

Автоматизация препродакшн-подготовки сценариев

Команда DiverCity

Сервис для автоматического создания препродакшн-таблиц из сценариев

Проблема

- Подготовка препродакшн-документации занимает многочасовые монотонные правки
- Ручное определение локаций, времени суток, персонажей, массовки, реквизита
- Ошибки приводят к сбоям в графике и дополнительным затратам на площадке
- Отсутствие автоматического распознавания производственных элементов
- Неоднозначные формулировки и опечатки в сценариях

Решение

Автоматический сервис, который загружает сценарий в формате PDF или DOCX и генерирует готовую препродакшн-таблицу со всеми производственными элементами.

- Автоматическое распознавание номеров сцен и сегментация текста
- Извлечение локаций, времени суток, персонажей, массовки, реквизита
- Определение спецэффектов, трюков, транспорта, животных
- Генерация структурированной таблицы с настраиваемыми столбцами
- Экспорт в CSV и XLSX с сохранением кодировок

Архитектура решения

Компонент	Технология	Назначение
Frontend	Streamlit	Интерфейс пользователя
Backend	FastAPI	REST API
Обработка данных	PyMuPDF, python-docx	Обработка PDF и DOCX
Анализ текста	Regex, Python	Анализ текста
Анализ данных	Pandas	Анализ данных

Корректная сегментация сценария

Система автоматически распознает номера сцен в различных форматах:

- Поддержка различных форматов: 'СЦЕНА 11', '11-N2', '1-11N2', '15-N6-04', '3/П'
- Автоматическое разбиение текста на отдельные сцены
- Корректная обработка документов в форматах PDF и DOCX
- Поддержка различных кодировок (UTF-8, UTF-16, CP1251, KOI8-R и др.)
- Обработка сценариев объемом до 120 страниц

Извлечение ключевых элементов

Элемент	Примеры
Идентификатор	Идентификатор: "ИДЕНТИФИКАТОР", "ИДЕНТИФИКАТОР", "ИДЕНТИФИКАТОР" ИДЕНТИФИКАТОР: "ИДЕНТИФИКАТОР-ИДЕНТИФИКАТОР"
Идентификатор имени	ИДЕНТИФИКАТОР, ИДЕНТИФИКАТОР, ИДЕНТИФИКАТОР, ИДЕНТИФИКАТОР
Идентификатор адреса	ИДЕНТИФИКАТОР ■ ИДЕНТИФИКАТОР ■ ИДЕНТИФИКАТОР
Идентификатор телефона	ИДЕНТИФИКАТОР, ИДЕНТИФИКАТОР, ИДЕНТИФИКАТОР
Идентификатор электронной почты	ИДЕНТИФИКАТОР@ИДЕНТИФИКАТОР, ИДЕНТИФИКАТОР, ИДЕНТИФИКАТОР
Идентификатор даты	ИДЕНТИФИКАТОР ■ ИДЕНТИФИКАТОР ■ ИДЕНТИФИКАТОР
Идентификатор времени	ИДЕНТИФИКАТОР, ИДЕНТИФИКАТОР
Идентификатор валюты	ИДЕНТИФИКАТОР ■ ИДЕНТИФИКАТОР
Идентификатор процента	ИДЕНТИФИКАТОР ■ ИДЕНТИФИКАТОР ■ ИДЕНТИФИКАТОР

Удобство загрузки и просмотра

- Интуитивная загрузка нескольких файлов одновременно (PDF и DOCX)
- Выбор пресета таблицы: базовый, расширенный или полный анализ
- Возможность настройки кастомных столбцов
- Наглядное представление результатов в виде интерактивной таблицы
- Прогресс-бар и отображение промежуточных этапов обработки
- Быстрое первичное отображение результатов
- Время обработки среднего сценария не превышает 5 минут

Работа с данными и экспорт

- Редактирование ячеек прямо в интерфейсе с моментальным сохранением
- Фильтрация по столбцам для быстрого поиска нужных сцен
- Поиск по содержимому таблицы
- Экспорт в CSV с кодировкой UTF-8-BOM для корректного отображения в Excel
- Экспорт в XLSX с сохранением структуры и форматирования
- Поддержка работы с несколькими сериями в одной таблице

Практическая применимость

Решение полностью соответствует реальным потребностям производственных команд:

- Все необходимые элементы для препродакшн-планирования присутствуют в таблице
- Результаты готовы к использованию без дополнительной ручной обработки
- Соответствие формату препродакшн-таблиц, используемых в индустрии
- Поддержка работы с несколькими сериями одновременно
- Автоматическое извлечение номера серии из имени файла
- Гибкая настройка структуры таблицы под специфику проекта

Точность и полнота извлечения

- Использование комбинации правил и ключевых слов для извлечения элементов
- Обработка неоднозначных формулировок и опечаток в исходном сценарии
- Корректное соответствие элементов к сценам
- Полнота распознавания всех ключевых элементов
- Отсутствие пропусков в определении локаций, персонажей и реквизита
- Возможность ручной корректировки результатов в интерфейсе

Технические характеристики

Параметр	Значение
Форматы документов	PDF, DOCX
Максимальное количество страниц	до 120 страниц
Максимальная глубина вложенности	до 5 уровней вложенности
Поддерживаемые кодировки	UTF-8, UTF-16, CP1251, KOI8-R, ISO-8859-5, MacRoman, ASCII
Форматы обмена данными	CSV (UTF-8-BOM), XLSX
Интерфейс API	REST API
Платформенность	Docker, Kubernetes

Преимущества решения

- Экономия времени: автоматизация многочасовой ручной работы
- Снижение ошибок: исключение человеческого фактора при обработке
- Стандартизация: единый формат препродакшн-таблиц
- Масштабируемость: обработка нескольких серий одновременно
- Гибкость: настройка структуры таблицы под проект
- Удобство: интуитивный интерфейс без необходимости обучения

Структура выходной таблицы

Таблица содержит следующие столбцы (в зависимости от выбранного пресета):

Серия, Сцена, Режим, Инт / нат, Объект, Подобъект, Синописис, Время года, Персонажи, Массовка, Грим, Костюм, Реквизит, Игровой транспорт, Трюк, Животные

Каждая строка соответствует отдельной сцене со всеми извлеченными элементами.

Планы развития

- Дообучение моделей машинного обучения на GPU для повышения точности
- Улучшение извлечения грима, костюмов и трюков с учетом контекста
- Интеграция с системами планирования съемок
- Расширение словарей и правил для различных жанров
- Автоматическое определение времени года по контексту
- Улучшение обработки сложных сцен с множеством элементов

Заключение

Решение полностью автоматизирует процесс создания препродакшн-таблиц, значительно сокращая время подготовки и снижая количество ошибок.

Сервис готов к использованию в реальных проектах и может быть легко развернут с помощью Docker или локальной установки.

Спасибо за внимание!

Команда DiverCity