

Интеллектуальная система автоматического извлечения знаний из документов

1. Цель системы

Система предназначена для преобразования неструктурированных и слабоструктурированных текстовых данных из разнородных документов в структурированные, машиночитаемые знания, пригодные для анализа, интеграции в базы знаний и поддержки принятия решений. Это позволит автоматизировать обработку документов, увеличить скорость обработки массивных документов, выявлять скрытые семантические связи, предоставлять пользователям интуитивно понятный доступ к извлечениям знаний.

2. Пользователи системы

1. *Специалист анализа и исследования* — ищет факты, анализирует рынок и научные исследования, строит отчеты.
2. *Специалист по данным* — наполняет датасеты, обучает ML-модели на тестовых данных.
3. *Руководители* — принимают стратегию развития системы, анализируют отчетность и риски.

3. Структура системы и её подсистемы

1. *Подсистема сбора данных* — принимает документы разных форматов (PDF, DOCX, HTML, JPG/PNG) через Web-интерфейс, почту, сканирование; интегрирует данные с облачными хранилищами (Google Drive, S3).
2. *Подсистема предобработки, OCR*, — конвертирует форматы, извлекает текст из сканов и PDF с помощью OCR (Tesseract, ABBYY), очищает текст (удаление шумов, нумерации страниц, колонтитулов), определяет язык и кодировки текста.
3. *Подсистема NLP-обработки* — распознает именованная сущностей, определяет отношения между сущностями, строит онтологии.
4. *Подсистема хранения данных* — хранит документы, сущности, связи между сущностями, факты.
5. *Подсистемы поиска и визуализации* — обеспечивает веб-интерфейс для пользователей, визуализирует знания в виде графов, итоговых отчетов.

4. Связи и взаимодействия между подсистемами

1. Данные поступают из подсистемы сбора в подсистему предобработки, где преобразуются в чистый текст.
2. Чистый текст поступает в подсистему NLP-обработки. Здесь он последовательно проходит этапы анализа, где из него извлекаются сущности, связи и другие метаданные.
3. Результаты сохраняются в подсистеме хранения данных.
4. Пользователи вносят изменения, которые учитываются в последующем цикле обработки и анализа.

5. Основные сценарии работы пользователей с системой

1. Сценарий поиска связей между компаниями. Аналитик получает все упоминания искомой компании, граф ее связей с другими компаниями и людьми, выделяя тип связи.
2. Анализ контракта. Юрист загружает проект договора, система выделяет потенциальные риски и визуализирует и визуализирует информацию в виде таблицы.
3. Наполнение базы знаний. Data scientist отправляет запрос на обработку потока научных статей, результаты возвращаются в рабочую среду для дальнейшего анализа.

6. Три направления развития системы

1. Углубление семантического понимания и причинно-следственных связей: переход от извлечения простых фактов к пониманию сложных утверждений, гипотез, причинно-следственных связей и логических умозаключений в тексте; использование более сложных языковых моделей (LLM, like GPT), обученных на предметных областях, и методов логического вывода.
2. Превращение в активного интеллектуального ассистента (Active AI Assistant): система не только отвечает на запросы, но и сама предлагает факты, генерирует краткие содержания, формирует гипотезы и предупреждает о важных изменениях или аномалиях в документах.
3. Мультимодальность и работа со сложными документами: извлечение знаний не только из текста, но и из таблиц, диаграмм, схем и изображений внутри документов с последующим

объединением этой информации в единую семантическую модель.