ПРАКТИЧЕСКОЕ ЗАНЯТИЕ №3

Тема: Основы работы в программах оптического распознавания информации, в справочно-правовых системах «Консультант-плюс», «Гарант

Цели:

– Получить представление об компьютерных справочных правовых системах.

– Получить навык работы с поиском информации в компьютерных справочных правовых системах.

– Получить представление об OCR - программах распознавания текста, познакомиться с возможностями данных программы, научить распознавать отсканированный текст, передавать и редактировать его в Word.

– Знать системы распознавания символов, форм и текста; уметь пользоваться программой распознавания текста.

Задание: *Время – 2 часа.*

Задание.

1. **Перед выполнением заданий, внимательно изучите теоретический материал!**
2. **Выполнить практическую часть. Скопировать результаты выполненных заданий в** отчет в текстовом редакторе WORD
3. **Ответить на контрольные вопросы по теме «СПС ГАРАНТ, СПС КонсультантПлюс».** Ответы оформить в текстовом редакторе WORD.
4. **Ответить на контрольные вопросы по теме «Системы оптического распознания символов».** Ответы оформить в текстовом редакторе WORD.

ТЕОРЕТИЧЕСКАЯ ЧАСТЬ

1. Системы оптического распознания символов - преобразуют элементы графического изображения в последовательности символов (FineReader, CuneiForm).

2. Системы оптического распознавания форм - распознаются рукопечатные тексты (данные вводятся в поля печатными буквами).

3. Системы распознавания рукописного текста - преобразуют текст, созданный на экране карманного компьютера специальной ручкой, в текстовый компьютерный документ.

С помощью сканера достаточно просто получить изображение страницы текста в графическом файле. Однако работать с таким текстом невозможно: как любое сканированное изображение, страница с текстом представляет собой графический файл — обычную картинку. Текст можно будет читать и распечатывать, но нельзя будет его редактировать и форматировать. Для получения документа в формате текстового файла необходимо провести распознавание текста, то есть преобразовать элементы графического изображения в последовательности текстовых символов.

Преобразованием графического изображения в текст занимаются специальные программы распознавания текста (**Optical Character Recognition - OCR**).

Современная OCR должна уметь многое: распознавать тексты, набранные не только определенными шрифтами (именно так работали OCR первого поколения), но и самыми экзотическими, вплоть до рукописных. Уметь корректно работать с текстами, содержащими слова на нескольких языках, корректно распознавать таблицы. И самое главное — корректно распознавать не только четко набранные тексты, но и такие, качество которых, мягко говоря, далеко от идеала. Например, текст с пожелтевшей газетной вырезки или третьей машинописной копии. Само собой, распознать текст — это еще полдела. Не менее важно обеспечить возможность сохранения результата в файле популярного текстового (или табличного) формата — скажем, формата Microsoft Word.

Как видим, для того чтобы получить электронную, готовую к редактированию копию любого печатного текста, программе OCR необходимо выполнить «цепочку» из множества отдельных операций. Сначала необходимо распознать структуру размещения текста на странице: выделить колонки, таблицы, изображения и так далее. Далее выделенные текстовые фрагменты графического изображения страницы необходимо преобразовать в текст. Если исходный документ имеет типографское качество (достаточно крупный шрифт, отсутствие плохо напечатанных символов или исправлений), то задача распознавания решается методом сравнения с растровым шаблоном. Сначала растровое изображение страницы разделяется на изображения отдельных символов. Затем каждый из них последовательно накладывается на шаблоны символов, имеющихся в памяти системы, и выбирается шаблон с наименьшим количеством отличных от входного изображения точек.

При распознавании документов с низким качеством печати (машинописный текст, факс и так далее) используется метод распознавания символов по наличию в них определенных структурных элементов (отрезков, колец, дуг и др.).

Любой символ можно описать через набор значений параметров, определяющих взаимное расположение его элементов. Например, буква «Н» и буква «И» состоят из трех отрезков, два из которых расположены параллельно друг другу, а третий соединяет эти отрезки. Различие между данными буквами — в величине углов, которые образует третий отрезок с двумя другими. При распознавании структурным методом в искаженном символьном изображении выделяются характерные детали и сравниваются со структурными шаблонами символов. В результате выбирается тот символ, для которого совокупность всех структурных элементов и их расположение больше всего соответствует распознаваемому символу.

Наиболее распространенные системы оптического распознавания символов, например, **ABBYY FineReader** и **CuneiForm** от **Cognitive**, используют как растровый, так и структурный методы распознавания. Кроме того, эти системы являются «самообучающимися» (для каждого конкретного документа они создают соответствующий набор шаблонов символов) и поэтому скорость и качество распознавания многостраничного документа постепенно возрастают.

При заполнении налоговых деклараций, при проведении переписей населения и так далее используются различного вида бланки с полями.

Рукопечатные тексты (данные вводятся в поля печатными буквами от руки) распознаются с помощью систем оптического распознавания форм и вносятся в компьютерные базы данных. Сложность состоит в том, что необходимо распознавать написанные от руки символы, довольно сильно различающиеся у разных людей. Кроме того, система должна определить, к какому полю относится распознаваемый текст.

Системы распознавания рукописного текста.

С появлением первого карманного компьютера Newton фирмы Apple в 1990 году начали создаваться системы распознавания рукописного текста.

Такие системы преобразуют текст, написанный на экране карманного компьютера специальной ручкой, в текстовый компьютерный документ. Программы для распознавания текста вы можете приобрести отдельно или получить бесплатно вместе с купленным вами сканером.

Возможно, самая известная программа для распознавания текстов — это FineReader от компании ABBYY. Именно эту программу чаще всего вспоминают, когда речь заходит о системах распознавания.

FineReader — **омнифонтовая** система оптического распознавания текстов. Это означает, что она позволяет распознавать тексты, набранные практически любыми шрифтами, без предварительного обучения. Особенностью программы FineReader является высокая точность распознавания и малая чувствительность к дефектам печати, что достигается благодаря применению технологии "целостного целенаправленного адаптивного распознавания".

FineReader имеет массы дополнительных функций, которые простому пользователю, возможно, и без надобности, но зато производят впечатление на определенные группы покупателей. Так, одним из козырей FineReader является поддержка неимоверного количества языков распознавания — 176, в числе которых вы найдете экзотические и древние языки, и даже популярные языки программирования.

Но далеко не все возможности включены в самую простую модификацию программы, которую вы можете получить бесплатно вместе со сканером. Пакетное сканирование, грамотная обработка таблиц и изображений — для всего этого стоит приобрести профессиональную версию программы.

Все версии FineReader, от самой простой до самой мощной, объединяет удобный интерфейс. Для запуска процесса распознавания вам достаточно просто положить документ в сканер и нажать единственную кнопку (мастер Scan & Read) на панели инструментов программы. Все дальнейшие операции — сканирование, разбивку изображения на «блоки» и, наконец, собственно распознавание программа выполнит автоматически. Пользователю останется только установить нужные параметры сканирования.

FineReader работает со сканерами через TWAIN-интерфейс. Это единый международный стандарт, введенный в 1992 году для унификации взаимодействия устройств для ввода изображений в компьютер (например, сканера) с внешними приложениями. Качество распознавания во многом зависит от того, насколько хорошее изображение получено при сканировании. Качество изображения регулируется установкой основных параметров сканирования:

типа изображения,

разрешения
и яркости.

Сканирование **в сером** является оптимальным режимом для системы распознавания. В случае сканирования в сером режиме осуществляется автоматический подбор яркости. Если Вы хотите, чтобы содержащиеся в документе цветные элементы (картинки, цвет букв и фона) были переданы в электронный документ с сохранением цвета, необходимо выбрать цветной тип изображения. В других случаях используйте серый тип изображения.

Оптимальным разрешением для обычных текстов является - **300 dpi** и **400-600 dpi** для текстов, набранных мелким шрифтом (9 и менее пунктов). После завершения распознавания страницы FineReader предложит пользователю выбор: сканировать и распознавать дальше (для многостраничного документа) или сохранить полученный текст в одном из множества популярных форматов — от документов Microsoft Office до HTML или PDF. Можно, впрочем, сразу же перебросить документ в Word или Excel, и уже там исправить все огрехи распознавания. При этом FineReader полностью сохраняет все особенности форматирования документа и его графическое оформление.

КОМПЬЮТЕРНАЯ СПРАВОЧНАЯ ПРАВОВАЯ СИСТЕМА

(СПС) – это программный комплекс, включающий в себя массив правовой информации и программные инструменты, позволяющие специалисту работать с этим массивом информации (производить поиск конкретных документов или их фрагментов, формировать подборки необходимых документов, выводить информацию на печать и т.д.).

Основными свойствами систем являются:

- возможность работы с огромными массивами текстовой информации с ежедневной актуализацией и хранением базы архивных документов;

- применение в СПС специальных поисковых программных средств, позволяющих осуществлять поиск во всем информационном массиве в режиме реального времени;

- гармоничное использование в работе СПС всех вариантов телекоммуникационных технологий, в частности, электронной почты, сети Интернет, режима on-line с эталонной базой данных, хранящихся на удаленном компьютере. В этом случае не расходуется дисковое пространство пользовательского компьютера. Хотя в настоящее время наиболее распространены варианты СПС с локализованными базами данных. Это объясняется пока что недостаточным качеством российских телефонных линий, необходимостью оплаты междугородных переговоров или использования трафика (канала связи) сети, малыми сервисными возможностями при работе в режиме on-line.

На рынке сегодня существуют следующие СПС:

- - ГАРАНТ (Научно-производственное предприятие «Гарант-Сервис», 1991 г.);
- - КонсультантПлюс (акционерное общество «КонсультантПлюс», 1992 г.);
- - Кодекс (ГП «Центр компьютерных разработок», г. Санкт-Петербург).

- Кроме них на рынке представлены и такие системы:
- - ЮСИС (фирма «Интралекс», 1989 г.);
- - Референт (ЗАО «Референт-Сервис»);
- - Юридический Мир (издательство «Дело и право»);
- - Ваше право и Юрисконсульт (фирма «Информационные системы и технологии»);
- - 1С: Кодекс, 1С: Гарант, 1С: Эталон (компания «1С», известная как лидер российских экономических, бухгалтерских, производственных, управленческих и иных программных продуктов);
- - Законодательство России (Ассоциация развития банковских технологий);

ОСНОВЫ РАБОТЫ СО СПРАВОЧНОЙ ПРАВОВОЙ СИСТЕМОЙ ГАРАНТ

Окно СПС ГАРАНТ содержит Верхнее меню, Панель инструментов, Панель навигации (в левой части экрана), Основное меню с разделами: Базовый поиск, Справочная информация, Последние открытые документы, Поиск. На Панели инструментов имеется кнопка  для вызова **Основного меню.**

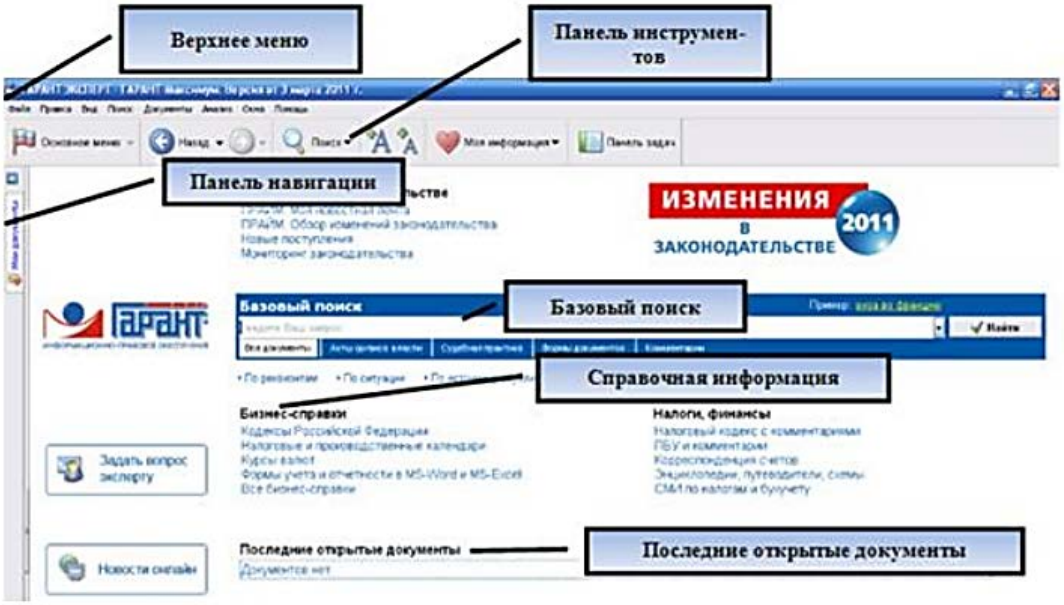


Рис.1. Внешний вид СПС ГАРАНТ

Для вызова **Панели навигатора** нажмите на Панели инструментов пикограмму

В зависимости от контекста работы на Панели навигатора (в левой части экрана) могут появляться вкладки: Основное меню, Мои документы, Документы на контроле, ПРАЙМ, Моя новостная лента, Мои консультации, Журнал работы, Толковый словарь.

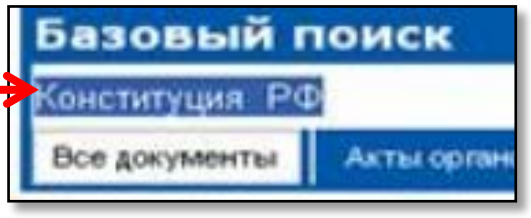
Навигатор представляет собой иерархический список разделов, в котором документы сгруппированы по нормам права или другим признакам (правая часть экрана).

Поиск документов

Базовый поиск

Пример 1. Требуется найти Конституцию РФ. В тексте закона найти статью, в которой говорится о государственной тайне.

Для поиска нажать вкладку Все документы и ввести в поле Базовый поиск ***«Конституция РФ»***



Далее кнопку Найти. Появится список документов, относящихся к этой теме. Конституция РФ в списке стоит на первом месте. Следует заметить, что Конституция – основной закон, ее можно найти почти в каждой папке.

Чтобы найти в тексте Конституции статью о государственной тайне, необходимо: открыть документ двойным щелчком мыши по названию. Далее в **Базовом поиске** ввести текст ***государственная тайна*** и нажать кнопку **Найти**. Найденное словосочетание в тексте будет выделено серым цветом. Закрыть окно **Поиск по контексту.** Выделить весь текст статьи 29 Конституции РФ и нажать кнопку **Экспорт в MS Word**. Документ сохранить в папке ГАРАНТ на диске С именем Статья 29 doc.

Поиск по реквизитам

Поиск по реквизитам является самым удобным и простым средством поиска в ГАРАНТЕ. Каждый документ характеризуется основными и дополнительными (расширенными) реквизитами, значения которых задаются в качестве условий поиска.

Чтобы выполнить поиск документов по реквизитам, нажмите клавишу F7 или кнопку Поиск по реквизитам на Панели инструментов. Система загрузит карточку запроса.

Основные сведения о справочно-правовой системе КонсультантПлюс.

Компания «КонсультантПлюс», образованная в 1992 г., является разработчиком компьютерной справочно-правовой системы Консультант Плюс.

СПС КонсультантПлюс обеспечивает доступ к различным типам правовой информации, как официальной, так и неофициальной: от нормативных актов, материалов судебной практики, комментариев, финансовых консультаций до бланков отчетности и узкоспециальных документов. Для удобства поиска информации все документы содержатся в ***Едином информационном массиве (ЕИМ)*** КонсультантПлюс. Это позволяет проводить поиск нужных документов, не заботясь о том, к какому типу информации они относятся. В то же время, поскольку документы, относящиеся к различным типам правовой информации, имеют свои специфические особенности, весь ЕИМ правовой и справочной информации системы КонсультантПлюс условно поделен на ***разделы***, каждый из

которых может содержать один или несколько близких по содержанию ***информационных банков***.

Таблица

Название раздела	Информационные банки, входящие в раздел
Законодательство	▪ **ВерсияПроф** (включая входящие в него ИБ Российское Законодательство, **Нормативные Документы** из системы **Консультант Бухгалтер Версия Проф**, **Налоги Бухучет**); ▪ **Эксперт Приложение**; ▪ **Региональный Выпуск**; ▪ **Документы СССР**.
Судебная практика	▪ **Решения высших судов**; ▪ **Подборки существенных решений**; ▪ **Окружной Выпуск**.
Финансовые консультации	▪ **Финансист** (включая входящий в него ИБ **Вопросы Ответы**); ▪ **Корреспонденция Счетов**; ▪ **Бухгалтерская пресса и книги**.
Комментарии законодательства	▪ **Постатейные комментарии и книги**; ▪ **Юридическая Пресса**.
Формы документов	▪ **Деловые Бумаги**.
Законопроекты	▪ **Законопроекты**.
Международные правовые акты	▪ **Международное Право**.
Правовые акты по здравоохранению	▪ **Медицина Фармацевтика**.
Технические нормы и правила	▪ **КонсультантПлюс: Строительство**.

Таким образом, при поиске документов пользователь получает наглядное представление, где какие документы находятся.

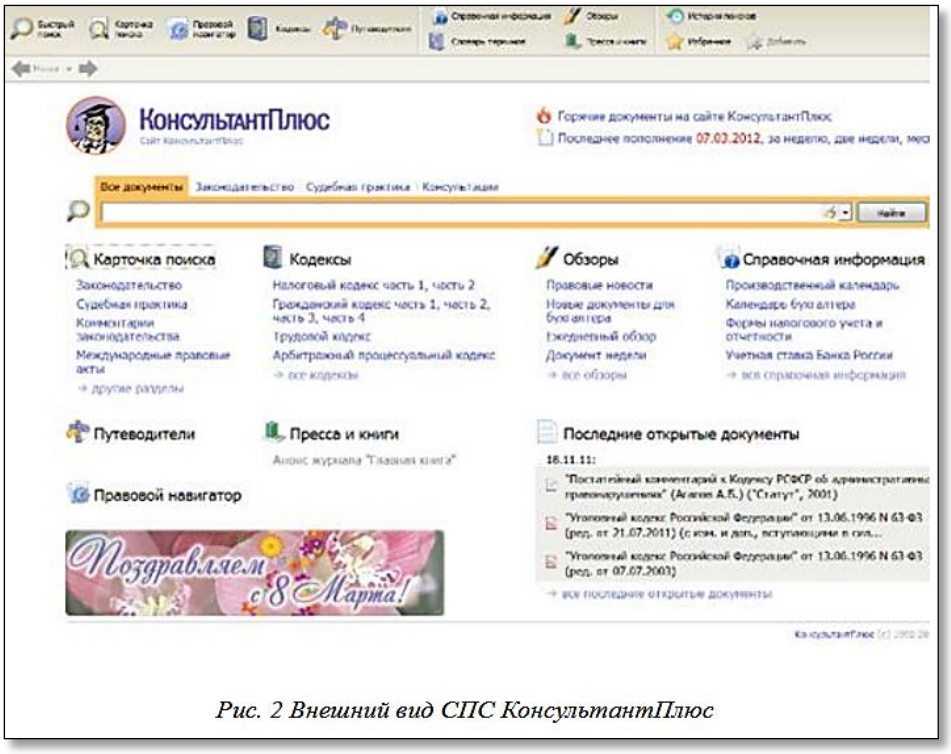
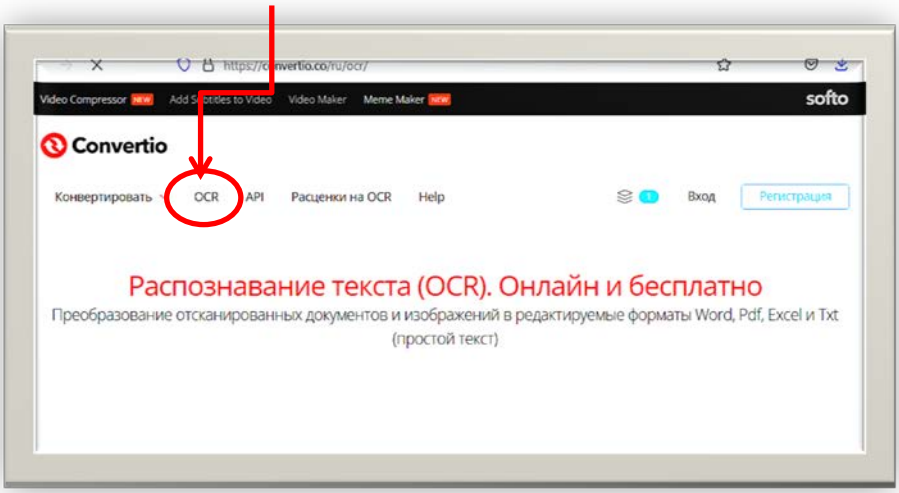


Рис. 2 Внешний вид СПС КонсультантПлюс

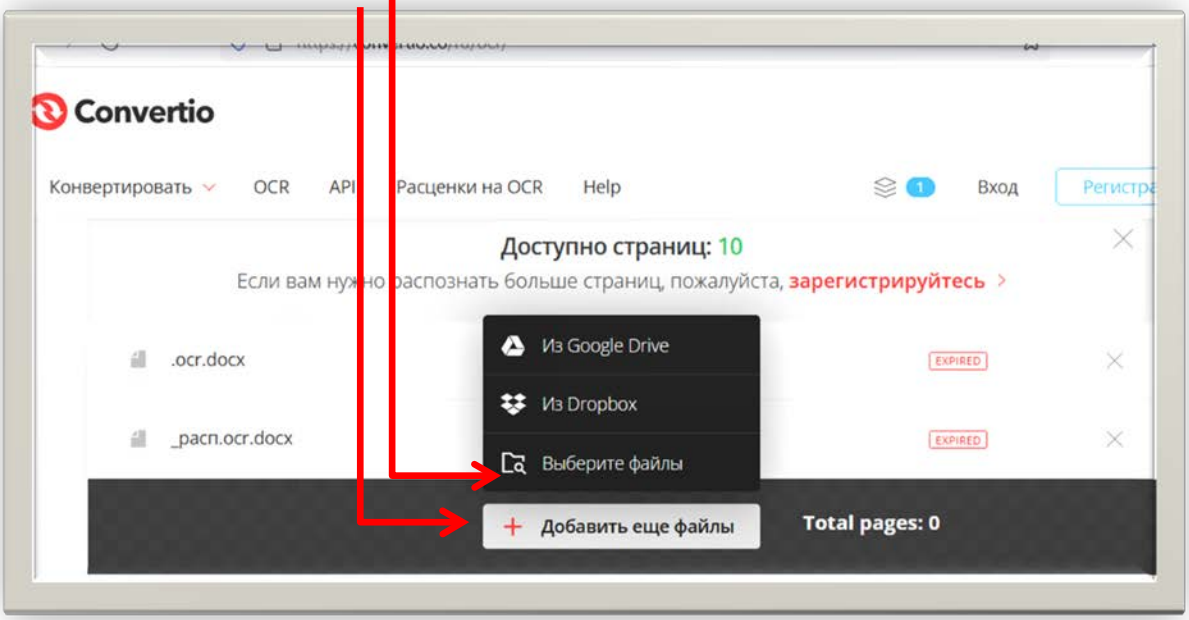
ПРАКТИЧЕСКАЯ ЧАСТЬ

Задание 1. Распознавание текста в онлайн-конвертере Convertio.Co

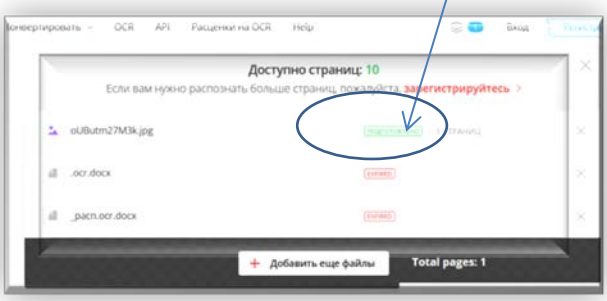
1. Заходим на сайт https://convertio.co/ru и выбираем раздел OCR

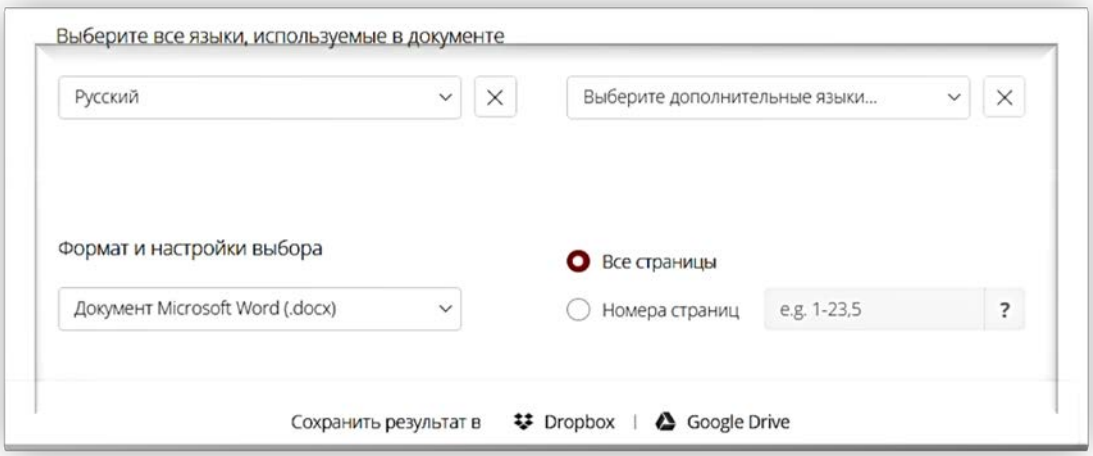


2. Загружаем отсканированный или сфотографированный файл текста, прикрепленного к этому заданию используя кнопку **+ Добавить еще файлы** и открываем в меню Выберите файлы. Затем загружаете с локального диска компьютера или смартфона данный файл.

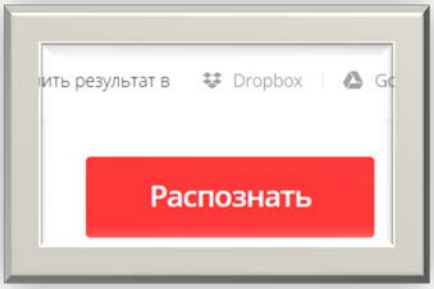


3. После того, как загрузится файл, проверьте параметры распознавания: Язык- русский, файл результата – формат DOCX

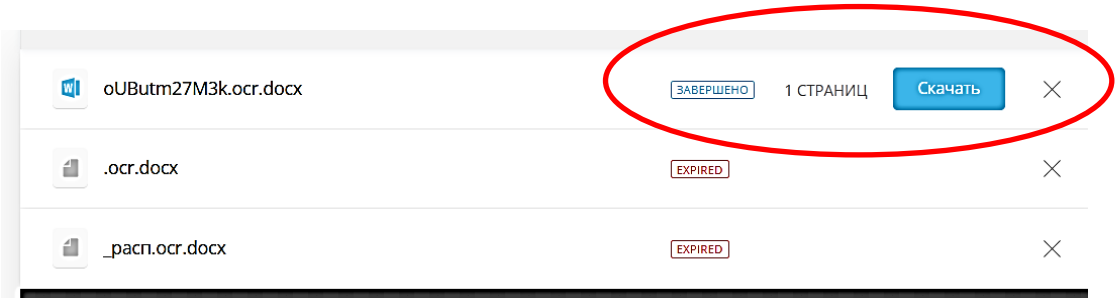




4. Ниже на странице нажмите красную кнопку РАСПОЗНАТЬ



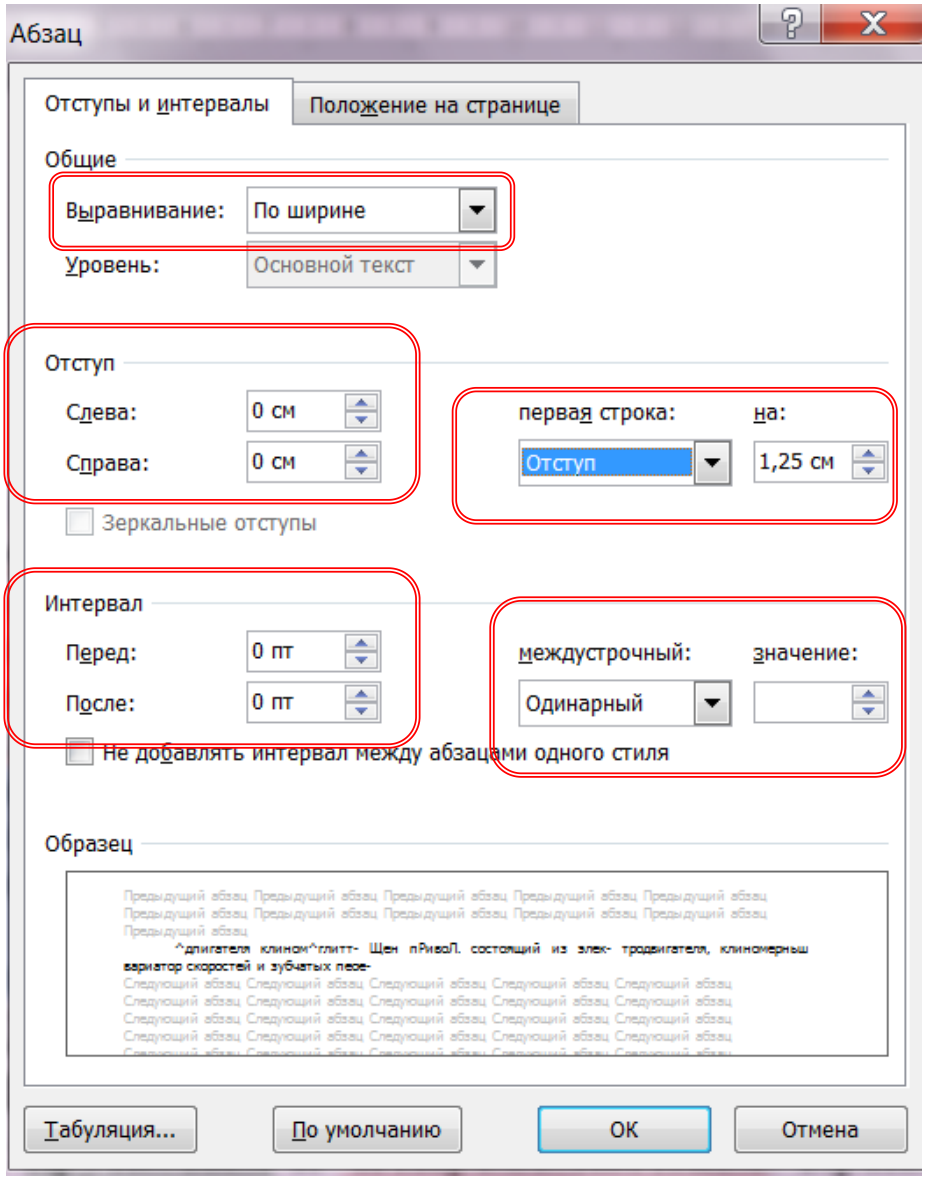
5. Скачайте файл результата



6. Откройте скачанный файл. Оцените качество распознавания текста, откорректируйте текст в программе MS Word. Исправьте ошибки распознавания и формат текста:
   1) Выделите фрагмент текста, в котором «скачут» слова
   2) Правой кнопкой мыши на выделенном фрагменте в контекстном меню выбираем «ШРИФТ»
   3) Снимаем фиксацию «надстрочный» и «подстрочный» текст
   4) Начертание и размер шрифта установите одинаковый для всего фрагмента
   5) Выделите обрабатываемый фрагмент текста
   6) Правой кнопкой мыши на выделенном фрагменте в контекстном меню выбираем «АБЗАЦ»
   7) Установите параметры для фрагмента (см. рисунок ниже)
      - выравнивание = по ширине
      - Отступы слева и справа = 0 см

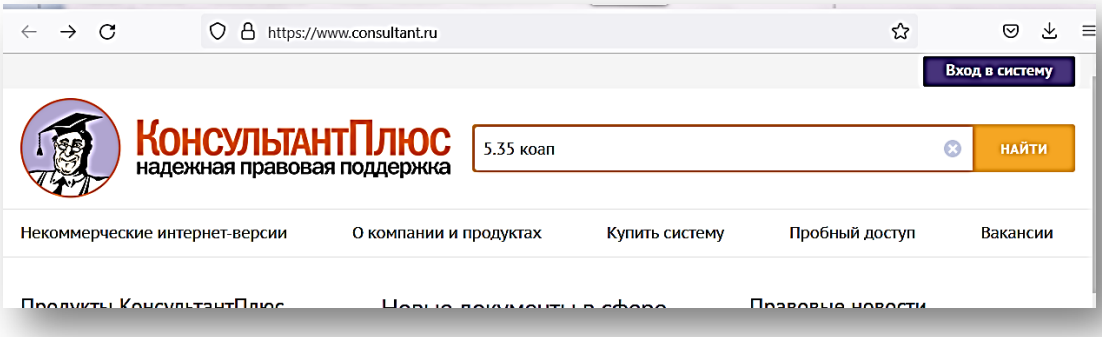
- Первая строка отступ = 1,25 см
- Интервалы перед и после =0
- Междустрочный = одинарный

8) Откорректируйте ошибки в словах и удалите лишние символы (иероглифы, значки, и т д).
9) Сверьте откорректированный текст с оригиналом (фото или отсканированный файл изображения)



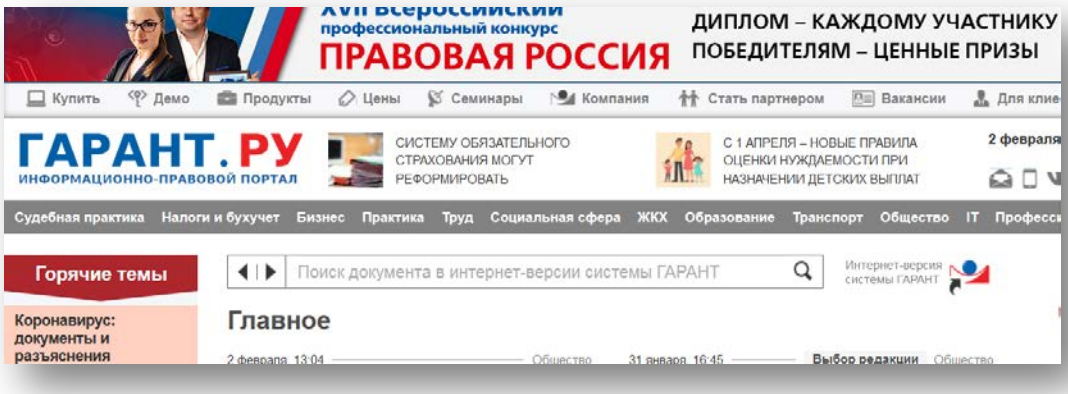
7. Сохраните отсканированный текст и скопируйте его в отчет.

Задание 2. Работа в СПС «Консультант Плюс»



1. Зайдите на сайт www.consultant.ru
2. В поисковой строке введите запрос ***5.35 КоАП***
3. Нажмите «Найти»
4. Скопируйте полный текст статьи 5.35 КоАП в отчет.

Задание 3. Работа в СПС «ГАРАНТ»



1. Зайдите на сайт www.garant.ru
2. В поисковой строке введите запрос ***Закон об образовании***
3. Нажмите «Найти» в виде 
4. Скопируйте полное название Закона об образовании с датой принятия и датой последних изменений.

Контрольные вопросы по теме:
СПС ГАРАНТ, СПС КонсультантПлюс:

1) Дайте определение понятию СПС
2) Перечислите основные свойства СПС
3) Перечислите основные продукты СПС
4) На какие структурные разделы разделяется информационный банк?
5) Перечислите, какое законодательство присутствует в банке данных СПС ГАРАНТ?
6) Какие основные возможности поиска информации присутствуют в СПС ГАРАНТ?

7) Основное назначение в СПС ГАРАНТ «Машина времени»?
8) Как в СПС ГАРАНТ найти различные схемы?
9) Основное назначение в СПС КонсультантПлюс?
10) Дайте определение понятию «информационный банк»?

Контрольные вопросы по теме:
Системы оптического распознания символов:

1) Зачем нужны программы распознавания текста?
2) Как происходит распознавание текста?
3) Какие программы распознания текста вы знаете? Какими пользовались?
4) Какое разрешение является оптимальным для сканирования текста, изображений?