

ML



**# Digital Fintech
Warriors**

TALENT

MATCH



Задача

Разработать и реализовать алгоритм, который будет разбивать информацию, содержащуюся в резюме кандидата по смысловым блокам:

- образование
- опыт работы
- навыки и т.д.

На выходе должен быть структурированный JSON



Руслан

DA / PM



Артур

DA / DS



Григорий

Backend /
DS



Александр

Backend

Команда

Форматы документов

- .docx
- .doc
- .pdf

Библиотеки

- Для текстовых файлов
 - docx2txt
 - docxcompose
 - docxtpl
- Для pdf
 - PyPDF2

Сущности

- Контактные данные
- Опыт, образование
- достижения

Подходы

- Регулярные выражения
 - re
- Готовые решения
 - pyresparser
- Метрика оценки
 - % заполненных полей в шаблоне
 - Чувство прекрасного

Обработка стандартных резюме

- Добавление других решений (ансамблевый подход)
- Уточнение существующих решений

Обработка нестандартных решений

- Обучение модели на размеченном датасете для отбраковки некорректных вариантов
- Дообучение модели на подобных примерах

Замечания по итогам хакатона

- По результатам удалось автоматизировать заполнение 50% полей шаблона (выбранная метрика результата) — есть пространство для модификаций
- Чтение и парсинг “сырых” резюме — отдельная большая задача
- Простые решения могут быть не только быстрее, но и эффективнее сложных моделей

[Репозиторий GitHub](#)

[Видео-демонстрация](#)

Спасибо за внимание!