thuvienso.net

Speech and Language Processing

An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition

Third Edition draft

Daniel Jurafsky
Stanford University

James H. Martin
University of Colorado at Boulder

Copyright ©2020. All rights reserved.

Draft of December 30, 2020. Comments and typos welcome!

thuvienso.net

Summary of Contents

1 Introduction 1
2 Regular Expressions, Text Normalization, Edit Distance 2
3 N-gram Language Models 29
4 Naive Bayes and Sentiment Classification 55
5 Logistic Regression 76
6 Vector Semantics and Embeddings 96
7 Neural Networks and Neural Language Models 127
8 Sequence Labeling for Parts of Speech and Named Entities 148
9 Deep Learning Architectures for Sequence Processing 173
10 Contextual Embeddings 202
11 Machine Translation and Encoder-Decoder Models 203

12 Constituency Grammars 231
13 Constituency Parsing 259
14 Dependency Parsing 280
15 Logical Representations of Sentence Meaning 305
16 Computational Semantics and Semantic Parsing 331
17 Information Extraction 332
18 Word Senses and WordNet 355
19 Semantic Role Labeling 373
20 Lexicons for Sentiment, Affect, and Connotation 393
21 Coreference Resolution 415
22 Discourse Coherence 442
23 Question Answering 464
24 Chatbots & Dialogue Systems 492
25 Phonetics 526
26 Automatic Speech Recognition and Text-to-Speech 548
Bibliography 575
Subject Index 607

thuvienso.net

Contents

1 Introduction **1**

2 Regular Expressions, Text Normalization, Edit Distance **2**
2.1 Regular Expressions . . . 3
2.2 Words . . . 11
2.3 Corpora . . . 13
2.4 Text Normalization . . . 14
2.5 Minimum Edit Distance . . . 22
2.6 Summary . . . 26
Bibliographical and Historical Notes . . . 27
Exercises . . . 28

3 N-gram Language Models **29**
3.1 N-Grams . . . 30
3.2 Evaluating Language Models . . . 35
3.3 Generalization and Zeros . . . 37
3.4 Smoothing . . . 41
3.5 Kneser-Ney Smoothing . . . 45
3.6 Huge Language Models and Stupid Backoff . . . 47
3.7 Advanced: Perplexity's Relation to Entropy . . . 49
3.8 Summary . . . 52
Bibliographical and Historical Notes . . . 52
Exercises . . . 53

4 Naive Bayes and Sentiment Classification **55**
4.1 Naive Bayes Classifiers . . . 56
4.2 Training the Naive Bayes Classifier . . . 59
4.3 Worked example . . . 61
4.4 Optimizing for Sentiment Analysis . . . 61
4.5 Naive Bayes for other text classification tasks . . . 63
4.6 Naive Bayes as a Language Model . . . 64
4.7 Evaluation: Precision, Recall, F-measure . . . 65
4.8 Test sets and Cross-validation . . . 67
4.9 Statistical Significance Testing . . . 69
4.10 Avoiding Harms in Classification . . . 72
4.11 Summary . . . 73
Bibliographical and Historical Notes . . . 73
Exercises . . . 74

5 Logistic Regression **76**
5.1 Classification: the sigmoid . . . 77
5.2 Learning in Logistic Regression . . . 81
5.3 The cross-entropy loss function . . . 82
5.4 Gradient Descent . . . 83
5.5 Regularization . . . 88
5.6 Multinomial logistic regression . . . 90
5.7 Interpreting models . . . 92
5.8 Advanced: Deriving the Gradient Equation . . . 93
5.9 Summary . . . 94

thuvienso.net

Bibliographical and Historical Notes 94
Exercises 95

6 Vector Semantics and Embeddings 96
6.1 Lexical Semantics 97
6.2 Vector Semantics 100
6.3 Words and Vectors 101
6.4 Cosine for measuring similarity 105
6.5 TF-IDF: Weighing terms in the vector 106
6.6 Pointwise Mutual Information (PMI) 109
6.7 Applications of the tf-idf or PPMI vector models 111
6.8 Word2vec 112
6.9 Visualizing Embeddings 118
6.10 Semantic properties of embeddings 118
6.11 Bias and Embeddings 120
6.12 Evaluating Vector Models 122
6.13 Summary 123
Bibliographical and Historical Notes 123
Exercises 126

7 Neural Networks and Neural Language Models 127
7.1 Units 128
7.2 The XOR problem 130
7.3 Feed-Forward Neural Networks 133
7.4 Training Neural Nets 137
7.5 Neural Language Models 142
7.6 Summary 146
Bibliographical and Historical Notes 147

8 Sequence Labeling for Parts of Speech and Named Entities 148
8.1 (Mostly) English Word Classes 149
8.2 Part-of-Speech Tagging 151
8.3 Named Entities and Named Entity Tagging 153
8.4 HMM Part-of-Speech Tagging 155
8.5 Conditional Random Fields (CRFs) 162
8.6 Evaluation of Named Entity Recognition 167
8.7 Further Details 167
8.8 Summary 169
Bibliographical and Historical Notes 170
Exercises 171

9 Deep Learning Architectures for Sequence Processing 173
9.1 Language Models Revisited 174
9.2 Recurrent Neural Networks 176
9.3 Managing Context in RNNs: LSTMs and GRUs 186
9.4 Self-Attention Networks: Transformers 190
9.5 Potential Harms from Language Models 198
9.6 Summary 200
Bibliographical and Historical Notes 200

10 Contextual Embeddings 202

thuvienso.net

11 Machine Translation and Encoder-Decoder Models **203**
11.1 Language Divergences and Typology . . . 205
11.2 The Encoder-Decoder Model . . . 208
11.3 Encoder-Decoder with RNNs . . . 209
11.4 Attention . . . 212
11.5 Beam Search . . . 215
11.6 Encoder-Decoder with Transformers . . . 217
11.7 Some practical details on building MT systems . . . 218
11.8 MT Evaluation . . . 221
11.9 Bias and Ethical Issues . . . 226
11.10 Summary . . . 227
Bibliographical and Historical Notes . . . 228
Exercises . . . 230

12 Constituency Grammars **231**
12.1 Constituency . . . 232
12.2 Context-Free Grammars . . . 232
12.3 Some Grammar Rules for English . . . 237
12.4 Treebanks . . . 244
12.5 Grammar Equivalence and Normal Form . . . 249
12.6 Lexicalized Grammars . . . 250
12.7 Summary . . . 255
Bibliographical and Historical Notes . . . 256
Exercises . . . 257

13 Constituency Parsing **259**
13.1 Ambiguity . . . 259
13.2 CKY Parsing: A Dynamic Programming Approach . . . 261
13.3 Span-Based Neural Constituency Parsing . . . 267
13.4 Evaluating Parsers . . . 269
13.5 Partial Parsing . . . 270
13.6 CCG Parsing . . . 271
13.7 Summary . . . 277
Bibliographical and Historical Notes . . . 278
Exercises . . . 279

14 Dependency Parsing **280**
14.1 Dependency Relations . . . 281
14.2 Dependency Formalisms . . . 283
14.3 Dependency Treebanks . . . 284
14.4 Transition-Based Dependency Parsing . . . 285
14.5 Graph-Based Dependency Parsing . . . 296
14.6 Evaluation . . . 301
14.7 Summary . . . 302
Bibliographical and Historical Notes . . . 302
Exercises . . . 304

15 Logical Representations of Sentence Meaning **305**
15.1 Computational Desiderata for Representations . . . 306
15.2 Model-Theoretic Semantics . . . 308
15.3 First-Order Logic . . . 311
15.4 Event and State Representations . . . 318

thuvienso.net

15.5 Description Logics . . . 323
15.6 Summary . . . 328
Bibliographical and Historical Notes . . . 329
Exercises . . . 330

16 Computational Semantics and Semantic Parsing 331

17 Information Extraction 332
17.1 Relation Extraction . . . 333
17.2 Relation Extraction Algorithms . . . 336
17.3 Extracting Times . . . 344
17.4 Extracting Events and their Times . . . 348
17.5 Template Filling . . . 350
17.6 Summary . . . 352
Bibliographical and Historical Notes . . . 353
Exercises . . . 354

18 Word Senses and WordNet 355
18.1 Word Senses . . . 356
18.2 Relations Between Senses . . . 358
18.3 WordNet: A Database of Lexical Relations . . . 360
18.4 Word Sense Disambiguation . . . 363
18.5 Alternate WSD algorithms and Tasks . . . 366
18.6 Using Thesauruses to Improve Embeddings . . . 369
18.7 Word Sense Induction . . . 369
18.8 Summary . . . 370
Bibliographical and Historical Notes . . . 371
Exercises . . . 372

19 Semantic Role Labeling 373
19.1 Semantic Roles . . . 374
19.2 Diathesis Alternations . . . 375
19.3 Semantic Roles: Problems with Thematic Roles . . . 376
19.4 The Proposition Bank . . . 377
19.5 FrameNet . . . 378
19.6 Semantic Role Labeling . . . 380
19.7 Selectional Restrictions . . . 384
19.8 Primitive Decomposition of Predicates . . . 389
19.9 Summary . . . 390
Bibliographical and Historical Notes . . . 390
Exercises . . . 392

20 Lexicons for Sentiment, Affect, and Connotation 393
20.1 Defining Emotion . . . 394
20.2 Available Sentiment and Affect Lexicons . . . 395
20.3 Creating Affect Lexicons by Human Labeling . . . 398
20.4 Semi-supervised Induction of Affect Lexicons . . . 399
20.5 Supervised Learning of Word Sentiment . . . 402
20.6 Using Lexicons for Sentiment Recognition . . . 406
20.7 Other tasks: Personality . . . 407
20.8 Affect Recognition . . . 408
20.9 Lexicon-based methods for Entity-Centric Affect . . . 410

thuvienso.net

20.10 Connotation Frames 411
20.11 Summary 413
Bibliographical and Historical Notes 413

21 Coreference Resolution **415**
21.1 Coreference Phenomena: Linguistic Background 418
21.2 Coreference Tasks and Datasets 423
21.3 Mention Detection 424
21.4 Architectures for Coreference Algorithms 427
21.5 Classifiers using hand-built features 429
21.6 A neural mention-ranking algorithm 430
21.7 Evaluation of Coreference Resolution 434
21.8 Winograd Schema problems 435
21.9 Gender Bias in Coreference 436
21.10 Summary 438
Bibliographical and Historical Notes 438
Exercises 441

22 Discourse Coherence **442**
22.1 Coherence Relations 444
22.2 Discourse Structure Parsing 447
22.3 Centering and Entity-Based Coherence 451
22.4 Representation learning models for local coherence 456
22.5 Global Coherence 458
22.6 Summary 460
Bibliographical and Historical Notes 461
Exercises 463

23 Question Answering **464**
23.1 Information Retrieval 465
23.2 IR-based Factoid Question Answering 473
23.3 Entity Linking 477
23.4 Knowledge-based Question Answering 482
23.5 Using Language Models to do QA 484
23.6 Classic QA Models 485
23.7 Evaluation of Factoid Answers 488
Bibliographical and Historical Notes 489
Exercises 491

24 Chatbots & Dialogue Systems **492**
24.1 Properties of Human Conversation 493
24.2 Chatbots 496
24.3 GUS: Simple Frame-based Dialogue Systems 504
24.4 The Dialogue-State Architecture 508
24.5 Evaluating Dialogue Systems 517
24.6 Dialogue System Design 520
24.7 Summary 522
Bibliographical and Historical Notes 523
Exercises 525

25 Phonetics **526**
25.1 Speech Sounds and Phonetic Transcription 526

thuvienso.net

25.2 Articulatory Phonetics 527
25.3 Prosody 532
25.4 Acoustic Phonetics and Signals 534
25.5 Phonetic Resources 544
25.6 Summary 545
Bibliographical and Historical Notes 546
Exercises 547

26 Automatic Speech Recognition and Text-to-Speech 548
26.1 The Automatic Speech Recognition Task 549
26.2 Feature Extraction for ASR: Log Mel Spectrum 551
26.3 Speech Recognition Architecture 555
26.4 CTC 557
26.5 ASR Evaluation: Word Error Rate 562
26.6 TTS 564
26.7 Other Speech Tasks 569
26.8 Summary 570
Bibliographical and Historical Notes 570
Exercises 573

Bibliography 575

Subject Index 607

thuvienso.net

CHAPTER

1 Introduction

La dernière chose qu'on trouve en faisant un ouvrage est de savoir celle qu'il faut mettre la première.
[The last thing you figure out in writing a book is what to put first.]
Pascal

thuvienso.net

CHAPTER

2 Regular Expressions, Text Normalization, Edit Distance

```
User:  I am unhappy.
ELIZA: DO YOU THINK COMING HERE WILL HELP YOU NOT TO BE UNHAPPY
User:  I need some help, that much seems certain.
ELIZA: WHAT WOULD IT MEAN TO YOU IF YOU GOT SOME HELP
User:  Perhaps I could learn to get along with my mother.
ELIZA: TELL ME MORE ABOUT YOUR FAMILY
User:  My mother takes care of me.
ELIZA: WHO ELSE IN YOU FAMILY TAKES CARE OF YOU
User:  My father.
ELIZA: YOUR FATHER
User:  You are like my father in some ways.
```

Weizenbaum (1966)

ELIZA

The dialogue above is from **ELIZA**, an early natural language processing system that could carry on a limited conversation with a user by imitating the responses of a Rogerian psychotherapist (Weizenbaum, 1966). ELIZA is a surprisingly simple program that uses pattern matching to recognize phrases like “I need X” and translate them into suitable outputs like “What would it mean to you if you got X?”. This simple technique succeeds in this domain because ELIZA doesn’t actually need to *know* anything to mimic a Rogerian psychotherapist. As Weizenbaum notes, this is one of the few dialogue genres where listeners can act as if they know nothing of the world. Eliza’s mimicry of human conversation was remarkably successful: many people who interacted with ELIZA came to believe that it really *understood* them and their problems, many continued to believe in ELIZA’s abilities even after the program’s operation was explained to them (Weizenbaum, 1976), and even today such **chatbots** are a fun diversion.

chatbots

Of course modern conversational agents are much more than a diversion; they can answer questions, book flights, or find restaurants, functions for which they rely on a much more sophisticated understanding of the user’s intent, as we will see in Chapter 24. Nonetheless, the simple pattern-based methods that powered ELIZA and other chatbots play a crucial role in natural language processing.

We’ll begin with the most important tool for describing text patterns: the **regular expression**. Regular expressions can be used to specify strings we might want to extract from a document, from transforming “I need X” in Eliza above, to defining strings like *$199* or *$24.99* for extracting tables of prices from a document.

text normalization

We’ll then turn to a set of tasks collectively called **text normalization**, in which regular expressions play an important part. Normalizing text means converting it to a more convenient, standard form. For example, most of what we are going to do with language relies on first separating out or **tokenizing** words from running text, the task of **tokenization**. English words are often separated from each other by whitespace, but whitespace is not always sufficient. *New York* and *rock ’n’ roll* are sometimes treated as large words despite the fact that they contain spaces, while sometimes we’ll need to separate *I’m* into the two words *I* and *am*. For processing tweets or texts we’ll need to tokenize **emoticons** like `:)` or **hashtags** like `#nlproc`.

tokenization