УДК [004.422.833+004.62]

СБОР, ОБРАБОТКА И АНАЛИЗ ДАННЫХ ДЛЯ ОБРАЗОВАТЕЛЬНОЙ СИСТЕМЫ

А. В. Саскевич
Аспирант БГУИР, инженер-программист Synesis

Белорусский государственный университет информатики и радиоэлектроники, факультет компьютерных систем и сетей, кафедра информатики, Республика Беларусь
ООО "Синезис", Республика Беларусь
E-mail: asaskevich@yandex.com

А. В. Саскевич
Окончил Белорусский Государственный Университет Информатики и Радиоэлектроники. Аспирант БГУИР. Работает в компании Synesis в должности инженера-программиста. Проводит научные исследования в области применимости методов машинного обучения и обработки больших объемов информации в сфере образования.

Аннотация. В данной статье рассматриваются возможности сбора и генерации обучающих материалов при помощи таких инструментов, как машинное обучение и скрапинг данных. Не меньшую роль в поставленной задаче имеют базы знаний, позволяющие структурировать и организовать большие объемы информации, которые могут быть использованы при построении информационной системы, применяемой в сфере образования. Полученные результаты могут быть применены для других сфер, требующих сбор и структурирование данных по определенным правилам.

Ключевые слова: обработка текста, обработка больших объемов данных, машинное обучение, NLP.

Введение. Бурный рост информационных технологий, связанных с хранением и обменом информации привел к падению цен на носители информации и увеличению скорости обмена этой информацией. Активный рост сети Интернет позволил множеству пользователей по всему миру получить доступ к знаниям, которые ранее находились вне доступа большинства. Это позволило модернизировать сферу образования, поставив применение информационных технологий в процессе обучения как необходимость.

Тем не менее, одной из проблем остается обработка информации таким образом, который позволил бы извлечь из нее нужные сведения, структурировать их в соответствии с определенными критериями, и на основе полученных результатов синтезировать базу знаний, которая могла бы поставляться в учреждения образования или индивидуальным учащимся, предоставляя им возможность обучаться выбранной специальности или профессии.

Сбор данных. Предварительно данные необходимо собрать и очистить от объективно лишних данных – тегов, комментариев, рекламных баннеров, нерелевантной информации и прочего. Для этого можно как воспользоваться готовыми данными – онлайн-библиотеками, файловыми хранилищами, репозиториями учебных заведений. Данные могут быть собраны как вручную, так и посредством специальных утилит, которые могут быть реализованы достаточно быстро при наличии даже небольшого опыта в программировании. Например, посредством библиотеки Scrapy [1] для языка программирования Python возможно с

легкостью написать скрипт, который загрузит данные с онлайн-ресурса, сохранив их в отдельный файл или выгрузив в онлайн-хранилище, которое позволит подключить облачные решения для анализа данных и машинного обучения, такие как Amazon Machine Learning [2] или Google Cloud Platform Machine Learning [3]. Основная задача этого этапа подготовить соответствующий необходимый объем данных, из которого в дальнейшем могут быть выделены фрагменты информации, которые могут нести ценность для создаваемой учебной базы данных.

В качестве источников следует выбирать те, которые обладают достаточным уровнем компетенции и доверия, так как итоговое качество построенной обучающей системы зависит не только от количества собранной информации, а также от ее качества. Для этого следует обратиться к репозиториям учреждений образования, литературе, представленной в крупных библиотеках, хранилищах научных публикаций и подобных архивов.

Подготовка базы данных. На следующем этапе необходимо выделить, как данные необходимо представить конечному пользователю, выделив набор сущностей, их свойств и группы связей между ними. Примером может являться группа вики-страниц, связанных между собой гиперссылками, объединенных в разделы по тематике. Необходимость данного этапа обусловлена тем, что минуя этап подготовки и проектирования базы данных можно упустить неочевидные свойства собранной информации и связи между этими свойствами. Кроме того, заранее спроектированная база данных облегчит применение определенных алгоритмов машинного обучения (в некоторых случаях алгоритмы машинного обучения требуют список свойств обрабатываемых наборов данных заранее), а также вывод информации конечному пользователю – а именно, поиск, отображение, изменение.

Очистка и обработка данных. Перед тем, как выгрузить данные в базу данных, придав им выбранную структуру, необходимо обработать данные на предмет лишней информации, которая не только не даст улучшения результата, но и может испортить качество этого результата. Для этого следует воспользоваться некоторыми методиками.

При выделении числовой информации, разнообразных статистик, результатов формул и подобных данных, поддающихся математической интерпретации, следует воспользоваться классическими подходами по обработке данных из машинного обучения [4]. В частности, следует исключить статистические выбросы или данные, противоречащие основной массе выборки. Работа с такими данными необходима в широком спектре научных направлений, например – физике, химии, математике, экономике, программировании, медицине, архитектуре и так далее.

При выделении графической информации следует определить, какие графические данные имеют наибольший вес, в какой форме они должны быть представлены. Работа с такими данными может потребоваться при подготовке данных для учащихся биологии, медицины, дизайна, архитектуры и других направлений, подразумевающих обилие графических данных.

При работе с текстом следует подготовить метрики, которые могут позволить оценить качество текста, его полноту и связность. В качестве таких метрик могут быть такие, как количество стоп-слов, количество научных терминов, эмоциональная окраска текста, средняя длина предложений и популярность используемых слов. Данные метрики могут помочь выделить тексты, которые были написаны для применения в научно-популярной сфере, в сфере массовой информации или в сфере науки и образования [5].

Подобрав необходимые критерии, выделив информацию с максимальной полезностью, можно перейти непосредственно к этапу обработки данных. На данном этапе посредством существующих решений либо посредством собственных структурировать данные и загрузить их в заранее спроектированную базу данных. После выполнения этого этапа будет доступна база данных, заполненная и готовая к использованию. Данная база данных может поставляться как есть – тогда процесс разработки и адаптации информационных систем

переходит на сторону конечного потребителя. Имея готовую базу данных, потребитель может заняться либо разработкой тестирующей системы, которая может помочь проверять уровень знаний учащихся, либо разработкой обучающей системы, которая с учетом потребностей учащихся может помочь им в изучении отдельного курса, предмета или специализации.

Заключение. Таким образом, представленные этапы позволяют собрать, обработать, адаптировать и подготовить достаточный объем данных, который может быть применен в учебной системе. Аналогичные подходы применяются в различных образовательных онлайн-приложениях типа Duolingo [6] – для конечного пользователя добавляется еще этап заключительной очистки данных, когда пользователь системы может сообщить, что тестовые задания или учебный материал содержит ошибку. Введение такого этапа позволяет быстро собрать базовый учебный набор, а затем последовательного улучшать его качество посредством помощи самих же пользователей. В рамках исследуемой темы разрабатывается система с учетом представленных результатов, которая позволит собирать данные любого формата и направления, генерировать тестовые задания для учащихся, а также выделять учебные фрагменты информации по выбранной теме, позволяя совместить процесс тестирования и обучения.

Список литературы

[1.] Scrapy Python [Электронный ресурс]. – Режим доступа: https://scrapy.org – Дата доступа: 20.02.2020.

[2.] Сервисы машинного обучения на базе Amazon Web Services [Электронный ресурс]. – Режим доступа: https://aws.amazon.com/machine-learning/ – Дата доступа: 24.02.2020.

[3.] Машинное обучение на базе платформы Google Cloud [Электронный ресурс]. – Режим доступа: https://cloud.google.com/products/ai – Дата доступа: 24.02.2020.

[4.] Han J., Kamber M., Pei J. Data Mining: Concepts and Techniques. Amsterdam, 2011. Вып. 3.

[5.] Ulicny B., Baclawski K., Magnus A. New metrics for blog mining // Data Mining, Intrusion Detection, Information Assurance, and Data Networks Security 2007. – International Society for Optics and Photonics, 2007. – Т. 6570. – С. 65700I.

[6.] Duolingo – мобильное приложение для изучения иностранных языков [Электронный ресурс]. – Режим доступа: https://www.duolingo.com/. – Дата доступа: 15.02.2020.

DATA COLLECTION AND PROCESSING FOR EDUCATIONAL SYSTEM

A. V. SASKEVICH

Postgraduate student of the BSUIR, software engineer Synesis

Belarusian State University of Informatics and Radioelectronics, Department of Computer Systems and Networks, Department of Informatics, Republic of Belarus
LLC "Synesis", Republic of Belarus
E-mail: asaskevich@yandex.com

Abstract. This article discusses the possibilities of collecting and generating training materials using tools such as machine learning and data scraping. No less important role in the task have knowledge bases that allow you to structure and organize large amounts of information that can be used to build the information system used in the field of education. The results can be applied to other areas that require the collection and structuring of data according to certain rules.

Keywords: text processing, large data processing, machine learning, NLP.