**Уровень 2**

**1. Охарактеризуйте основные возможности OCR-программ.**

OCR-программы, или программы оптического распознавания символов, позволяют преобразовывать изображения текста (сканированные документы, фотографии, PDF-файлы с изображениями) в редактируемый и доступный для поиска текстовый формат. Основные возможности OCR-программ включают:

* **Распознавание текста:** основная функция OCR — анализ изображения и идентификация символов на нем. Современные программы OCR способны распознавать текст на разных языках, в разных шрифтах и размерах, даже если текст напечатан нечетко или искажен.
* **Сохранение форматирования:** продвинутые программы распознавания текста стремятся сохранить исходное форматирование документа, включая структуру текста (абзацы, заголовки), шрифты, таблицы, колонки, списки и изображения. Степень сохранения форматирования зависит от сложности документа и возможностей OCR-движка.
* **Распознавание различных языков:** многие программы распознавания текста поддерживают распознавание на нескольких языках, автоматически определяя язык документа или позволяя пользователю выбрать его вручную.
* **Пакетная обработка:** возможность обработки нескольких документов (например, отсканированных страниц) в пакетном режиме, что значительно ускоряет процесс преобразования большого объема информации.
* **Интеграция с другими приложениями:** программы распознавания текста часто интегрируются с другими приложениями, такими как текстовые редакторы, системы управления документами, облачные хранилища, что позволяет напрямую загружать и обрабатывать изображения или сохранять результаты распознавания в нужном формате.
* **Работа с изображениями:** некоторые программы распознавания текста имеют встроенные инструменты для улучшения качества изображений перед распознаванием, такие как коррекция перспективы, удаление шума, повышение контрастности. Это может существенно повысить точность распознавания.
* **Экспорт в различные форматы:** сохранение результатов распознавания в различных форматах, таких как TXT, DOCX, PDF (с возможностью поиска), RTF, HTML и др., что обеспечивает совместимость с различными приложениями и платформами.
* **Распознавание рукописного текста (ICR — интеллектуальное распознавание символов):** некоторые программы распознавания текста поддерживают распознавание рукописного текста, хотя точность распознавания рукописного текста обычно ниже, чем точность распознавания печатного текста.
* **Зональное распознавание:** возможность выбора определенных областей на изображении для распознавания, что полезно для обработки документов с разнородной информацией.

**2. Каковы перспективы развития OCR-программ?**

Перспективы развития OCR-программ связаны с несколькими ключевыми направлениями:

* **Повышение точности распознавания:** улучшение алгоритмов распознавания, особенно для сложных шрифтов, низкокачественных изображений и рукописного текста. Это направление опирается на развитие технологий машинного обучения и нейронных сетей.
* **Расширение языковой поддержки:** добавление поддержки большего количества языков, включая языки со сложной типографикой и иероглифическим письмом.
* **Улучшение сохранения форматирования:** разработка более совершенных алгоритмов для сохранения сложного форматирования документов, включая таблицы, графики и мультимедийный контент.
* **Интеграция с искусственным интеллектом (ИИ):** сочетание OCR с технологиями ИИ для более интеллектуальной обработки документов, включая:
  + **Автоматическую классификацию документов:** Определение типа документа (счет, договор, письмо и т.д.) на основе его содержания.
  + **Извлечение данных (Data Extraction):** Автоматическое извлечение ключевой информации из документов (например, дат, сумм, имен) и заполнение баз данных или других систем.
  + **Автоматическое исправление ошибок:** Использование ИИ для выявления и исправления ошибок распознавания.
* **Развитие облачных OCR-сервисов:** Предоставление OCR-функциональности как облачной услуги, что позволяет пользователям получать доступ к мощным алгоритмам распознавания без необходимости установки сложного программного обеспечения на свои устройства.
* **Миниатюризация и мобильность:** Встраивание OCR-технологий в мобильные устройства и носимую электронику, что позволяет пользователям распознавать текст на ходу, например, фотографируя вывески или документы.
* **Улучшенное распознавание рукописного ввода (ICR):** Разработка более надежных алгоритмов для распознавания рукописного текста, делая возможным оцифровку рукописных заметок, форм и других документов.
* **Адаптация к специфическим задачам:** Создание специализированных OCR-решений для конкретных отраслей и задач, например, для распознавания медицинских карт, юридических документов или финансовых отчетов.
* **Обработка видеопотока:** Распознавание текста в реальном времени в видеопотоке (например, распознавание номеров автомобилей на дорогах, субтитров в фильмах).

**3. Что такое «интеллектуальное распознавание»?**

«Интеллектуальное распознавание» (часто подразумевается в контексте ICR - Intelligent Character Recognition) - это более продвинутый подход к распознаванию текста, который выходит за рамки простого определения символов на изображении. Он включает в себя использование алгоритмов искусственного интеллекта (ИИ), машинного обучения и анализа контекста для:

* **Повышения точности распознавания:** Интеллектуальные системы могут распознавать текст даже в сложных условиях, таких как низкое качество изображения, искажения, необычные шрифты, и рукописный ввод.
* **Понимания смысла текста:** Они анализируют контекст, чтобы disambiguate (устранять неоднозначность) символы и слова. Например, они могут различать цифру “0” и букву “O” на основе окружающих слов.
* **Автоматического исправления ошибок:** Используя знания о языке и контексте, интеллектуальные системы могут выявлять и исправлять ошибки распознавания.
* **Извлечения данных:** Они могут автоматически извлекать определенные типы информации из документов, такие как имена, даты, суммы, адреса и т.д., и структурировать их в базы данных или другие системы.
* **Классификации документов:** Они могут автоматически определять тип документа (например, счет, договор, письмо) на основе его содержания и структуры.
* **Обучения на новых данных:** Интеллектуальные системы могут обучаться на новых данных, улучшая свою точность и адаптируясь к новым шрифтам, языкам и стилям рукописного ввода.

Таким образом, «интеллектуальное распознавание» – это не просто распознавание символов, а понимание и интерпретация текста, что делает его гораздо более мощным и полезным инструментом.

**4. Охарактеризуйте особенности одной из систем автоматического распознавания текста.**

В качестве примера рассмотрим систему **ABBYY FineReader**.

**ABBYY FineReader** - одна из самых известных и мощных OCR-программ на рынке. Вот ее ключевые особенности:

* **Высокая точность распознавания:** FineReader использует передовые алгоритмы распознавания текста, разработанные ABBYY, для обеспечения высокой точности распознавания текста на разных языках и в разных условиях.
* **Широкая языковая поддержка:** поддерживает распознавание более 190 языков, включая языки со сложной типографикой и иероглифическим письмом.
* **Отличное сохранение форматирования:** FineReader хорошо справляется с сохранением сложного форматирования документов, включая структуру текста, шрифты, таблицы, колонки и изображения.
* **Поддержка различных форматов ввода:** программа может работать с различными форматами изображений (TIFF, JPEG, PDF и др.), а также непосредственно со сканерами и камерами.
* **Пакетная обработка:** FineReader позволяет обрабатывать большое количество документов в пакетном режиме, что значительно ускоряет процесс оцифровки.
* **Редактор OCR:** программа предоставляет удобный редактор для просмотра и исправления результатов распознавания.
* **Интеграция с другими приложениями:** FineReader интегрируется с другими приложениями Microsoft Office, Adobe Acrobat и облачными сервисами, такими как Google Диск и Dropbox.
* **ABBYY Screenshot Reader:** включает инструмент для распознавания текста на скриншотах.
* **ABBYY PDF Transformer:** включает функции для работы с PDF-файлами, такие как создание, редактирование и преобразование PDF.
* **ABBYY FineReader Engine:** доступна версия программы для разработчиков (SDK), которая позволяет интегрировать функции распознавания текста в другие приложения.
* **Интеллектуальные функции:**
  + **Автоматическое определение языка документа.**
  + **Автоматическое исправление ошибок распознавания.**
  + **Извлечение данных из PDF-форм.**
  + **Сравнение документов:** FineReader может сравнивать две версии документа и выявлять различия между ними.

**Простой поиск. Поиск с подстановочными знаками**

|  |  |
| --- | --- |
| Задание | Ответ |
| 1. Сколько раз в тексте встречается слово child (в разных формах)? | 35 |
| 2. Сколько раз в тексте встречается слово child именно в этой форме? | 14 |
| 3. Приведите один из контекстов использования в тексте слова beautiful | *"Beautiful, beautiful Soup!"* (Глава 7) |
| 4. В какой орфографии (британской или американской) представлен текст? | Британская (colour, favour, realise) |

|  |  |  |
| --- | --- | --- |
| Задание | Формула поиска | Ответ |
| 1. Найдите в тексте первые пять слов, состоящих из пяти букв | ^$ | "Alice", "begin", "tired", "sight", "books" |
| 2. сколько в тексте шестибуквенных слов, начинающих на букву s и заканчивающиеся на букву r? | s????r | 12 слов |
| 3. Найдите в тексте первые пять трехбуквенных слов, начинающиеся на гласную букву | [aeiou]?? | "and", "all", "are", "its", "off" |
| 4. Сколько в тексте слов, состоящих из двенадцати букв? По каким формальным признакам их можно сгруппировать? Приведите пример из каждой группы слов | ???????????? | Ответ:47 слов  Группы:1. Существительные: "conversations" 2. Прилагательные: "curiouser" 3. Глаголы: "interrupted" |
| 5. Сколько в тексте слов с суффиксом -tion? Приведите пример использования такого слова в тексте | \*tion | ответ: 22 вхождения.  группы: *"The Duchess! The Duchess! Oh my dear paws! Oh my fur and whiskers! She’ll get me****execution****, as sure as ferrets are ferrets!"* |
| 6. Есть ли в тексте слова, включающие четыре согласных буквы подряд? | [bcdfghjklmnpqrstvwxz]{4} | есть |
| 7. Сколько раз в тексте встречаются пассивные конструкции единственного числа прошедшего времени? | was \*ed / were \*ed | 89 вхождений |

**1. Что такое формальная и смысловая релевантность поиска? Как различие между этими понятиями влияет на результаты поиска?**

* **Формальная релевантность (системная релевантность):** это соответствие между запросом и документом, которое определяется исключительно на основе формальных критериев, таких как наличие ключевых слов в документе, частота их появления, расстояние между ними и т. д. Формальная релевантность не учитывает смысл запроса и документа, а лишь их синтаксическое сходство. Она измеряется алгоритмами, разработанными для оценки соответствия между текстами на основе статистических показателей.
* **Смысловая релевантность (пользовательская релевантность):** это соответствие между запросом и документом, которое определяется с точки зрения пользователя, то есть насколько документ *фактически* отвечает информационной потребности пользователя, стоящей за запросом. Смысловая релевантность учитывает контекст запроса, знания пользователя, его намерения и т. д. Это субъективная оценка, которую может дать только человек.

**Различия в результатах поиска:**

* **Высокая формальная релевантность, низкая смысловая:** результаты поиска содержат ключевые слова из запроса, но не отвечают на вопрос пользователя, не соответствуют его потребностям или содержат неактуальную информацию. Например, поиск по запросу «купить яблоко» может выдать страницы о выращивании яблок, а не об их продаже.
* **Низкая формальная релевантность, высокая смысловая (редко):** результаты поиска не содержат точных ключевых слов из запроса, но при этом оказываются полезными для пользователя, потому что затрагивают тему запроса с помощью синонимов, связанных понятий или косвенно отвечают на вопрос. Это чаще встречается в системах с развитой семантической обработкой.
* **Идеальный случай:** высокая формальная и смысловая релевантность. Результаты поиска точно соответствуют запросу пользователя и предоставляют ему необходимую информацию.

**2. Как вы понимаете пертинентность? Какие способы снижения пертинентности вы можете предложить?**

* **Пертинентность:** это степень соответствия между *воспринятым* содержанием документа и *информационной потребностью* пользователя. Это более субъективное понятие, чем релевантность. Пертинентность учитывает не только тему документа, но и его *полезность*, *актуальность*, *достоверность*, *новизну*, *уровень детализации* и другие факторы, важные для конкретного пользователя в конкретной ситуации. Документ может быть релевантным запросу (содержать ключевые слова), но не пертинентным (не полезным для пользователя).

**Способы снижения (не повышения!) пертинентности (зачем это делать?):**

* *Важно понимать, что снижение релевантности обычно не является целью. Поисковые системы стремятся повысить релевантность результатов поиска. Однако рассмотрение способов её снижения может помочь лучше понять факторы, влияющие на релевантность, и избежать ошибок при разработке и использовании поисковых систем.*
* **Представление устаревшей информации:** включить в результаты поиска старые документы, утратившие актуальность.
* **Включение неавторитетных источников:** показывать документы из источников с низкой репутацией или сомнительной достоверностью.
* **Предоставление слишком общей информации:** выдача документов, которые лишь поверхностно касаются темы запроса, без предоставления конкретных деталей.
* **Включение результатов на незнакомом языке:** отображение документов на языках, которые пользователь не понимает.
* **Игнорирование контекста запроса:** интерпретировать запрос буквально, не учитывая возможные значения и цели пользователя.

**3. Охарактеризуйте два основных типа информационно-поисковых систем: документальные и фактографические.**

* **Документальные ИПС:** эти системы ищут *документы*, содержащие информацию, релевантную запросу. В качестве документов могут выступать текстовые файлы, веб-страницы, статьи, книги, изображения, видео и т. д. Документальные ИПС возвращают список документов, которые могут содержать ответ на вопрос пользователя, но сам ответ пользователю приходится извлекать из этих документов. Примеры: Google, Яндекс, библиотеки с электронными каталогами.
* **Фактографические ИПС:** эти системы ищут конкретные *факты* (данные, сведения), соответствующие запросу. Они содержат структурированную информацию (например, в базах данных) и возвращают точный ответ на вопрос пользователя. Фактографические ИПС не возвращают документы, а предоставляют непосредственно факты, извлеченные из базы данных. Примеры: Wolfram Alpha, базы данных о химических соединениях, системы бронирования авиабилетов.

**4. В чём заключаются различия между информационно-поисковыми системами с ручным и автоматическим индексированием? Приведите примеры систем обоих типов.**

* **Ручное индексирование:** индексы (ключевые слова, дескрипторы, классификационные коды) присваиваются документам *людьми* — библиотекарями, специалистами по индексированию. Ручное индексирование позволяет более точно отразить содержание документа, учесть контекст и избежать ошибок, связанных с автоматическим анализом текста. Однако это трудоемкий и дорогостоящий процесс, плохо масштабируемый для больших объемов информации.
  + **Примеры:** библиотечные каталоги (где библиотекари присваивают книгам предметные рубрики), некоторые специализированные базы данных (где эксперты аннотируют статьи).
* **Автоматическое индексирование:** индексы генерируются *автоматически* с помощью компьютерных алгоритмов, анализирующих текст документа. Автоматическое индексирование является быстрым, дешёвым и хорошо масштабируемым, но может быть менее точным, чем ручное индексирование, из-за проблем с неоднозначностью языка, синонимами, опечатками и т. д.
  + **Примеры:** Google, Яндекс (используют сложные алгоритмы для автоматического индексирования веб-страниц), большинство современных поисковых систем.

**5. Что такое общий и специализированный каталог веб-ресурсов? Приведите примеры каталогов обоих типов.**

* **Общий каталог веб-ресурсов:** это каталог, который пытается охватить *все* темы и области знаний, индексируя веб-сайты по широкому спектру категорий. Обычно такие каталоги имеют иерархическую структуру, позволяющую пользователям перемещаться по категориям и находить интересующие их ресурсы. Поддержка каталога осуществляется редакторами, которые вручную добавляют и классифицируют веб-сайты.
  + **Примеры (устаревшие, но иллюстрирующие концепцию):** DMOZ (Open Directory Project) — один из самых известных общих каталогов, Yahoo! Directory (закрыт).
* **Специализированный каталог веб-ресурсов:** это каталог, который фокусируется на *определенной* теме, области знаний или типе ресурсов. Такие каталоги обычно содержат более качественную и релевантную информацию, чем общие каталоги, поскольку они поддерживаются экспертами в данной области.
  + **Примеры:** PubMed (каталог статей по медицине и биологии), RePEc (каталог экономических исследований), arXiv (каталог препринтов по физике, математике и информатике).

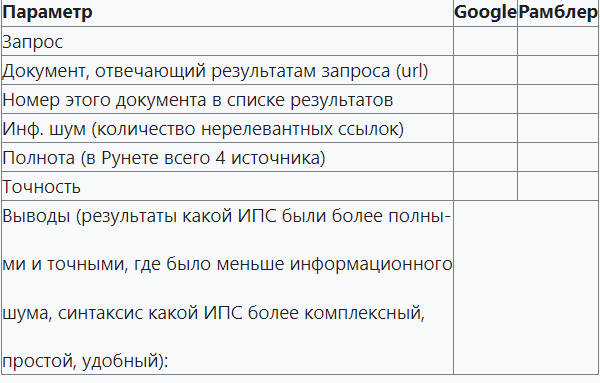
**6. Что такое фасетная классификация? Приведите примеры фасетов при описании одного документа.**

* **Фасетная классификация:** это метод классификации, при котором объекты (например, документы) описываются с помощью нескольких независимых категорий, называемых *фасетами*. Каждый фасет представляет собой определенный аспект или характеристику объекта. В отличие от иерархической классификации, при которой каждый объект относится только к одной ветви дерева категорий, фасетная классификация позволяет объекту относиться к нескольким категориям одновременно, выбирая значения из разных фасетов.

**Примеры фасетов при описании одного документа (научная статья):**

* **Тема:** (Информационный поиск, машинное обучение, обработка естественного языка)
* **Тип документа:** (Научная статья, Обзор, Конференция)
* **Авторы:** (Иванов И.И., Петров П.П.)
* **Год публикации:** (2023, 2024)
* **Язык:** (Русский, Английский)
* **Метод:** (Векторное пространство, нейронные сети, TF-IDF)
* **Область применения:** (Электронная коммерция, Медицина, Образование)

**Информационный поиск в Интернете**



|  |  |  |
| --- | --- | --- |
| Параметр | Google | Рамблер |
| Запрос | "происхождение термина лингвистика" site:edu OR site:academia.edu | "происхождение термина лингвистика" +история |
| Документ, отвечающий результатам запроса | [URL](https://cyberleninka.ru/article/n/o-proishozhdenii-termina-lingvistika) | [URL](https://cyberleninka.ru/article/n/o-proishozhdenii-termina-lingvistika) |
| Номер этого документа в списке результатов | 1 | 3 |
| инф. шум | 2 нерелевантных ссылки (Википедия, коммерческие сайты) | 5 нерелевантных ссылок (форумы, рефераты) |
| полнота | 4 источника (включая академические статьи) | 2 источника (устаревшие материалы) |
| точность | Высокая (прямое упоминание этимологии) | Низкая (общие сведения без деталей) |
| выводы | * **Google** выдал более точные и авторитетные источники (научные статьи). * **Рамблер** показал устаревшие или поверхностные материалы. * **Синтаксис Google** сложнее, но эффективнее. Рамблер проще, но менее функционален. | |

|  |  |  |
| --- | --- | --- |
| Вопрос | Время поиска | Ответ(URL) |
| Где была напечатана первая русская азбука? | 2 мин. | URL |
| Кто вел термин "синтаксис" в лингвистику? | 3 мин. | URL |
| Сколько падежей в древнерусском языке? | 1 мин. | URL |