DOI 10.37882/2223–2966.2022.03–2.18

АНАЛИЗ ТЕХНОЛОГИИ КЛАССИФИКАЦИИ ТЕКСТА

ANALYSIS OF TEXT CLASSIFICATION TECHNOLOGY

Zheng Jing
Wei Xiaoyu

Summary. The article provides an overview of the main approaches to the analysis of textual information. Particular attention is paid to Text Mining technologies and a bag (or cloud) of words. An algorithm for the complex classification of texts is considered. The software solution for constructing the semantic core of the text in Python based on the collections library is described.

Keywords: classification of texts, semantic core, signs of documents, key phrases, distributive semantics, text mining, bag (or cloud) of words.

Чжэн Цзини
МГТУ им. Н.Э. Баумана
sofiazjy@gmail.com
Вэй Сяоюй
МГТУ им. Н.Э. Баумана
569006420@mail.ru

Аннотация. В статье дается обзор основных подходов к анализу текстовой информации. Особое внимание уделяется технологиям TextMining и мешка (или облака) слов. Рассмотрен алгоритм комплексной классификации текстов. Описывается программное решения построения семантического ядра текста на языке Python на базе библиотеки collections.

Ключевые слова: классификация текстов, семантическое ядро, признаки документов, ключевые фразы, дистрибутивная семантика, text mining, мешок (или облако) слов.

Сырые неструктурированные данные составляют не менее 90% информации, с которой имеют дело пользователи. Однако все эти огромные запасы данных бесполезны при отсутствии удобных инструментов поиска и/или быстрой классификации информации.

Вопросы классификации текстовой информации изучаются давно и плодотворно [см., например, 1–7]. Существующие сегодня системы классификация применяется, например, в таких задачах как группировка документов в intranet-сетях и на Web-сайтах, размещение документов в определенных папках, сортировка сообщений электронной почты, персонифицированная доставка новостей подписчикам.

Задача классификации (или рубрикации, или кластеризации, или распознавания) имеет две основные постановки — бинарную и мультиклассовую.

Бинарная классификация даёт ответ на какой-то один вопрос исследования в стиле «да/нет», например, является ли данное электронное письмо спамом или нет, или представляет ли литературный источник научный интерес или нет. В этом случае говорят о байесовской или наивной классификации информации.

Мультиклассовая классификация ставит перед исследователем задачу отнесения тематики документа к одному из тематических классов предметной области. Количество классов может в некоторых случаях достигать нескольких десятков, количество объектов и их атрибутов может быть очень большим; поэтому должны быть предусмотрены интеллектуальные механизмы оптимизации процесса классификации.

Мультиклассовая классификация может быть статической, в которой выходные категории заранее предопределены, или динамической, когда рубрики классификации формируются динамически в процессе обработки информации с использованием математических, лингвистических или онтологичеких моделей [10].

Технология эффективного анализа текста Text Mining способна выступить в роли концентратора, который делает экстракт из наиболее ключевой и значащей информации предметной области. К основным элементам Text Mining относятся суммаризация (summarization), выделение феноменов, понятий (feature extraction), кластеризация (clustering), классификация (classification), ответ на запросы (question answering), тематическое индексирование (thematic indexing) и поиск по ключевым словам (keyword searching). Также в некоторых случаях набор дополняют средства поддержки и создания таксономии (oftaxonomies) и тезаурусов (thesauri).

Результат — таксономия или визуальная карта, которая обеспечивает эффективный охват больших объемов данных. Семантические сети или анализ связей, которые определяют появление дескрипторов (ключевых фраз) в документе для обеспечения и навигации. Извлечение фактов предназначено для получения некоторых фактов из текста с целью улучшения классификации, поиска и кластеризации.

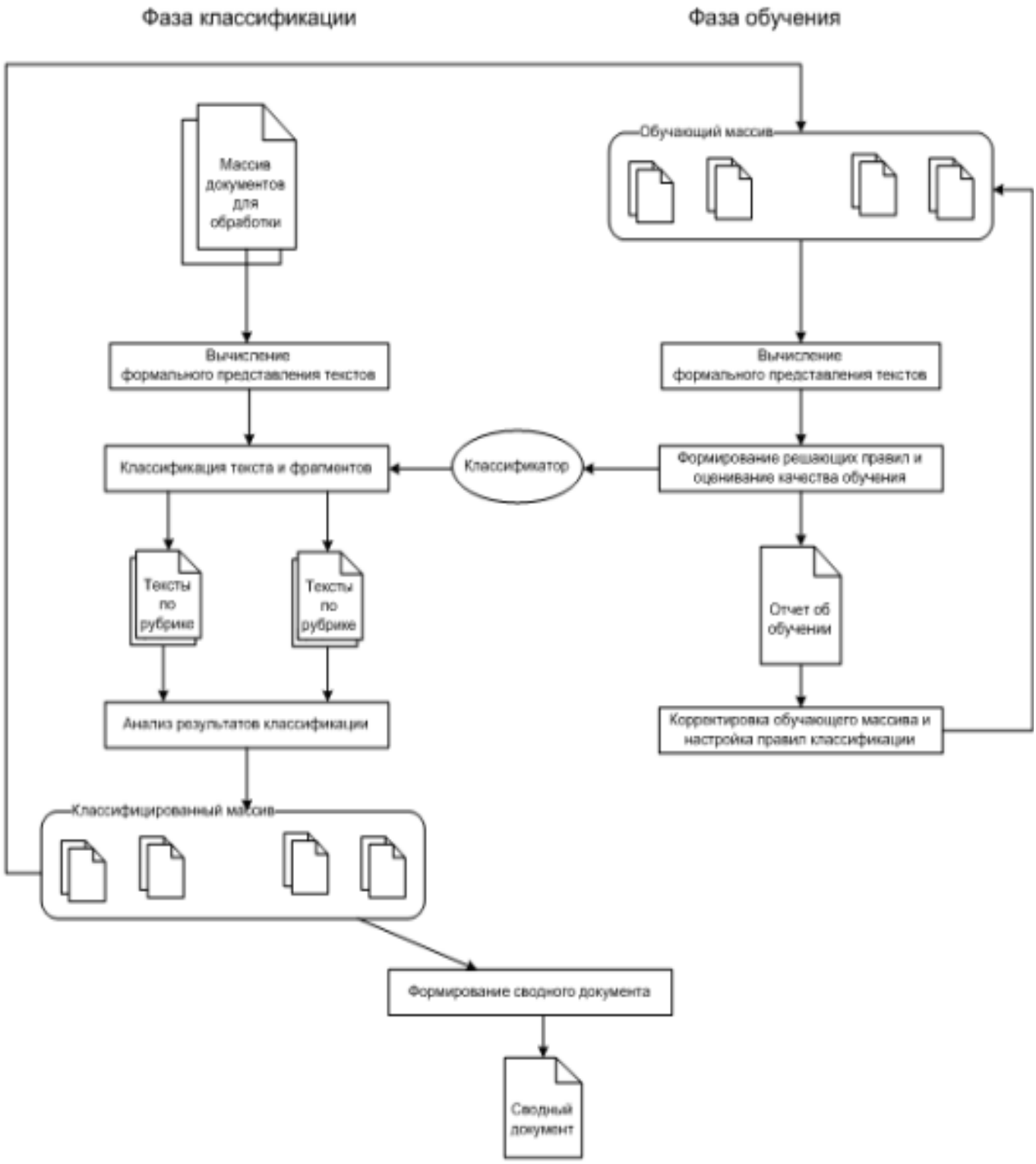

Рис. 1. Алгоритм комплексной классификации тестов [6, с.83]

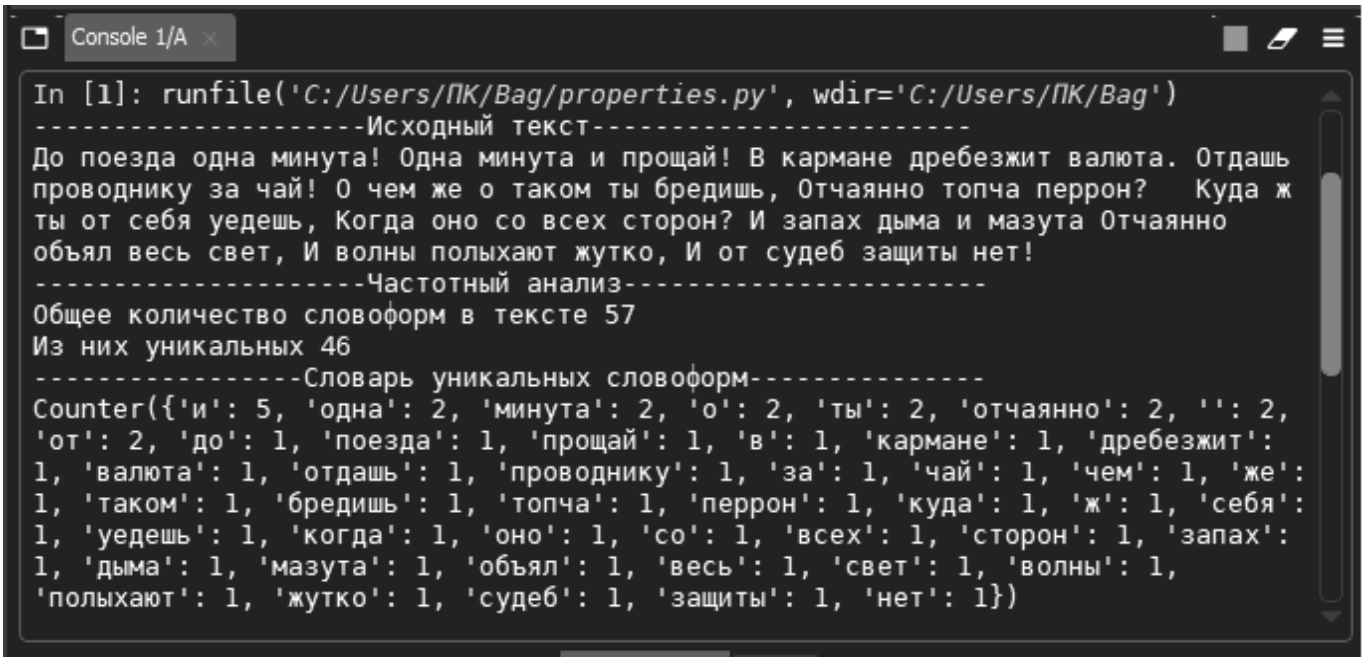

Рис. 2. Пример реализации частотного анализа

Вторая задача — кластеризация — выделение компактных подгрупп объектов с близкими свойствами. Система должна самостоятельно найти признаки и разделить объекты по подгруппам. Она, как правило, предшествует задаче классификации, поскольку позволяет определить группы объектов.

Можно назвать еще несколько задач технологии Text Mining, например, прогнозирование, которое состоит в том, чтобы предсказать по значениям одних признаков объекта значения остальных.

Еще одна задача — нахождение исключений, то есть поиск объектов, которые своими характеристиками сильно выделяются из общей массы. Для этого сначала выясняются средние параметры объектов, а потом исследуются те объекты, параметры которых наиболее сильно отличаются от средних значений. Как известно, поиск исключений широко применяется, например, в работе спецслужб. Подобный анализ часто проводится после классификации, для того чтобы выяснить, насколько последняя была точна.

Несколько отдельно от задачи кластеризации стоит задача поиска связанных признаков (полей, понятий) отдельных документов. От предсказания эта задача отличается тем, что заранее не известно, по каким именно признакам реализуется взаимосвязь; цель именно в том и состоит, чтобы найти связи признаков. Эта задача сходна с кластеризацией, но не по множеству документов, а по множеству присущих им признаков.

Общий алгоритм комплексной классификации текстов представлен на рисунке 1.

Технология Text Miner позволяет определять, насколько правдив тот или иной текстовый документ. Обнаружение лжи в документах производится путем анализа текста и выявления изменений стиля письма, которые могут возникать при попытке исказить или скрыть информацию [8].

Дистрибутивная семантика — модели построения метрик подобия текстов — существует несколько программных реализаций: Word2Vec, GloVe, AdaGram, Text2Vec, Seq2Vec и другие [9]. Анализ взаимозаменяемости слов (модель Skipgram), анализ ассоциированных слов (модель BagOfWords).

Технология мешка (или облака) слов заключается в следующем. Из текста исключаются знаки препинания и стоп-символы (предлоги, союзы, междометия). Текст переводится в какой-то один регистр, как правило, нижний. Все слова исходного текста приводятся к своей основной форме — для существительных — именительный падеж единственного числа в мужском роде [10].

Научная новизна заключается в том, что предлагается актуальное программное решения на языке Python на базе библиотеки collections, которое строит семантическое ядро текста в виде словаря уникальных слов и проводит его частотный анализ (рисунок 2).

Язык Python обладает большим спектром инструментов, делающим его наиболее перспективной платформой для анализа текстов. К ним, в частности, относятся библиотека NLTK для анализа текста, библиотека wordcloud для построения облака слов и модуль PyEnchant, который не только позволяет проверять правописание слов, но и предлагает варианты исправления ошибок, библиотека matplotlib для визуализации результатов.

ЛИТЕРАТУРА

1. Агеев М.С. и др. Автоматическая рубрикация текстов: методы и проблемы // Учен. зап. Казан. ун-та. Сер. Физ.-матем. науки. — 2008. — № 4 [Электронный ресурс] Режим доступа: https://cyberleninka.ru/article/n/avtomaticheskaya-rubrikatsiya-tekstov-metody-i-problemy
2. Амиева А.М. и др. Основные методики исследования структуры текста [Электронный ресурс] Режим доступа: https://elar.urfu.ru/bitstream/10995/31691/1/conf_rtf_2015_28.pdf
3. Андреев А.М. и др. Автоматическая классификация текстовых документов с использованием нейросетевых алгоритмов и семантического анализа [Электронный ресурс] Режим доступа: http://www.ixlab.ru/pub/docs/RCDL_2003.pdf
4. Батура Т.В. Методы автоматической классификации текстов. // Программные продукты и системы. — 2017. — № 1 [Электронный ресурс] Режим доступа: https://cyberleninka.ru/article/n/metody-avtomaticheskoy-klassifikatsii-tekstov.
5. Васильев В.Г. Автоматическое выделение значимых фрагментов в текстах. // Обозрение прикладной и промышленной математики. Выпуск 3, том 14. — М., 2007. — 518 с.
6. Васильев В.Г. Комплексная технология автоматической классификации текстов. // Труды международной конференции «Диалог 2008» [Электронный ресурс] Режим доступа: https://docplayer.com/55296280-Kompleksnaya-tehnologiya-avtomaticheskoy-klassifikacii-tekstov-complex-technology-of-automatic-text-classification.html
7. Добров А.В. Автоматическая рубрикация текстов средствами комплексного лингвистического анализа [Электронный ресурс] Режим доступа: http://aiire.org/pubs/2012…
8. Кузнецов И.В. Введение в анализ текстовой информации с помощью Python и методов машинного обучения [Электронный ресурс] Режим доступа: https://habr.com/ru/post/205360/
9. Ландэ Д. Глубинный анализ текстов. Технология эффективного анализа текстовых данных [Электронный ресурс] Режим доступа: http://www.visti.net/~dwl/art/dz/
10. Сидорова Е.А. Подход к моделированию процесса извлечения информации из текста на основе онтологии. // Онтология проектирования. — 2018. — № 1 (27) [Электронный ресурс] Режим доступа: https://cyberleninka.ru/article/n/podhod-k-modelirovaniyu-protsessa-izvlecheniya-informatsii-iz-tekstana-osnove-ontologii.
11. Частотный анализ русского текста и облако слов на Python [Электронный ресурс] Режим доступа: https://habr.com/ru/post/517410/?amp&amp

© Чжэн Цзини (sofiazjy@gmail.com), Вэй Сяоюй (569006420@mail.ru).

Журнал «Современная наука: актуальные проблемы теории и практики»