Инженерный вестник Дона, №2 (2023)
ivdon.ru/ru/magazine/archive/n2y2023/8215

Инструменты решения проблем распознавания и кластеризации данных из документов методами машинного обучения

О.В. Золотарев, В.А. Юрчак

Российский новый университет, Москва

Аннотация: В статье описываются возможности, достоинства и отличия систем машинного обучения без учителя от обучения по шаблонам. Также дается определение понятию кластеризации с указанием основных методов и задач, решаемых данным алгоритмом машинного обучения. Подробно расписывается алгоритм распознавания данных из документов посредством технологии OCR, формируются цели и задачи использования технологии OCR в бизнес – процессах IT – компаний. Далее приводятся инструменты решения проблемы распознавания и кластеризации данных из сканов документов PDF посредством библиотек машинного обучения Nanonets и Tesseract. В заключении к данной статье описываются достоинства и недостатки использования данных библиотек в решении проблемы распознавания и кластеризации данных из сканов документов.

Ключевые слова: машинное обучение, кластеризация, распознавание данных, библиотека Nanonets, библиотека Tesseract, формат файла, документ, алгоритм, оцифровка.

Введение

В бизнес-процессах наблюдается такая проблема, как большое количество времени, потраченное на рутинную работу, и невозможность извлечь данные из файлов формата PDF. В статье предложены инструменты решения данных проблем.

Цель работы состоит в описании инструментов решения проблем распознавания и кластеризации данных из документов методами машинного обучения.

Основная часть (Материалы и методы)

Системы машинного обучения дают возможность оперативно использовать знания, которые получены при обучении на больших объемах данных [1]. В свою очередь, алгоритмы машинного обучения, в отличие от программ со встроенными вручную инструкциями, самостоятельно учатся распознавать шаблоны документов.

При обучении без учителя машина учится сама, используя данные, и не прибегая к вмешательству извне. При этом машина не имеет правильного ответа, но выявляет закономерности, основываясь на данных, что позволяет находить решение.

Одним из наиболее популярных и точных методов машинного обучения по поиску решения является метод кластеризации данных. Кластеризация данных – это модель обучения, которое происходит без учителя и содержит в себе группировку точек данных. Она часто применяется для выявления мошеннических действий, структуризации документов и сегментации пользователей.

Можно сказать, что кластеризация актуальна только для тех задач, в которых известны описания множества объектов, и нужно выявить внутренние связи, закономерности и зависимости между объектами [2]. Кроме того, кластеризация содержит в себе группировку заданных немаркированных данных. Вкратце рассмотрим существующие методы кластеризации:

1. Иерархические методы. Кластеры образуют древовидную структуру, представленную в виде иерархии. Новые кластеры, которые появляются на дереве, возникают из прежде образованных комков. При этом можно выделить следующие категории:

- разделяющий – подход сверху вниз. Все данные, включенные в один кластер, постепенно разбиваются, пока все точки не будут поделены;
- агломерационный – подход снизу вверх. Каждая точка – это единый кластер, они сливаются, и постепенно образуется новый кластер;
- методы на основе сетки позволяют сформировать пространство данных в конкретном количестве ячеек, образуя структуру в виде простой сетки. Каждый процесс кластеризации независим и оперативен;

- методы на основе плотности исследуют кластеры в качестве более плотных регионов, имеющих сходства и различия, в сравнении с менее плотными регионами. Посредством данных методов гарантирована точность результата [3];
- методы разбиения позволяют разделить объекты и превратить их в k-кластеры.

2. K – образные кластеры более узнаваемый метод, его реализация более проста:

- распознавание фальшивых новостей – кластеры позволяют алгоритму распознать истинные и неистинные фрагменты;
- продажи и маркетинг позволяет компаниям ориентироваться на конкретную аудиторию. Алгоритмы смогут сгруппировать людей с похожими чертами и определить, купят ли они разрабатываемый на предприятии продукт. Формирование групп позволит компании проводить тестирование, чтобы выявить аспекты, которые поднимут продажи;
- фэнтези-спорт – алгоритмы помогут определить похожих игроков, которые применяют некоторые атрибуты;
- определение преступления – посредством кластеризации можно анализировать GPS-журналы и создать группу схожего поведения преступника. Далее исследовать характерные черты группы и структурировать мошенническое и истинное поведение;
- фильтрация спама – на предприятиях такие письма исключаются при использовании алгоритмов для идентификации спама и пометки его флажками.

Бизнес-процессы часто требуют распознавания данных из документов, извлечения текста из файлов. Множество решений, которые помогут в этом, на сегодняшний день используют возможности оптического распознавания

символов OCR [4]. Данная технология может применяться для распознавания и извлечения данных из картинок, файлов, документов [5].

К примеру, чтобы извлечь данные из PDF файлов можно использовать конвертеры или специальные инструменты. Из небольших документов можно получать данные посредством простого форматирования. Но, если документов много, и они имеют сложное форматирование, графики, картинки, таблицы, сделать это вручную или с помощью инструментов окажется проблематичным. В таком случае пригодится программное обеспечение OCR.

Владельцы компаний активно применяют ПО для оцифровки бумажной документации, сканируя их для получения информации, что позволяет работать более эффективно и экономить время и силы.

После преобразования ПО физического документа или изображения документа в цифровые данные, которые могут корректировать процессоры или программы, пользователи могут проводить поиск посредством простых редакторов [6].

Рассмотрим наиболее популярные библиотеки, позволяющие распознавать данные из документов. Первая из них – Nanonets, отличается более высокой точностью и масштабом (см. рис.1).

Nanonets создан на основе искусственного интеллекта и автоматизирует процесс извлечения данных из документов. Извлеченные данные можно экспортировать в форматы XML, CSV, JSON, Excel.

Данная библиотека отлично справляется с оцифровкой документа, получением данных из «коробки». Актуальна для использования в бизнес-процессах в целях автоматизации ряда рабочих операций, связанных с документацией [3].

Библиотека Nanonets может читать все виды документов, на различных языках.

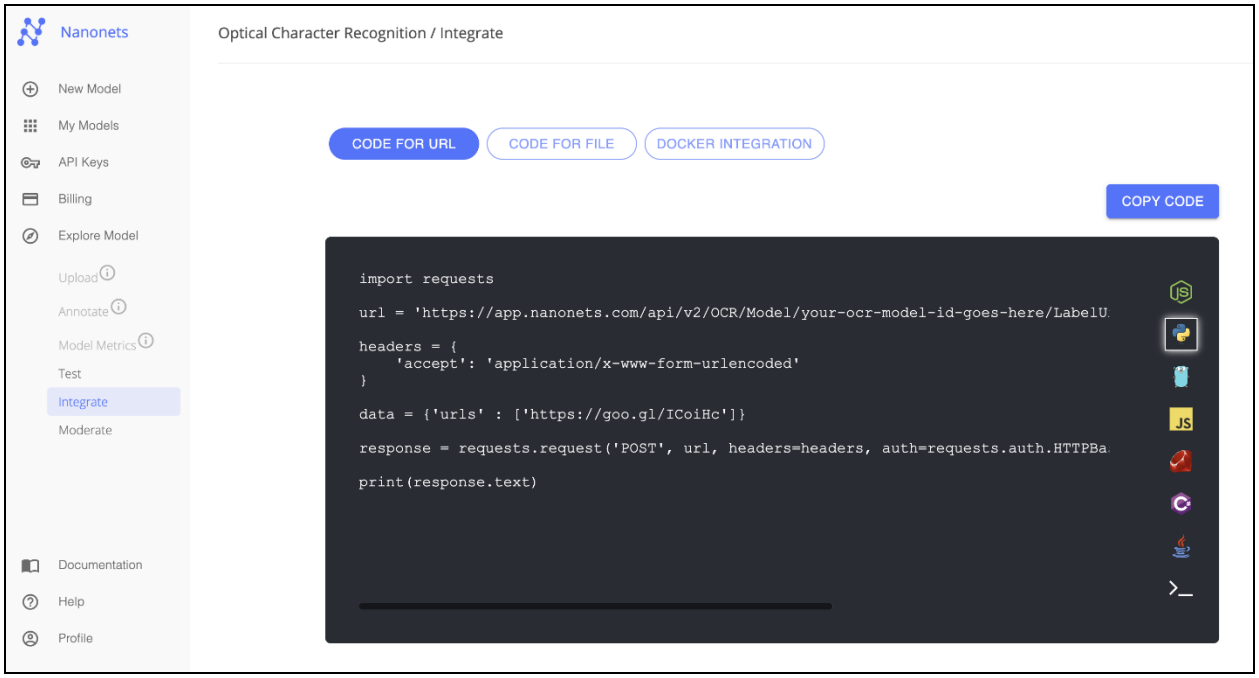

Рис. 1. – Библиотека Nanonets

Искусственной интеллект постоянно обновляется, растет точность распознавания, извлечения данных из документов. К преимуществам библиотеки можно отнести следующее:

- интуитивно простой и доступный интерфейс;
- простая в применении;
- существует бесплатная пробная версия;
- функционирует в автономном режиме, если пользователь приобретет премиум-версию;
- возможность работать с PDF;
- позволяет увеличить производительность труда;
- соответствует требованиям GDPR;
- быстрая скорость отклика API [7, 8].

Из минусов библиотеки можно отметить то, что на аннотирование может уйти много времени.

Следующая библиотека, которая будет рассмотрена – Tesseract (см. рис.2).

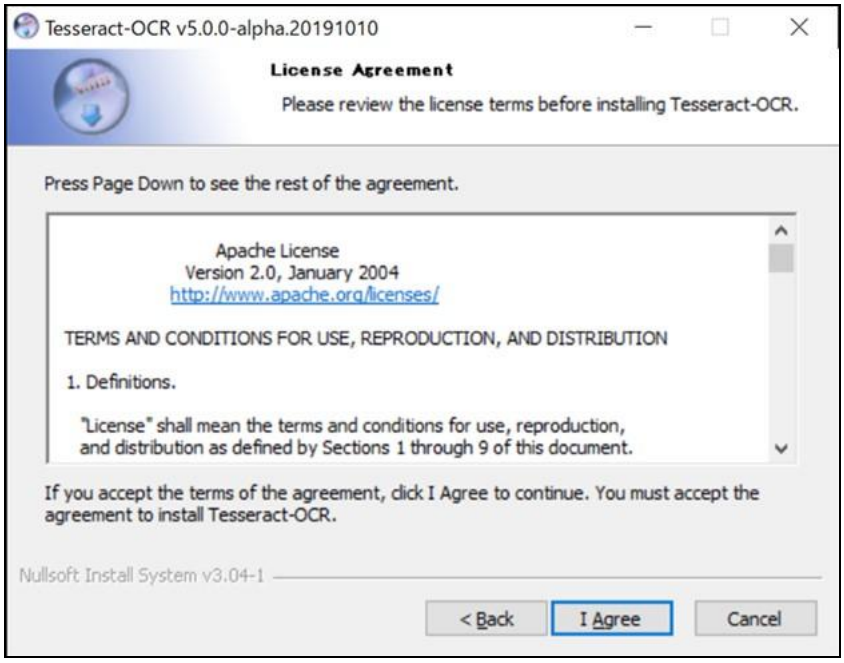

Рис. 2. – Библиотека Tesseract

Библиотека имеет открытый исходный код, который поддерживает свыше 100 языков. Для более простого использования разработчиками в своих проектах библиотека имеет интерфейсы API и GUI. Данная библиотека с открытым кодом развивается, при этом не существует ограничений для ее использования [9, 10]. К очевидным преимуществам стоит отнести:

- бесплатное использование;
- поддержка свыше 100 языков;
- открытый исходный код;
- интерфейсы API и GUI.

Заключение

Кластеризация является мощным методом машинного обучения, который содержит группировку по точкам данных. При наборе разных точек данных можно применять алгоритм кластеризации для классификации каждой отдельной точки в отдельную группу.

Для работы с текстом, прежде всего, необходимо извлечь его из картинки, ввиду чего важно использовать OCR. Одни из самых популярных

библиотек Tesseract и Nanonets позволяют распознавать данные из документов с высоким процентом распознавания.

Литература

1. Красников И.А., Никуличев Н.Н. Гибридный алгоритм классификации текстовых документов на основе анализа внутренней связности текста // Инженерный вестник Дона, 2013, №3. URL: ivdon.ru/ru/magazine/archive/n3y2013/1773.

2. Различия между искусственным интеллектом, машинным обучением и глубоким обучением. URL: habr.com/ru/post/526984/ (дата обращения: 22.12.2020).

3. Tesseract Open Source OCR Engine. URL: github.com/tesseract-ocr/tesseract (дата обращения: 26.11.2022).

4. Шепелев А.Н., Букатов А.А., Пыхалов А.В., Березовский А.Н. Анализ подходов и средств обработки сервисных журналов // Инженерный вестник Дона, 2013, №4. URL: ivdon.ru/ru/magazine/archive/n4y2013/1966.

5. Акулич М. Кластерный подход. Экономический рост и инновационные кластеры. М: Издательские решения, 2017. 886 с.

6. Елисеева И.И., Рукавишников В.О. Группировка, корреляция, распознавание образов (статистические методы классификации и измерения связей). М.: РГГУ, 2014. 144 с.

7. Vadapalli P. Clustering in Machine Learning. URL: upgrad.com/blog/clustering-in-machine-learning/ (дата обращения: 26.11.2022).

8. Nigar N., Faisal H.M., Shahzad M.K., Islam Sh., Oki O. An Offline Image Auditing System for Legacy Meter Reading Systems in Developing Countries: A Machine Learning Approach // Journal of Electrical and Computer Engineering. 2022. V. 2022. URL: doi.org/10.1155/2022/4543530.

9. Mahajan A., Samvelyan M., Mao L., Makoviychuk V., Garg A., Kossaifi J., Whiteson Sh., Zhu Y., Anandkumar A. Tesseract: Tensorised actors for multi-

agent reinforcement learning // International Conference on Machine Learning. PMLR, 2021. С. 7301-7312.

10. Hegghammer T. OCR with Tesseract, Amazon Textract, and Google Document AI: a benchmarking experiment // Journal of Computational Social Science. 2022. V. 5. №1. pp. 861-882.

References

1. Krasnikov I.A., Nikulichev N.N. Inzhenernyj vestnik Dona, 2013, №3. URL: ivdon.ru/ru/magazine/archive/n3y2013/1773.

2. Razlichiya mezhdu iskusstvennym intellektom, mashinnym obucheniem i glubokim obucheniem. [Differences between artificial intelligence, machine learning and deep learning]. URL: habr.com/ru/post/526984/ (accessed: 22.12.2020).

3. Tesseract Open Source OCR Engine. URL: github.com/tesseract-ocr/tesseract (accessed: 26.11.2022).

4. Shepelev A.N., Bukatov A.A., Pykhalov A.V., Berezovsky A.N. Inzhenernyj vestnik Dona, 2013, №4. URL: ivdon.ru/ru/magazine/archive/n4y2013/1966

5. Akulitch M. Klasternyj podkhod. Ekonomicheskiy rost i innovatsionnye klastery [Cluster approach. Economic growth and innovation clusters]. M.: Izdatel'skie resheniya, 2017. 886 p.

6. Eliseeva I.I., Rukavishnikov V.O. Gruppirovka, korrelyatsiya, raspoznavanie obrazov (statisticheskie metody klassifikatsii i izmereniya svyazey) [Grouping, correlation, pattern recognition (statistical methods of classification and measurement of connections)]. Moskva: RSUH, 2014. 144 p.

7. Vadapalli P. Clustering in Machine Learning. URL: upgrad.com/blog/clustering-in-machine-learning/ (accessed: 26.11.2022).

8. Nigar N. Faisal H.M., Shahzad M.K., Islam Sh., Oki O. Journal of Electrical and Computer Engineering. 2022. V. 2022. URL: doi.org/10.1155/2022/4543530.

9. Mahajan A., Samvelyan M., Mao L., Makoviychuk V., Garg A., Kossaifi J., Whiteson Sh., Zhu Y., Anandkumar A. Tesseract: Tensorised actors for multi-agent reinforcement learning. International Conference on Machine Learning, PMLR, 2021. pp. 7301-7312.

10. Hegghammer T. Journal of Computational Social Science. 2022. V. 5. №1. Pp. 861-882.