

## ОБЗОР СРЕДСТВ ОПТИЧЕСКОГО РАСПОЗНАВАНИЯ ТЕКСТА НА ПРИМЕРЕ ЗАДАЧИ ОЦИФРОВКИ РУССКОЯЗЫЧНЫХ АРХИВНЫХ ЖУРНАЛОВ

© 2021 г. А. А. КАПРАНЧИКОВА, А. С. КЛЕВЦОВА

**Аннотация.** Особенности задачи оптического распознавания русскоязычного текста со страниц архивных журналов и обзор существующих средств оптического распознавания текста.

**Ключевые слова:** OCR, оптическое распознавание текста, русскоязычные журналы, библиотека, облачный сервис, .NET.

### ВВЕДЕНИЕ

Оптическое распознавание символов (англ. optical character recognition, OCR) – процесс классификации оптических шаблонов, содержащихся в цифровом изображении. Оно используется для перевода изображений рукописного, машинописного или печатного текста в текстовые данные [2]. Распознавание символов имеет широкую сферу применения - от создания электронных версий книг и документов до решения задач автоматизации систем учета данных.

Задаче оптического распознавания символов было уделено много внимания в последние несколько десятилетий, в результате чего появилось большое количество библиотек, сервисов и приложений, решающих данную задачу с разной степенью успешности. В данной статье рассмотрены некоторые из них на примере задачи оцифровки русскоязычных архивных газет и журналов.

Несмотря на то, что задачи распознавания текста широко распространены, данная задача имеет свои особенности, обусловленные спецификой источника информации. Типографика газет и журналов обладает набором особенностей и правил, отличающих ее от типографики книг. Эти особенности обуславливают ряд сложностей, с которыми могут столкнуться существующие сервисы и библиотеки для распознавания текста.

Данная задача актуальна для учреждений, имеющих в своем распоряжении архивы газет и журналов и желающих предоставить полноценный доступ к их цифровым версиям. Примером такого учреждения является «Воронежская областная детская библиотека».

### 1. СПЕЦИФИКА ТИПОГРАФИКИ ГАЗЕТ И ЖУРНАЛОВ

Типографика – искусство расположения композиции из наборного материала на плоскости листа [1]. Именно оно обуславливает особенности расположения текста на страницах газет и журналов, выбор шрифтов и множество других нюансов дизайна печатных изданий. При этом, «принятые в типографике правила – обобщение опыта, накопленного за века существования печатного дела» [1]. Типографика не является строгой наукой, поэтому любое печатное издание имеет уникальный дизайн. Однако можно проследить ряд



Рис. 1. Пример страниц журналов с разным количеством колонок

особенностей, которыми обладает типографика разных газет и журналов, и которые могут оказать существенное влияние на процесс оптического распознавания текста.

Расположение текста в газетах и журналах значительно отличается от расположения текста в книгах. Такой текст часто располагается в колонках, на которые разделена страница, при этом их количество обычно одинаково для каждой страницы конкретного номера газеты или журнала, но может отличаться для разных изданий. Наиболее часто используются 2 – 4 колонки. Возможны и другие значения количества колонок, так как они определяются дизайном конкретного издания. На рис. 1 представлены примеры страниц журналов с разным количеством колонок.

Ширина колонки зависит прежде всего от характера текста [1], и даже в рамках одного номера журнала эта ширина может изменяться. Вертикальный пробел между двумя колонками текста – межколонник – тоже может быть разной ширины [1]. При этом, строки текста в соседних колонках принято располагать на одном уровне. Эти особенности также можно увидеть на рис. 1.

В процессе распознавания текста перечисленные выше особенности могут стать проблемой: система распознавания текста может некорректно разметить блоки текста, например, восприняв строку второй колонки продолжением строки первой колонки. В таком случае распознанные текстовые данные потеряют правильную последовательность, и появится задача ее восстановления. Подобную задачу нельзя назвать простой, так как нельзя утверждать, что проблема будет возникать всегда или определенным образом. Некоторые системы распознавания текста предоставляют возможность пользователю разметить блоки текста самостоятельно, но автоматизировать данный процесс – тоже достаточно сложная задача с учетом того, что разные издания могут обладать уникальными особенностями расположения текста. Ручная же разметка подобных блоков является достаточно трудоемким процессом.

Можно выделить и другие особенности типографики газет и журналов. Например, часто используется букваца – первая прописная буква текста, главы и т.п. (обычно увеличенного размера), служащая элементом художественного оформления издания [3]. В газетах и журналах она появляется чаще, чем в книгах, порой несколько раз на одной странице. Ее

В.М. ХАРЬКОВА, заведующая Библиотечной МБОУ «Гимназия № 11», г. Елец, Липецкая область.

Школьная библиотека и безопасный Интернет

Овладение способами взаимодействия с информацией Библиотечный урок для учащихся 5-х классов

Ключевые ресурсы развития современного мира – знания и люди, их обладающие. ... богатый человек мира, Билл Гейтс, не владеет ничем осязаемым – ни землей, ни золотом или нефтью, ни фабриками, ни промашинными процессорами, ни армиями. Владельцем богатейший человек мира владеет только знаниями. (Листер К. Турбу).

Мир知识. Единица информации – бит. Развитие каждого человека.

Информация есть всеобщее свойство материи во все формы ее проявления она требует к себе внимательного и ответственного отношения.

Знание, умение овладеть способами взаимодействия с информацией – целесообразны, законны, полезны. Это, ребята, дастся вам, ребята, трудом, тренировкой.

Что такое информация? Каждый день вы узнаете что-то, чего не знали раньше, получаете новую информацию. Информация – это знания, которые выки в школе, это сведения, которые чертятся в тетради или от людей, с которыми общаетесь. Информацию, значит, вы изучаете. Кто владеет информацией, тот владеет миром.

В ходе повседневной жизни каждый из нас получает (воспринимает) самую разнообразную информацию. Но умеем ли мы правильно использовать, осмысливать, ответственно использовать разного рода знания? Рассмотрим информацию об окружающем мире – температуру, цвет, запах, вкус, качества, физические свойства предметов – люди и другие живые существа получают через органы зрения, слуха, вкуса, осязания, обоняния, через внутреннюю память и нервные системы. Но можете ли вы полностью доверять своим органам чувств? Большая часть информации вы получаете с помощью зрения.

Для получения более точной информации и дополнения к органам чувств человек издавна использует различные устройства и приборы: линейку, транспортир, термометр, барометр, весы, компас, телескоп, микроскоп и т.д. Полученную информацию человек может представить в виде записей, изображений, звуков и т.д. С давних времен люди стремились облегчить свой труд. Они придумали подвешенный экран, автомобиль, самолеты и множество других машин, механизмов и приспособлений, усиливающих их физические возможности.

Давайте примерами с вами, ребята, путешествовать во времени и узнаем, как хранится информация раньше, и сравним это с тем, как ее хранят теперь. Ведь никто теперь не будет оспаривать, что крупнейшая и старейшая книга, датируемая жироисполь на самом современном полиграфическом оборудовании альбом репродукций картин или фотоальбом – это тоже носители информации.

Знания, накопленные в течение человеческой жизни, не могут сохраниться гомогенно (неспособным) путем. Память – самый первый инструмент хранения информации. Вспомнить и сказать, сказать и передать в виде бытия, сказаний, рассказов (того, что мы называем фольклором) передавали знания от одного поколения к другому. Но стихийные бедствия, войны и эпидемии годиче уничтожали целые племена, а вместе с ними терялись и все знания, накопленные поколениями людей.

Люди всегда понимали необходимость человеческой памяти и с давних времен стремились доступными им способами зафиксировать наиболее важную информацию на каких-либо внешних носителях. По насильственному расколу и разрыву вы можете судить о том, как охотились древние люди, по ним ученые определяли не только все знания охоты, но и ритуалы, ее сопровождавшие. К сожалению, у раскопанных пис-

сформирован так: «Книга книга поет? Ребенку достаточно нажать кнопку видеорекамера и увидеть, а затем услышать звуковую подставленную герметик.

Мотивация авторских эскизов связана с субъективными факторами. Так возникает, например, выставка-настроение. Сложность ее не только в том, чтобы выразить в визуальных средствах субъективное настроение автора-библиотекаря.

Одной авторской выставки может стать проект, центральным элементом которого выступит оригинальное дизайнерское решение. Например, у сотрудников отдела «Отрочество, Юность» Ивановской области возникла идея создать необычную, привлекающую внимание выставку под названием «1 апреля».

Ее читательский адрес – дети старшего возраста, которые требуют необычного, острого или, как говорят ребята, «крикливо» решения. Она была найдена. Выставка под названием «Первоапрельский курьер» представляла собой своеобразный поэтический образовательный проект.

Важно, чтобы оно совпало с неосознанными ожиданиями посетителей, выходящих из соответствующей эмоциональной атмосферы. Библиотека в этом случае становится своего рода своеобразным мостиком-посредником.

Характер в этом смысле экспозиции цветных фотографий «Архитектура осени, подоконная ваза в одной из библиотечных областей. В ней было ярко выражено личностное начало авторского видения ее автора-библиотекаря, при этом превалировала эстетическая составляющая. Выставка сопровождалась поэзией в исполнении учащихся, а также выставкой взаимодополняющих с фотографическим. Работы библиотекаря разместили

В.М. Харькова, заведующая Библиотечной МБОУ «Гимназия № 11», г. Елец, Липецкая область.

Видеосериале «Видеосериале» была проведена программа под названием «Литературная кругосветка». Во всех залах были оформлены выставки, тематические уголки, создан интернет-старый московского суда. Романы 19-го столетия – это роман-кавалерия-нава-примать на борт первых пассажиров. Срочно

после отплытия состоялась флеш-моб. Поддержка и участие приняли школьники на водных шарах дублируя показание веса жителей городов, страны, планеты Земля. Все участники сошлись в центре и одновременно отпустили свои шары в свободный полет. Культурно-массовые мероприятия, организованное коллегами в Мещинском районе, началось также в 19.00. Выбранные для этого посетителя. Начальным гостем были представлены развлекательные программы, литературные игры, семейные вечера отдухи, конкурсы эссе, конкурсы, шашечные турниры. Открыты видеосериале и компьютерные игры.

Традиционной группой Библиотечки семейного чтения г. Чапаевска в рамках «Библиотечки» была проведена программа под названием «Литературная кругосветка». Во всех залах были оформлены выставки, тематические уголки, создан интернет-старый московского суда. Романы 19-го столетия – это роман-кавалерия-нава-примать на борт первых пассажиров. Срочно

после отплытия состоялась флеш-моб. Поддержка и участие приняли школьники на водных шарах дублируя показание веса жителей городов, страны, планеты Земля. Все участники сошлись в центре и одновременно отпустили свои шары в свободный полет. Культурно-массовые мероприятия, организованное коллегами в Мещинском районе, началось также в 19.00. Выбранные для этого посетителя. Начальным гостем были представлены развлекательные программы, литературные игры, семейные вечера отдухи, конкурсы эссе, конкурсы, шашечные турниры. Открыты видеосериале и компьютерные игры.

Видеосериале «Видеосериале» была проведена программа под названием «Литературная кругосветка». Во всех залах были оформлены выставки, тематические уголки, создан интернет-старый московского суда. Романы 19-го столетия – это роман-кавалерия-нава-примать на борт первых пассажиров. Срочно

после отплытия состоялась флеш-моб. Поддержка и участие приняли школьники на водных шарах дублируя показание веса жителей городов, страны, планеты Земля. Все участники сошлись в центре и одновременно отпустили свои шары в свободный полет. Культурно-массовые мероприятия, организованное коллегами в Мещинском районе, началось также в 19.00. Выбранные для этого посетителя. Начальным гостем были представлены развлекательные программы, литературные игры, семейные вечера отдухи, конкурсы эссе, конкурсы, шашечные турниры. Открыты видеосериале и компьютерные игры.

Видеосериале «Видеосериале» была проведена программа под названием «Литературная кругосветка». Во всех залах были оформлены выставки, тематические уголки, создан интернет-старый московского суда. Романы 19-го столетия – это роман-кавалерия-нава-примать на борт первых пассажиров. Срочно

после отплытия состоялась флеш-моб. Поддержка и участие приняли школьники на водных шарах дублируя показание веса жителей городов, страны, планеты Земля. Все участники сошлись в центре и одновременно отпустили свои шары в свободный полет. Культурно-массовые мероприятия, организованное коллегами в Мещинском районе, началось также в 19.00. Выбранные для этого посетителя. Начальным гостем были представлены развлекательные программы, литературные игры, семейные вечера отдухи, конкурсы эссе, конкурсы, шашечные турниры. Открыты видеосериале и компьютерные игры.

Видеосериале «Видеосериале» была проведена программа под названием «Литературная кругосветка». Во всех залах были оформлены выставки, тематические уголки, создан интернет-старый московского суда. Романы 19-го столетия – это роман-кавалерия-нава-примать на борт первых пассажиров. Срочно

после отплытия состоялась флеш-моб. Поддержка и участие приняли школьники на водных шарах дублируя показание веса жителей городов, страны, планеты Земля. Все участники сошлись в центре и одновременно отпустили свои шары в свободный полет. Культурно-массовые мероприятия, организованное коллегами в Мещинском районе, началось также в 19.00. Выбранные для этого посетителя. Начальным гостем были представлены развлекательные программы, литературные игры, семейные вечера отдухи, конкурсы эссе, конкурсы, шашечные турниры. Открыты видеосериале и компьютерные игры.

Видеосериале «Видеосериале» была проведена программа под названием «Литературная кругосветка». Во всех залах были оформлены выставки, тематические уголки, создан интернет-старый московского суда. Романы 19-го столетия – это роман-кавалерия-нава-примать на борт первых пассажиров. Срочно

после отплытия состоялась флеш-моб. Поддержка и участие приняли школьники на водных шарах дублируя показание веса жителей городов, страны, планеты Земля. Все участники сошлись в центре и одновременно отпустили свои шары в свободный полет. Культурно-массовые мероприятия, организованное коллегами в Мещинском районе, началось также в 19.00. Выбранные для этого посетителя. Начальным гостем были представлены развлекательные программы, литературные игры, семейные вечера отдухи, конкурсы эссе, конкурсы, шашечные турниры. Открыты видеосериале и компьютерные игры.

Видеосериале «Видеосериале» была проведена программа под названием «Литературная кругосветка». Во всех залах были оформлены выставки, тематические уголки, создан интернет-старый московского суда. Романы 19-го столетия – это роман-кавалерия-нава-примать на борт первых пассажиров. Срочно

после отплытия состоялась флеш-моб. Поддержка и участие приняли школьники на водных шарах дублируя показание веса жителей городов, страны, планеты Земля. Все участники сошлись в центре и одновременно отпустили свои шары в свободный полет. Культурно-массовые мероприятия, организованное коллегами в Мещинском районе, началось также в 19.00. Выбранные для этого посетителя. Начальным гостем были представлены развлекательные программы, литературные игры, семейные вечера отдухи, конкурсы эссе, конкурсы, шашечные турниры. Открыты видеосериале и компьютерные игры.

Видеосериале «Видеосериале» была проведена программа под названием «Литературная кругосветка». Во всех залах были оформлены выставки, тематические уголки, создан интернет-старый московского суда. Романы 19-го столетия – это роман-кавалерия-нава-примать на борт первых пассажиров. Срочно

после отплытия состоялась флеш-моб. Поддержка и участие приняли школьники на водных шарах дублируя показание веса жителей городов, страны, планеты Земля. Все участники сошлись в центре и одновременно отпустили свои шары в свободный полет. Культурно-массовые мероприятия, организованное коллегами в Мещинском районе, началось также в 19.00. Выбранные для этого посетителя. Начальным гостем были представлены развлекательные программы, литературные игры, семейные вечера отдухи, конкурсы эссе, конкурсы, шашечные турниры. Открыты видеосериале и компьютерные игры.

Видеосериале «Видеосериале» была проведена программа под названием «Литературная кругосветка». Во всех залах были оформлены выставки, тематические уголки, создан интернет-старый московского суда. Романы 19-го столетия – это роман-кавалерия-нава-примать на борт первых пассажиров. Срочно

после отплытия состоялась флеш-моб. Поддержка и участие приняли школьники на водных шарах дублируя показание веса жителей городов, страны, планеты Земля. Все участники сошлись в центре и одновременно отпустили свои шары в свободный полет. Культурно-массовые мероприятия, организованное коллегами в Мещинском районе, началось также в 19.00. Выбранные для этого посетителя. Начальным гостем были представлены развлекательные программы, литературные игры, семейные вечера отдухи, конкурсы эссе, конкурсы, шашечные турниры. Открыты видеосериале и компьютерные игры.

Видеосериале «Видеосериале» была проведена программа под названием «Литературная кругосветка». Во всех залах были оформлены выставки, тематические уголки, создан интернет-старый московского суда. Романы 19-го столетия – это роман-кавалерия-нава-примать на борт первых пассажиров. Срочно

после отплытия состоялась флеш-моб. Поддержка и участие приняли школьники на водных шарах дублируя показание веса жителей городов, страны, планеты Земля. Все участники сошлись в центре и одновременно отпустили свои шары в свободный полет. Культурно-массовые мероприятия, организованное коллегами в Мещинском районе, началось также в 19.00. Выбранные для этого посетителя. Начальным гостем были представлены развлекательные программы, литературные игры, семейные вечера отдухи, конкурсы эссе, конкурсы, шашечные турниры. Открыты видеосериале и компьютерные игры.

Видеосериале «Видеосериале» была проведена программа под названием «Литературная кругосветка». Во всех залах были оформлены выставки, тематические уголки, создан интернет-старый московского суда. Романы 19-го столетия – это роман-кавалерия-нава-примать на борт первых пассажиров. Срочно

после отплытия состоялась флеш-моб. Поддержка и участие приняли школьники на водных шарах дублируя показание веса жителей городов, страны, планеты Земля. Все участники сошлись в центре и одновременно отпустили свои шары в свободный полет. Культурно-массовые мероприятия, организованное коллегами в Мещинском районе, началось также в 19.00. Выбранные для этого посетителя. Начальным гостем были представлены развлекательные программы, литературные игры, семейные вечера отдухи, конкурсы эссе, конкурсы, шашечные турниры. Открыты видеосериале и компьютерные игры.



Рис. 2. Пример буквиц в печатных изданиях

выделяющееся художественное оформление может привести к некорректному распознаванию символа. На рис. 2 представлены примеры буквиц.

Кроме того, частота появления изображений на страницах газет и журналов гораздо выше, чем на страницах книг. Это тоже может осложнить процесс распознавания текста.

## 2. ОБЗОР СУЩЕСТВУЮЩИХ РЕШЕНИЙ

На рис. 3 представлен пример страницы журнала, который был использован для тестирования всех решений, представленных ниже.

**2.1. Tesseract.** Tesseract – средство для оптического распознавания символов с открытым исходным кодом, которое было разработано между 1984 и 1994 годами. С 2006 года Tesseract принадлежит компании Google, которая купила его и открыла исходный код под лицензией Apache 2.0 [4].

На данный момент, наиболее актуальной версией является версия 4.1.1. Tesseract можно использовать как из командной строки, так и через API, чтобы распознавать текст с цифровых изображений [8]. Он предоставляет возможность использовать в качестве основы для распознавания один из трех наборов тренировочных данных, предоставленных компанией Google, а также обучить модель на основе собственного набора данных чтобы добавить поддержку новых языков или повысить качество распознавания уже представленных. При использовании Tesseract для распознавания на русском языке достаточно использовать базовые наборы данных.

Ниже представлен пример использования Tesseract для распознавания текста с отсканированной страницы журнала в .NET Core приложении.

```
using (var engine = new TesseractEngine(@"tesdata-master", "rus",
EngineMode.Default))
{
    using (Pix img = Pix.LoadFromFile(imageFilename))
    {
        using (Page recognizedPage = engine.Process(img))
```

мья демонстрировал свое традиционное предназначение и одновременно новый «вариант использования». Издания на перевернутом стеллаже (вспомним игру в перевертыши), стояли не чинно в ряд, как на обычной полке, но в виде стопок книг. Одна из них напоминала по форме винтовую лестницу, другая — построенный из книг домик. Журналы — с закладками «по месту прописки» — интересных литературных произведений, а также статей и вовсе были беспорядочно разбросаны по подиуму.

Визуальным центром выставки стал портрет «грустного клоуна» с серьезным усталым лицом — Юрия Никулина. Рядом стояли аксессуары из фильмов с его участием: разноцветный женский зонтик, шарфик, пустая бутылка. Библиотекари — творцы выставки — рассматривали их как «приманку», направленную на привлечение непрофильного внимания.

Несомненно, беспорядок как и вроде бы случайный отбор книг по одинаковому формату, были кажущимися. Поскольку грустное и веселое всегда соседствуют в жизни, в экспозиции «участвовали» книги разных отраслей и тем, объединенные не столько жанром — юмористической литературой, книгами об искусстве комедии и пр., сколько другим признаком: их незаслуженно редко спрашивали читатели. Описанная выставка представляла собой книжно-журнальный развал, который, по данным исследований еще 1970—1980-х гг., особенно любят молодые читатели. Поскольку развалы по-своему противостоят привычной и нередко наводящей скуку «упорядоченности» традиционных форм библиотечной рекомендации.

Неординарный подход к оформлению развала в день 1 апреля, самозовонно библиотечной, которые «покусились на святое», не побоялись нарушить устоявшийся порядок и даже перевернуть стеллаж и книги вверх тормашками, подростки оценили адекватно. Развал пользовался громадным успехом, и возведенные библиотекарем книжные пирамиды рушились и исчезали на глазах создателей. Сотрудники

не устали подкладывать заранее подготовленные «резервные» издания.

Данный пример хорошо иллюстрирует довод о самостоятельной роли оформления. В авторских выставках оно является частью их содержательно-смысловой основы. Детей, растущих у экранов (телевизоров, видеоплееров, компьютеров), нелегко убедить в необходимости взять ту или иную книгу, если она напрямую не связана с процессом обучения. Их надо заинтересовать, более того — УДИВИТЬ. Не случайно ивановские библиотекари приписали свою первоапрельскую шутку к разряду выставок-сортировок. «Замешанные» на веселой фантазии, эти выставки, пользуясь фразеологией их создателей, можно причислить и к так называемым «фигурным», в основе которых находится игра объемами. Последние ориентированы на необычную форму представления. Вот, например, как можно из обычной выставки новых поступлений для дошкольников создать «фигурную». Выставку «К нам новая книга пришла» размещают на любимом детьми подиуме «лесенкой»: от нижних ступенек — сказок и рассказов для самых маленьких — до верхней, где представлены энциклопедии для дошкольников. По замыслу авторов, такое расположение книг по-своему отражает ступени читательского роста.

Игра объемами в детской библиотеке нередко сочетается с игрой цветом. Например, необычно выглядит выставка «Зеркало женской души». Разработанная талантливыми библиотекарями из Кировской областной детской библиотеки имени А. С. Грина, она была адресована девочкам-подросткам\*. Читательницы могли увидеть на столе задрапированные белыми и черными шелковыми шторами коробки разной величины с книгами русских поэтов разных эпох.

\* Информационная среда детской библиотеки Вып. 2. Книжная выставка: традиции и инновации. Из опыта работы ОДБ имени А. С. Грина. Киров, 2003. 39 с.

Попеременно чередовавшийся белый и черный цвета позволяли визуально подчеркнуть различия поколений поэтов и одновременно преемственность их творчества. Данный прием также подчеркивал неоднозначный, часто трагический жизненный путь женщин-поэтов. Сделанные тушью на белом ватмане цитаты, рисунки, виньетки хорошо смотрелись на фоне «черного шикра». На белой драпировке соседнего раздела аналогичный изобразительный ряд был представлен белыми буквами на черном листе.

Данная выставка отражала и игру линий, которые выступали своеобразным знаком эпохи. Например, портретам поэтов XIX в. соответствовали виньетки. Серебряный век был представлен черными женскими головками-профилями на белом листе. Наши современницы — черными фотографиями. Книжный ряд был также проиллюстрирован с помощью игры знаками. В качестве последних выступали «старинный» подсвечник, цветок в бокале, перчатки, шаталка и другие аксессуары, создающие романтический женский образ.

К сожалению, авторы выставки обошли вниманием выигрышное изобразительное решение, которое позволяло уточнить смыслы ее заглавия. Мы имеем в виду зеркало: при оформлении такого рода экспозиций им умело «играют» профессиональные дизайнеры. Семиотика свидетельствует, что зеркало «отсылает вперед» и одновременно замыкает пространство, поглощая его интимный характер\*.

Желание отступить от стандарта, помочь читателям развить воображение, творческую фантазию приводит к появлению таких игровых по своей природе книжно-иллюстративных экспозиций, как «выставка-ситуация». Ее методику также описывает частинский библиотекарь. Организуемая для подростков, она моделирует необычную, и иногда даже фантастическую ситуацию, о которой мечтают тинэйджеры:

\* Бодрийяр Ж. Система вещей. М.: Рудомино, 1995. С. 18.

Рис. 3. Страница журнала, использованная для тестирования решений

```
{
    string recognizedText = recognizedPage.GetText();
}
}
```

Даже в конфигурации по умолчанию Tesseract достаточно хорошо справляется с распознаванием текста с подобной страницы. Он обрабатывает блоки текста в правильной последовательности и корректно распознает большую часть символов. На рис. 4 представлены распознанные текстовые данные.

Можно заметить, что Tesseract не удалось корректно распознать буквицы, что обусловлено их внешним отличием от остального текста. Также стоит отметить, что итоговые текстовые данные нуждаются в обработке перед дальнейшим использованием: удалении лишних символов переноса слов и переноса строки. Только после подобной обработки текстовые данные можно будет использовать, например, для полнотекстового поиска.

**2.2. Iron OCR.** IronOCR — библиотека на языке C# для оптического распознавания символов с цифровых изображений и PDF документов. Она использует последнюю версию Tesseract, а также предоставляет возможность использовать предыдущие версии [6]. Для данной библиотеки существует пакет для распознавания текста на русском языке.

Помимо функциональности распознавания текста, данная библиотека предоставляет доступ к ряду методов предварительной обработки изображения, которые могут улучшить

<p>мя демонстрировал свое традиционное предназначение и одновременно новый «вариант использования». Издания на перевернутом стеллаже (вспомним игру в перевертыши), стояли не чинно в ряд, как на обычной полке, но в виде стогок книг. Одна из них напоминала по форме винтовую лестницу, другая – построенный из книг домик. Журналы – с закладками «по месту прописки» интересных литературных произведений, а также статей и вовсе были беспорядочно разбросаны по подиуму.</p>	<p>Неординарный – подход к оформлению развала в день 1 апреля, самоирония библиотекарей, которые «покусились на святое», не боялись нарушить устоявшийся порядок и даже перевернуть стеллаж и книги вверх тормашками, подростки оценили адекватно. Развал пользовался громадным успехом, и возведенные библиотекарем книжные пирамиды рушились и исцезали на глазах создателей. Сотрудники</p>	<p>для самых маленьких – до верхней, где представлены энциклопедии для дошкольников. По замыслу авторов, такое расположение книг по-своему отражает ступени читательского роста. Игра объемами в детской библиотеке нередко сочетается с игрой цветом. Например, необычно выглядит выставка «Зеркало женской души». Разработанная талантливыми библиотекарями из Кировской областной детской – библиотеки – имени А. С. Грина, она была адресована девочкам-подросткам*. – Читательницы могли увидеть на столе задрапированные белыми и черными шелковыми шторами коробки разной величины с книгами русских поэтов разных эпох.</p>	<p>XX в. соответствовали виньетки. Серебряный век был представлен черными женскими головками-профилями на белом листе. Наши современницы – чернотелыми фотографиями. Книжный ряд был также проиллюстрирован с помощью игры знаками. В качестве последних выступали «старинный» подсвечник, цветок в бокале, перчатка, шкатулка и другие аксессуары, создающие романтический женский образ.</p>
<p>Визуальным центром выставки стал портрет «грустного клоуна» с серьезным усталым лицом – Юрия Никулина. Рядом стояли аксессуары из фильмов с его участием: разноцветный женский зонтик, шарфик, пустая бутылка. Библиотекари – творцы выставки – рассматривали их как «приманку», направленную на привлечение непроизвольного внимания.</p>	<p>не устали подкладывать зараннее подготовленные «резервные» издания.</p>	<p>* Информационная среда детской библиотеки Вып. 2: Книжная выставка: традиции и инновации. Из опыта работы ОДБи имени А. С. Грина. Киров, 2003. 39 с.</p>	<p>К сожалению, авторы выставки обошли вниманием выигрышное изобразительное решение, которое – позволяло – уточнить смыслы ее заглавия. Мы имеем в виду зеркало: при оформлении такого рода экспозиций им умело «играют» профессиональные дизайнеры. Семиотика свидетельствует, что зеркало «отсылает вперед» и одновременно замыкает пространство, воплощая его интимный характер*.</p>
<p>Несомненно, беспорядок, как и вроде бы случайный отбор книг по одинаковому формату, были кажущимися. Поскольку грустное и веселое всегда соседствуют в жизни, в экспозиции «участвовали» книги разных отраслей и тем, объединенные не столько жанром – юмористической литературой, книгами об искусстве комедии и пр., сколько другим признаком: их незаслуженно редко спрашивали читатели. Описанная выставка представляла собой книжно-журнальный развал, который, по данным исследований еще 1970–1980-х гг., особенно любят молодые читатели. Поскольку развалы по-своему противостоят привычной и нередко навязывающей скуку «упорядоченности» традиционных форм библиотечной рекомендации.</p>	<p>ранный пример хорошо иллюстрирует довод о самодостаточной роли оформления. В авторских выставках оно является частью их содержательно-смысловой основы. Детей, растущих у экранов (телевизоров, видеоплееров, компьютеров), нелегко убедить в необходимости взять ту или иную книгу, если она напрямую не связана с процессом обучения. Их надо заинтересовать, более того – УДИВИТЬ. Не случайно ивановские библиотекари причислили свою первоапрельскую шутку к разряду выставок-сюрпризов. «Замешанные» на веселой Фантазии, эти выставки, пользуясь фразеологией их создателей, можно причислить и к так называемым «фигурным», в основе которых находится игра объемами. Последние ориентированы на необычную форму представления. Вот, например, как можно из обычной выставки новых поступлений для дошкольников создать «фигурную». Выставку «К нам новая книга пришла» размещают на любимом детьми подиуме «лесенкой»: от нижних ступенек – сказок и рассказов</p>	<p>Ола Попеременно чередовавшийся – белый и черный цвета позволяли визуально подчеркнуть различия поколений поэтов и одновременно преемственность их творчества. Данный прием также подчеркивал – неоднозначный, часто трагический жизненный путь женщин-поэтов. Сделанные только на белом ватмане цитаты, рисунки, виньетки хорошо смотрелись на фоне «черного ящика». На белой драпировке соседнего раздела аналогичный изобразительный ряд был представлен белыми буквами на черном листе.</p>	<p>«елание отступить от стандарта, помочь читателям развить воображение, творческую фантазию приводит к появлению таких игровых по своей природе – книжно-иллюстративных экспозиций, как «выставка-ситуация». Ее методику также описывает частинский библиотекарь. Организуемая для подростков, она моделирует необычную, и иногда даже фантастическую ситуацию, о которой мечтают тинэйджеры:</p>
		<p>Данная выставка отражала и игру линий, которые выступали своеобразным знаком эпохи. Например, портретам поэтов</p>	<p>* Бодрийяр Ж. Система вещей. М.: Рудомино. 1995. С. 18.</p>

Рис. 4. Результат распознавания с помощью Tesseract

результаты распознавания текста, а также к методам конвертации файлов разных форматов.

Данная библиотека бесплатна для разработки и тестирования, но для коммерческого использования необходимо приобрести лицензию.

Ниже представлен пример использования IronOCR для распознавания текста с отсканированной страницы журнала в .NET Core приложении.

```
var ocr = new IronTesseract();
ocr.Language = OcrLanguage.Russian;
string Text = ocr.Read(imageFilename).Text;
```

Можно заметить, что результаты распознавания данной библиотеки близки к результатам распознавания Tesseract. Блоки текста также обработаны в правильном порядке, а большинство символов распознано корректно. У данной библиотеки тоже возникли трудности с распознаванием букв, и можно заметить, что результаты их распознавания отличаются от результатов распознавания Tesseract. На рис. 5 представлены распознанные текстовые данные.

Для использования итоговых текстовых данных, например, для полнотекстового поиска, тоже потребуется их обработка – удаление лишних символов переноса строк и слов, а также некоторых лишних символов, появившихся в итоговых данных.

<p>мя демонстрировал свое традиционное предназначение и одновременно новый «вариант использования». Издания на перевернутом стеллаже (вспомним игру в перевертыши), стояли не чинно в ряд, как на обычной полке, но в виде стопок книг. Одна из них напоминала по форме винтовую лестницу, другая – построенный из книг домик. Журналы – с закладками «по месту прописки» – интересных литературных произведений, а также статей и вовсе были беспорядочно разбросаны по подиуму.</p> <p>Визуальным центром выставки стал портрет «грустного клоуна» с серьезным усталым лицом – Юрия Никулина. Рядом стояли аксессуары из фильмов с его участием: разноцветный женский зонтик, шарфик, пустая бутылка. Библиотекари – творцы выставки – рассматривали их как «приманку», направленную на привлечение непроизвольного внимания.</p> <p>Несомненно, беспорядок, как и вроде бы случайный отбор книг по одинаковому формату, были кажущимися. Поскольку грустное и веселое всегда соседствуют в жизни, в экспозиции «участвовали» книги разных отраслей и тем, объединенные не столько жанром – юмористической литературой, книгами об искусстве комедии и пр., сколько другим признаком: их незаслуженно редко спрашивали читатели. Описанная выставка представляла собой книжно-журнальный развал, который, по данным исследований еще 1970–1980-х гг., особенно любят молодые читатели. Поскольку развалы по-своему противостоят привычной и нередко навязывающей скуку «упорядоченности» традиционных форм библиотечной рекомендации.</p>	<p>Неординарный – подход к оформлению развала в день 1 апреля, самоиронии библиотечкарей, которые «покусились на святое», не побоялись нарушить устоявшийся порядок и даже перевернуть стеллаж и книги вверх тормашками, подростки оценили адекватно. Развал пользовался громадным успехом, и возведенные библиотекарем книжные пирамиды рушились и исчезали на глазах создателей. Сотрудники не устали подкладывать заранее подготовленные «резервные» издания.</p> <p>@эзнный пример хорошо иллюстрирует довод о самодостаточной роли оформления. В авторских выставках оно является частью их содержательно-смысловой основы. Детей, растущих у экранов (телевизоров, видеоплееров, компьютеров), нелегко убедить в необходимости взять ту или иную книгу, если она напрямую не связана с процессом обучения. Их надо заинтересовать, более того – УДИВИТЬ. Не случайно ивановские библиотекари причислили свою первоапрельскую шутку к разряду выставок-сюрпризов. «Замешанные» на веселой фантазии, эти выставки, пользуясь фразеологией их создателей, можно причислить и к так называемым «фигурным», в основе которых находится игра объемами. Последние ориентированы на необычную форму представления. Вот, например, как можно из обычной выставки новых поступлений для дошкольников создать «фигурную». Выставку «К нам новая книга пришла» размещают на любимом детьми подиуме «лесенкой»: от нижних ступенек – сказок и рассказов для самых маленьких – до верхней, где представлены энцикло-</p>	<p>педии для дошкольников. По замыслу авторов, такое расположение книг по-своему отражает ступени читательского роста. Игра объемами в детской библиотеке нередко сочетается с игрой цветом. Например, необычно выглядит выставка «Зеркало женской души». Разработанная талантливыми библиотекарями из Кировской областной детской – библиотеки – имени А. С. Грина, она была адресована девочкам-подросткам*. Читательницы могли увидеть на столе задрапированные белыми и черными шелковыми шторами коробки разной величины с книгами русских поэтов разных эпох.</p> <p>* Информационная среда детской библиотеки Вып. 2: Книжная выставка: традиции и инновации. Из опыта работы ОДБ имени А. С. Грина. Киров, 2003. 39 с.</p> <p>РЕ</p> <p>Попеременно чередовавшийся белый и черный цвета позволяли визуально подчеркнуть различия поколений поэтов и одновременно преемственность их творчества. Данный прием также подчеркивал – неоднозначный, часто трагический жизненный путь женщин-поэтов. Сделанные тушью на белом ватмане цитаты, рисунки, виньетки хорошо смотрелись на фоне «черного ящика». На белой драпировке соседнего раздела аналогичный изобразительный ряд был представлен белыми буквами на черном листе.</p> <p>Данная выставка отражала и игру линий, которые выступали своеобразным знаком эпохи. Например, портретам поэтов XX в. соответствовали виньетки. Серебряный век был представ-</p>	<p>лен черными женскими головками-профилями на белом листе. Наши современницы – черными фотографиями. Книжный ряд был также проиллюстрирован с помощью игры знаками. В качестве последних выступали «старинный» подсвечник, цветок в бокале, перчатки, шкатулка и другие аксессуары, создающие романтический женский образ.</p> <p>К сожалению, авторы выставки обошли вниманием выигранные изобразительные решения, которое позволяло – уточнить смыслы ее заглавия. Мы имеем в виду зеркало: при оформлении такого рода экспозиций им умело «играют» профессиональные дизайнеры. Семантика свидетельствует, что зеркало «отсылает вперед» и одновременно замыкает пространство, воплощая его интимный характер*.</p> <p>*Желание отступить от стандарта, помочь читателям развить воображение, творческую фантазию приводит к появлению таких игровых по своей природе – книжно-иллюстративных экспозиций, как «выставка-ситуация». Ее методику также описывает частинский библиотекарь. Организуемая для подростков, она моделирует необычную, и иногда даже фантастическую ситуацию, о которой мечтают тинэйджеры:</p> <p>* Бодрийяр Ж. Система вещей. М.: Рудомино. 1995. С. 18.</p>
---	--	--	---

Рис. 5. Результат распознавания с помощью IronOCR

**2.3. Google Cloud Vision API.** Google Cloud Vision API – набор сервисов в сфере компьютерного зрения от компании Google, доступ к которым предоставляется по подписке. Среди них есть OCR сервис – сервис для оптического распознавания текста. Доступ к ним осуществляется через REST API.

Для использования данного сервиса необходима регистрация аккаунта в Google Cloud Platform. OCR сервис можно использовать бесплатно при ограничении в 1000 изображений в месяц [5].

Данный сервис автоматически распознает язык текста и может работать с русским языком.

Ниже представлен пример использования Google Cloud Vision API для распознавания текста с отсканированной страницы журнала в .NET Core приложении.

```
Image image = Image.FromFile(imageFilename);
var client = ImageAnnotatorClient.Create();
TextAnnotation text = client.DetectDocumentText(image);
string detectedText = text.Text;
```

Результаты распознавания с помощью Google Cloud Vision API представлены на рис. 6.



отличаются корректным порядком слов и большим процентов правильно распознанных символов, что значительно упрощает необходимую обработку результатов.

Облачное решение Google Cloud Vision API в ряде случаев испытывает сложности с сохранением корректного порядка слов, что значительно усложнит необходимую обработку текстовых данных, полученных в результате распознавания, по сравнению с другими решениями. А облачное решение от Microsoft на данный момент не предоставляет возможность работы с текстом на русском языке в актуальной версии сервиса.

### СПИСОК ЛИТЕРАТУРЫ

1. Королькова А. Живая типографика / А. Королькова. — Москва : IndexMarket, 2012. — 224 с.
2. Chaudhuri A. Optical Character Recognition Systems / A. Chaudhuri, K. Mandaviya, P. Badelia, S. K. Ghosh. — Cham : Springer International Publishing AG, 2017. — 248 p.
3. Словари и энциклопедии на Академике [Электронный ресурс] / Академик — Режим доступа: <https://dic.academic.ru/dic.nsf/es/97444/>
4. An Overview of the Tesseract OCR Engine [Электронный ресурс] / Ray Smith, Google Inc. — Режим доступа: <https://tesseract-ocr.github.io/docs/tesseractidar2007.pdf>
5. Cloud Vision documentation [Электронный ресурс] / Google Inc. — Режим доступа: <https://cloud.google.com/vision/docs>
6. IronOCR for .NET [Электронный ресурс] / Iron Software LLC — Режим доступа: <https://ironsoftware.com/csharp/ocr/docs/>
7. Microsoft Computer Vision Documentation [Электронный ресурс] / Microsoft — Режим доступа: <https://docs.microsoft.com/en-us/azure/cognitive-services/computer-vision/overview-ocr>
8. Tesseract User Manual [Электронный ресурс] / Tesseract OCR — Режим доступа: <https://tesseract-ocr.github.io/tessdoc/>

Капранчикова Алисия Александровна

Воронежский государственный университет, факультет прикладной математики и механики, кафедра программного обеспечения и администрирования информационных систем, магистрант (ВГУ, ПММ, ПОиАИС)

E-mail: [kapranchikovaal@gmail.com](mailto:kapranchikovaal@gmail.com)

Клевцова Александра Сергеевна

Воронежский государственный университет, факультет прикладной математики и механики, кафедра программного обеспечения и администрирования информационных систем, магистрант (ВГУ, ПММ, ПОиАИС)

E-mail: [alexandra.klevtsova59@gmail.com](mailto:alexandra.klevtsova59@gmail.com)