming, 64
stop words, 46, 67
removing, 67
subnational matching, 112
subtweeting, 49
supervised learning models, 75
Sweden, 45, 117
political tweets, 83

Tahrir Square, 1
television, 2, 5
terms of use, 150
text analysis
emotion recognition, 69
languages, 40
LDA modeling, 70
lemmatization, 64
neural networks, 77
polarity, 69
processing tweet text, 66
sentiment analysis, 67
stemming, 65
stop words, 67
supervised learning, 75
tokenization, 64
topic modeling, 70
text messages, 12, 42
third-party sources of Tweets, 15
Tiananmen Square, 1
TikTok, 10
time zones, 20
tokenization, 64
topic modeling, 70
evaluation, 72
number of topics, 73, 74

transaction costs, 5
trauma, 153
tweets
aggregating for analysis, 75
corpus, 19
data limitations, 18
downloading, 21, 25
entities object, 20
Excel importing, 18
extended mode, 27
full text, 19
goecoding, 85
JSON structure, 16
language, 19
large integer problem, 18
parsing, 28
polarity, 69
processing, 29
text of tweet, 18
time stamp, 20
time zone field, 20
unique identifier, 17, 25, 32, 36

Twitter
140-character limit, 12
3200 tweet limit on timeline scraping, 123
connection limits, 27
endpoints, 16
goecoding history, 89
history, 19
language code inconsistencies, 20
languages, 41
Library of Congress archive, 16
place definitions, 87
places, 111
reasons for academic use, 10
terms of use, 150
timeline limitations, 123
timelines, 124
tweets per day, 36

Ukraine, 14, 19, 43, 157
time zones, 20
URLs
removing, 66
user objects, 18

videos, 60
downloading, 62
JSON object, 20
VKontakte, 150
Volapuk, 42

web scraping, 54, 151

YouTube, 10, 149

zipping data, 38