

Przetwarzanie języka naturalnego w praktyce : przewodnik po budowie rzeczywistych systemów NLP / Sowmya Vajjala, Bodhisattwa Majumder, Anuj Gupta, Harshit Surana. – Gliwice, © 2023

Spis treści

Słowo wstępne	15
Przedmowa	17
CZĘŚĆ I. Podstawy	27
1. NLP – elementarz	29
NLP w rzeczywistym świecie	30
Zadania NLP	32
Czym jest język?	33
Podstawowe elementy języka	34
Dlaczego NLP jest trudnym wyzwaniem?	37
Uczenie maszynowe, uczenie głębokie i NLP — przegląd	38
Podejścia do NLP	40
NLP oparte na heurystyce	40
Uczenie maszynowe w NLP	42
Uczenie głębokie w NLP	46
Dlaczego uczenie głębokie nie jest jeszcze „srebrną kulą” NLP?	51
Przewodnik po NLP — agenty konwersacyjne	54
Podsumowanie	55
Bibliografia	56
2. Potok NLP	59
Pozyskiwanie danych	61
Ekstrakcja i oczyszczanie tekstu	64
Parsowanie i oczyszczanie HTML-a	64
Normalizacja Unikodu	66
Poprawianie pisowni	67
Poprawianie błędów specyficzne dla systemu	68
Przetwarzanie wstępne	70
Czynności wstępne	70
Częste czynności	73
Inne czynności	76
Przetwarzanie zaawansowane	77
Inżynieria cech	80
Klasyczny potok NLP/ML	80
Potok DL	82
Modelowanie	82

Zacznij od prostej heurystyki	82
Budowanie modelu	84
Budowanie ostatecznego modelu	84
Ewaluacja	87
Ewaluacja wewnętrzna	88
Ewaluacja zewnętrzna	90
Fazy następujące po modelowaniu	91
Wdrażanie	91
Monitorowanie	92
Aktualizowanie modelu	92
Praca z innymi językami	93
Studium przypadku	94
Podsumowanie	95
Bibliografia	95
3. Reprezentacja tekstu	99
Modele przestrzeni wektorowej	102
Proste metody wektoryzacji	102
Kodowanie one-hot	103
Worek słów	104
Worek n-gramów	106
TF-IDF	107
Reprezentacje rozproszone	109
Osadzenia słów	110
Ponad słowa	118
Reprezentacje rozproszone na poziomach wyższych niż słowa i znaki	120
Uniwersalne reprezentacje tekstu	121
Wizualizacja osadzeń	123
Ręcznie utworzone reprezentacje cech	126
Podsumowanie	128
Bibliografia	128
CZĘŚĆ II. Niezbędnik	131
4. Klasyfikacja tekstu	133
Zastosowania	134
Potok budowania systemów klasyfikacji tekstu	137
Prosty klasyfikator bez potoku klasyfikacji tekstu	138
Używanie istniejących interfejsów API do klasyfikacji tekstu	139
Jeden potok, wiele klasyfikatorów	139
Naiwny klasyfikator bayesowski	140
Regresja logistyczna	144
Maszyna wektorów nośnych	145
Osadzenia neuronowe w klasyfikacji tekstu	147
Osadzenia słów	147
Osadzenia podsłów i fastText	148

Osadzenia dokumentów	150
Uczenie głębokie w klasyfikacji tekstu	152
Sieci CNN do klasyfikacji tekstu	155
Sieci LSTM do klasyfikacji tekstu	156
Klasyfikacja tekstu z wykorzystaniem dużych, wstępnie wytrenowanych modeli językowych	157
Interpretacja modeli klasyfikacji tekstu	158
Wyjaśnianie prognoz klasyfikatora za pomocą Lime'a	159
Uczenie się bez danych lub na mniejszej ilości danych i adaptowanie modeli do nowych dziedzin	160
Brak danych treningowych	160
Mało danych treningowych — nauka aktywna i adaptacja dziedzinowa	161
Studium przypadku — obsługa zgłoszeń problemów	163
Praktyczne rady	165
Podsumowanie	167
Bibliografia	167
5. Ekstrakcja informacji	170
Zastosowania IE	171
Zadania IE	172
Ogólny potok IE	174
Ekstrakcja fraz kluczowych	175
Implementowanie KPE	176
Praktyczne rady	177
Rozpoznawanie nazwanych encji	177
Budowanie systemu NER	179
NER z wykorzystaniem istniejącej biblioteki	183
NER z wykorzystaniem nauki aktywnej	183
Praktyczne rady	184
Ujednoznacznianie i łączenie nazwanych encji	185
NEL z wykorzystaniem Azure API	186
Ekstrakcja relacji	188
Podejścia do RE	188
RE z wykorzystaniem Watson API	190
Inne zaawansowane zadania IE	191
Ekstrakcja informacji temporalnych	192
Ekstrakcja zdarzeń	193
Uzupełnianie szablonów	194
Studium przypadku	196
Podsumowanie	198
Bibliografia	199
6. Czatboty	203
Zastosowania	204
Prosty bot FAQ	205
Taksonomia czatbotów	206

Dialog ukierunkowany na cel	208
Pogawędki	208
Potok budowania systemów dialogowych	208
Szczegóły systemu dialogowego	210
Czatbot PizzaStop	211
Szczegółowa analiza komponentów systemu dialogowego	220
Klasyfikacja aktu dialogowego	221
Identyfikacja slotów	222
Generowanie odpowiedzi	223
Systemy dialogowe z przykładami kodu	223
Inne potoki dialogowe	228
Podejście kompleksowe	228
Generowanie dialogu poprzez uczenie głębokie ze wzmocnieniem	228
Człowiek w pętli	229
Rasa NLU	231
Studium przypadku — polecenie przepisów	233
Korzystanie z istniejących platform	234
Czatboty generatywne o strukturze otwartej	236
Podsumowanie	237
Bibliografia	238
7. Tematy w skrócie	240
Wyszukiwanie i zwracanie informacji	242
Komponenty wyszukiwarki	244
Typowy potok wyszukiwarki korporacyjnej	246
Budowanie wyszukiwarki — przykład	247
Studium przypadku — wyszukiwarka dla księgarni	248
Modelowanie tematyczne	249
Trenowanie modelu tematycznego — przykład	253
Co dalej?	254
Streszczanie tekstu	255
Zastosowania streszczania	256
Konfigurowanie narzędzia streszczającego — przykład	257
Praktyczne rady	257
Systemy rekomendujące dane tekstowe	259
Tworzenie systemu rekomendacji książek — przykład	260
Praktyczne rady	261
Tłumaczenie maszynowe	262
Używanie interfejsu API do tłumaczenia maszynowego — przykład	262
Praktyczne rady	263
Systemy odpowiadania na pytania	264
Budowanie własnego systemu odpowiadania na pytania	266
Poszukiwanie głębszych odpowiedzi	266
Podsumowanie	267
Bibliografia	267

CZĘŚĆ III. Praktyka	271
8. Media społecznościowe	273
Aplikacje	274
Unikatowe wyzwania	276
Przetwarzanie języka naturalnego w danych społecznościowych	281
Chmura wyrazowa	281
Jonizator SMTD	282
Popularne tematy	283
Odczucia użytkowników Twittera	285
Wstępne przetwarzanie danych SMTD	287
Reprezentacja tekstu w SMTD	290
Obsługa klienta w kanałach społecznościowych	293
Memy i fake newsy	295
Identyfikowanie memów	295
Fake newsy	296
Podsumowanie	298
Bibliografia	298
9. E-commerce i handel detaliczny	302
Katalog e-commerce	302
Analiza recenzji	303
Wyszukiwanie produktów	303
Rekomendacje produktów	304
Wyszukiwanie w e-commerce	304
Budowanie katalogu e-commerce	306
Ekstrakcja atrybutów	306
Kategoryzacja i taksonomia produktów	311
Wzbogacanie produktów	313
Deduplikacja i dopasowywanie produktów	315
Analiza recenzji	317
Analiza odczuć	318
Aspektowa analiza odczuć	319
Łączenie ocen ogólnych z aspektami	321
Rozumienie aspektów	322
Rekomendacje w e-commerce	324
Studium przypadku — produkty substytucyjne i komplementarne	325
Podsumowanie	328
Bibliografia	328
10. Opieka zdrowotna, finanse i prawo	330
Opieka zdrowotna	330
Dokumentacja zdrowotna i medyczna	331
Ustalanie priorytetów i rozliczanie pacjentów	331
Nadzór farmakologiczny	332
Systemy wspomaganie decyzji klinicznych	332

Asystenty zdrowotne	333
Elektroniczna dokumentacja medyczna	334
Monitorowanie zdrowia psychicznego	342
Ekstrakcja i analiza informacji medycznych	344
Finanse i prawo	346
Zastosowania NLP w finansach	348
NLP w krajobrazie prawnym	350
Podsumowanie	352
Bibliografia	354
CZĘŚĆ IV. Synteza	357
11. Kompleksowy proces NLP	359
Powrót do potoku NLP — wdrażanie oprogramowania NLP	360
Przykładowy scenariusz	361
Budowanie i utrzymywanie dojrzałego systemu	363
Znajdowanie lepszych cech	363
Iteracyjne rozwijanie istniejących modeli	364
Odtwarzalność kodu i modelu	365
Rozwiązywanie problemów i interpretowalność	366
Monitorowanie	368
Minimalizowanie długu technicznego	369
Automatyzacja uczenia maszynowego	370
Proces data science	374
Proces KDD	374
Proces Microsoft Team Data Science	375
Droga do sukcesu AI w Twojej organizacji	377
Zespół	377
Właściwy problem i właściwe oczekiwania	378
Dane i czas	379
Dobry proces	380
Inne aspekty	381
Spojrzenie poza horyzont	382
Ostatnie słowa	385
Bibliografia	386
Skorowidz	391