Practical Natural Language Processing

A Comprehensive Guide to Building Real-World NLP Systems

Sowmya Vajjala, Bodhisattwa Majumder, Anuj Gupta, and Harshit Surana

Beijing • Boston • Farnham • Sebastopol • Tokyo
O'REILLY®

Table of Contents

Foreword.. xv

Preface.. xvii

Part I. Foundations

1. NLP: A Primer.. 3
   - NLP in the Real World 5
     - NLP Tasks 6
   - What Is Language? 8
     - Building Blocks of Language 9
     - Why Is NLP Challenging? 12
   - Machine Learning, Deep Learning, and NLP: An Overview 14
   - Approaches to NLP 16
     - Heuristics-Based NLP 16
     - Machine Learning for NLP 19
     - Deep Learning for NLP 22
     - Why Deep Learning Is Not Yet the Silver Bullet for NLP 28
   - An NLP Walkthrough: Conversational Agents 31
   - Wrapping Up 33

2. NLP Pipeline.. 37
   - Data Acquisition 39
   - Text Extraction and Cleanup 42
     - HTML Parsing and Cleanup 44
     - Unicode Normalization 45
     - Spelling Correction 46

System-Specific Error Correction 47
Pre-Processing 49
Preliminaries 50
Frequent Steps 52
Other Pre-Processing Steps 55
Advanced Processing 57
Feature Engineering 60
Classical NLP/ML Pipeline 62
DL Pipeline 62
Modeling 62
Start with Simple Heuristics 63
Building Your Model 64
Building THE Model 65
Evaluation 68
Intrinsic Evaluation 68
Extrinsic Evaluation 71
Post-Modeling Phases 72
Deployment 72
Monitoring 72
Model Updating 73
Working with Other Languages 73
Case Study 74
Wrapping Up 76

3. Text Representation. 81
Vector Space Models 84
Basic Vectorization Approaches 85
One-Hot Encoding 85
Bag of Words 87
Bag of N-Grams 89
TF-IDF 90
Distributed Representations 92
Word Embeddings 94
Going Beyond Words 103
Distributed Representations Beyond Words and Characters 105
Universal Text Representations 107
Visualizing Embeddings 108
Handcrafted Feature Representations 112
Wrapping Up 113

Part II. Essentials

4. Text Classification.. 119
Applications 121
A Pipeline for Building Text Classification Systems 123
A Simple Classifier Without the Text Classification Pipeline 125
Using Existing Text Classification APIs 126
One Pipeline, Many Classifiers 126
Naive Bayes Classifier 127
Logistic Regression 131
Support Vector Machine 132
Using Neural Embeddings in Text Classification 134
Word Embeddings 134
Subword Embeddings and fastText 136
Document Embeddings 138
Deep Learning for Text Classification 140
CNNs for Text Classification 143
LSTMs for Text Classification 144
Text Classification with Large, Pre-Trained Language Models 145
Interpreting Text Classification Models 147
Explaining Classifier Predictions with Lime 148
Learning with No or Less Data and Adapting to New Domains 149
No Training Data 149
Less Training Data: Active Learning and Domain Adaptation 150
Case Study: Corporate Ticketing 152
Practical Advice 155
Wrapping Up 157

5. Information Extraction.. 161
IE Applications 162
IE Tasks 164
The General Pipeline for IE 165
Keyphrase Extraction 166
Implementing KPE 167
Practical Advice 168
Named Entity Recognition 169
Building an NER System 171
NER Using an Existing Library 175
NER Using Active Learning 176
Practical Advice 177
Named Entity Disambiguation and Linking 178
NEL Using Azure API 179

Relationship Extraction 181
  Approaches to RE 182
  RE with the Watson API 184
Other Advanced IE Tasks 185
  Temporal Information Extraction 186
  Event Extraction 187
  Template Filling 189
Case Study 190
Wrapping Up 193

6. Chatbots.. 199
Applications 200
  A Simple FAQ Bot 201
A Taxonomy of Chatbots 202
  Goal-Oriented Dialog 204
  Chitchats 204
A Pipeline for Building Dialog Systems 205
Dialog Systems in Detail 206
  PizzaStop Chatbot 208
Deep Dive into Components of a Dialog System 218
  Dialog Act Classification 219
  Identifying Slots 219
  Response Generation 220
  Dialog Examples with Code Walkthrough 221
Other Dialog Pipelines 226
  End-to-End Approach 227
  Deep Reinforcement Learning for Dialogue Generation 227
  Human-in-the-Loop 228
Rasa NLU 229
A Case Study: Recipe Recommendations 232
  Utilizing Existing Frameworks 233
  Open-Ended Generative Chatbots 235
Wrapping Up 236

7. Topics in Brief.. 241
Search and Information Retrieval 243
  Components of a Search Engine 245
  A Typical Enterprise Search Pipeline 248
  Setting Up a Search Engine: An Example 249
  A Case Study: Book Store Search 251
Topic Modeling 252
  Training a Topic Model: An Example 256

What's Next? 257
Text Summarization 258
Summarization Use Cases 258
Setting Up a Summarizer: An Example 259
Practical Advice 260
Recommender Systems for Textual Data 262
Creating a Book Recommender System: An Example 263
Practical Advice 264
Machine Translation 265
Using a Machine Translation API: An Example 266
Practical Advice 267
Question-Answering Systems 268
Developing a Custom Question-Answering System 270
Looking for Deeper Answers 270
Wrapping Up 271

Part III. Applied

8. **Social Media** **277**
Applications 279
Unique Challenges 280
NLP for Social Data 286
Word Cloud 286
Tokenizer for SMTD 288
Trending Topics 288
Understanding Twitter Sentiment 290
Pre-Processing SMTD 292
Text Representation for SMTD 296
Customer Support on Social Channels 299
Memes and Fake News 301
Identifying Memes 301
Fake News 302
Wrapping Up 304

9. **E-Commerce and Retail** **309**
E-Commerce Catalog 310
Review Analysis 310
Product Search 311
Product Recommendations 311
Search in E-Commerce 311
Building an E-Commerce Catalog 314

Attribute Extraction 314
Product Categorization and Taxonomy 319
Product Enrichment 323
Product Deduplication and Matching 325
Review Analysis 326
Sentiment Analysis 327
Aspect-Level Sentiment Analysis 329
Connecting Overall Ratings to Aspects 331
Understanding Aspects 332
Recommendations for E-Commerce 334
A Case Study: Substitutes and Complements 335
Wrapping Up 338

10. Healthcare, Finance, and Law.. 341
Healthcare 341
Health and Medical Records 343
Patient Prioritization and Billing 344
Pharmacovigilance 344
Clinical Decision Support Systems 344
Health Assistants 344
Electronic Health Records 346
Mental Healthcare Monitoring 355
Medical Information Extraction and Analysis 357
Finance and Law 360
NLP Applications in Finance 362
NLP and the Legal Landscape 365
Wrapping Up 368

Part IV. Bringing It All Together

11. The End-to-End NLP Process.. 373
Revisiting the NLP Pipeline: Deploying NLP Software 374
An Example Scenario 376
Building and Maintaining a Mature System 378
Finding Better Features 379
Iterating Existing Models 380
Code and Model Reproducibility 381
Troubleshooting and Interpretability 381
Monitoring 384
Minimizing Technical Debt 385
Automating Machine Learning 386

The Data Science Process 390
The KDD Process 390
Microsoft Team Data Science Process 392
Making AI Succeed at Your Organization 394
Team 394
Right Problem and Right Expectations 395
Data and Timing 396
A Good Process 397
Other Aspects 398
Peeking over the Horizon 400
Final Words 403

Index. 409