

ЛУЧШЕЕ РЕШЕНИЕ НА NOCODE

СИСТЕМА СТРУКТУРИРОВАНИЯ И ИЗВЛЕЧЕНИЯ ИНФОРМАЦИИ

Команда "Meows and Paws"

Капитан Игорь Шаталин
+7 987 655 67 79

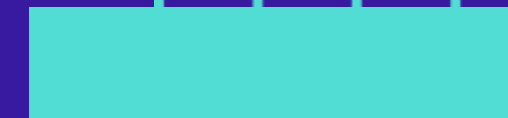
Хакатон труда

0101111110111111101011100100
0110100010000100000000111
101110110001111010111110111

ЗАДАЧА ХАКАТОНА

РАЗРАБОТКА

Система структурирования и извлечения информации
Разработать систему структурирования и извлечения информации по каждой из составляющих компетенции текстов резюме и вакансий (профессиональные компетенции - знания, умения, навыки/ «мягкие» компетенции - опыт, личностные характеристики, достижения) для уточнения резюме и вакансий.



РЕШЕНИЕ

ДАННАЯ ЗАДАЧА ОЧЕНЬ ХОРОШО РЕШАЕТСЯ, НО ТОЛЬКО С ПРЕДВАРИТЕЛЬНОЙ ПОДГОТОВКОЙ. МЫ ГОТОВЫ ВЗЯТЬСЯ ЗА РЕАЛИЗАЦИЮ УСПЕШНОГО РЕШЕНИЕ ДАННОЙ ЗАДАЧИ, ЕСЛИ ВО ВРЕМЯ ХАКАТОНА НИКТО НЕ РЕАЛИЗОВАЛ ОПИСАННЫЙ НИЖЕ ПАЙПЛАЙН РАЗРАБОТКИ.

ЗАДАЧА ВЫЯВЛЕНИЯ СКИЛЛОВ СВОДИТСЯ К КЛАССИЧЕСКОЙ И ХОРОШО РЕШЁННОЙ NLP-ЗАДАЧЕ "NAMED ENTITY RECOGNITION" ("ВЫЯВЛЕНИЕ ИМЕНОВАННЫХ СУЩНОСТЕЙ"). ВЕСЬ ЭТАП РЕАЛИЗАЦИИ ПРОЕКТА МОЖНО РАЗДЕЛИТЬ НА ТРИ БЛОКА:

1. ПОДГОТОВКА ДАТАСЕТА.

2. ОБУЧЕНИЕ МОДЕЛИ.

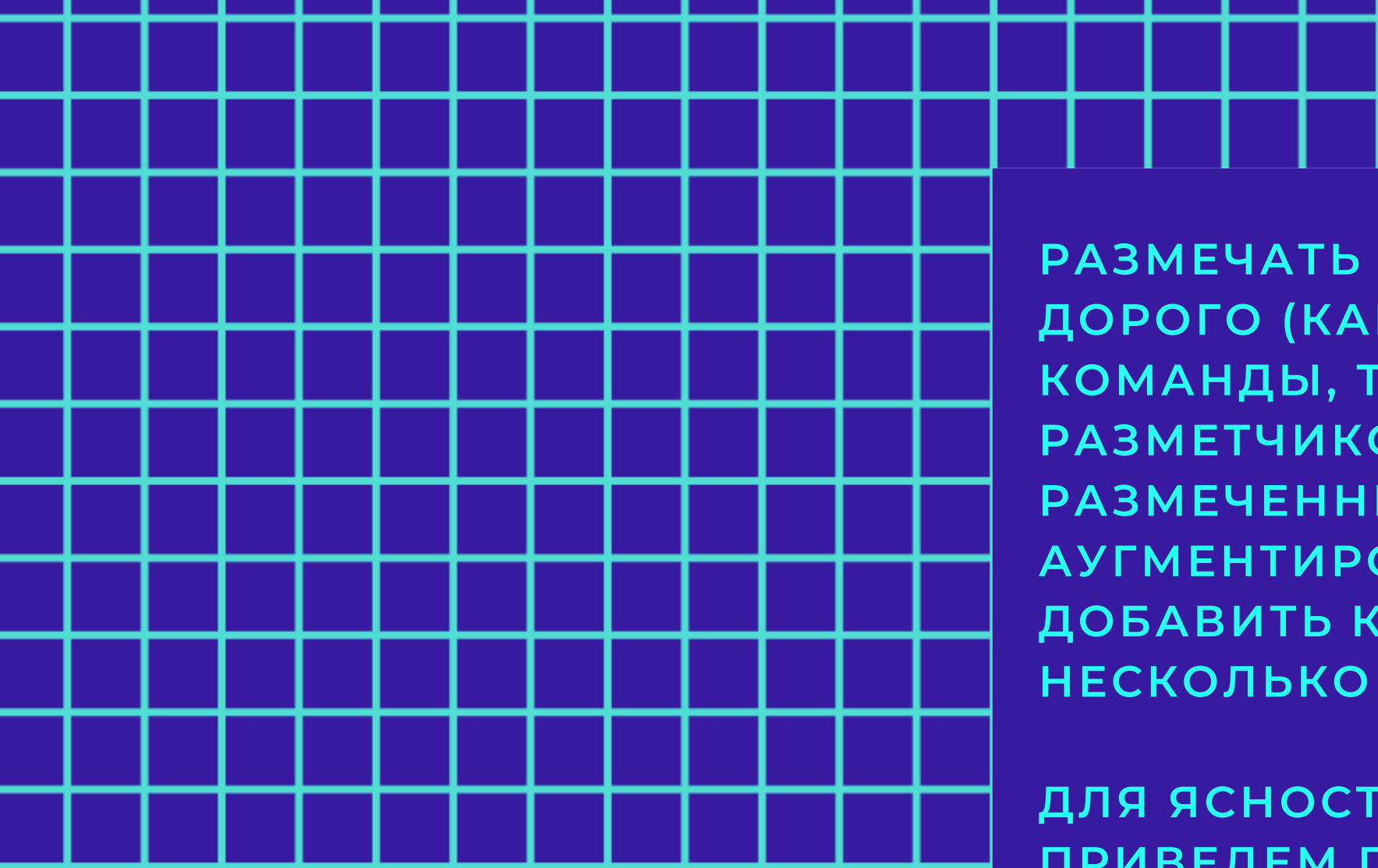
3. СОЗДАНИЕ PRODUCTION-ВЕРСИИ МОДЕЛИ.

Подготовка датасета

ДЛЯ ОБУЧЕНИЯ МОДЕЛИ НАМ НЕОБХОДИМО БУДЕТ РАЗМЕТИТЬ СЫРЫЕ ДАННЫЕ (РЕЗЮМЕ И ВАКАНСИИ). В ТЕКСТЕ НУЖНО БУДЕТ ПОМЕТИТЬ НЕОБХОДИМЫЕ НАВЫКИ НУЖНЫМИ МЕТКАМИ, НАПРИМЕР SOFT_SKILLS, HARD_SKILLS, EXPERIENCE. ПРИ ЖЕЛАНИИ ЧТО-ТО ЕЩЁ ИЗВЛЕЧЬ, МЕТКИ МОЖНО ДОБАВИТЬ.

"Обязанности: - встречи HARD_SKILL, обзвон клиентской базы HARD_SKILL, выезд на стройплощадки HARD_SKILL; - работа с клиентской базой HARD_SKILL; - выполнение плана продаж и задач HARD_SKILL, поставленных руководством; - ежедневный отчет о проделанной работе; - прямые продажи HARD_SKILL Подвесных потолков. - подготовка коммерческих предложений HARD_SKILL, проведение переговоров HARD_SKILL, заключение договоров HARD_SKILL; - контроль документооборота HARD_SKILL; - контроль отгрузок HARD_SKILL. Требования: Целеустремленность SOFT_SKILL, мотивация на успех SOFT_SKILL, активная жизненная позиция SOFT_SKILL. Навык ведения успешных переговоров HARD_SKILL. опрятный внешний вид SOFT_SKILL. Опыт работы в продаже подвесных потолков обязателен. Встречи в офисе и выезд на встречу к клиентам. Требования: Уверенный пользователь ПК, 1C:8.2 HARD_SKILL, Word HARD_SKILL, Excel HARD_SKILL, (и тд.) Своя база будет преимуществом! Условия: Крупная компания по продажам стройматериалов, приглашает на работу амбициозного специалиста продаж подвесных потолков. Мы предлагаем комфортный офис рядом с ТРК Парк Хаус. Молодой активный коллектив. Возможен карьерный рост. Стабильная выплата зарплаты без задержек: оклад +премиальная система. Всегда в наличии на складе самый "ходовой товар по потолкам!"- прямые контракты с производителями."

Картинка разметки



РАЗМЕЧАТЬ ВРУЧНУЮ ДАННЫЕ ДОВОЛЬНО ТАКИ ДОРОГО (КАК С ПРИВЛЕЧЕНИЕМ ЧЛЕНОВ КОМАНДЫ, ТАК И НАНИМАЯ СТОРОННИХ РАЗМЕТЧИКОВ). ПОЭТОМУ ПРЕДЛАГАЕТСЯ РАЗМЕЧЕННЫЙ ВРУЧНУЮ КОРПУС АУГМЕНТИРОВАТЬ, Т.Е. АВТОМАТИЧЕСКИ ДОБАВИТЬ К НЕМУ НОВЫЕ ДАННЫЕ. СУЩЕСТВУЕТ НЕСКОЛЬКО МЕТОДОВ АУГМЕНТАЦИИ.

ДЛЯ ЯСНОСТИ ТОГО, ЧТО ТАКОГО АУГМЕНТАЦИЯ ПРИВЕДЕМ ПРИМЕР ЗАМЕНЫ СЛОВ ИСХОДНОГО ТЕКСТА “СИНОНИМАМИ”:

"ТРЕБУЕТСЯ ЗНАНИЕ WORD" =>
ТРЕБУЕТСЯ ЗНАНИЕ ВОРД
ТРЕБУЕТСЯ ЗНАНИЕ EXCEL
ТРЕБУЕТСЯ ЗНАНИЕ POWERPOINT
ТРЕБУЕТСЯ УМЕНИЕ WORD

ТАКИМ ОБРАЗОМ, БЛАГОДАРЯ АУГМЕНТАЦИИ МЫ ПОЛУЧАЕМ НЕ ОДИН, А СРАЗУ 5 ТЕКСТОВ!

ЕСТЬ ЕЩЁ ОДИН ЭКСПЕРИМЕНТАЛЬНЫЙ МЕТОД, КОТОРЫЙ МОЖЕТ НАМ ПОМОЧЬ. О НЁМ БЫЛО РАССКАЗАНО НА НЕДАВНО ПРОШЕДШЕЙ КОНФЕРЕНЦИИ ПО NLP "ДИАЛОГ-21". ТАМ ИССЛЕДОВАТЕЛИ БРАЛИ НЕ NER-ЗАДАЧУ, А QA ("ЭКСТРАКТНЫЙ ВОПРОС-ОТВЕТ"), ЧТО НЕСКОЛЬКО СХОЖЕ С НАШЕЙ ЗАДАЧЕЙ (МЫ ТОЖЕ БУДЕМ УЧИТЬ МОДЕЛЬ ИЗВЛЕКАТЬ НЕКИЕ КУСКИ ТЕКСТА). ОНИ ВЗЯЛИ НЕБОЛЬШОЙ РУССКИЙ ДАТАСЕТ И АНГЛИЙСКИЙ ДАТАСЕТ ДЛЯ АНАЛОГИЧНОЙ ЗАДАЧИ. ПОСЛЕ ОБУЧЕНИЯ МОДЕЛИ НА ТАКОМ СИНТЕЗИРОВАННОМ ДАТАСЕТЕ РЕЗУЛЬТАТ ОКАЗАЛСЯ РАВНЫМ РЕЗУЛЬТАТУ ОБУЧЕНИЯ МОДЕЛИ НА БОЛЬШОМ РУССКОЯЗЫЧНОМ ДАТАСЕТЕ. Т.Е. НЕСМОТРА НА ЯЗЫК МОДЕЛЬ ПОНЯЛА ПОСТАВЛЕННУЮ ПЕРЕД НЕЙ ЗАДАЧУ И РЕШИЛА ЕЁ.



НОВЫЙ МЕТОД

ПРЕДЛАГАЕТСЯ В КАЧЕСТВЕ ЭКСПЕРИМЕНТА ПРИМЕНИТЬ АНАЛОГИЧНЫЙ МЕТОД: ОТЫСКАТЬ АНГЛОЯЗЫЧНЫЕ ОТКРЫТЫЕ ДАТАСЕТЫ ДЛЯ РЕШЕНИЯ ПОХОЖЕЙ ЗАДАЧИ И ДОБАВИТЬ ЕЁ К НАШЕМУ РУССКОЯЗЫЧНОМУ ДАТАСЕТУ, И ПОСМОТРЕТЬ, ПРИРАСТЕТ ЛИ КАЧЕСТВО ОБУЧАЕМОЙ МОДЕЛИ.

010
110
111
1111010111001001110110100010000
00011100101110110001111010111110

Обучение модели

ИЗНАЧАЛЬНО МЫ ПРЕДЛАГАЕМ ОБУЧИТЬ МОДЕЛИ, КОТОРЫЕ СЕГОДНЯ ИМЕЮТ SOTA-РЕЗУЛЬТАТЫ В РЕШЕНИИ NER-ЗАДАЧИ. ЭТО МОДЕЛИ НА ОСНОВЕ БЕРТА. ПРЕДЛАГАЕТСЯ ДЛЯ ЭКСПЕРИМЕНТА ВЗЯТЬ 3 ПРЕДОБУЧЕННЫХ МОДЕЛИ:

1.RUBERT

2.BERT BASE

3.SPANBERT

В ХОДЕ ДАННОГО ЭТАПА ПЛАНИРУЕТСЯ ПРОВЕСТИ РЯД ЭКСПЕРИМЕНТОВ И ВЫБРАТЬ НАИБОЛЕЕ ЛУЧШИЙ РЕЗУЛЬТАТ.

101100011
011010001
010010111
00000001
010010111
011100100
011010001
1111011110
1111011110

СОЗДАНИЕ PRODUCTION- ВЕРСИИ МОДЕЛИ

Отобрав лучшую из обученных моделей, нужно будет подготовить её к работе в "боевом режиме". На этом этапе работы предполагается провести 2 типа экспериментов, а их результаты предоставить заказчику для выбора наиболее подходящего для него технологического решения.

01

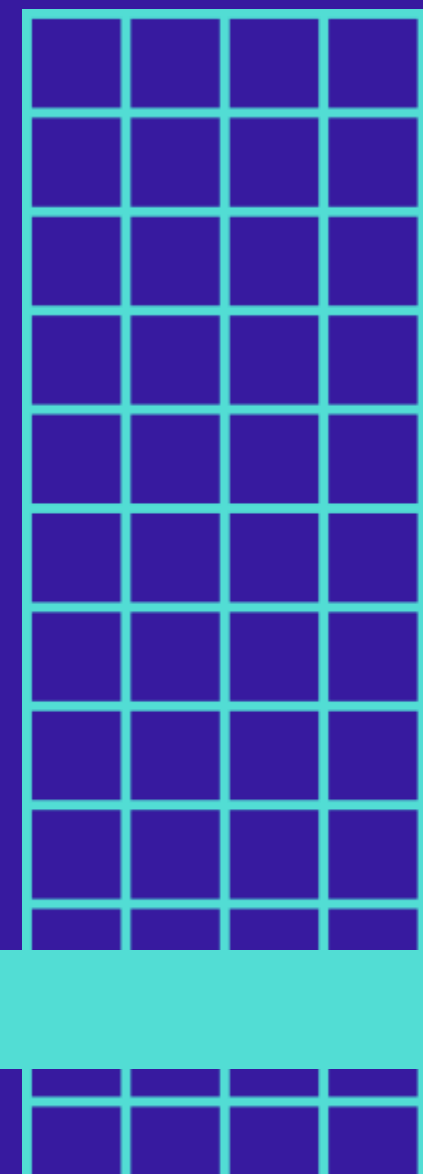
УМЕНЬШЕНИЕ ВЕСОВ МОДЕЛИ

Модели на основе архитектуры Берта достаточно большие, и ресурсов на них требуется, естественно, больше, чем, на FF, CNN- или RNN архитектуры. Здесь мы предлагаем провести прунинг отобранной на предыдущем этапе модели.

02

ДИСТИЛЛЯЦИЯ ЗНАНИЙ

У нас будет: модель учитель (наша большая модель, выбранная на предыдущем этапе); модель ученик (модель, имеющая сравнительно меньше слоёв, как правило это классическая архитектура - CNN или RNN).



111011101111101111101110111011111011101011
101110101100111111010101100011001000110
101100011001000111110101111101010111111
01011111010101111101011111110111111101011

Все эти методы сравнительно новые и не всегда удаётся с первого раза получить приемлемый результат. Поэтому планируется провести несколько экспериментов, а их результаты предоставить заказчику для утверждения окончательной версии используемой модели.





Яндекс Облако

Внедрение

ДЛЯ УДОБНОЙ И БЫСТРОЙ ИНТЕГРАЦИИ МОДЕЛИ НА СТОРОНЕ ЗАКАЗЧИКА МЫ ПРЕДЛАГАЕМ СОЗДАТЬ REST-API-ПРИЛОЖЕНИЕ. ЭТО НАИМЕНЕЕ ЗАТРАТНЫЙ ВАРИАНТ ВНЕДРЕНИЯ.

ПРИЛОЖЕНИЕ МОЖЕТ БЫТЬ ОФОРМЛЕНО В ВИДЕ PYTHON-ПАКЕТА, ЛИБО В ВИДЕ DOCKER-КОНТЕЙНЕРА, И БУДЕТ РАБОТАТЬ В КАЧЕСТВЕ МИКРОСЕРВЕРА.

УСТАНОВКА И ЗАПУСК - ЭТО БУКВАЛЬНО 2-3 BASH-КОМАНДЫ СИСТЕМНОГО АДМИНИСТРАТОРА.

ПРИЛОЖЕНИЕ МОЖЕТ БЫТЬ УСТАНОВЛЕНО КАК НА СТОРОНЕ ЗАКАЗЧИКА, ТАК И НА СТОРОННИХ РЕСУРСАХ, ПРЕДОСТАВЛЯЮЩИХ ТЕХНОЛОГИИ БЕЗ СЕРВЕРНОЙ АРХИТЕКТУРЫ (НАПРИМЕР GOOGLE CLOUD И ЯНДЕКС.ОБЛАКО).

1011000111101011111101111000010101
001011110101011011010001011011001
1001001110101011100100000110010
0101111101111111010111001001110110
1000100001000000011100101110110
001111010111111011110000101010010
111101010110110100010110110011001
00111010101110010000011001001011
111101111110101110010011101101000
1000010000000111001011101100011
1101011111101111000010101001011110
101011

НАША КОМАНДА

РАЗРАБОТЧИК

Игорь Шаталин
shatalin.ip@gmail.com

ДИЗАЙНЕР

Андрей Лукин
andrew.luckin2015@yandex.ru

НАДЕЕМСЯ НА ТО, ЧТО ВМЕСТЕ С ВАМИ
МЫ СМОЖЕМ УСПЕШНО РЕШИТЬ ЭТУ
ЗАДАЧУ, И СДЕЛАТЬ ЖИЗНЬ В НАШЕЙ
СТРАНЕ НЕМНОГО ЛУЧШЕ!

СПАСИБО!

КОМАНДА "MEOWS AND PAWS"

01111010
111101010
010011101
011111101
10001000
00011110
010111101
10010011
00101111

11010001000010000000111

0110001111010111110111100

10010111101010111011000111

