Big Data for Labour Market Intelligence

День 1
Презентация системы и Результаты

Алессандро Ваккарино – Мауро Пелукки

Июнь 2021 г.

Разделы

1. Цель и контекст
2. Задачи
    1. Участники
    2. Функциональная архитектура
    3. Способы приема данных
    4. Конвейер обработки данных
    5. Методы классификации

Разделы

1. **Цель и контекст**
2. Задачи
    1. Участники
    2. Функциональная архитектура
    3. Способы приема данных
    4. Конвейер обработки данных
    5. Методы классификации

Контекст

Постоянно развивающийся рынок труда:

- Цифровизация профессий
- Востребованность Гибких навыков
- Интернационализация
- Появление новых профессий и навыков
- Работа в режиме гибкого рабочего времени и дистанционная работа
- Влияние пандемии Covid-19
- …

Существует потребность в *инструменте*, который будет помогать нам анализировать и наблюдать за тем, **как** развивается Рынок труда, **а также способствовать тому, чтобы** Лица, ответственные за принятие решений, принимали **нужные решения в нужное время**

Что у нас есть / что нам нужно

У нас уже есть **данные официальной статистики**, которые являются:

- *Репрезентативными*
- *Строгими* в части значений

Но мы можем использовать и **дополнительную информацию**, которая может быть:

- *Оперативной* – для отслеживания текущих событий (например, анализ влияния Covid-19)
- *Подробной* и *соответствующей* реальным и текущим рыночным условиям – для выявления новых тенденций и анализа реальных интересов компаний

Как найти подобный дополнительный источник информации?

Воспользуйтесь **онлайн-рынком труда**

Преимущества онлайн-рынка труда

Здесь представлены точные сведения о том, что именно интересует компании в указанный период времени:

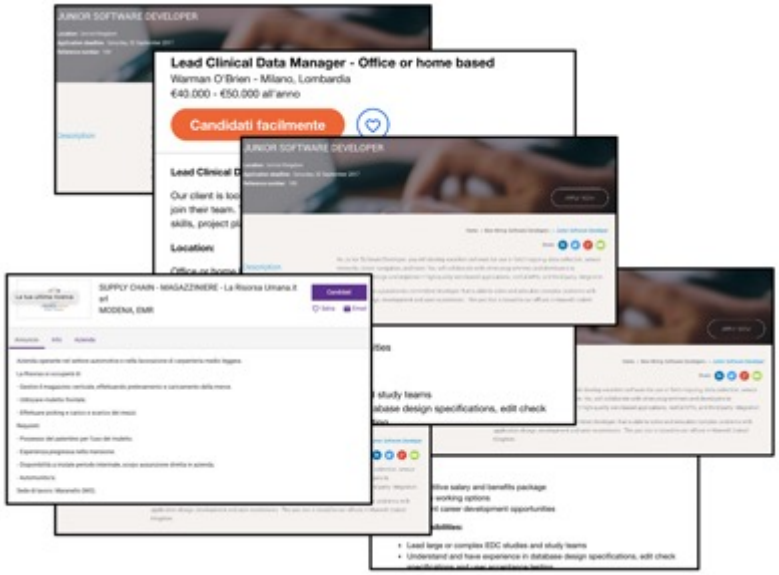
- Актуально: компании размещают объявления тогда, когда они действительно ищут новых сотрудников
- Подробно: в объявлении максимально четко описываются следующие потребности:
  - Необходимая специальность
  - Требования (навыки, опыт, образовательный уровень,...)
  - Условия работы (местоположение, договор, отрасль, рабочее время и пр.)
- Соответствует реальности: рыночные условия применяются как к роду занятий, так и к навыкам. Это позволяет разбираться в новых терминах, используемых на Рынке

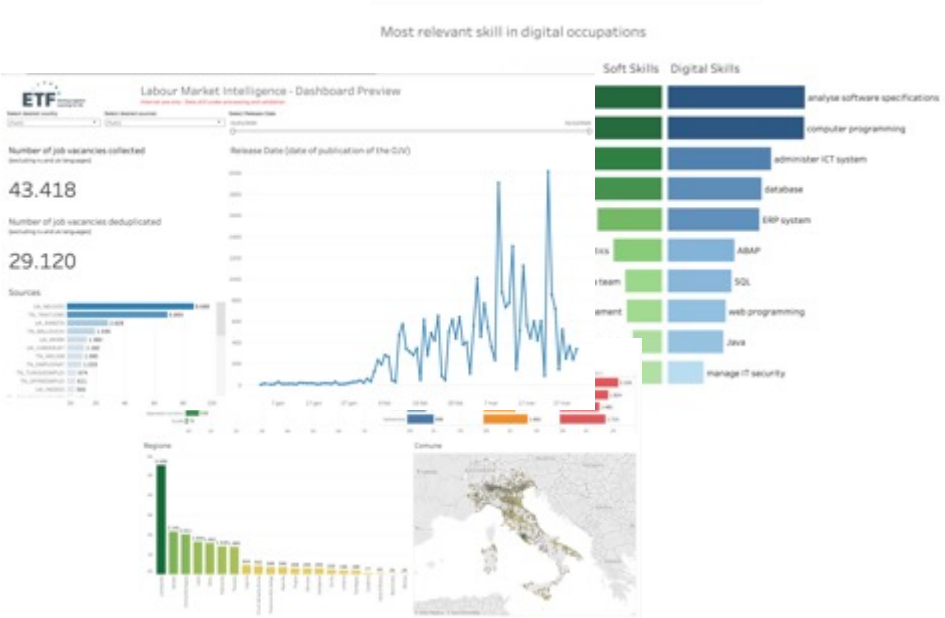
Использование этой дополнительной информации позволит лучше и глубже понять, как развивается Рынок труда в представленной стране и в сравнении с другими странами

Наша цель

Превратить объявления в Интернете о приеме на работу…

…в статистику и аналитику



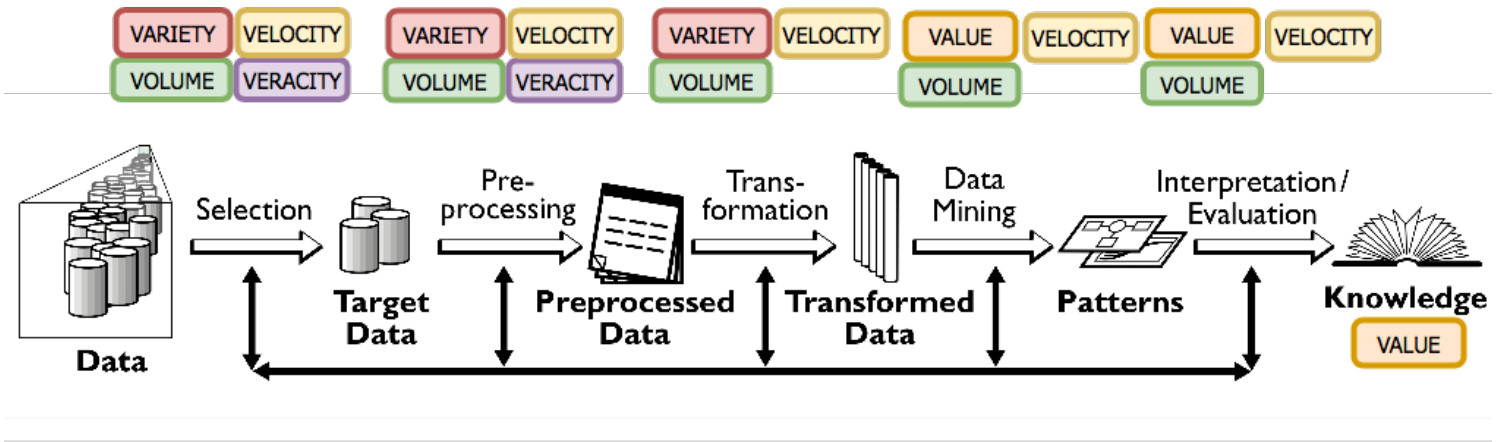


Задачи

- Обработка огромного **объема** оперативных данных
- Данные из сети Интернет → Необходимо выявить и снизить **шум**
- **Многоязычная** среда
- Необходимо соответствовать **стандартам классификации**
- Найти способ **сформулировать и представить** масштабный и комплексный сценарий

Методологическая основа

ОЗБД (Обнаружение знаний в базах данных) – Файяд, 1997 г.



VARIETY

Неструктурированные данные (открытый текст, подлежащий обработке)

VELOCITY

Данные в режиме реального времени

Огромный объем данных (терабайты)

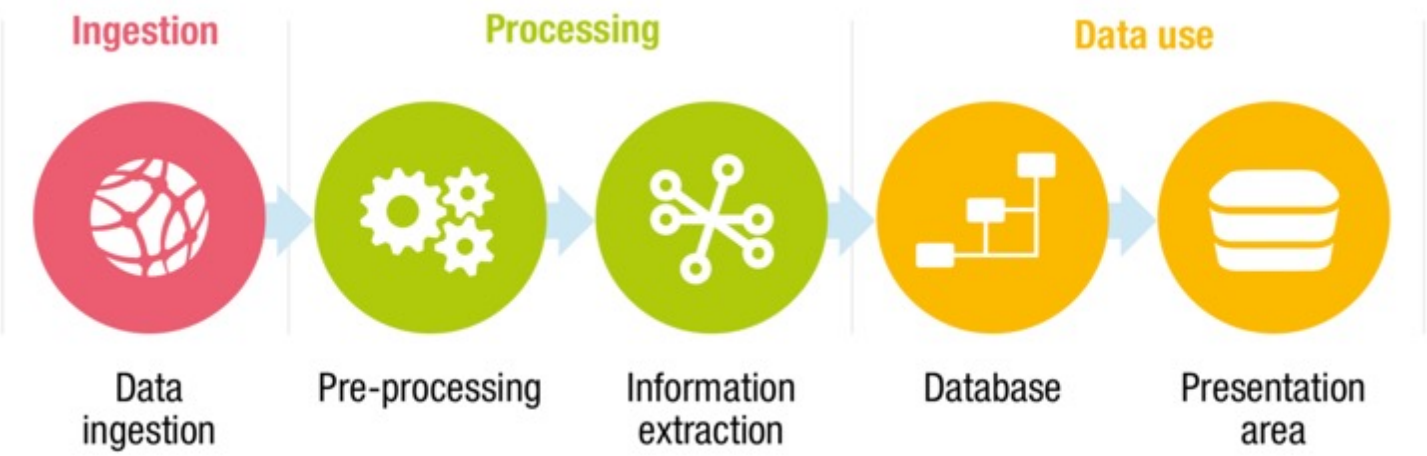
VOLUME

Данные зашумленные, неконтролируемые

VERACITY

Наш подход

ОЗБД для ИРТ (Исследования рынка труда)



Некоторые результаты

- Skillspanorama – Навыки по вакансиям в Интернете
  - https://skillspanorama.cedefop.europa.eu/en/indicators/skills-online-vacancies
- Skills OVATE (средство анализа навыков по вакансиям в Интернете для Европы)
  - https://www.cedefop.europa.eu/en/data-visualisations/skills-online-vacancies
- ЕФО (Европейский фонд образования) – Большие данные для ИРТ
  - Тунис
  - Украина

Разделы

1. Цель и контекст
2. Задачи
    1. **Участники**
    2. Функциональная архитектура
    3. Способы приема данных
    4. Конвейер обработки данных
    5. Методы классификации

Участники

Руководитель проекта

Ключевые пользователи

Специалист предметной области

Конечные пользователи

Руководитель проекта

- ЕФО
  - Руководить проектом с координационным комитетом
  - Определить объем работ по проекту
  - Определить ключевые организации
  - Поддерживать отношения с участниками проекта из ЕС
  - Консультировать

Ключевые пользователи

- ЕФО, Burning Glass
  - Определить требования
  - Следить за качеством выполнения проекта
  - Предоставлять входные данные для разработки проекта
  - Управлять архитектурой
  - Проверять весь поток данных и методологию

Специалист предметной области

- Международные специалисты по страноведению
  - Предоставлять специальные знания
  - Реализовывать архитектуру
  - Понимать язык/условия в конкретном контексте
  - Оценивать точность результатов
  - Тестировать продукцию
  - Предоставлять обратную связь

Конечные пользователи

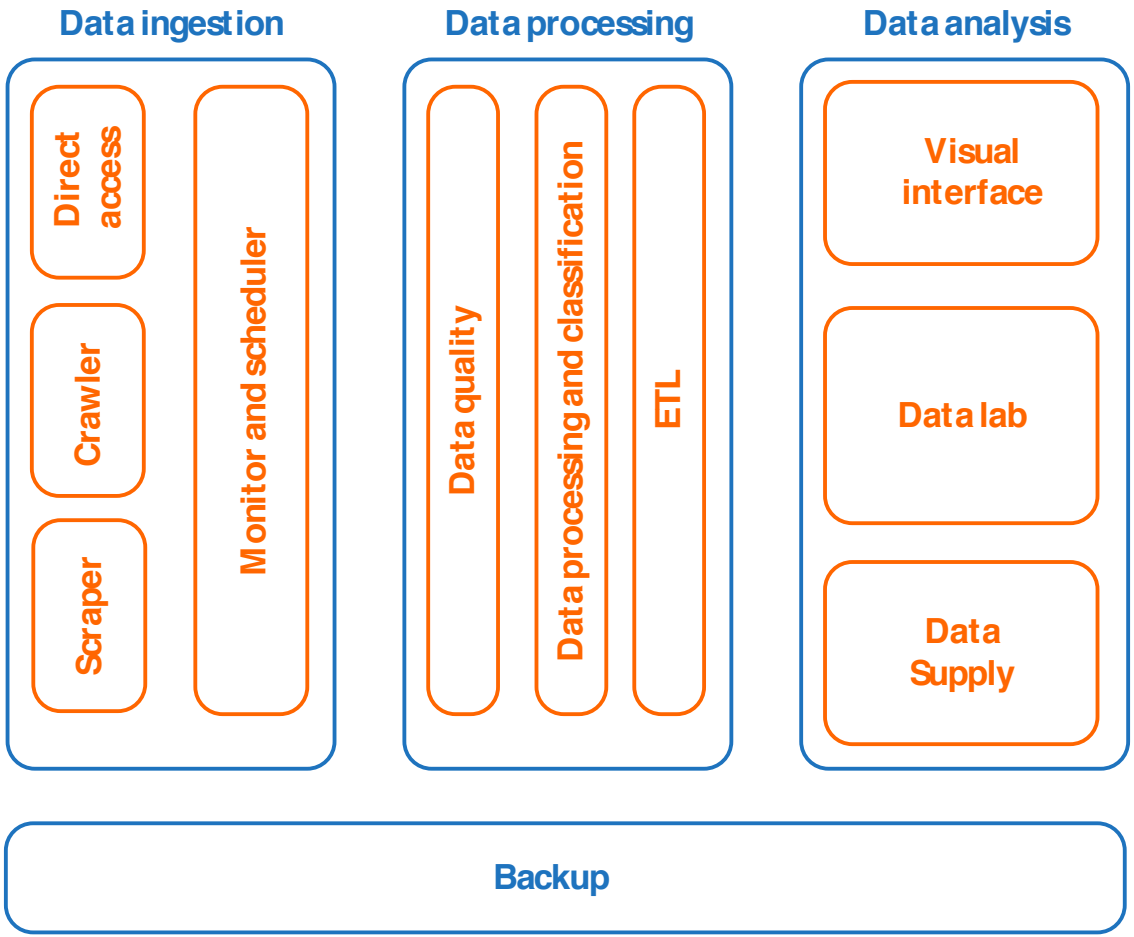
- Лица, ответственные за принятие решений, и Корпоративные пользователи
  - (Визуально) Исследовать наборы данных, данные анализа и сводные данные
  - Определить новые процессы анализа
  - Осуществлять сторителлинг данных
  - Принимать решения на основании изучения данных
- Исследователи данных
  - Применять новые модели машинного обучения и технологии ИИ
  - Извлекать из данных новые идеи
  - Применять продвинутые способы моделирования данных к наборам данных
- Аналитики данных
  - Толковать данные и превращать их в информацию
  - Определять модели и тенденции
  - Извлекать и анализировать сводные данные
  - Публиковать и распространять результаты своего анализа

Разделы

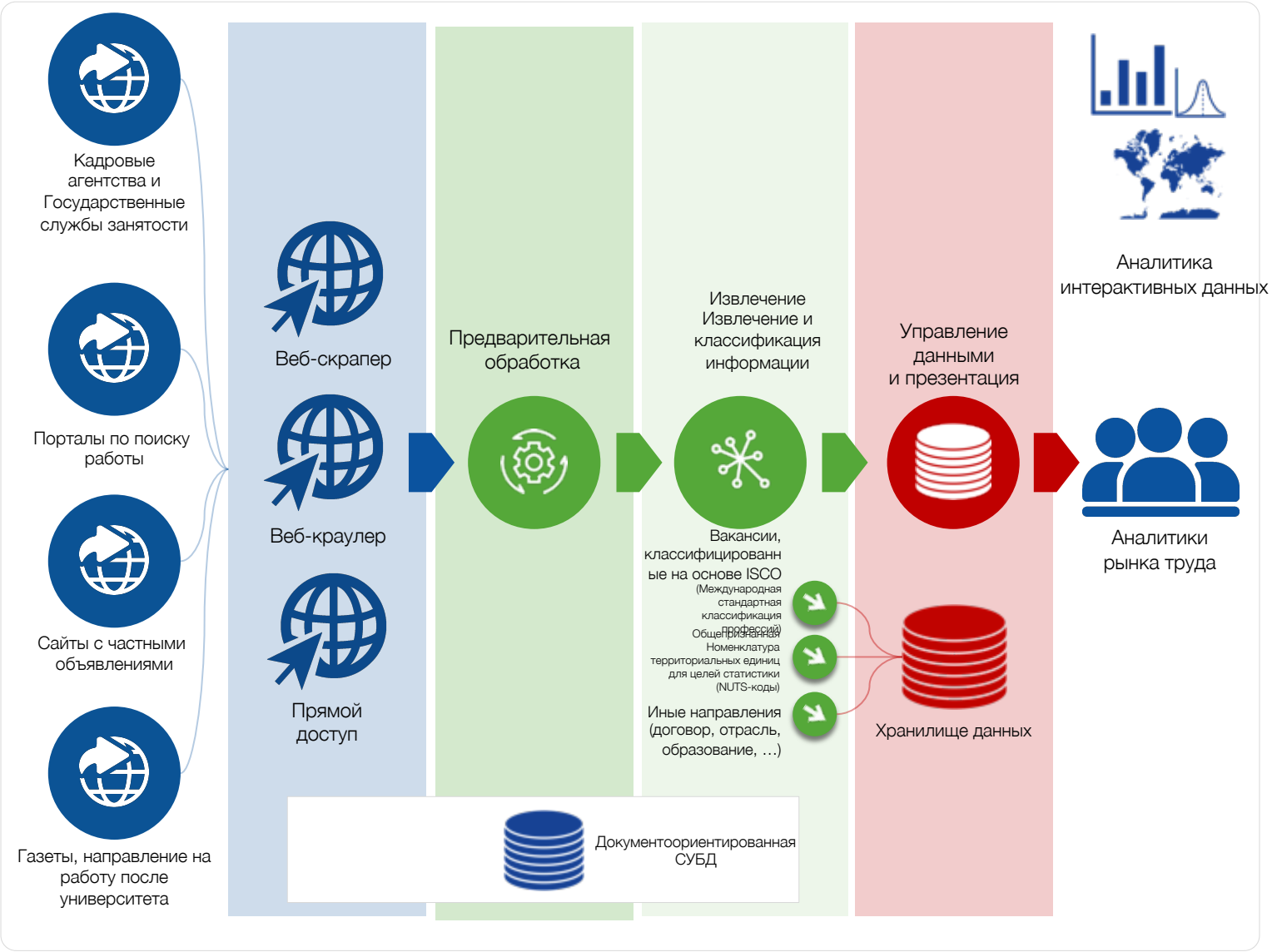
1. Цель и контекст
2. Задачи
    1. Участники
    2. **Функциональная архитектура**
    3. Способы приема данных
    4. Конвейер обработки данных
    5. Методы классификации

Общий поток данных

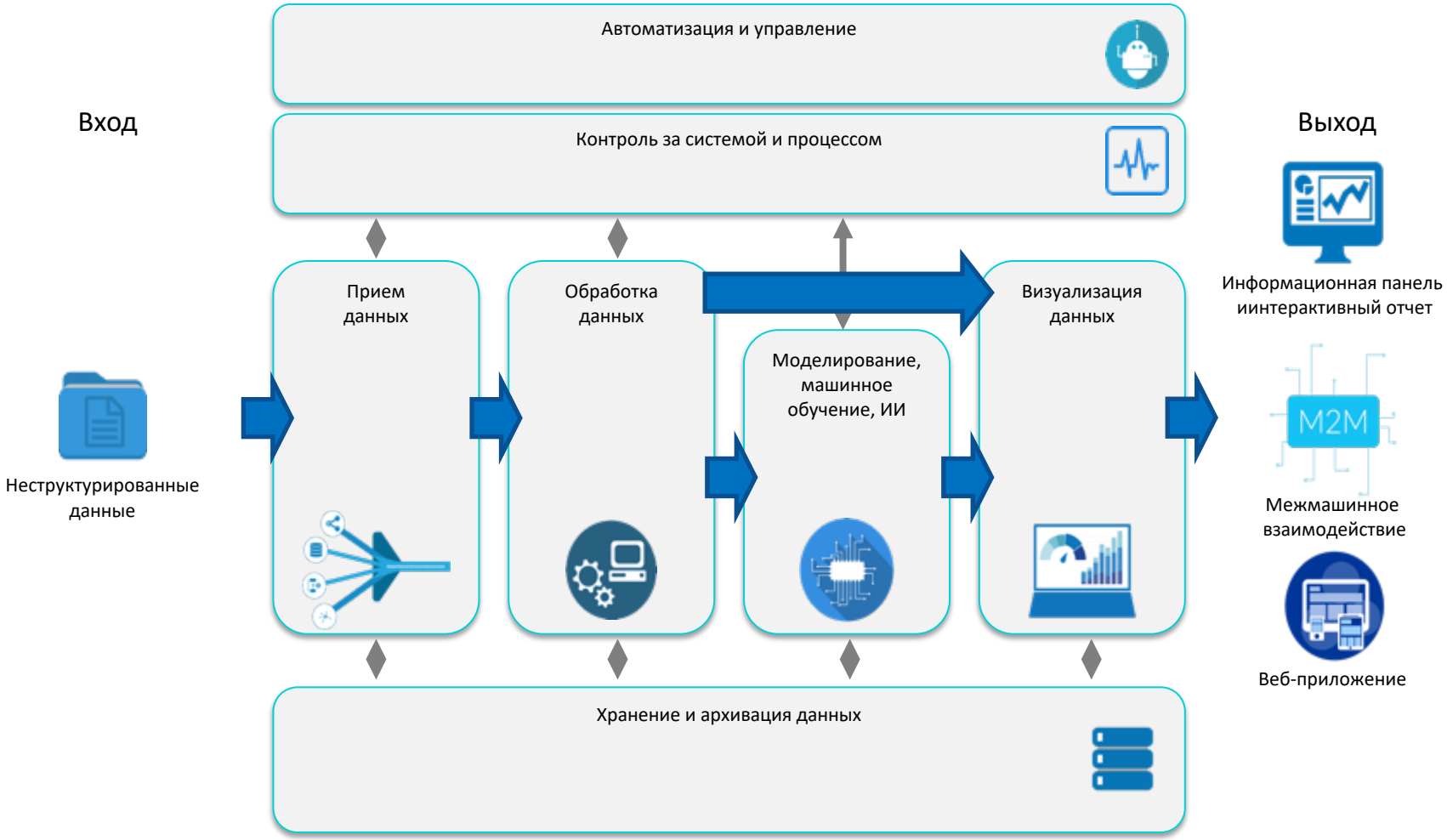
Концептуальная архитектура



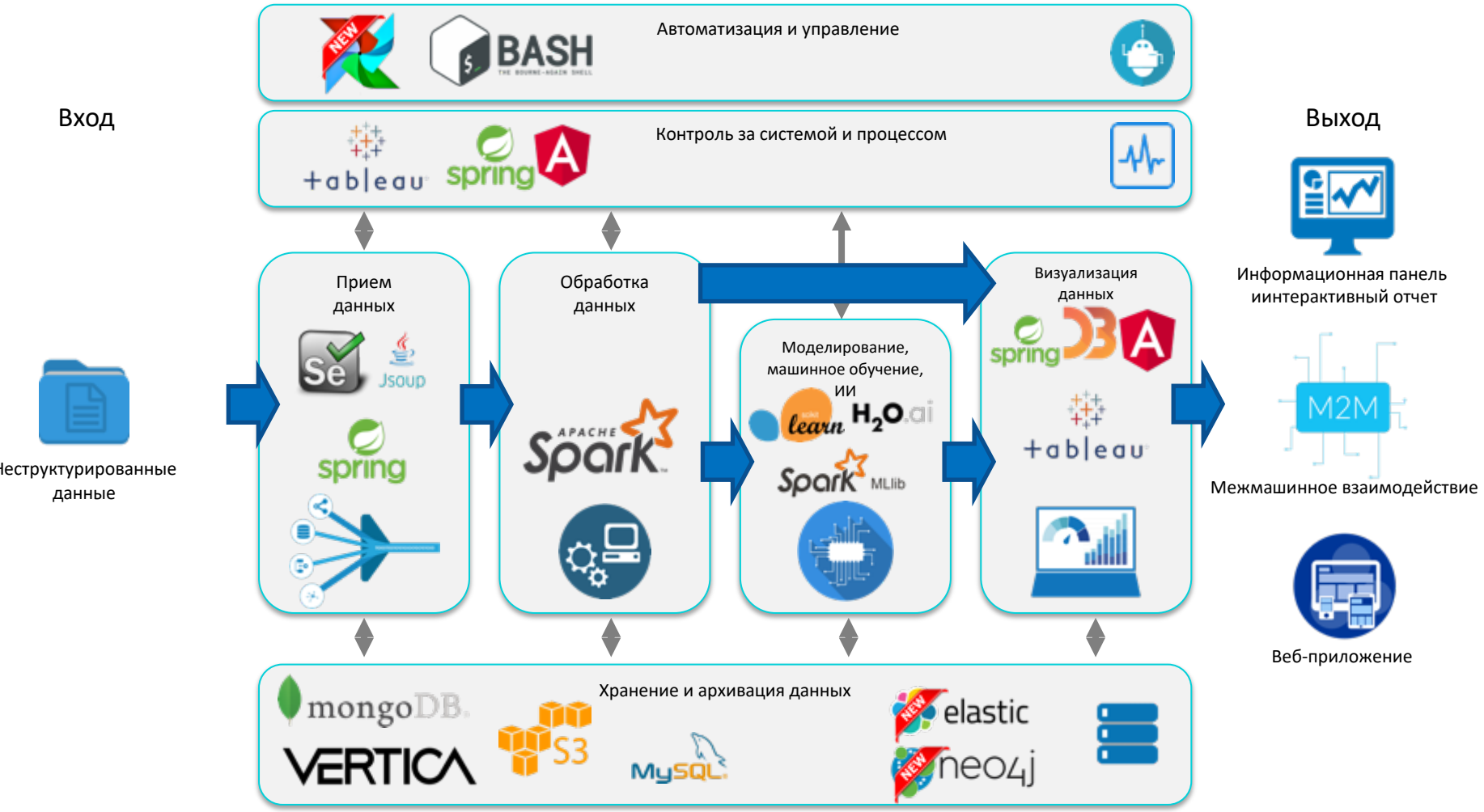
Логическое представление данных



Физическое представление данных



Технологическое представление данных



Ключевые проекты разработки

- Микросервисы
- Компонентизация
  - Специализация компонентов
  - Небольшие приложения
  - Портируемость
  - Повторное использование
  - Техническое обслуживание
- Горизонтальное масштабирование
  - Производительность

Ключевые компоненты

- Прием данных: сбор первичных данных из вакансий в Интернете как в структурированном, так и в неструктурированном (необработанный текст) виде
- Обработка данных: классификация данных посредством техник машинного обучения
- Анализ данных: извлечение информации из данных и ее распространение посредством визуализации
- Резервное копирование: хранение данных в безопасной среде для обеспечения "теплого" и "холодного" восстановления данных

Проблемы инфраструктуры

- Управлять множеством **параллельных процессов приема данных**
- **Быстрый** доступ к **высокопроизводительной** вычислительной инфраструктуре
- Требования **к верхней памяти**
- Большие объемы **хранилища** для хранения исходных и промежуточных данных
- Среда больших данных
- **Масштабируемая** архитектура

Поток больших данных

Требования к качеству

Разработка микросервисов

Компоненты по своей сути

Проблемы инфраструктуры

Контекст

Обслуживание

Мониторин
г

Масштабируемость

Обновления

Внедрение

Микросервисы по предварительной обработке

Определитель языка	Спам-фильтр	Фильтр на вакансии	Стеммер
Компонент дедупликации	Компонент для N-граммы	Очиститель текста	Объединение вакансий
Трансформер TF-IDF (Частота слова и обратная частота документа)	Алгоритм Document2Vec	Токенизатор	Удаление стоп-слов

Микросервисы по классификации

Классификатор навыков

Классификатор профессий

Классификатор отраслей

Классификатор требований к образованию

Определитель рабочего времени

Определитель типа договора

Определитель местоположения

Фукнция извлечения размера зарплаты

Фукнция извлечения опыта

Фукнция извлечения дат

Требования к технологии

1. Сервисы по запросу
2. Доступ к сети
3. Объединение ресурсов
    1. Управление
4. Быстрая эластичность
5. Измерение сервисов
    1. Качество данных
    2. Производительность
6. Портируемость (в локальной среде и на разных облачных сервисах)
7. Полиглот
    1. Языки программирования
    2. Технологии

Органическое представление данных

Приложения и компоненты по потоку данных
(Программное обеспечение как сервис)

Среда больших данных
(Платформа как сервис)

Вычислительные ресурсы
(Инфраструктура как сервис)

Хранение и СУБД
(Данные как сервис)

Ядро программного обеспечения

Пул общих эластичных аппаратных ресурсов
(Аппаратные средства как сервис)

Резюме и ключевые слова

- Ключевые компоненты и поток данных
  - Прием, обработка, классификация, презентация
- Компонентизация и микросервисы
- Стек неоднородных и больших данных
  - Selenium, Hadoop, Spark, Sklearn, Spark
- Масштабируемая среда
  - Облако

Вопросы?

Разделы

1. Цель и контекст
2. Задачи
    1. Участники
    2. Функциональная архитектура
    3. **Способы приема данных**
    4. Конвейер обработки данных
    5. Методы классификации

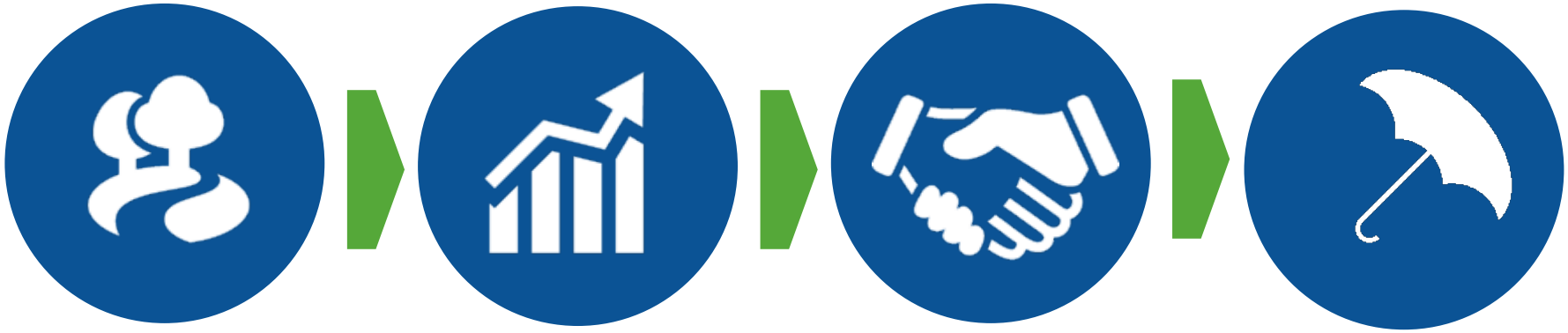
Архитектура

Деятельность по созданию архитектуры заключается в написании списка **источников** (веб-порталов), которые отражают текущее состояние Рынка труда в Интернете в указанной стране.

Специалист по данной стране **утверждает** этот список, что становится начальным этапом создания Системы ИРТ

Стратегия выбора источников

4 этапа обработки



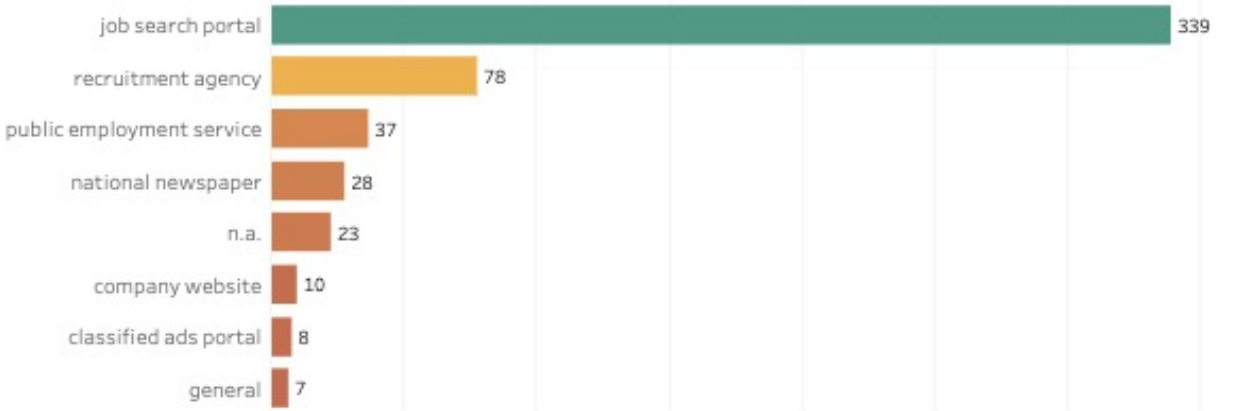
Выбор источника в рамках архитектуры

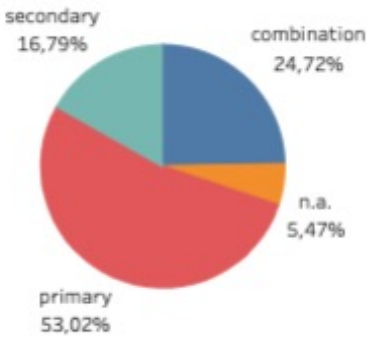
Прирост

Соглашения

Покрытие

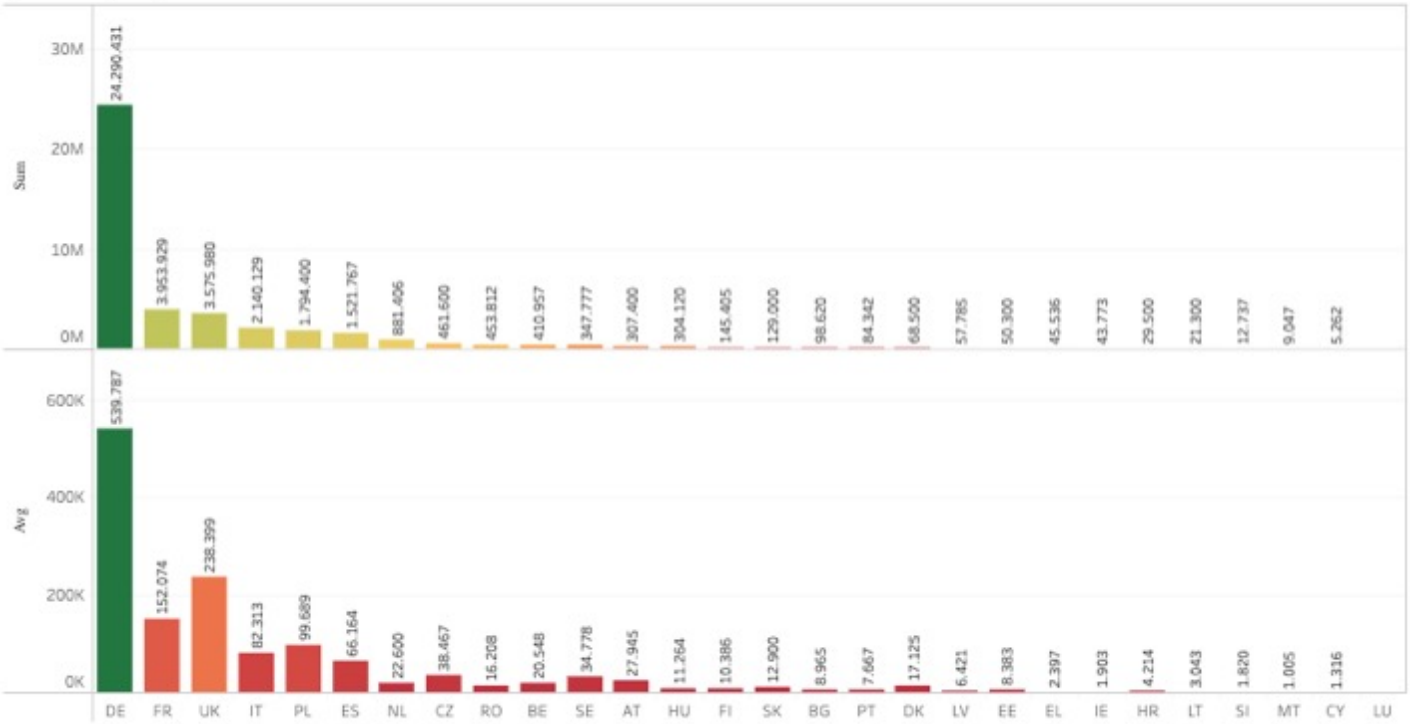
Sites by type of operator

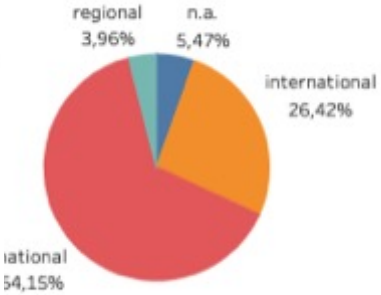


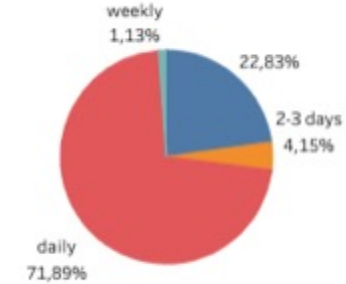


Vacancy volume by country

(estimated by ICE)







Прирост

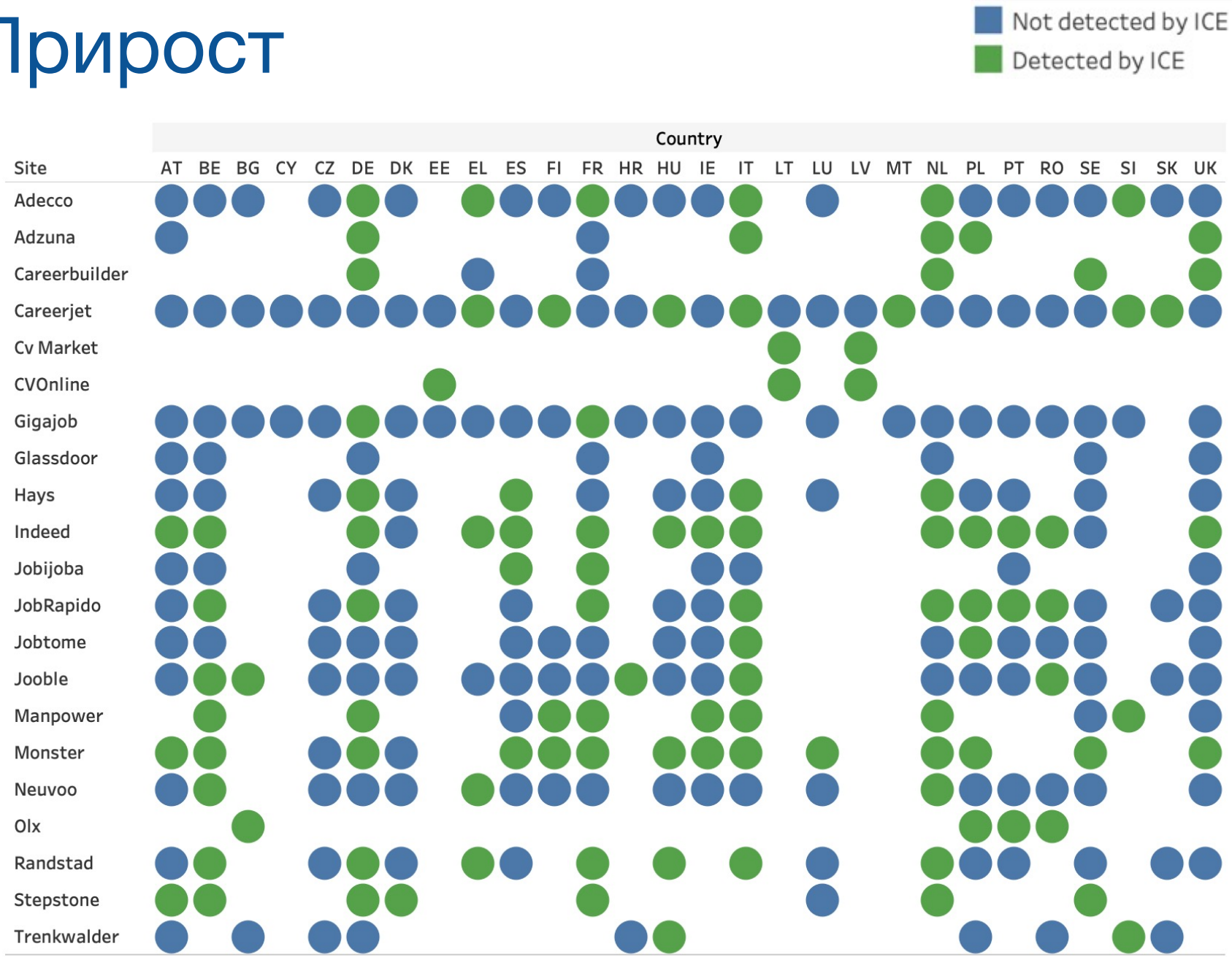
Мы проанализировали результаты деятельности по созданию архитектуры

- Нанесли на карту транснациональные источники
- Добавили дополнительные транснациональные источники
- Добавили полный набор источников EURES (Европейская служба занятости)

С целью

- составить список очередности для определения соглашений
- определить порядок значимости для реализации каналов приема данных

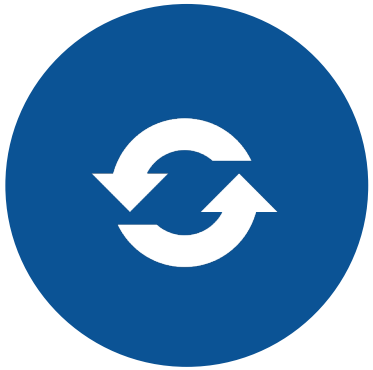
Прирост


Not detected by ICE
Detected by ICE
Country
Site
AT BE BG CY CZ DE DK EE EL ES FI FR HR HU IE IT LT LU LV MT NL PL PT RO SE SI SK UK
Adecco
Adzuna
Careerbuilder
Careerjet
Cv Market
CVOnline
Gigajob
Glassdoor
Hays
Indeed
Jobijoba
JobRapido
Jobtome
Jooble
Manpower
Monster
Neuvoo
Olx
Randstad
Stepstone
Trenkwalder

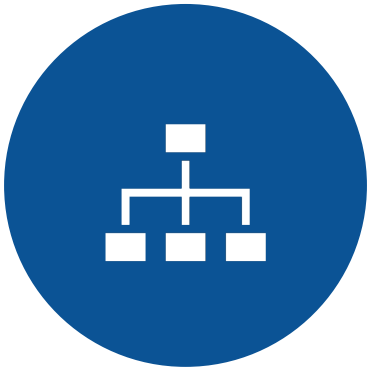
Значимость и ранжировка источников

Объем

Тип веб-портала



Обновление данных



Структурированные данные

Фаза приема данных

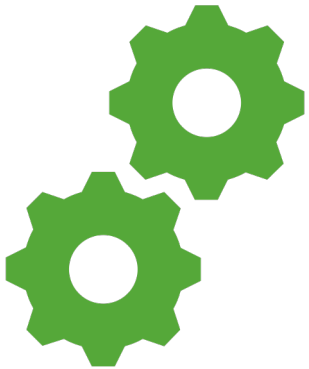
Процесс получения и импорта данных с веб-порталов и их хранение в Базе данных

Внимание к объемам

Рост и максимизация покрытия

Прямые соглашения с самыми значимыми источниками

Проблемы приема

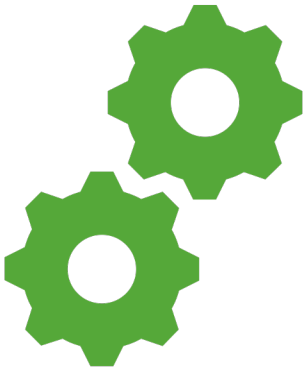


Устойчивость процесса

Качество собранных данных

Масштабируемость и управление

Проблемы приема



1. Устойчивость

Проблема: потенциальные технические проблемы при сборе данных из источника (недоступность, блокировка, изменения в структуре данных)

Риск: потеря данных

Решение: избыточность

- Самые важные сайты (по объему и/или покрытию) с приемом данных из двух или более источников
- Избегать потери данных при возникновении проблем с источником
- Собирать данные из первичных и вторичных источников

Проблемы приема

2. Качество

Проблема: необходимо получить как можно более достоверные данные, по возможности выявляя структурированные данные

Риск: потеря качества

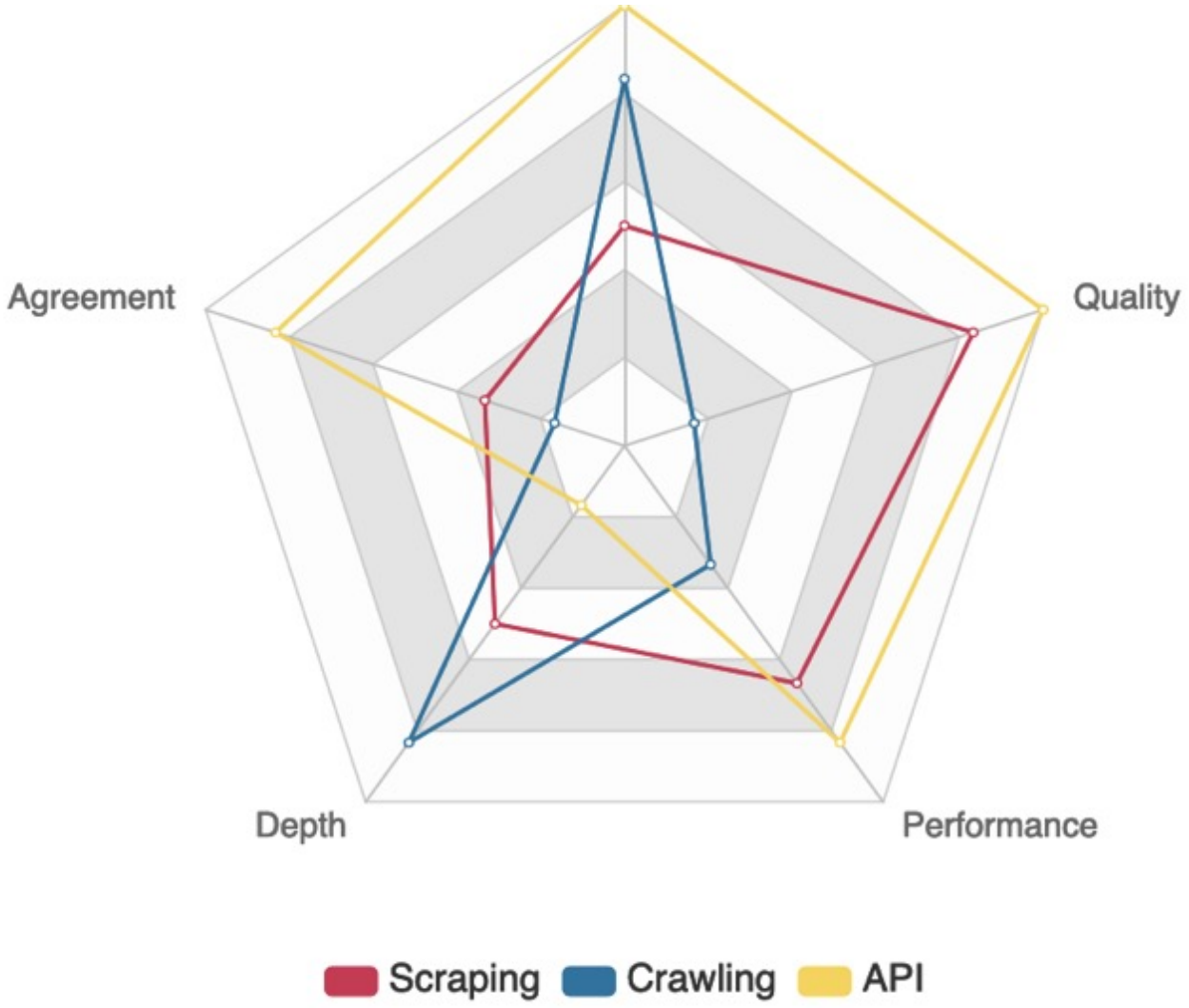
Решение: оптимизированный прием данных. Мы собираем данные, применяя определенный подход на основании одного источника:

- API
- Скрапинг
- Краулинг

Проблемы приема – Качество

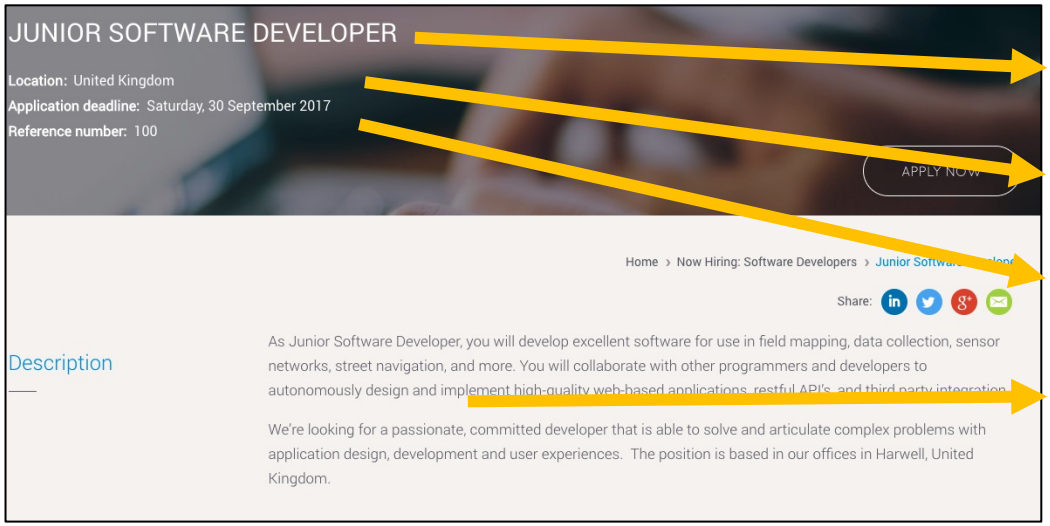
- API: при наличии возможности (соглашений) мы в основном собираем структурированные данные с Веб-порталов.
  - За: Очень высокое качество (большинство полей структурированы)
  - Против: Необходимо соглашение, не всегда возможно его получить
- Скрапинг: если использование API не эффективно, а структура веб-портала единообразна, мы можем разработать специальный скрапер, который будет извлекать структурированные/неструктурированные данные из страниц
  - За: Высокое качество (множество структурированных полей)
  - Против: особая разработка веб-портала
- Краулинг: если структура страницы веб-портала не единообразна, мы принимаем данные посредством многоцелевого краулинга
  - За: Качество ниже (нет структурированных полей)
  - Против: Быстрый и универсальный подход

Техническое обслуживание



Скрапинг – Пример

Веб-скрапинг – это скрапинг данных, используемый для извлечения структурированных данных из веб-сайтов

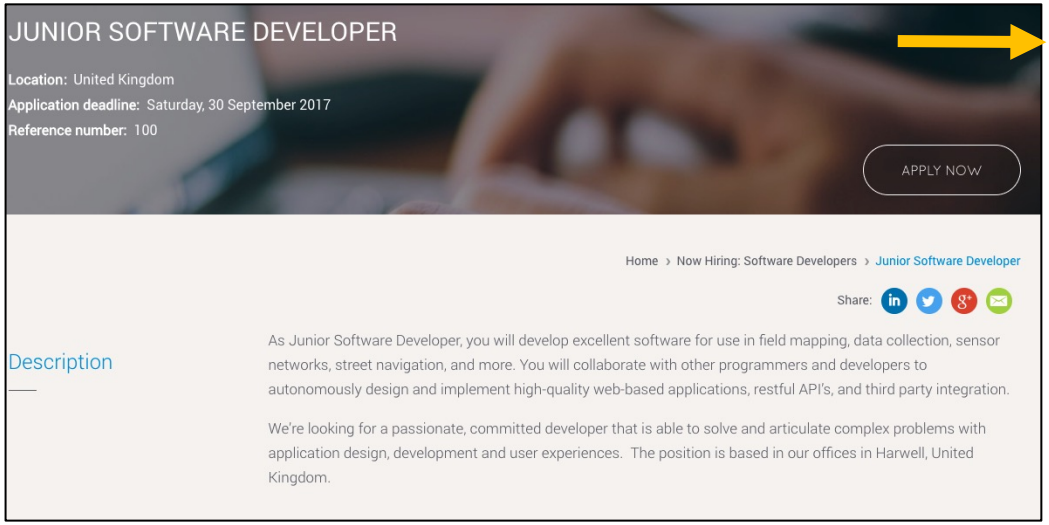


Должность:	Младший разработчик программного обеспечения
Местоположение:	Великобритания
Время:	Суббота, 30 сентября 2017 года
Описание:	В качестве Младшего разработчика ПО вы будете заниматься разработкой отличного ПО для использования …

Краулинг – Пример

Веб-краулер – это бот, который систематически просматривает веб-порталы с целью загрузки всех их страниц.

Краулинг – это стандартный способ массового сбора информации из Интернета с помощью поисковых ботов (например GoogleBot)



Веб-страница:

```
<!DOCTYPE html>
  <head>
    <meta name="title" content="Junior Software
Developer" />
  </head>
  <body>
    <header>
      <h2>Junior Software Developer</h2>
      <div><div>Location</div>United
Kingdom</div>
    …
    </header>
    <div><div>Description</div>
    <span>As Junior Software Developer, you will
develop excellent software for use…
```

Проблемы приема

3. Масштабируемость и управление

Проблема: необходимость обработки сложной среды Больших данных с одновременным подключением к тысячам веб-сайтов

Риск: Потеря контроля за процессом и потеря данных по веб-вакансиям из-за медлительности процесса

Решение:

- Масштабируемая инфраструктура
- Специальный инструмент по мониторингу и управлению

Проблемы приема - Масштабирование

Мы разработали решение на основе микросервисов, с помощью которого можно при необходимости создавать и удалять "виртуальные поисковые компьютеры". На каждом компьютере устанавливается множество браузеров, которые могут имитировать навигацию человека на веб-сайтах.

Основные отличия от настоящего компьютера:

1. Отсутствие мониторов, сохранение страниц в нашем Озере данных
2. Масштабируемость в обоих направлениях по мере необходимости

Резюме и ключевые слова

- Архитектура, выбор источников и прирост
- Оптимизированный подход
  - Компоненты API, скрапинга, краулинга
- Внимание к количеству
  - Масштабирование и сбор данных в режиме реального времени
- Мониторинг собранных данных в режиме реального времени

Вопросы?

Разделы

1. Цель и контекст
2. Задачи
    1. Участники
    2. Функциональная архитектура
    3. Способы приема данных
    4. **Конвейер обработки данных**
    5. Методы классификации

Предварительная обработка данных – Проблемы и определения

- Цель:
  - Вводить надлежащие данные во время фазы извлечения информации
- Задачи:
  - Измерять, мониторить и повышать Качество данных, чтобы максимизировать полноту, единообразие, комплексность, своевременность и повторяемость
- Подход:
  - Разработать многофазный конвейер обработки данных с акцентом на:
    - Выявление вакансий: анализировать страницу веб-сайта, чтобы выделить только то содержимое, которое относится к вакансиям
    - Дедупликация: выявить повторные публикации о вакансиях, чтобы получить общую информацию по одной вакансии
    - Определение даты: установить даты выпуска и истечения срока посредством анализа описания вакансии
    - Срок действия вакансии: метод определения даты истечения срока, если она прямо не указана
- Особенности:
  - Гарантированное качество данных в течение всех фаз обработки

Предварительная обработка данных – Проблемы и определения

Процесс очистки принятых данных и дедупликация веб-вакансий гарантирует, что анализ данных пройдет на максимально высоком уровне качества

Определение языка

Снижение шума

Дедупликация веб-вакансий

Этапы предварительной обработки

Объединение

Очистка

Обработка и резюмирование текста

Предварительная обработка данных Определение языка

- Зачем:
  - В каждом языке есть различные ключевые слова, стоп-слова,…
  - Язык отражает разные культуры и варианты развития Рынка труда…
  - … Таким образом, крайне важно определить язык веб-вакансий, чтобы использовать наиболее подходящий способ классификации
- Как:
  - По каждому языку (60+) мы разработали специальный классификатор на основе корпуса текстов Википедии
  - Полученные модели отличаются высокой точностью (~99 % точности) и оперативным внедрением в конвейер обработки данных
- Что мы получаем:
  - Быструю и надежную классификацию языка, используемого в определенной веб-вакансии
  - Способ архивировать веб-вакансии, по которым у нас нет подходящей классификации

Предварительная обработка данных
Как справляться с шумом?

- В среде Больших данных нам приходится иметь дело с шумом
  - Почему? Потому что сбор информации осуществляется в Интернете, самом зашумленном месте из когда-либо известных
- Прежде всего, нам нужно определить, с каким типом шума нам предстоит столкнуться…:
  - Веб-страницы, очевидно не относящиеся к веб-вакансиям:
    - Страницы социальных сетей
    - Новости
    - Страницы с информацией о конфиденциальности
    - …
  - Веб-страницы, которые выглядят как веб-вакансии:
    - Обучающие курсы
    - Резюме
    - Консультационные услуги
    - …

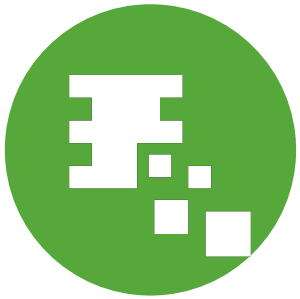
- …Затем нам необходимо выявить и обработать повторные веб-вакансии:
  - Обычно одна вакансия публикуется на множестве порталов
  - Если мы не будем различать такие публикации, то мы можем переоценить Спрос на рабочую силу
  - Таким образом, мы должны выявить повторные веб-вакансии и объединить информацию из них в одну единственную вакансию

Предварительная обработка данных Выявление шума – Как?

- Подход в 2 этапа:
  - Метод машинного обучения
    - По каждому языку мы разработали Наивный байесовский классификатор и включили в него более 20 тысяч веб-страниц:
      - » 10 тысяч страниц, которые действительно относятся к веб-вакансиям
      - » 10 тысяч веб-страниц, которые не относятся к веб-вакансиям
    - Точность ~99 %
    - Быстрое обучение и использование
    - Подход похож на систему обнаружения спама в электронной почте
  - Метод неточных совпадений
    - Используется для выявления веб-страниц, похожих на веб-вакансии, но относящихся к предложениям обучения, консультационных услуг,….
    - Метод работает через просмотр заголовка и основного текста страницы с объявлением и выявление ключевых слов (в зависимости от языка), а в результате мы сможем пометить страницу как "не относящуюся к веб-вакансии"

Но до начала фазы дедупликации веб-вакансий нам нужно очистить текст, упростить его и объединить…

Предварительная обработка данных
Фаза дедупликации



Объявления о работе

Физическая дедупликация или поиск неточных совпадений

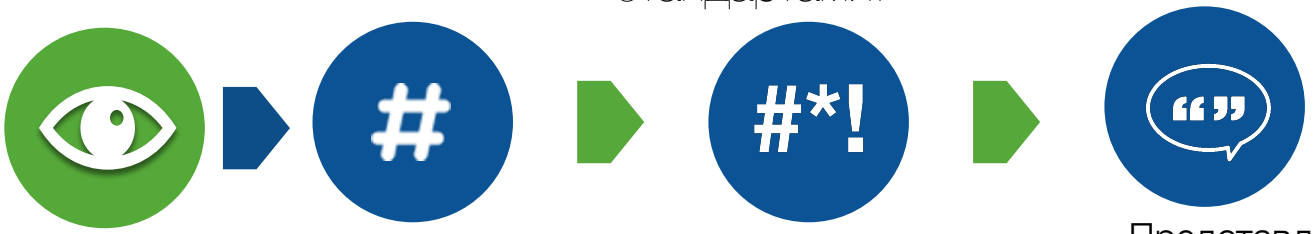
Применяется к описанию (или содержанию) вакансии.

Сопоставление метаданных

Использование метаданных, полученных из порталов с объявлениями о работе, для удаления повторяющихся вакансий на сайтах-агрегаторах (например, идентификационный номер, URL-адрес страницы)

Обработка и резюмирование текста

Фаза обработки и резюмирования текста нацелена на сокращение текста для улучшения процесса классификации вакансий в соответствии с европейскими стандартами.



Определитель языка → Текст объявления о работе → Очистка от шумов и обработка → Представление Векторно-пространственной модели

JUNIOR SOFTWARE DEVELOPER

Location: United Kingdom
Application deadline: Saturday, 30 September 2017
Reference number: 100

Description

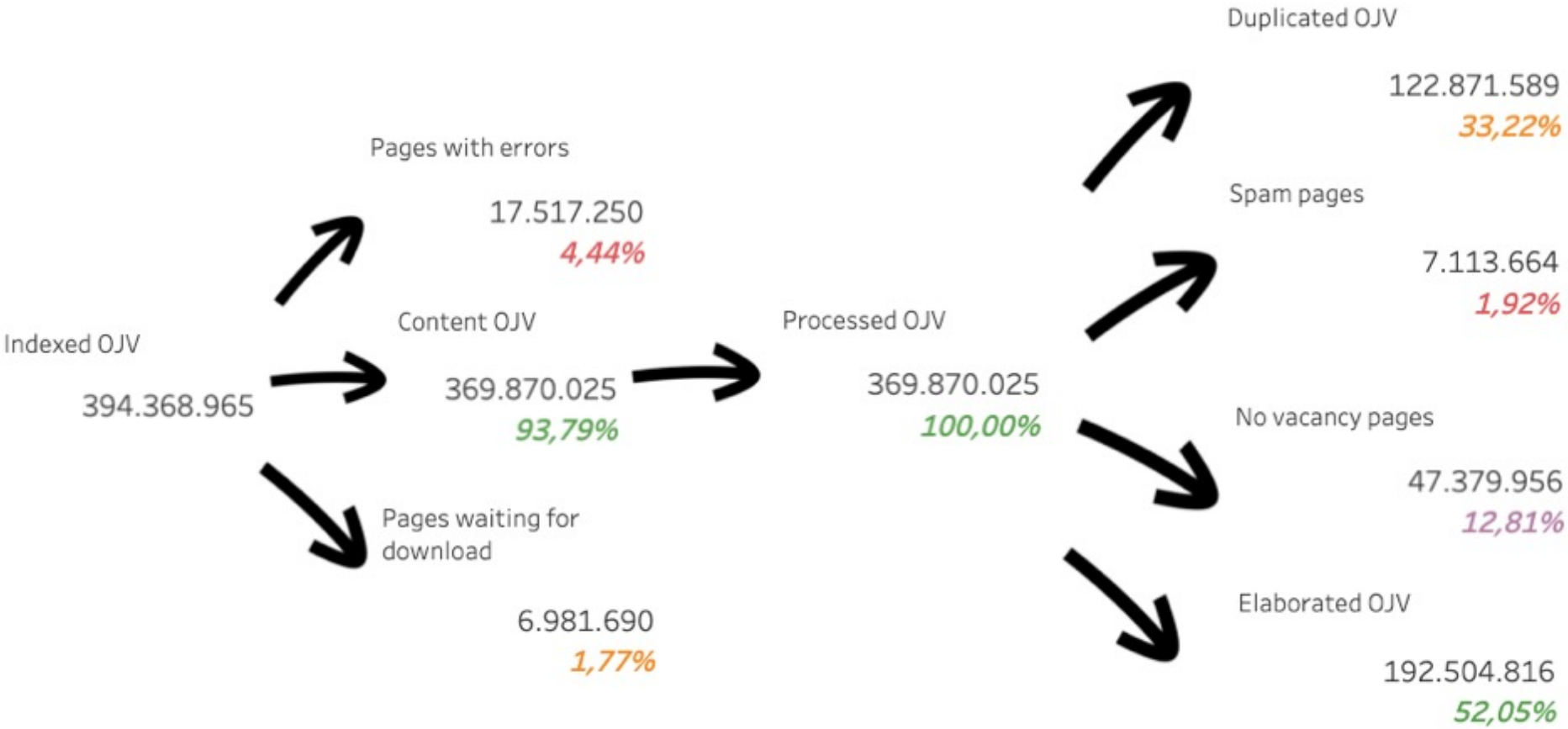
As Junior Software Developer, you will develop excellent software for use in field mapping, data collection, sensor networks, street navigation, and more. You will collaborate with other programmers and developers to autonomously design and implement high-quality web-based applications, restful API's, and third party integration.

We're looking for a passionate, committed developer that is able to solve and articulate complex problems with application design, development and user experiences. The position is based in our offices in Harwell, United Kingdom.

В качестве Младшего ⟨разработчика ПО⟩ вы будете разрабатывать отличное ⟨программное обеспечение⟩ для использования при ⟨сопоставлении полей⟩, ⟨сборе данных⟩, в ⟨сенсорных сетях⟩, ⟨уличной навигации⟩и так далее. Вы будете ⟨сотрудничать⟩ с другими ⟨программистами⟩ и ⟨разработчиками⟩ для ⟨автоматической⟩ разработки и внедрения высококачественных ⟨веб-приложений⟩, с учетом ⟨API⟩ и ⟨интеграцией⟩ третьей стороны.

Мы ищем полного энтузиазма и преданного делу ⟨разработчика⟩, который сможет ⟨решать⟩ и формулировать ⟨сложные проблемы⟩ посредством ⟨создания приложений⟩, ⟨разработки⟩ и ⟨взаимодействия с пользователями⟩. Вакансия открыта для нашего офиса в ⟨Харвелле⟩, ⟨Великобритания⟩.

Предварительная обработка данных – Результаты
Шум



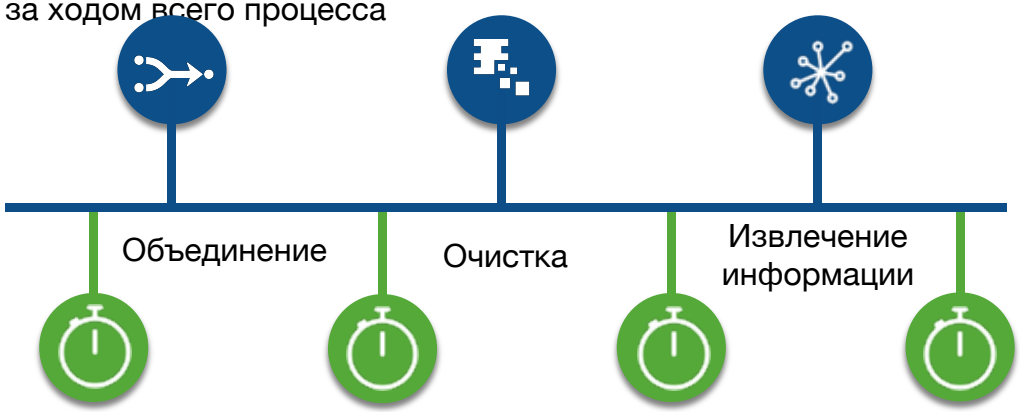
Предварительная обработка данных
Что делать с шумом?

Мы не удаляем шум физически

Мы собираем его, чтобы следить за состоянием всего процесса и выявлять:

- Тип шума → Для определения необходимости разработки процесса более глубокой проверки качества
- Тенденции шума → Для выявления источников, которые повышают/понижают шум, и работы с ними
- Аналитические цели → Анализировать культурные среды определенных стран, например, использование порталов с веб-вакансиями для продвижения обучающих курсов
- Мониторинг →Следить за ходом всего процесса

Как мы следим за качеством данных?



Постоянная проверка качества

Резюме и ключевые слова

- Внимание к качеству
  - Как устранить шум?
  - Действия по дедупликации
- Языковые задачи
  - Компонент, подобранный к определенному языку
- Отслеживание качества данных
  - Постоянная проверка качества

Вопросы?

Разделы

1. Цель и контекст
2. Задачи
    1. Участники
    2. Функциональная архитектура
    3. Способы приема данных
    4. Конвейер обработки данных
    5. **Методы классификации**

Content

379.794.151

Processed

379.794.151

Elaborated

199.008.930

Contract

Structured fields collected
(% over total elaborated OJV)

23,75%

Total fields extracted
(% over total elaborated OJV)

49,00%

Method

Educational level

Structured fields collected
(% over total elaborated OJV)

5,45%

Total fields extracted
(% over total elaborated OJV)

77,63%

Method

Experience

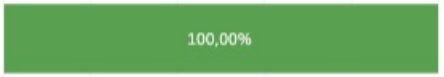
Structured fields collected
(% over total elaborated OJV)

3,86%

Total fields extracted
(% over total elaborated OJV)

35,49%

Method



Occupation

Structured fields collected
(% over total elaborated OJV)

5,21%

Total fields extracted
(% over total elaborated OJV)

76,09%

Method

Industry

Structured fields collected
(% over total elaborated OJV)

25,36%

Total fields extracted
(% over total elaborated OJV)

78,37%

Method

Salary

Structured fields collected
(% over total elaborated OJV)

13,71%

Total fields extracted
(% over total elaborated OJV)

18,24%

Method

Skill

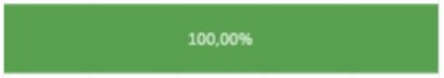
Structured fields collected
(% over total elaborated OJV)

49,94%

Total fields extracted
(% over total elaborated OJV)

62,52%

Method



Working hours

Structured fields collected
(% over total elaborated OJV)

18,21%

Total fields extracted
(% over total elaborated OJV)

43,11%

Method

Feature extraction (Equal)
Feature extraction (Similarity)

Классификация данных

- Цель:
  - Извлечь и структурировать информацию из данных, чтобы предоставить ее на уровне представления
- Задачи:
  - Обработать большой массив неоднородных данных на различных языках
- Подход:
  - Разработать адаптируемую схему с учетом языка, подстроенную к различным особенностям информации. Некоторые актуальные задачи:
    - Классификация по **роду деятельности**: комбинированные методы, такие как Машинное обучение, Тематическое моделирование и Обучение без учителя
    - Классификация по **навыкам**: другие различные комбинированные методы, такие как Анализ текста с учетом сходства на основе корпуса текста или знаний
- Особенности:
  - Гарантированное извлечение объясняемой информации, методы классификации сбора данных и прочие соответствующие особенности.

Классификация данных - Пример

Вакансия

Извлечение информации

- Род деятельности
- Навыки
- Время
- Местоположение
- Отрасль
- …

Младший разработчик программного обеспечения

В качестве Младшего разработчика ПО вы будете разрабатывать отличное программное обеспечение для использования при сопоставлении полей, сборе данных, в сенсорных сетях, уличной навигации и так далее. Вы будете сотрудничать с другими программистами и разработчиками для автоматической разработки и внедрения высококачественных веб-приложений, с учетом API и интеграцией третьей стороны.

Мы ищем полного энтузиазма и преданного делу разработчика, который сможет решать и формулировать сложные проблемы посредством создания приложений, разработки и взаимодействия с пользователями. Вакансия открыта для нашего офиса в Харвелле, Великобритания.

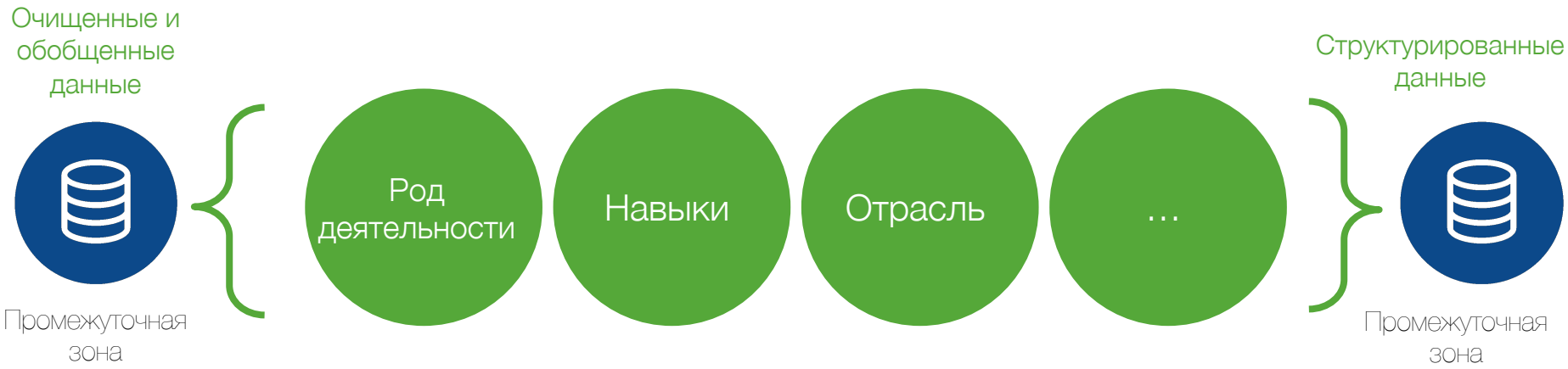
Извлечение информации

- 2512 – Разработчик ПО
- Навыки: разрабатывать программное обеспечение, внедрять веб-приложения, решать проблемы, развивать взаимодействие с пользователями
- Харвелл, Великобритания
- …

Извлечение и классификация информации
Исследование рынка труда в режиме реального времени

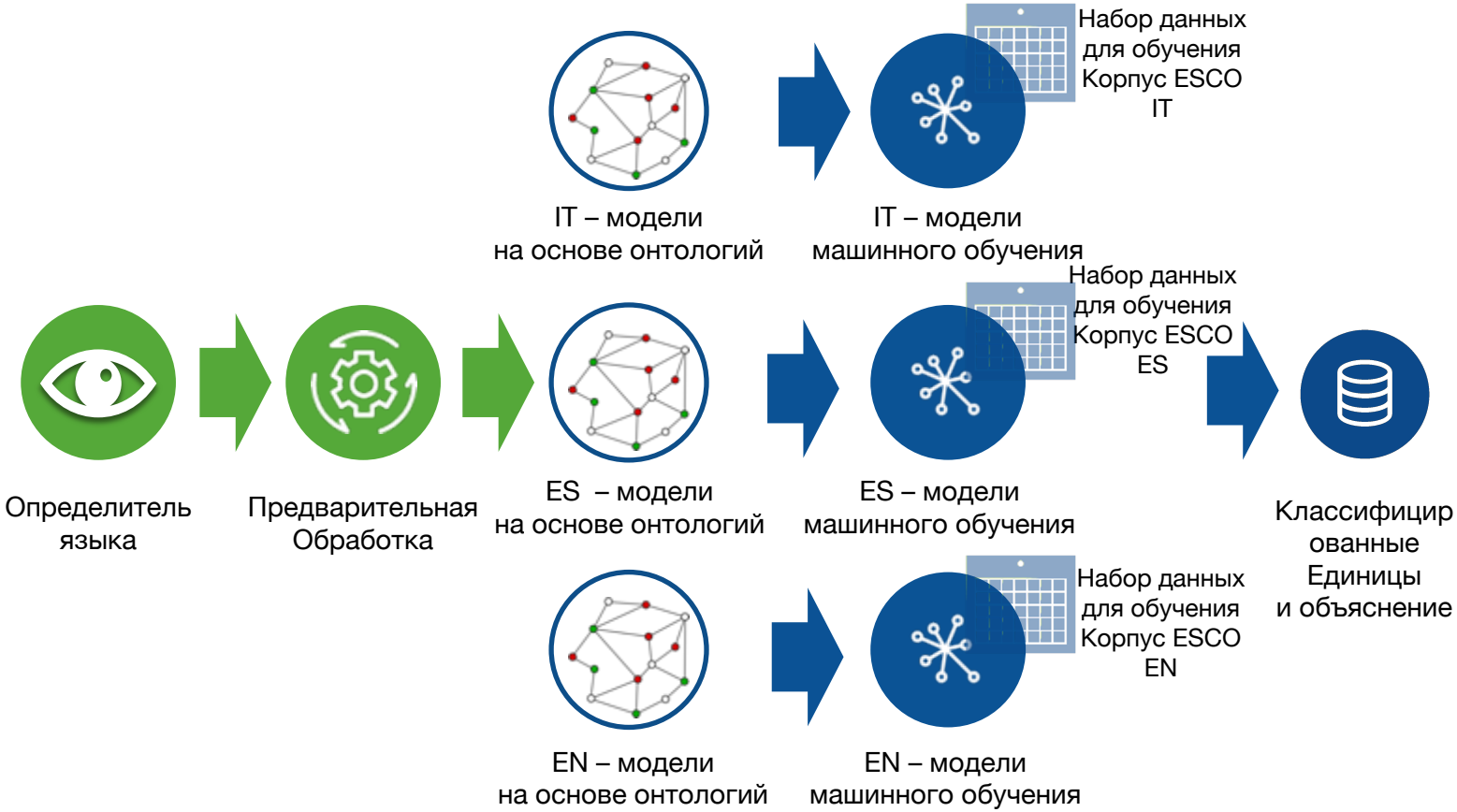
Извлечение информации – это область естественной обработки языка, которая связана с поиском фактической информации в открытом тексте.

Для данной задачи используются методы машинного обучения (обучение, основанное на онтологии, обучение с учителем и без учителя), чтобы сопоставлять объявления о работе со стандартными классификациями.



Машинное обучение → Обучение, основанное на онтологии, обучение с учителем и без учителя и т.д.

Классификация



Что означает "Модели на основе онтологий"?
Как мы можем использовать онтологии для классификации?

Обработка данных по роду деятельности

Вопросы по классификатору по роду деятельности

- Обучение на основе онтологий + Обучение с учителем
  - Онтология Esco
  - Новые наименования из Тематического моделирования
- Одна модель на каждый язык
- Данные, отмеченные специалистами из каждой страны
  - ~100 тысяч объявлений о работе (набор очищенных данных для обучения с помощью нашей онтологии)
  - 436 возможных целей
- Оценка 20 % объявлений о работе с наборами данных золотого стандарта
  - Взвешенная точность ~86 %
  - ~430 выявленных профессий

Подходы к сопоставлению текстов

Основанный на строках

Сопоставление строк оперирует последовательностям и строк и комбинациями символов.

Сходство Джаро — Винклера, Коэффициент Жаккара, Коэффициент Отиаи

Основанный на корпусе

Сопоставление по корпусу текстов – мера семантической близости, по которой определяется сходство между словами в соответствии с информацией, полученной из большого корпуса текста.

Латентно-семантический анализ, явный семантический анализ, инструмент DISCO для поиска схожих слов

Основанный на знаниях

Сопоставление на основе знаний заключается в определении степени схожести между словами с использованием информации, полученной из семантических сетей

Precision of occupation (overall)

86,66%

Validation Set (overall)

317.864

Validation Set by language

bg	ca	cs	da	de	el	en	es	et	eu	fi	fr	gl	hr	hu	it	lt	lv	nl	pl	pt	ro	sk	sl	sv
5.050	14.210	31.290	6.022	17.420	7.173	35.019	21.680	8.414	196	11.972	39.146	811	4.637	13.813	17.228	7.447	4.443	8.687	10.554	14.678	10.226	3.089	4.576	20.083

Precision of occupation by language

bg	ca	cs	da	de	el	en	es	et	eu	fi	fr	gl	hr	hu	it	lt	lv	nl	pl	pt	ro	sk	sl	sv
85,96%	86,45%	88,96%	96,89%	83,17%	85,02%	82,83%	87,39%	83,63%	68,59%	98,80%	83,89%	83,07%	89,45%	97,40%	81,30%	92,86%	96,95%	93,31%	80,02%	83,01%	95,12%	98,50%	87,78%	78,43%

Precision of occupation (lv1)

Clerical support workers	85,77%
Craft and related trades ..	86,10%
Elementary occupations	86,19%
Managers	86,32%
Plant and machine operat..	86,29%
Professionals	86,61%
Service and sales workers	89,38%
Skilled agricultural, fores..	88,79%
Technicians and associate..	85,54%

Precision of occupation (lv2)

Administrative and comm..	85,06%
Agricultural, forestry and ..	80,82%
Assemblers	84,87%
Building and related trad..	92,30%
Business and administrati..	85,66%
Business and administrati..	80,06%
Chief executives, senior o..	91,36%
Cleaners and helpers	85,11%
Customer services clerks	82,21%
Drivers and mobile plant ..	86,49%
Electrical and electronic t..	74,60%
Food preparation assista..	89,08%
Food processing, wood w..	82,61%
General and keyboard cler..	97,20%
Handicraft and printing w..	89,65%

Precision of occupation (lv3)

Administration professio..	86,21%
Administrative and specia..	84,92%
Agricultural, forestry and ..	80,82%
Animal producers	83,13%
Architects, planners, surv..	87,56%
Artistic, cultural and culin..	91,74%
Assemblers	84,87%
Authors, journalists and li..	90,72%
Blacksmiths, toolmakers ..	86,70%
Building and housekeepin..	90,33%
Building finishers and rel..	95,47%
Building frame and relate..	90,00%
Business services agents	89,57%
Business services and ad..	79,10%
Car, van and motorcycle d..	90,40%

Precision of occupation (lv4)

Accountants	83,60%
Accounting and bookkeepi..	58,14%
Accounting associate prof..	85,65%
Actors	93,41%
Administrative and execu..	84,32%
Advertising and marketin..	65,30%
Advertising and public rel..	71,63%
Aged care services manag..	78,81%
Agricultural and forestry ..	94,55%
Agricultural and industria..	76,49%
Agricultural technicians	81,32%
Air conditioning and refri..	85,95%
Air traffic controllers	84,43%
Air traffic safety electroni..	95,52%
Aircraft engine mechanics..	79,61%

Резюме и ключевые слова

- Внимание к резюмированию
  - Как резюмировать данные и улучшить результаты наших аналитиков данных?
- Связь со стандартными принципами таксономии
  - Сравнить данные веб-вакансий с другими источниками
- Задачи, связанные с наборами данных золотого стандарта (мощность множества, качество и разнообразие)
- Смешанные подходы
  - Машинное обучение
  - Обучение на основе онтологии
  - Методы сопоставления текста и извлечения информации
- Жизненный цикл модели

Вопросы?