

УДК 004.738.52

ББК 32.972.53

И59

**Ингерсолл, Грант С.**

**И59**      Обработка неструктурированных текстов. Поиск, организация и манипулирование / Г. С. Ингерсолл, Т. С. Мортон, Э. Л. Фэррис ; пер. с англ. А. А. Слинкина. — 2-е изд., эл. — 1 файл pdf : 416 с. — Москва : ДМК Пресс, 2023. — Систем. требования: Adobe Reader XI либо Adobe Digital Editions 4.5 ; экран 10". — Текст : электронный.

ISBN 978-5-89818-308-0

В книге описаны инструменты и методы обработки неструктурированных текстов. Прочитав ее, вы научитесь пользоваться полнотекстовым поиском, распознавать имена собственные, производить кластеризацию, пометку, извлечение информации и автореферирование. Знакомство с фундаментальными принципами сопровождается изучением реальных применений.

Издание предназначено для читателей без подготовки в области математической статистики и обработки естественных языков. Примеры написаны на Java, но сами идеи могут быть реализованы на любом языке программирования.

УДК 004.738.52

ББК 32.972.53

**Электронное издание на основе печатного издания:** Обработка неструктурированных текстов. Поиск, организация и манипулирование / Г. С. Ингерсолл, Т. С. Мортон, Э. Л. Фэррис ; пер. с англ. А. А. Слинкина. — Москва : ДМК Пресс, 2015. — 414 с. — ISBN 978-5-97060-144-0. — 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 978-5-89818-308-0

© 2013 by Manning Publications Co.

© Оформление, перевод на русский язык  
ДМК Пресс, 2015



# ОГЛАВЛЕНИЕ

<b>Предисловие .....</b>	<b>11</b>
<b>Вступление .....</b>	<b>12</b>
<b>Благодарности .....</b>	<b>16</b>
<b>Об этой книге .....</b>	<b>19</b>
Предполагаемая аудитория .....	19
Структура книги .....	20
Автор в сети .....	21
<b>Об иллюстрации на обложке .....</b>	<b>23</b>
<b>Глава 1. Готовимся к приручению текста .....</b>	<b>24</b>
1.1. Почему так важна задача обработки текста .....	25
1.2. Предварительный обзор фактографической вопросно-ответной системы .....	28
1.2.1. Здравствуй, доктор Франкенштейн .....	29
1.3. Понять смысл текста трудно .....	32
1.4. Прирученный текст .....	35
1.5. Текст и интеллектуальные приложения: поиск и не только .....	38
1.5.1. Поиск и сопоставление .....	39
1.5.2. Извлечение информации .....	40
1.5.3. Группировка информации .....	41
1.5.4. Интеллектуальное приложение .....	41
1.6. Резюме .....	42
1.7. Ресурсы .....	42
<b>Глава 2. Основы приручения текста .....</b>	<b>44</b>
2.1. Основы лингвистики .....	45
2.1.1. Категории слов .....	46
2.1.2. Словосочетания и части предложения .....	48
2.1.3. Морфология .....	50
2.2. Популярные инструменты для обработки текста .....	51
2.2.1. Инструменты для манипуляций со строками .....	52
2.2.2. Лексемы и лексический анализ .....	52
2.2.3. Частеречная разметка .....	55

2.2.4. Стемминг .....	57
2.2.5. Распознавание предложений .....	59
2.2.6. Грамматика и грамматический анализ .....	61
2.2.7. Моделирование последовательности .....	63
2.3. Предобработка и выделение содержимого из файлов в распространенных форматах .....	65
2.3.1. Важность предобработки .....	65
2.3.2. Извлечение содержимого с помощью Apache Tika .....	68
2.4. Резюме .....	71
2.5. Ресурсы .....	72
<b>Глава 3. Поиск .....</b>	<b>73</b>
3.1. Пример фасетного поиска: Amazon.com .....	74
3.2. Введение в концепции поиска .....	77
3.2.1. Индексирование содержимого .....	78
3.2.2. Ввод запроса пользователем .....	81
3.2.3. Ранжирование документов с помощью векторной модели .....	85
3.2.4. Отображение результатов .....	89
3.3. Введение в поисковый сервер Apache Solr .....	92
3.3.1. Первый запуск Solr .....	93
3.3.2. Основные концепции Solr .....	95
3.4. Индексирование содержимого с помощью Apache Solr ...	100
3.4.1. Индексирование данных в формате XML .....	101
3.4.2. Извлечение и индексирование содержимого с помощью Solr и Apache Tika .....	103
3.5. Поиск по содержимому в Apache Solr .....	107
3.5.1. Параметры запроса к Solr .....	108
3.5.2. Построение фасетов по извлеченному содержимому .....	112
3.6. Факторы, влияющие на производительность поиска .....	115
3.6.1. Оценка качественных показателей .....	116
3.6.2. Оценка количественных показателей .....	121
3.7. Повышение производительности поиска .....	122
3.7.1. Совершенствование на уровне оборудования .....	123
3.7.2. Повышение качества анализа .....	124
3.7.3. Повышение качества обработки запросов .....	127
3.7.4. Альтернативные модели оценивания .....	130
3.7.5. Способы повышения производительности Solr .....	131
3.8. Альтернативные поисковые системы .....	134
3.9. Резюме .....	136
3.10. Ресурсы .....	136
<b>Глава 4. Неточное сравнение строк .....</b>	<b>138</b>
4.1. Различные подходы к неточному сравнению строк .....	140
4.1.1. Меры, основанные на множестве общих символов .....	141

4.1.2. Редакционные расстояния .....	144
4.1.3. <i>N</i> -граммное редакционное расстояние .....	148
4.2. Нахождение строк, неточно совпадающих с данной .....	150
4.2.1. Использование префиксного сравнения в Solr .....	151
4.2.2. Использование префиксных деревьев для префиксного сравнения .....	152
4.2.3. Сравнение с помощью <i>l</i> -грамм .....	158
4.3. Использование неточного сравнения строк в приложениях .....	159
4.3.1. Добавления механизма автозаполнения к поиску .....	160
4.3.2. Проверка орфографии запроса .....	164
4.3.3. Сопоставление записей .....	170
4.4. Резюме .....	177
4.5. Ресурсы .....	177

## **Глава 5. Распознавание имен людей, географических названий и других сущностей .... 178**

5.1. Различные подходы к распознаванию именованных сущностей .....	180
5.1.1. Применение правил для распознавания имен и названий ...	181
5.1.2. Применение статистических классификаторов для распознавания имен и названий .....	182
5.2. Основы распознавания сущностей в OpenNLP .....	184
5.2.1. Нахождение имен и названий с помощью OpenNLP .....	185
5.2.2. Интерпретация имен, распознанных OpenNLP .....	187
5.2.3. Фильтрация имен на основе вероятности .....	188
5.3. Подробнее о распознавании сущностей в OpenNLP .....	189
5.3.1. Распознавание разнородных сущностей в OpenNLP .....	189
5.3.2. Под капотом: как в OpenNLP распознаются имена .....	193
5.4. Качество работы OpenNLP .....	196
5.4.1. Качество результатов .....	196
5.4.2. Производительность .....	197
5.4.3. Потребление памяти в OpenNLP .....	198
5.5. Настройка OpenNLP для распознавания сущностей в новой предметной области .....	200
5.5.1. Зачем и как обучают модель .....	200
5.5.2. Обучение модели OpenNLP .....	202
5.5.3. Изменение входных данных для модели .....	204
5.5.4. Другой способ моделирования имен .....	206
5.6. Резюме .....	209
5.7. Ресурсы .....	210

## **Глава 6. Кластеризация текста..... 211**

6.1. Кластеризация документов в Google News .....	212
---	-----

6.2. Основы кластеризации .....	213
6.2.1. Три типа текстов, поддающихся кластеризации .....	214
6.2.2. Выбор алгоритма кластеризации .....	216
6.2.3. Определение сходства .....	218
6.2.4. Пометка результатов .....	219
6.2.5. Как оценивать результаты кластеризации .....	220
6.3. Подготовка к созданию простого приложения кластеризации .....	222
6.4. Кластеризация результатов поиска с помощью Carrot <sup>2</sup> ....	223
6.4.1. Использование Carrot <sup>2</sup> API .....	224
6.4.2. Кластеризация результатов поиска Solr с помощью Carrot <sup>2</sup> .....	226
6.5. Кластеризация наборов документов с помощью Apache Mahout.....	229
6.5.1. Подготовка данных к кластеризации .....	230
6.5.2. Кластеризация методом К-средних .....	234
6.6. Тематическое моделирование с помощью Apache Mahout.....	239
6.7. Качество кластеризации .....	243
6.7.1. Отбор и уменьшение числа признаков.....	243
6.7.2. Производительность и качество Carrot2 .....	246
6.7.3. Тесты производительности кластеризации в Mahout.....	247
6.8. Благодарности .....	254
6.9. Резюме .....	254
6.10. Ресурсы .....	255

## **Глава 7. Классификация, категоризация и пометка ..... 257**

7.1. Введение в классификацию и категоризацию .....	260
7.2. Процесс классификации .....	263
7.2.1. Выбор схемы классификации .....	265
7.2.2. Отбор признаков для категоризации .....	266
7.2.3. Важность обучающих данных .....	268
7.2.4. Оценка качества классификатора.....	271
7.2.5. Внедрение классификатора в эксплуатацию .....	274
7.3. Построение классификаторов документов с помощью Apache Lucene .....	276
7.3.1. Классификация текстов с помощью Lucene .....	276
7.3.2. Подготовка обучающих данных для классификатора MoreLikeThis .....	279
7.3.3. Обучение классификатора MoreLikeThis .....	281
7.3.4. Классификация документов с помощью классификатора MoreLikeThis .....	285

7.3.5. Тестирование классификатора MoreLikeThis .....	288
7.3.6. Классификатор MoreLikeThis в производственной системе .....	291
7.4. Обучение наивного байесовского классификатора в Apache Mahout .....	292
7.4.1. Наивная байесовская классификация текста .....	293
7.4.2. Подготовка обучающих данных .....	294
7.4.3. Резервирование тестовых данных .....	298
7.4.4. Обучение классификатора .....	299
7.4.5. Тестирование классификатора .....	300
7.4.6. Усовершенствованный процесс бутстрапинга .....	302
7.4.7. Интеграция байесовского классификатора Mahout с Solr .....	304
7.5. Классификация документов с помощью OpenNLP .....	308
7.5.1. Регрессионные модели и классификация документов методом максимальной энтропии .....	309
7.5.2. Подготовка обучающих данных для классификатора документов на основе алгоритма максимальной энтропии .....	313
7.5.3. Обучение классификатора документов на основе алгоритма максимальной энтропии .....	314
7.5.4. Тестирование классификатора документов на основе алгоритма максимальной энтропии .....	320
7.5.5. Классификатор документов на основе алгоритма максимальной энтропии в производственной системе .....	322
7.6. Построение рекомендателя меток с помощью Apache Solr .....	323
7.6.1. Подготовка обучающих данных для рекомендателя меток .....	327
7.6.2. Подготовка обучающих данных .....	329
7.6.3. Обучение рекомендателя меток на основе Solr .....	330
7.6.4. Создание рекомендаций меток .....	332
7.6.5. Оценивание рекомендателя меток .....	335
7.7. Резюме .....	338
7.8. Ресурсы .....	340

## Глава 8. Пример вопросно-ответной системы..... 341

8.1. Основы вопросно-ответной системы .....	343
8.2. Установка и запуск QA-системы .....	345
8.3. Архитектура демонстрационной вопросно-ответной системы .....	347
8.4. Установление смысла вопроса и порождение ответов ....	351
8.4.1. Обучение классификатора типов ответов .....	351
8.4.2. Разбиение вопроса на блоки .....	356
8.4.3. Вычисление типа ответа .....	357
8.4.4. Генерация запроса .....	361
8.4.5. Ранжирование фрагментов-кандидатов .....	362

8.5. Усовершенствование системы .....	365
8.6. Резюме .....	365
8.7. Ресурсы .....	366

## **Глава 9. Неприрученный текст: на переднем крае ..... 367**

9.1. Семантика, дискурс и прагматика: высшие уровни NLP .....	368
9.1.1. Семантика.....	369
9.1.2. Дискурс.....	371
9.1.3. Прагматика .....	373
9.2. Реферирование документов и наборов документов .....	375
9.3. Извлечение отношений.....	377
9.3.1. Обзор имеющихся подходов .....	379
9.3.2. Оценка .....	383
9.3.3. Инструменты для извлечения отношений .....	383
9.4. Выявление важного содержимого и людей .....	384
9.4.1. Глобальная важность и авторитетность .....	386
9.4.2. Персональная важность .....	386
9.4.3. Ресурсы и ссылки на тему важности .....	387
9.5. Распознавание эмоций с помощью анализа тональности .....	388
9.5.1. Исторический обзор.....	389
9.5.2. Инструменты и данные.....	391
9.5.3. Базовый алгоритм определения тональности.....	392
9.5.4. Дополнительные темы .....	394
9.5.5. Библиотеки с открытым исходным кодом для анализа тональности .....	396
9.6. Межъязыковой информационный поиск .....	397
9.7. Резюме .....	399
9.8. Ресурсы .....	400

## **Предметный указатель ..... 403**