УДК 004 DOI: 10.18698/2541-8009-2021-06-708

СБОР И ПОДГОТОВКА ТЕКСТОВЫХ ДАННЫХ ДЛЯ ЗАДАЧ ОБРАБОТКИ ЕСТЕСТВЕННОГО ЯЗЫКА

А.Ф. Ладонцев alexander.ladontsev@yandex.ru
SPIN-код: 7449-9699

МГТУ им. Н.Э. Баумана, Москва, Российская Федераци

Аннотация

Изучение компьютерной репрезентации и анализа естественного языка является одним из актуальных направлений исследования современной науки в условиях цифровизации общества. В статье описан один из возможных вариантов сбора и подготовки данных в целях применения методов машинного обучения с учителем для создания классификатора тональностей текстов. В качестве практического материала были выбраны и проанализированы отзывы интернет-пользователей на зарубежную литературу и соответствующие им оценки. В результате получены переменная с текстами отзывов и переменная с соответствующими им оценками, что позволит в дальнейшем осуществить предобработку и использовать эти данные для обучения модели автоматического распознавания тональности текста.

Ключевые слова

Компьютерная лингвистика, естественные языки, обработка, анализ тональности, машинное обучение, язык программирования, Python

Поступила в редакцию 07.06.2021

Введение. NLP (*natural language processing*), или ОЕЯ (обработка естественного языка) — направление, посвященное исследованию методов компьютерного представления и анализа естественного языка [1]. В задачи этого направления входит разработка алгоритмов классификации текстов (например, по тематике или тональности), информационный поиск, извлечение информации из тестов, машинный перевод, создание диалоговых систем и чат-ботов, систем распознавания и синтеза речи. В целом это направление посвящено созданию эффективных методов взаимодействия человека и ЭВМ.

Одна из набирающих популярность областей — анализ тональности текста (англ. *sentiment analysis*), позволяющий оценивать его эмоциональный компонент. «Анализ тональности не предполагает извлечения фактографической информации, он занимается только степенью эмоциональной окраски сообщений» [2, с. 247]. Например, в результате анализа эмоциональной составляющей сообщений пользователей социальной сети Twitter можно получить данные о настроении пользователей относительно важных общественно-политических событий, в частности, выборов разных уровней. Данный инструмент можно использовать также для анализа сообщений пользователей о компаниях и их цен-

ных бумагах, что может быть полезным работникам финансового сектора для прогнозирования спроса на акции (и, как следствие, цен).

«Как и в других задачах прикладной лингвистики, основные подходы к автоматическому определению тональности текста можно разделить на две большие группы. Алгоритмы первой группы основаны на правилах (*rule-based*), а алгоритмы второй группы используют методы машинного обучения (*machine learning*)» [2, с. 250].

«Машинное обучение с учителем (*supervised learning*) включает моделирование признаков данных и соответствующих данным меток. После выбора модели ее можно использовать для присвоения меток новым, неизвестным ранее данным» [3, с. 380]. Например, если мы хотим обучить классификатор, который бы автоматически определял тематику новостных текстов, нам необходимо каждому тексту сопоставить его тематику (экономика, политика, экология, образование и т. д.). После обучения модели ее можно использовать для присвоения меток (в рассматриваемом случае — категорий новостных текстов) новым, неизвестным ранее данным. Так же организован и классификатор тональностей.

Данная статья посвящена описанию одного из вариантов сбора и подготовки данных для применения методов машинного обучения с учителем для создания классификатора тональностей текстов. Источником данных послужил сайт [4], с которого были скачаны более 17 тысяч отзывов на зарубежную литературу и соответствующие им оценки.

Автоматический сбор текстов. Для сбора данных было использовано браузерное расширение (т. е. программа, расширяющая стандартные функциональные возможности браузера) webscraper.io [5]. После скачивания с сайта это расширение становится доступно в панели разработчика, которая открывается нажатием клавиши F12 в браузере (рис. 1). Важно переместить панель в нижнюю часть экрана, если изначально она находится в другом месте (например, слева), тогда в правом верхнем углу появится надпись Web Scraper (см. рис. 1).

Рис. 1. Web Scraper

Перейдя в этот раздел, начнем создавать алгоритм для автоматического сбора рецензий: Create new sitemap → Create Sitemap (рис. 2).

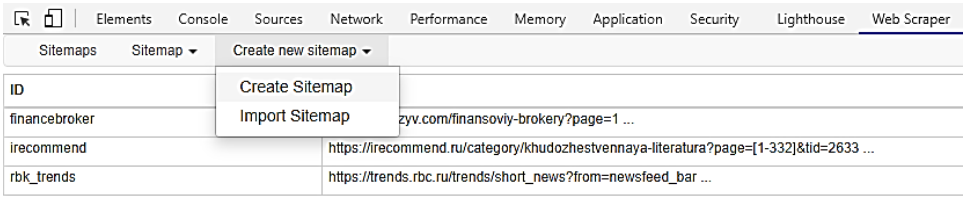

Рис. 2. Создание алгоритма для автоматического сбора рецензий

В появившемся окне введем название алгоритма и ссылку, с которой будет начинаться сбор текстов (рис. 3).

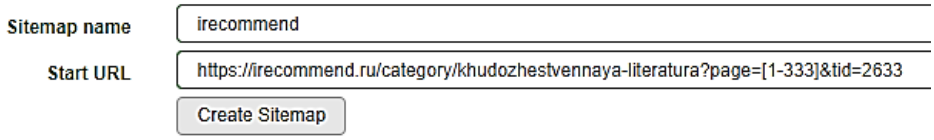

Рис. 3. Название алгоритма и ссылка на него

Для решения задачи пагинации (перехода между страницами) укажем в ссылке диапазон от 1 до 333 — именно столько страниц содержит данный раздел сайта без первой страницы (рис. 4). Адрес первой страницы (https://irecommend.ru/category/khudozhestvennaya-literatura?tid=2633) отличается от адреса второй и следующих страниц (https://irecommend.ru/category/khudozhestvennaya-literatura?page=1&tid=2633) отсутствием параметра page (второй странице соответствует page=1, третьей – page=2 и т. д.), поэтому решено было начать сбор данных со второй страницы.

Рис. 4. Указание диапазона страниц сайта для сбора данных

При необходимости данные с первой страницы можно собрать, добавив дополнительный адрес при создании алгоритма.

После добавления метаданных выберем складку Create new Sitemap. Теперь нужно «показать» программе, по каким ссылкам необходимо переходить, какие данные и в каком виде требуется собирать. Кнопка Add new selector позволяет добавить это действие (рис. 5).

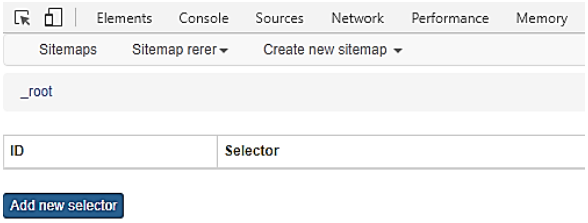

Рис. 5. Определение зоны поиска

При нажатии открывается окно, где необходимо ввести параметры (рис. 6):

– Id (название селектора);

– Type (тип селектора; выбирается в зависимости от того, какое действие необходимо совершить алгоритму: перейти по ссылке, пролистать вниз, считать данные в виде текста на естественном языке, считать данные в виде html-тегов и пр.);

– Selector (инструмент для выбора необходимого элемента сайта);

– Multiple (поле, которое необходимо пометить галочкой, если селектор должен обработать несколько элементов);

– Regex (регулярное выражение, шаблон для поиска информации);

– Parent Selectors (родительские селекторы, т. е. селекторы более высокого уровня абстракции, которым принадлежит данные селектор).

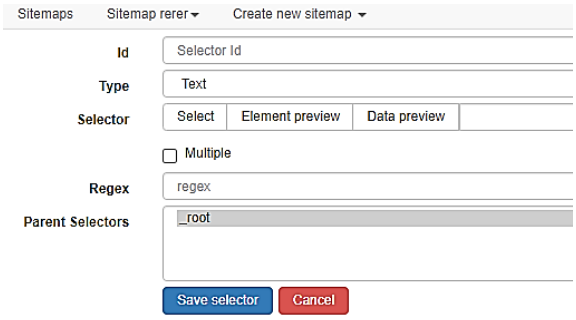

Рис. 6. Установление параметров поиска

Рассмотрим структуру сайта. Главная страница содержит множество ссылки на книги: «Скотный двор» Джорджа Оруэлла; «12 правил жизни: противоядие от хаоса» Джордана Питерсона и т. п., всего 20 ссылок на странице (рис. 7).

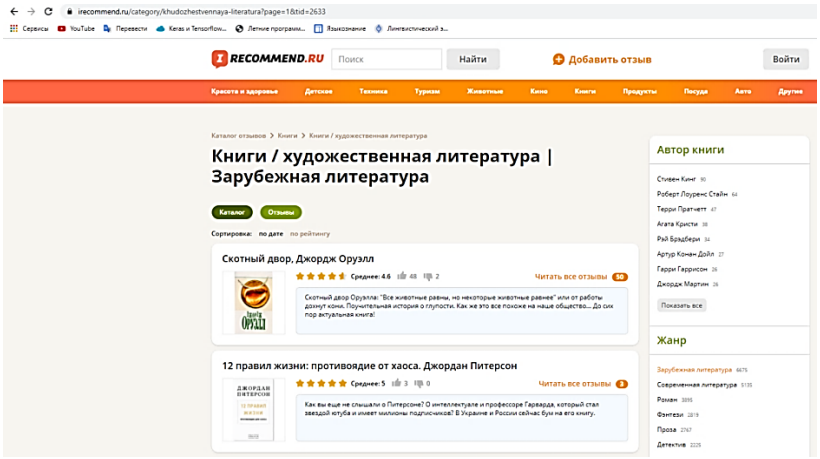

Рис. 7. Структура сайта recommend.ru

Перейдя по одной из ссылок и пролистав вниз, увидим множество комментариев (рис. 8). Вне зависимости от их количества они расположены на одной странице. Чтобы прочитать комментарий полностью, нужно перейти по ссылке, нажав на кнопку «Читать весь отзыв».

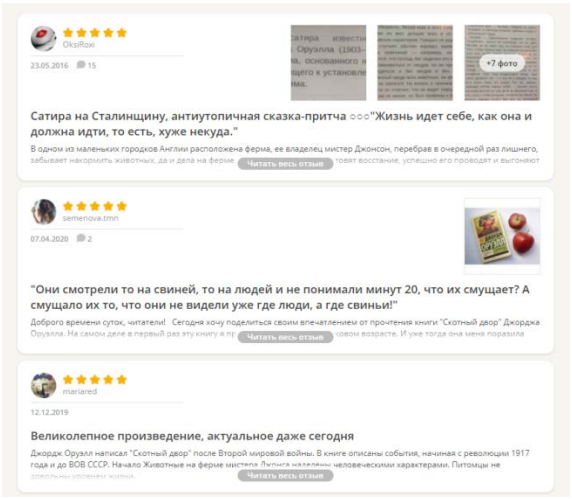

Рис. 8. Интерфейс страницы сайта recommend.ru с отзывами пользователей

Наконец, можно увидеть сам комментарий и оценку, находящуюся в верхнем правом углу в виде звезд (рис. 9).

Рис. 9. Пример комментария к книге и ее оценки на сайте recommend.ru

Проанализировав структуру интересующих нас разделов сайта, создадим селектор, который будет заходить на страницу каждой книги (рис. 10).

Затем создадим селектор, который будет заходить на страницу каждого комментария (рис. 11).

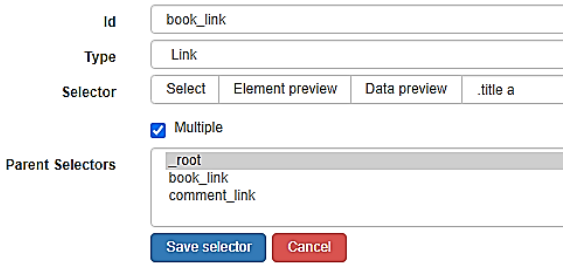

Рис. 10. Создание селектора для каждой страницы книги

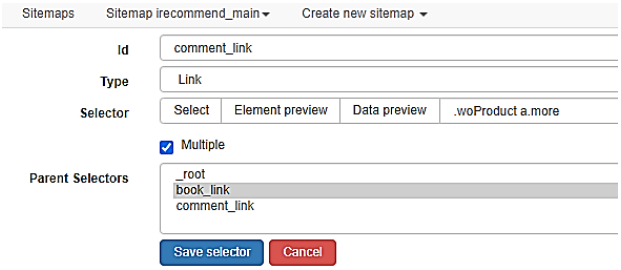

Рис. 11. Создание селектора для каждого комментария

Наконец, создадим селектор, который будет извлекать текст и оценку со страницы комментария (рис. 12, 13).

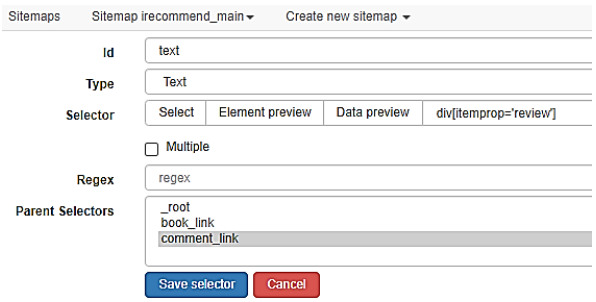

Рис. 12. Создание селектора для извлечения текста

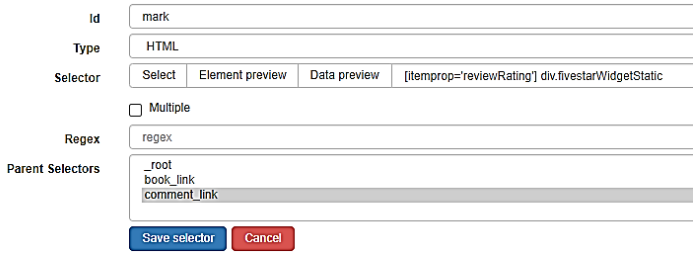

Рис. 13. Создание селектора для извлечения оценки

В результате получили следующий граф (рис. 14), показывающий механизм работы алгоритма: из корневой ссылки (_root), переходим по ссылкам книг (book_link), оттуда — по ссылкам комментариев (commentary_link), из которых извлекаем текст отзыва (text) и оценку (mark).

Рис.14. Граф, показывающий механизм работы алгоритма

Алгоритм создан, теперь соберем данные. Нажмем Sitemap irrecommend → Browse (рис.15) → Start scraping (рис.16).

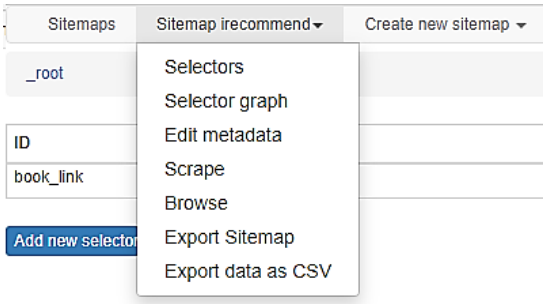

Рис. 15. Сбор данных. Шаг 1: Sitemap irrecommend → Browse

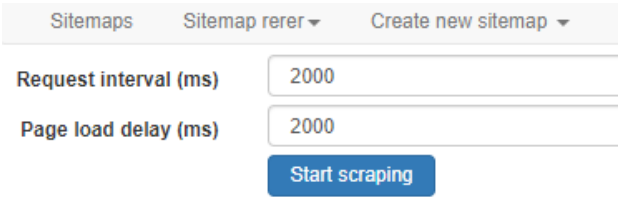

Рис. 16. Сбор данных. Шаг 2: Start scraping

Когда данные будут собраны, нужно нажать refresh (рис. 17).

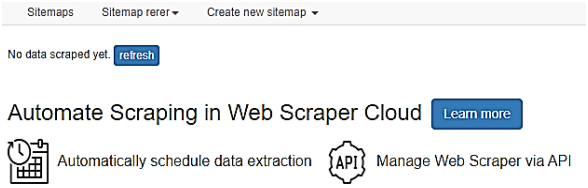

Рис. 17. Сбор данных. Шаг 3: refresh

На рис. 18, 19 показано, как будут выглядеть данные после скачивания. Тексты хранятся в столбце text, оценки — в столбце mark.

Рис. 18. Данные после скачивания: общий вид

text	mark
Nata ref Опубликовано 24 Февраль, 2021 - 04:40 Сложная, невероятная, поразительная история о том, на что способен человек в ужасающих условиях реальности. Разгадка - книга, которую нужно понять, а не просто прочитать. Получено с условием написания отзыва (бесплатно или со скидкой) Hi! Считаю, что мне очень повезло, поскольку моя личная библиотека полнится сильными, мощными книгами, способными вызвать бурю эмоций, способными заставить задуматься или же не позволяющими оторваться от чтива. Одной из такой по праву можно считать произведение Кальи Рид - Разгадка, ведь здесь собрано столько всего, а концовка настолько удивительная, что я отходила от прочитанного весь день и к моменту написания отзыва все еще не отошла.Скажу сразу, я не просто прочитала эту книгу, а углубилась в прочитанное, автор умело втянула в поток событий так, что просто не могла оторваться и мне очень хотелось помочь главной героине или хотя бы понять почему она оказалась именно там, где оказалась. В общем, давайте ближе к сути...Итак, сначала немного о том, что за книгу я фактически держала в руках. Об издании Книга Разгадка Кальи Рид является новинкой января месяца 2021-го года, выпущенной Inspiria (Эксмо) и знаете, если автор пойдет дальше, будет писать больше и перестанет так часто обращаться к постельным сценам, то я еще что-нибудь у нее прочту.Цена книги колеблется в зависимости от электронного или печатного варианта. Электронная книга стоит 250 руб., а такая как у меня (бумажная) 410 руб.Заказать книгу Разгадка (Калья Рид) можно на сайте Эксмо, Бук24 (вот прямая ссылка на произведение), а так же на других сайтах, где книга представлена. Внешний вид Книга не тонкая, в ней около 400 страниц. Есть небольшое оглавление в конце произведения и, к слову, названия глав не всегда напрямую говорят о том, что рассказывается в самой главе. Так же есть благодарности автора и прочее.На обложке изображена веточка дерева или куста и застывшая на ней капля. Не знаю, возможно это воспаленный мозг главной героини все время фокусируется на этой самой капле, видя ее в окне, а может это ее константа, способ держаться, но, так или иначе, о капле в книге будет сказано не раз и под конец уже привыкаешь к ней что ли.В общем, изображение капли здесь не спроста и даже сама автор книги за это оформление благодарит создателя, читай художника/дизайнера. Обложка книги твердая, листы внутри чуть шершавые, белые, шрифт достаточно крупный, чтобы удобно было читать и чтобы не казаться слишком большим на странице.Сами страницы хорошо прошиты и проклеены, а к обложке изначально была прицеплена синяя бумага, которую я постоянно вижу на книгах Inspiria. У книги есть возрастные ограничения 16+ и это правильно, поскольку в самом произведении много эротики или даже порнухи, много насилия и есть рукоприкладство. Я бы даже поставила ей 18+, все-таки там очень жесткие моменты есть, которые даже подростку лучше не знать. О чем повествует "Разгадка", книга Кальи Рид Полгода назад я была счастлива. Обычная девушка с обычной судьбой, Наоми Кэррадайн. Что-то важное случилось после. Месяц назад я оказалась в психбольнице. Вчера ко мне приходил Лахлан. Он поцеловал меня и сказал, что я теряю рассудок. А несколько часов спустя я вспомнила о Максе. Его голос звучал в моей голове, он повторял, что я вовсе не сумасшедшая и мне нужна помощь. Пару минут назад я решила сбежать из реальности – лишь бы разгадать, что со мной случилось в прошлом. Я не больна. Я	

Рис. 19. Данные после скачивания: вид комментариев и оценок

Скачаем данные в формате CSV, нажав Sitemap irrecommend → Export Data as CSV.

Извлечение данных с помощью языка Python. Для решения этой задачи используется язык программирования Python [6]. Более подробно о языке Python можно узнать, например, в литературе [7, 8], среда разработки PyCharm описана в [9]. Отметим, что код может быть написан в любой другой среде или любом другом редакторе кода. В нашем случае необходимо поместить скачанный файл irecommend.csv в папку проектов PyCharm, после чего он будет доступен для обработки (рис. 20).

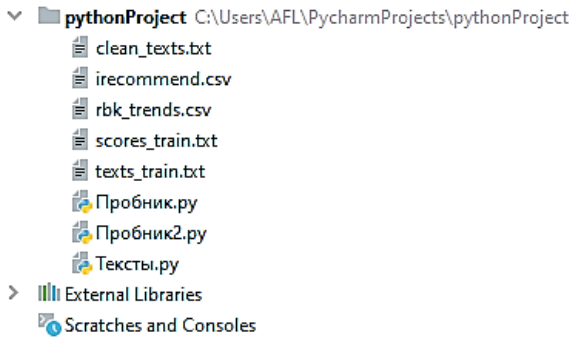

Рис. 20. Размещение скачанного файла irecommend.csv в папке проектов PyCharm

Напишем код, затем выполним его (рис. 21). Мы импортировали библиотеку pandas [10] и указали, что будем обращаться к ней через pd (строка 1). С помощью метода read_csv этой библиотеки мы считали информацию из скачанного файла (строка 3). В переменную texts были отобраны все элементы столбца text (строка 4), в переменную marks — элементы из столбца mark (строка 5).

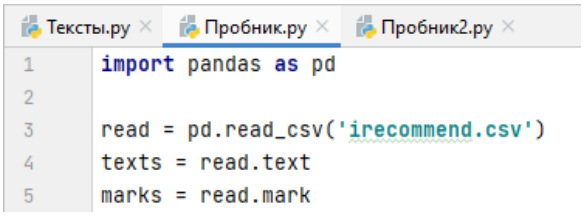

Рис. 21. Написание и выполнение кода

Таким образом, мы получили переменную, содержащую тексты отзывов, и переменную, содержащую соответствующие оценки. Далее эти данные будут предобработаны и использованы для обучения модели автоматического распознавания тональности текста.

Литература

[1] Большакова Е.И., Клышинский Э.С., Ландэ Д.В. и др. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика. М., МИЭМ, 2011.
[2] Николаев И.С., Митренина О.В., Ландо Т.М., ред. Прикладная и компьютерная лингвистика. М., URSS, 2017.
[3] ВандерПлас Дж. Python для сложных задач: наука о данных и машинное обучение. СПб., Питер, 2018.
[4] Отзывы читателей о книгах Джорджа Мартина. irecommend.ru: веб-сайт. URL: https://irecommend.ru/category/khudozhestvennaya-literatura?tid=2633&tid1=106869 (дата обращения: 25.04.2021).
[5] Webscraper: веб-сайт. URL: https://webscraper.io/ (дата обращения: 25.04.2021).
[6] Python: веб-сайт. URL: https://www.python.org/ (дата обращения: 25.04.2021).
[7] Лутц М. Изучаем Python. М., Вильямс, 2015.
[8] Силен Д., Мейсман А., Али М. Основы Data Science и Big Data. Python и наука о данных. СПб., Питер, 2017.
[9] PyCharm. jetbrains.com: веб-сайт. URL: https://www.jetbrains.com/ru-ru/pycharm/ (дата обращения: 25.04.2021).
[10] Pandas. devdocs.io: веб-сайт. URL: https://devdocs.io/pandas~0.25/ (дата обращения: 25.04.2021).

Ладонцев Александр Филиппович — студент кафедры «Романо-германские языки», МГТУ им. Н.Э. Баумана, Москва, Российская Федерация.

Научный руководитель — Куровская Юлия Геннадьевна, доктор педагогических наук, профессор кафедры «Романо-германские языки», МГТУ им. Н.Э. Баумана, Москва, Российская Федерация.

Ссылку на эту статью просим оформлять следующим образом:
Ладонцев А.Ф. Сбор и подготовка текстовых данных для задач обработки естественного языка. *Политехнический молодежный журнал*, 2021, № 06(59). http://dx.doi.org/ 10.18698/2541-8009-2021-06-708

COLLECTING AND PREPARING TEXT DATA FOR NATURAL LANGUAGE PROCESSING TASKS

A.F. Ladontsev
alexander.ladontsev@yandex.ru
SPIN-code: 7449-9699

Bauman Moscow State Technical University, Moscow, Russian Federation

Abstract

The study of computer representation and analysis of natural language is one of the topical research areas in modern science in the context of the digitalization of society. The article describes one of the possible options for collecting and preparing data in order to use supervised machine learning methods to create a text sentiment classifier. As a practical material, we selected and analyzed the responses of Internet users to foreign literature and the corresponding assessments. As a result, a variable with feedback texts and a variable with the corresponding estimates were obtained, which will allow further preprocessing and use of this data for training the automatic sentiment recognition model.

Keywords

Computational linguistics, natural languages, processing, sentiment analysis, machine learning, programming language, Python

Received 07.06.2021

References

[1] Bol'shakova E.I., Klyshinskiy E.S., Lande D.V., et al. Avtomaticheskaya obrabotka tekstov na estestvennom yazyke i komp'yuternaya lingvistika [Automated text processing natural language and computer linguistics]. Moscow, MIEM Publ., 2011 (in Russ.).

[2] Nikolaev I.S., Mitrenina O.V., Lando T.M., eds. Prikladnaya i komp'yuternaya lingvistika [Applied and computer linguistics]. Moscow, URSS Publ., 2017 (in Russ.).

[3] VanderPlas J. Python data science handbook. Essential tools for working with data. O'Reilly Media, 2016. (Russ. ed.: Vander Python dlya slozhnykh zadach: nauka o dannykh i mashinnoe obuchenie. Sankt-Petersburg, Piter Publ., 2018.)

[4] Otzyvy chitateley o knigakh Dzhordzha Martina [Reader reviews on George Martin's books]. irecommend.ru: website (in Russ.). URL: https://irecommend.ru/category/ khudozhestvennaya-literatura?tid=2633&tid1=106869 (accessed: 25.04.2021).

[5] Webscraper: website. URL: https://webscraper.io/ (accessed: 25.04.2021).

[6] Python: website. URL: https://www.python.org/ (accessed: 25.04.2021).

[7] Lutz M. Learning Python. O'Reilly Media , 2013. (Russ. ed.: Izuchaem Python. Moscow, Vil'yams Publ., 2015.)

[8] Cielen D., Meysman A., Ali M. Introducing data science. Big data, machine learning, and more, using Python tools. Manning Publications, 2016. (Russ. ed.: Osnovy Data Science i Big Data. Python i nauka o dannykh. Sankt-Petersburg, Piter Publ., 2017.)

[9] PyCharm. jetbrains.com: website (in Russ.). URL: https://www.jetbrains.com/ru-ru/pycharm/ (accessed: 25.04.2021).

[10] Pandas. devdocs.io: website. URL: https://devdocs.io/pandas~0.25/ (accessed: 25.04.2021).

Ladontsev A.F. — Student, Department of Romano-Geranic Languages, Bauman Moscow State Technical University, Moscow, Russian Federation.

Scientific advisor — Kurovskaya Yu.G., Dr. Sc. (Pedag.), Professor, Department of Romano-Geranic Languages, Bauman Moscow State Technical University, Moscow, Russian Federation.

Please cite this article in English as:
Ladontsev A.F. Collecting and preparing text data for natural language processing tasks. *Politekhnicheskiy molodezhnyy zhurnal* [Politechnical student journal], 2021, no. 06(59). http://dx.doi.org/ 10.18698/ 2541-8009-2021-06-708.html (in Russ.).