

УДК 004.438Java

ББК 32.973.2

P49

**Риз, Ричард.**

P49      Обработка естественного языка на Java / Р. Риз ; пер. с англ. А. В. Снастина. — 2-е изд., эл. — 1 файл pdf : 266 с. — Москва : ДМК Пресс, 2023. — Систем. требования: Adobe Reader XI либо Adobe Digital Editions 4.5 ; экран 10". — Текст : электронный.

ISBN 978-5-89818-333-2

Обработка естественного языка (Natural Language Procession — NLP) представляет собой важную область разработки прикладного ПО и, с учетом современных задач ИТ, в будущем эта важность будет только возрастать. Уже сейчас наблюдается рост потребности в приложениях, работающих с естественными языками на основе NLP-методик.

В данной книге рассматриваются способы организации автоматической обработки текста с применением таких методик, как полнотекстовый поиск, правильное распознавание имен, кластеризация, классификация, извлечение информации и составление аннотаций. Концепции обработки естественного языка излагаются таким образом, что даже читатели, не обладающие знаниями об этой технологии и о методах статистического анализа, смогут понять их.

УДК 004.438Java

ББК 32.973.2

**Электронное издание на основе печатного издания:** Обработка естественного языка на Java / Р. Риз ; пер. с англ. А. В. Снастина. — Москва : ДМК Пресс, 2016. — 264 с. — ISBN 978-5-97060-331-4. — Текст : непосредственный.

Все права защищены. Любая часть этой книги не может быть воспроизведена в какой бы то ни было форме и какими бы то ни было средствами без письменного разрешения владельцев авторских прав.

Материал, изложенный в данной книге, многократно проверен. Но поскольку вероятность технических ошибок все равно существует, издательство не может гарантировать абсолютную точность и правильность приводимых сведений. В связи с этим издательство не несет ответственности за возможные ошибки, связанные с использованием книги.

В соответствии со ст. 1299 и 1301 ГК РФ при устранении ограничений, установленных техническими средствами защиты авторских прав, правообладатель вправе требовать от нарушителя возмещения убытков или выплаты компенсации.

ISBN 978-5-89818-333-2

© 2015 Packt Publishing

© Оформление, перевод, ДМК Пресс, 2016

# Содержание

<b>Об авторе .....</b>	<b>10</b>
<b>О рецензентах.....</b>	<b>11</b>
<b>Предисловие .....</b>	<b>13</b>

## **Глава 1. Основы обработки естественного языка ..... 18**

Что такое обработка естественного языка.....	19
Для чего используется обработка естественного языка.....	21
Трудности обработки естественного языка .....	23
Обзор инструментальных средств обработки естественного языка .....	25
Apache OpenNLP.....	27
Stanford NLP.....	28
LingPipe.....	30
GATE.....	31
UIMA .....	31
Обзор задач обработки текста .....	32
Поиск фрагментов текста.....	33
Поиск предложений.....	35
Поиск людей и прочих именованных объектов .....	37
Определение частей речи.....	40
Классификация текстов и документов.....	41
Выделение взаимоотношений.....	42
Комплексные методики обработки .....	44
О моделях обработки естественного языка .....	45
Определение задачи (типа задачи).....	45
Выбор модели.....	46
Создание и обучение модели.....	46
Проверка модели.....	47
Практическое использование модели .....	47
Подготовка данных .....	47
Резюме .....	50

## **Глава 2. Поиск фрагментов текста ..... 52**

Части или фрагменты текста .....	53
Что такое токенизация.....	53
Использование токенизаторов.....	56
Простые токенизаторы языка Java .....	57
Использование класса Scanner .....	57
Определение разделителя .....	58
Использование метода split().....	59

Использование класса BreakIterator.....	60
Использование класса StreamTokenizer.....	61
Использование класса StringTokenizer .....	63
Проблемы производительности при выполнении токенизации штатными средствами Java.....	64
Прикладные программные интерфейсы NLP для токенизации.....	64
Использование класса Tokenizer из библиотеки OpenNLP .....	65
Использование класса SimpleTokenizer .....	65
Использование класса WhitespaceTokenizer.....	66
Использование класса TokenizerME.....	66
Использование токенизатора из библиотеки Stanford.....	67
Использование класса PTBTokenizer.....	68
Использование класса DocumentPreprocessor .....	69
Использование конвейера.....	70
Использование токенизаторов из библиотеки LingPipe .....	71
Обучение токенизатора поиску заданных элементов текста .....	72
Сравнение токенизаторов.....	76
Нормализация.....	76
Преобразование букв в нижний регистр .....	77
Удаление шумовых слов.....	78
Создание класса StopWords.....	78
Использование библиотеки LingPipe для удаления шумовых слов .....	80
Использование стемминга.....	82
Использование инструмента стемминга Porter Stemmer.....	82
Стемминг с использованием библиотеки LingPipe.....	83
Использование лемматизации .....	84
Использование класса StanfordLemmatizer .....	85
Поддержка лемматизации в библиотеке OpenNLP .....	86
Нормализация с применением конвейера .....	88
Резюме .....	89

### **Глава 3. Поиск предложений.....91**

Процесс разрешения границ предложений.....	91
Затруднения при разрешении границ предложений.....	92
Правила разрешения границ предложений в классе HeuristicSentenceModel библиотеки LingPipe .....	95
Простые средства разрешения границ предложений в языке Java .....	96
Использование регулярных выражений .....	97
Использование класса BreakIterator.....	99
Использование библиотек NLP API .....	101
Использование библиотеки OpenNLP.....	101
Использование класса SentenceDetectorME.....	101

Использование метода sentPosDetect .....	103
Использование библиотеки Stanford API.....	104
Использование класса PTBTokenizer.....	104
Использование класса DocumentPreprocessor.....	108
Использование класса StanfordCoreNLP.....	111
Использование библиотеки LingPipe.....	112
Использование класса IndoEuropeanSentenceModel.....	113
Использование класса SentenceChunker.....	115
Использование класса MedlineSentenceModel .....	116
Обучение модели SentenceDetector .....	117
Использование обученной модели .....	120
Вычисление характеристик модели с помощью класса SentenceDetectorEvaluator .....	120
Резюме .....	122

## **Глава 4. Поиск людей и именованных объектов..... 123**

Трудности, возникающие при распознавании и идентификации именованных объектов .....	124
Методики распознавания именованных объектов.....	125
Списки и регулярные выражения.....	127
Статистические классификаторы.....	127
Использование регулярных выражений для распознавания и идентификации именованных объектов .....	128
Использование регулярных выражений в языке Java для поиска объектов .....	128
Использование класса RegExChunker из библиотеки LingPipe .....	131
Использование библиотек NLP .....	132
Использование библиотеки OpenNLP для поиска именованных объектов .....	133
Вычисление точности идентификации именованного объекта .....	135
Использование других типов именованных объектов.....	136
Одновременная обработка нескольких типов объектов.....	137
Использование библиотеки Stanford API для поиска именованных объектов .....	138
Использование библиотеки LingPipe для поиска именованных объектов .....	140
Использование моделей именованных объектов из библиотеки LingPipe.....	140
Использование класса ExactDictionaryChunker .....	142
Обучение модели .....	145
Оценка характеристик модели .....	147
Резюме .....	148

## Глава 5. Определение частей речи ..... 150

Процесс разметки.....	150
Важное значение инструментов разметки по частям речи .....	154
Трудности в идентификации частей речи.....	155
Использование библиотек NLP API .....	157
Использование инструментов разметки по частям речи из библиотеки OpenNLP .....	158
Использование класса POSTaggerME для разметки по частям речи.....	159
Использование средств поверхностного синтаксического анализа из библиотеки OpenNLP .....	161
Использование класса POSDictionary.....	164
Использование инструментов разметки по частям речи из библиотеки Stanford.....	168
Использование класса MaxentTagger.....	168
Использование класса MaxentTagger для разметки текста на смс-языке .....	172
Использование конвейера, поддерживаемого библиотекой Stanford, для POS-разметки .....	172
Использование инструментов разметки по частям речи из библиотеки LingPipe .....	175
Использование класса HmmDecoder с тегами Best_First.....	176
Использование класса HmmDecoder с тегами NBest .....	177
Определение степени достоверности назначенного тега с помощью класса HmmDecoder .....	179
Обучение модели POSModel из библиотеки OpenNLP .....	180
Резюме .....	182

## Глава 6. Классификация текстов и документов ..... 184

Как используется классификация текста.....	185
Особенности анализа эмоциональной окраски текста .....	187
Методики классификации текста .....	189
Использование библиотек NLP API для классификации текста .....	190
Использование библиотеки OpenNLP.....	190
Обучение классификационной модели из библиотеки OpenNLP ....	190
Использование класса DocumentCategorizerME для классификации текста .....	192
Использование библиотеки Stanford API.....	194
Использование класса ColumnDataClassifier для классификации текста .....	195
Использование конвейера, поддерживаемого библиотекой Stanford для анализа эмоциональной окраски текста .....	198
Использование библиотеки LingPipe для классификации текста .....	200

Подготовка обучающего текста с помощью класса Classified .....	200
Использование других обучающих категорий .....	202
Классификация текста с помощью библиотеки LingPipe .....	203
Анализ эмоциональной окраски текста с помощью библиотеки LingPipe .....	204
Определение языка документа с помощью библиотеки LingPipe .....	206
Резюме .....	208

## **Глава 7. Использование синтаксического анализатора (парсера) для выделения взаимосвязей ..... 209**

Типы взаимосвязей .....	211
Деревья синтаксического анализа .....	212
Использование полученных взаимосвязей .....	214
Извлечение взаимосвязей из текста .....	217
Использование библиотек NLP API .....	217
Использование библиотеки OpenNLP .....	218
Использование библиотеки Stanford API .....	221
Использование класса LexicalizedParser .....	221
Использование класса TreePrint .....	222
Поиск зависимостей между словами с помощью класса GrammaticalStructure .....	223
Поиск референциального тождества между объектами .....	225
Извлечение взаимосвязей для системы «вопрос–ответ» .....	228
Поиск взаимосвязей (зависимостей) между словами .....	228
Определение типа вопроса .....	230
Поиск ответа на вопрос .....	231
Резюме .....	233

## **Глава 8. Комплексные методики ..... 235**

Подготовка данных .....	236
Использование библиотеки Boilerpipe для извлечения текста из HTML-документов .....	236
Использование библиотеки POI для извлечения текста из документов в формате Word .....	239
Использование библиотеки PDFBox для извлечения текста из документов в формате PDF .....	242
Конвейеры .....	243
Использование конвейера, поддерживаемого библиотекой Stanford .....	244
Использование нескольких ядер процессора для конвейера библиотеки Stanford .....	249
Создание конвейера для текстового поиска .....	251
Резюме .....	256

## **Предметный указатель ..... 258**