УДК 004.4:004.93

СРАВНИТЕЛЬНЫЙ АНАЛИЗ РАЗЛИЧНЫХ СИСТЕМ ОПТИЧЕСКОГО РАСПОЗНАВАНИЯ СИМВОЛОВ ПРИ РАБОТЕ С ТЕКСТОМ, НАПИСАННЫМ С ПОМОЩЬЮ КИРИЛЛИЧЕСКОГО АЛФАВИТА

Качалин В.С., Панов Ю.Н., Попов Н.-Л.Э.

ФГБОУ ВО «Московский авиационный институт (национальный исследовательский университет)», Москва, e-mail: vasiliy.kachalin@gmail.com

В современном мире все процессы подвергаются цифровизации, даже процесс переноса текста из бумажных носителей в цифровой вид. В этом помогают системы оптического распознавания символов (OCR). Их существует довольно большое количество, однако если необходимо работать с текстом, написанным кириллическими символами, то часть систем не рассматривается ввиду отсутствия возможности распознавания таких символов. При этом необходимость сделать выбор в пользу той или иной системы остается. Выбор конкретной системы оптического распознавания символов должен основываться на объективных характеристиках, таких как точность, скорость и используемая память. Исследования, оценивавшие эти параметры для OCR, уже существуют, однако они были проведены на некириллических символах. Цель этой статьи – провести сравнительный анализ некоторых систем OCR (ABBYY FineReader, CuneiForm, OCRopus, Tesseract, Transym OCR) при работе с текстами, написанными с помощью кириллического алфавита. Исследование проводилось на 20 отсканированных страницах без применения предварительной обработки, чтобы учесть наличие таковой в системах OCR. В результате сравнительного анализа было установлено, что лучшей из рассматриваемых OCR по точности распознавания кириллических символов является ABBYY FineReader, худшие показатели по точности имеет Transym OCR. Самой быстрой системой оказалась Tesseract, самой медленной – OCRopus. Лучший результат по используемой памяти показала система CuneiForm. Худший же – ABBYY FineReader.

Ключевые слова: OCR, оптическое распознавание символов, кириллица, ABBYY FineReader, CuneiForm, OCRopus, Tesseract, Transym OCR

COMPARATIVE ANALYSIS OF VARIOUS OPTICAL CHARACTER RECOGNITION SYSTEMS WHEN WORKING WITH TEXT WRITTEN USING THE CYRILLIC ALPHABET

Kachalin V.S., Panov Yu.N., Popov N.-L.E.

Moscow Aviation Institute (National Research University), Moscow, e-mail: vasiliy.kachalin@gmail.com

In the modern world, all processes are digitalized, even the process of transferring text from paper to digital form. Optical character recognition systems (OCR) help in this. There are quite a large number of them, however, if it is necessary to work with text written in Cyrillic characters, then some of the systems are not considered due to the lack of recognition of such characters. At the same time, the need to make a choice in favor of one or another system remains. The choice of a specific optical character recognition system should be based on objective characteristics such as accuracy, speed and memory used. Studies evaluating these parameters for OCR already exist, but they were conducted on non-Cyrillic characters. The purpose of this article is to conduct a comparative analysis of some OCR systems (ABBYY FineReader, CuneiForm, OCRopus, Tesseract, Transym OCR) when working with texts written using the Cyrillic alphabet. The study was conducted on 20 scanned pages without the use of preprocessing to take into account the presence of such in OCR systems. As a result of a comparative analysis, it was found that ABBYY FineReader is the best OCR in terms of Cyrillic character recognition accuracy, while Transym OCR has the worst accuracy indicators. The fastest system turned out to be Tesseract, the slowest – OCRopus. The CuneiForm system showed the best result in terms of memory used. The worst one is ABBYY FineReader.

Keywords: OCR, optical character recognition, Cyrillic, ABBYY FineReader, CuneiForm, OCRopus, Tesseract, Transym OCR

В мире наступила эпоха цифровизации, когда бумажные документы переводятся в цифровой вид. Огромное количество текста не позволяет использовать человеческий труд из-за низкой скорости печатания. Исследование, проведенное группой ученых в 1999 г., показало, что средняя скорость ввода слов на клавиатуре человеком – 19 слов в минуту [1]. Если взять за среднее число символов в слове равное 5,1 [2], то в среднем человек вводит 5,1×19=96,9 символов в минуту. Так, чтобы перепечатать документ с 40000 символами (один авторский лист) потребуется 40000/96,9/60=6,88 ч, а таких документов может быть очень много. Поэтому практичным решением будет переложить процесс переноса текста из бумажного вида в цифровой с человека на компьютер, а человеку оставить только редактуру введенного программой текста, однако и тут можно значительно снизить участие человека, если использовать системы корректировки грамотности слов.

Преимущество передачи роли наборщика текста компьютеру также заключается в том, что компьютер не знает уста-

лости, и если человек при долгом наборе текста устает и начинает делать ошибки, то компьютер не подвержен этому и точность вводимого текста не зависит от времени работы.

Примером, когда может понадобиться данный подход, может служить следующий случай: имеется большой объем документации, представленной в бумажном виде, при этом нет цифровых оригиналов. В этой ситуации автоматическое введение текста может сильно упростить жизнь человеку. Такой подход полезен, когда надо оцифровать содержимое старой книги. Ведь постоянное перелистывание страниц может повредить листы старого документа.

В связи с этим встает вопрос: какую систему оптического распознавания символов (OCR) стоит использовать при работе с документами, написанными с помощью кириллического алфавита? Чтобы ответить на этот вопрос, необходимо провести исследование, в ходе которого будут установлены характеристики (точность, скорость, потребление памяти) различных систем оптического распознавания символов при работе с кириллическим алфавитом. В данной статье будет приведено описание такого исследования и результаты его проведения.

В мире существуют исследования, сравнивающие различные системы оптического распознавания символов [3, 4]. Однако все они были проведены на некириллических символах.

В большинстве OCR процесс распознавания символов обязательно содержит следующие этапы [5]:

− определение потенциальной области интереса;

− обнаружение признаков символов в области интереса;

− определение символа по обнаруженным признакам.

В отдельно взятых системах оптического распознавания символов могут присутствовать и дополнительные действия как, например, в Tesseract, который в процессе распознавания текста дополнительно обучается [6].

Материалы и методы исследования

Существуют различные системы оптического распознавания символов, однако часть из них может работать только с текстами, состоящими из одной латиницы. Такие OCR в данной работе рассматриваться не будут. В качестве исследуемых OCR выступают следующие системы оптического распознавания символов, в скобках указана версия: ABBYY FineReader (15.0.117.9681), CuneiForm (1.1.0), OCRopus (1.3.2), Tesseract (5.1.0.20220510) и Transym OCR (5.1). Существуют и другие OCR, которые работают с кириллическими символами, например OmniPage, однако они будут рассмотрены в последующих работах.

ABBYY FineReader – коммерческое программное решение, созданное российской компанией ABBYY. Работа над первой версией программы началась в 1992 г. из-за возникшей потребности при разработке комплекса программ Lingvo Systems [7]. Внутренняя технология распознавания держится в секрете [8]. Программа поддерживается и по сей день, периодически выходят обновления и новые версии.

CuneiForm – бесплатная система OCR, которая изначально разрабатывалась российской компанией Cognitive Technologies. Первая версия программы появилась в 1993 г. и изначально распространялась как коммерческий продукт, однако в 2008 г. Cognitive Technologies решила открыть исходный код CuneiForm всему миру. Система оптического распознавания символов в своей работе использует следующие технологии: адаптивное распознавание, нейронные сети, когнитивный анализ альтернатив распознавания, меридианная сегментация таблиц [9]. Программа CuneiForm способна распознавать тексты на более чем 17 языках, при этом программа может обрабатывать кириллицу и латиницу в одном тексте [10]. На текущий момент эта OCR не развивается.

OCRopus – набор программ с открытым исходным кодом для анализа документов, в том числе для распознавания текста. Первая версия системы анализа документов была выпущена в 2007 г. OCRopus является расширяемой системой, что позволяет изменять ее под свои нужды, стоит отметить, что модульность в его архитектуру закладывалась ещё в самом начале разработки [11]. OCRopus написан на устаревшей на сегодняшний день версии Python – 2.7.

Tesseract – система оптического распознавания символов, изначально разработанная компанией HP для использования в принтерах своего производства. Примечательно, что Tesseract начинался как PhD проект [12]. Первое публичное упоминание датируется 1995 г. на конференции, посвященной системам OCR, после чего Tesseract на какое-то время исчез из информационного поля [12]. В 2005 г. HP открыла исходный код Tesseract [6], после этого с 2006 по 2018 г. система поддерживалась и разрабатывалась компанией Google. Стоит отметить, что Tesseract проходит по тексту два раза, в первый раз он распознает

то, что получится, при этом он одновременно дополнительно обучается; и во второй раз Tesseract распознает, то что не было распознано в первый раз [6]. Также Tesseract предлагает три натренированные модели: быстрая, точная и стандартная, которая поддерживается старыми версиями программы.

Transym OCR – коммерческая система оптического распознавания символов, разработанная компанией Transym Computer Services и выпущенная в 2002 г. TOCR разработан специально для встраиваемых систем и различных интеграций. По заявлениям разработчиков для обучения TOCR используется более 108000 файлов изображений документов. Для проверки правильного распознавания слов в системе используется соответствие распознанного слова со списком часто употребляемых слов [10]. Перед распознаванием TOCR проводит предварительную подготовку изображения текста в виде перевода изображения в оттенки серого [13]. При работе с изображением документа программа сама определяет язык, на котором нужно производить распознавание текста [14].

В качестве тестовых данных для исследования использовались отсканированные страницы из книги Джека Лондона «Белый клык» в количестве 20 экземпляров. Для более объективных результатов сравнительного анализа предварительная подготовка изображений страниц книги не проводилась. Сделано это было для того, чтобы на результате распознавания отразилось наличие или отсутствие предварительной обработки изображений, непосредственно встроенной в конкретную систему оптического распознавания символов. Фрагмент одной отсканированной страницы представлен на рис. 1.

Стоит отметить, что никакие настройки в системах оптического распознавания символов не изменялись, за исключением установки русского языка, однако в Transym OCR это не применялось из-за заверений разработчиков, что система может сама определять язык. Вся работа происходила с установленными разработчиком настройками. Некоторые системы OCR являются коммерческим продуктом, в этом случае для сравнительного анализа использовались их пробные версии.

Само исследование проводилось на персональном компьютере, обладающем следующими характеристиками: процессор – Intel Core i5 8500 с частотой 3 ГГц; оперативная память – DDR4 24 Гб с частотой 1 ГГц.

Сравнение различных систем оптического распознавания символов проводилось по определенному алгоритму. Были подготовлены отсканированные страницы книги в формате PNG. Если OCR представлял собой программный фреймворк, то была написана программа, использующая его. Далее одновременно замерялось время выполнения распознавания и расход памяти в диспетчере задач. После распознавания символов полученный текст сравнивался с эталонным, то есть оригинальным; подсчитывалось число ошибок и исходя из этого вычислялась точность распознавания символов. Вычисление точности распознавания текста с кириллическими символами производилось по формуле

$$precision = \left(1 - \frac{error\ \ number}{character\ \ number}\right) \times 100\ \%,$$

где *precision* – точность распознавания символов; *error number* – число ошибок; *character number* – число символов в эталонном тексте.

В качестве еще одной меры точности распознавания использовалось расстояние Левенштейна, которое представляет собой минимальное количество операций, производимых над символами и необходимых для преобразования одной строки в другую. Чем меньшее значение имеет расстояние Левенштейна, тем больше совпадение распознанного текста и эталонного.

В качестве результирующих данных для каждой OCR были взяты средние арифметические значения результатов, полученных для всех 20 изображений текста.

Графическое представление вышеописанного алгоритма изображено на рис. 2.

Товарищ посмотрел на него с любопытством.

– Первый раз слышу, чтобы ты сомневался в их уме.

– Генри, – сказал Билл, медленно разжёвывая бобы, – а ты не заметил, как собаки грызлись, когда я кормил их?

– Действительно, возни было больше, чем всегда, – подтвердил Генри.

– Сколько у нас собак, Генри?

– Шесть.

Рис. 1. Фрагмент отсканированной страницы

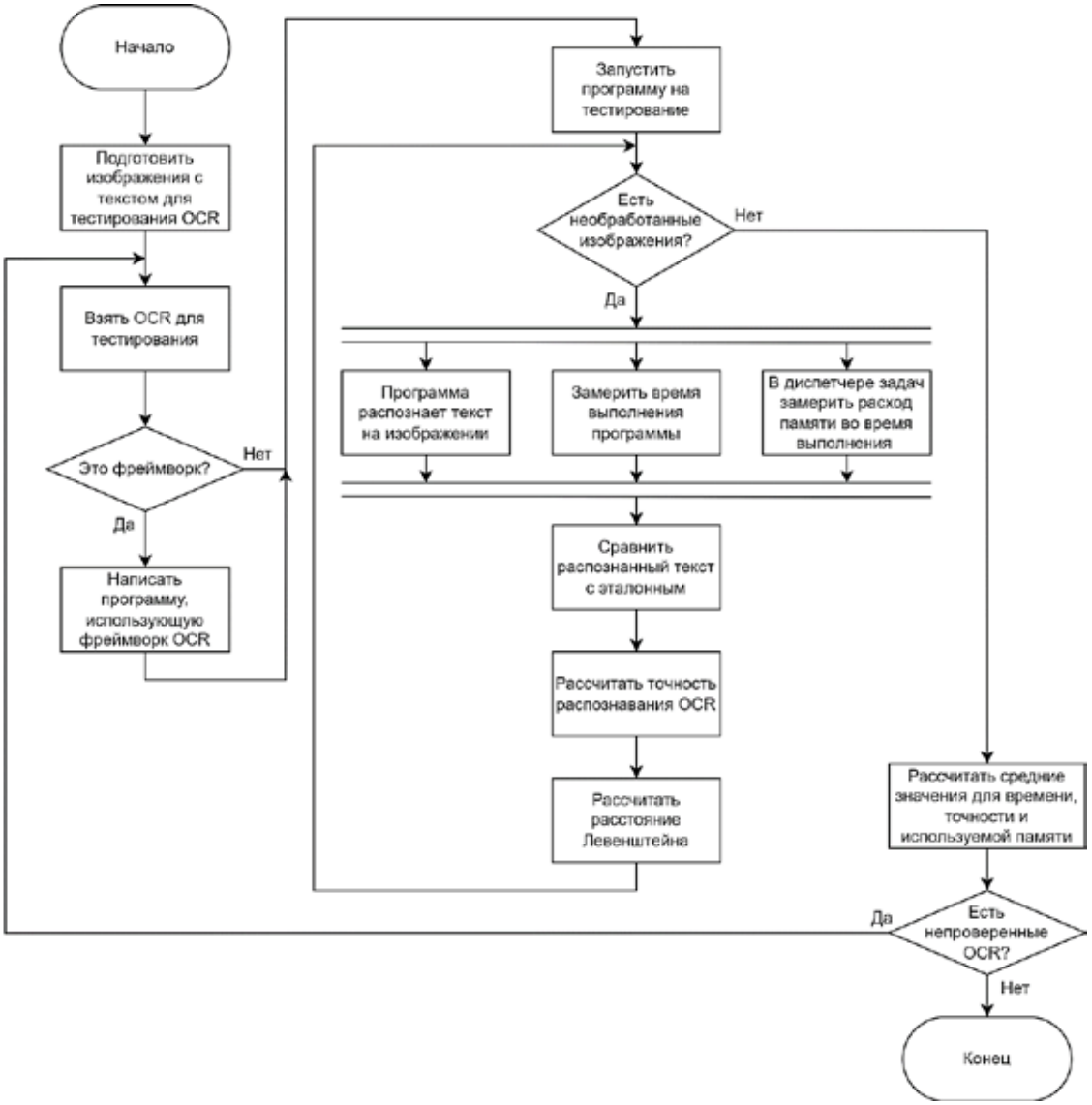

Рис. 2. Алгоритм сравнения систем оптического распознавания символов

Результаты исследования и их обсуждение

В процессе распознавания системами OCR были допущены различные ошибки. Некоторые из них представлены в табл. 1.

Результаты сравнительного анализа представлены в табл. 2.

В процессе проведения сравнительного анализа были отмечены различные особенности некоторых систем OCR. Система OCRopus требует выполнения сложной установки относительно всех остальных систем оптического распознавания символов. Помимо этого, система написана на старой версии языка программирования Python, что накладывает некоторые ограничения при работе с ней. Еще одним выделяющимся аспектом является то, что необходимо вручную запускать множество скриптовых сценариев. Весьма важную особенность также продемонстрировала система Tesseract, при выполнении распознавания текста данная OCR добавляет множество символов переноса в результат своей работы, что следует учитывать при работе с Tesseract.

В процессе распознавания символов OCR допускали различные ошибки, которые были продемонстрированы в табл. 1. Можно предположить, что при проведении аналогичного этой работе исследования при наличии гораздо большей выборки можно будет выделить некоторые частотные характеристики для пар «Правильный символ – Ошибка», наличие этой инфор-

мации позволит построить систему корректировки слов, опирающуюся на вероятностные значения. Однако стоит отметить, что для каждой OCR будет своя частотная таблица с парами «Правильный символ – Ошибка», связано это с тем, что различные OCR задействуют различные технологии при своём целевом использовании.

Таблица 1

Некоторые ошибки, допущенные различными OCR

OCR	Оригинал	Ошибка
ABBYY FineReader	И	11
	О	()
	Л	-1
CuneiForm	Он	Ип
	й	н
	Н	П
OCRopus	П	1г
	У	1
	И	П
Tesseract	Д	Ц
	М	З
	Г	Т
Transym OCR	О	()
	Н	11
	И	U

Затрагивая результаты сравнительного анализа, приведенные в табл. 2, стоит отметить некоторые моменты, которые сильно выделяются на фоне показателей остальных OCR. Так, худший результат по скорости показала система OCRopus, связано это с тем, что в процессе распознавания требуется запуск множества скриптовых сценариев. Среди моделей Tesseract, к удивлению, самой быстрой оказалась стандартная модель, она обошла быструю модель (Fast) почти на целую секунду. Transym OCR показала себя наихудшим образом по качеству распознавания, это связано с тем, что система, несмотря на заверения разработчиков, не смогла правильно определить язык, на котором написан текст. Больше всех использует оперативную память ABBYY FineReader, можно предположить, что это связано с тем, что сама система является весьма громоздкой и обладает обширным графическим пользовательским интерфейсом.

Результаты данного сравнительного анализа несут в себе следующую практическую ценность – на основе данных, приведенных в табл. 2, можно делать выбор в пользу той или иной системы OCR для применения в конкретном программном проекте.

Научная новизна исследования заключается в следующем: впервые проведен сравнительный анализ для некоторых систем оптического распознавания символов при работе с кириллическим текстом, а также установлены численные характеристики: точность, скорость, потребление памяти – различных OCR при работе с кириллическим алфавитом.

В будущем планируется проведение аналогичного исследования для других систем оптического распознавания символов. Также планируется проведение исследования, в ходе которого будут установлены частотные характеристики для пар «Правильный символ – Ошибка» для рассмотренных в этой работе систем OCR, а также для других систем оптического распознавания символов, которые будут затрагиваться в будущих исследованиях.

Таблица 2

Сводная таблица с результатами сравнительного анализа

OCR	Бесплатно	Время выполнения, с	Точность	Используемая память, Мб	Расстояние Левенштейна
ABBYY FineReader	Нет	4,15	99,14 %	150,2	16
CuneiForm	Да	2,08	78,05 %	25,1	906
OCRopus	Да	84,70	71,73 %	89,3	1048
Tesseract	Да	1,84	96,41 %	51,9	55
Tesseract Best	Да	3,11	96,52 %	39,9	53
Tesseract Fast	Да	2,96	96,25 %	43,3	66
Transym OCR	Нет	2,06	4,07 %	54,4	3825

Заключение

Опираясь на результаты исследования, можно сказать, что лучшей из рассматриваемых OCR по точности распознавания кириллических символов является ABBYY FineReader, однако данная система является коммерческой, и если стоит вопрос об использовании бесплатной системы, то тут стоит остановить свой выбор на Tesseract. По качеству распознавания хуже всего показала себя Transym OCR. С точки зрения скорости распознавания лучший результат показал Tesseract со стандартной моделью. Хуже всего результаты скорости оказались у OCRopus. Лучший результат по используемой памяти показала система CuneiForm. Худший же – ABBYY FineReader.

Стоит отметить, что среди рассматриваемых систем оптического распознавания символов нельзя однозначно выбрать лучшую, так как по каким-то критериям одна OCR лучше другой, по другим – результат противоположный. И выбор конкретной системы для проекта стоит делать исходя из налагаемых ограничений. Так, например, если критическим фактором является скорость выполнения распознавания, то, опираясь на результаты этого исследования, можно в качестве кандидатов на использование в проекте рассматривать Tesseract или CuneiForm.

Список литературы

1. Karat C.-M., Halverson C., Horn D., Karat J. Patterns of Entry and Correction in Large Vocabulary Continuous Speech Recognition System. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. 1999. P. 568–575. DOI: 10.1145/302979.303160.

2. Bochkarev V.V., Shevlyakova A.V., Solovev V.D. The average word length dynamics as an indicator of cultural changes in society. Social Evolution & History. 2015. Vol. 14. No. 2. [Электронный ресурс]. URL: https://www.sociostudies.org/journal/files/seh/2015_2/153-175.pdf (дата обращения: 11.02.2022).

3. Vijayarani S., Sakila A. Performance Comparison of OCR Tools. International Journal of UbiComp. 2015. Vol. 6. No. 3. P. 19–30. DOI: 10.5121/iju.2015.6303.

4. Wick C., Reul C., Puppe F. Comparison of OCR Accuracy on Early Printed Books using the Open Source Engines Calamari and OCRopus. Journal for Language Technology and Computational Linguistics. 2018. Vol. 33. No 1. [Электронный ресурс]. URL: https://jlcl.org/content/2-allissues/2-heft1-2018/jlcl_2018-1_4.pdf (дата обращения: 16.02.2022).

5. Mittal R., Garg A. Text extraction using OCR: A Systematic Review. 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA). 2020. P. 357–362. DOI: 10.1109/ICIRCA48905.2020.9183326.

6. Smith R. History of the Tesseract OCR engine: what worked and what didn't. Proc. SPIE 8658, Document Recognition and Retrieval XX. 2013. URL: https://www.spiedigitallibrary.org/conference-proceedings-of-spie/8658/1/History-of-the-Tesseract-OCR-engine--what-worked-and/10.1117/12.2010051.full. (дата обращения: 16.02.2022). DOI: 10.1117/12.2010051.

7. Как мы сделали ABBYY FineReader, или история, произошедшая 20 лет назад. [Электронный ресурс]. URL: https://www.abbyy.com/ru/blog/2015/11/kak-myi-sdelali-abbyy-finereader-ili-istoriya-proizoshedshaya-20-let-nazad/ (дата обращения: 16.02.2022).

8. Tafti A.P., Baghaie A., Assefi M., Arabnia H.R., Yu Z., Peissig P. OCR as a Service: An Experimental Evaluation of Google Docs OCR, Tesseract, ABBYY FineReader, and Transym. International Symposium on Visual Computing. 2016. Vol. 10072. P. 735-746. DOI: 10.1007/978-3-319-50835-1_66.

9. Технологии. 2009. [Электронный ресурс]. URL: https://web.archive.org/web/20090401061559/http://www.cuneiform.ru/tech/index.html (дата обращения: 16.02.2022).

10. Jain P., Taneja K., Taneja H. Which OCR toolset is good and why: A comparative study. Kuwait Journal of Science. 2021. Vol. 48. No. 2. P. 1–12. DOI: 10.48129/kjs.v48i2.9589.

11. Kainz O., Dujava M., Petija R., Michalko M., Jakab F. Measurement of Water Consumption based on Image Processing. 2021 IEEE 19th World Symposium on Applied Machine Intelligence and Informatics (SAMI). 2021. P. 33–38. DOI: 10.1109/SAMI50585.2021.9378611.

12. Pawar N., Shaikh Z., Shinde P., Warke Y.P. Image to Text Conversion Using Tesseract. International Research Journal of Engineering and Technology. 2019. Vol. 6. No. 2. [Электронный ресурс]. URL: https://www.irjet.net/archives/V6/i2/IRJET-V6I299.pdf (дата обращения: 16.02.2022).

13. Patel C., Patel A., Patel D. Optical Character Recognition by Open source OCR Tool Tesseract: A Case Study. International Journal of Computer Applications. 2012. Vol. 55. No. 10. P. 50–56. DOI: 10.5120/8794-2784.

14. FAQ. [Электронный ресурс]. URL: https://transym.com/faq/ (дата обращения: 16.02.2022).