

Хакатон {ML TALENT MATCH}

Решение команды SlovarikDB

Постановка задачи

Исходная задача:

Сопоставление резюме и вакансии кандидатов, отбор наиболее подходящих.

Формулировка в терминах ML

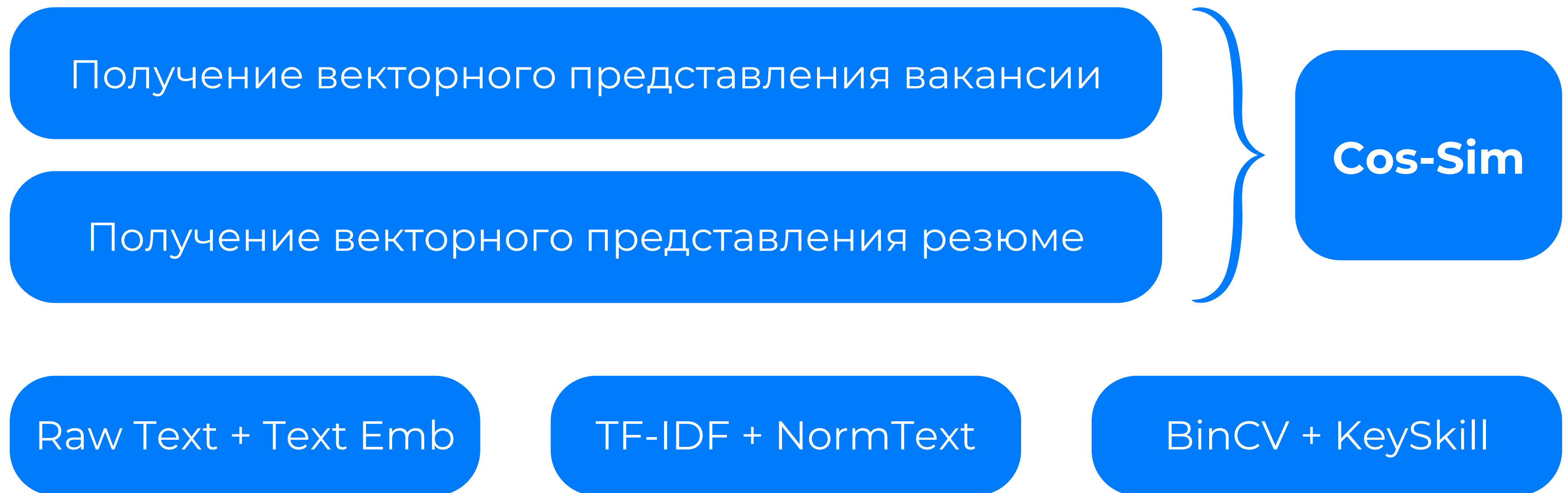
Ранжирование кандидатов на основании соответствия их резюме условиям вакансии.

Продуктовое видение:

Алгоритм получает **на вход** вакансию и резюме.
Резюме ранжируются на основании их подобия условиям вакансии.
Можно **выбрать топ-n кандидатов** по версии алгоритма.

Идея решения

Векторное представление вакансии и “идеального” резюме будут максимально близки.



Raw Text + Text Emb

Получение эмбединга для полного текста вакансии и получение эмбединга для навыков из резюме и суммаризированного опыта.

Используемые модели:

Суммаризация:
bart-large-cnn-samsum-rescom-finetuned-resume-summarizer

Векторизация:
RabotaRu/HRBert-mini

Решенные проблемы:

Мультиязычность:
перевод через API Google Translate

Агрегация опыта:
суммаризация

Смысл? Анализ общего сходства опыта кандидата и вакансии.

TF-IDF + NormText

Этап 1: Расширение датасета

Получение 3401 вакансии с HH.ru. Необходимы для обучения векторайзера.

Этап 2: Нормализация текста

Перевод (**Google**), лемматизация (**SOTA**: MyStem), очистка от мусорных слов и разметки.

Этап 3: Обучение TF-IDF векторайзера

Выделение 3000 ключевых слов и отбрасывание мусорных.

Смысл? Анализ полнотекстового сходства по значимым ключевым словам.

BinCV + KeySkill

Выделение ключевых компетенций для отдельной вакансии и проверка их наличия в резюме.

Этап 1: Выделение ключевых компетенций

Запрос к Сбер GigaChat с просьбой выделить ключевые навыки

Этап 2: Проверка наличия в тексте

Обучение бинарного Count Vectors на ключевых компетенциях и их поиск в нормализованном тексте вакансии

Смысл? Проверка по тех. стеку

СПАСИБО ЗА ВНИМАНИЕ!