Занятие 5

Изучение знаний и опыта, полученных в результате проекта ЕФО «Большие данные для аналитики рынка труда»

Сжатая презентация технического построения системы сбора и анализа данных. Акцент на сборе, классификации и визуализации

Мауро Пелуччи

22—24 ноября 2021

Темы

1. Что такое машинное обучение?
2. Databricks (введение)
3. Создание конвейеров обработки данных
    1. Как извлечь данные онлайн-вакансий при помощи скрейпинга и как построить наш конвейер обработки данных на основе Spark
    2. Акцент на категоризации профессии

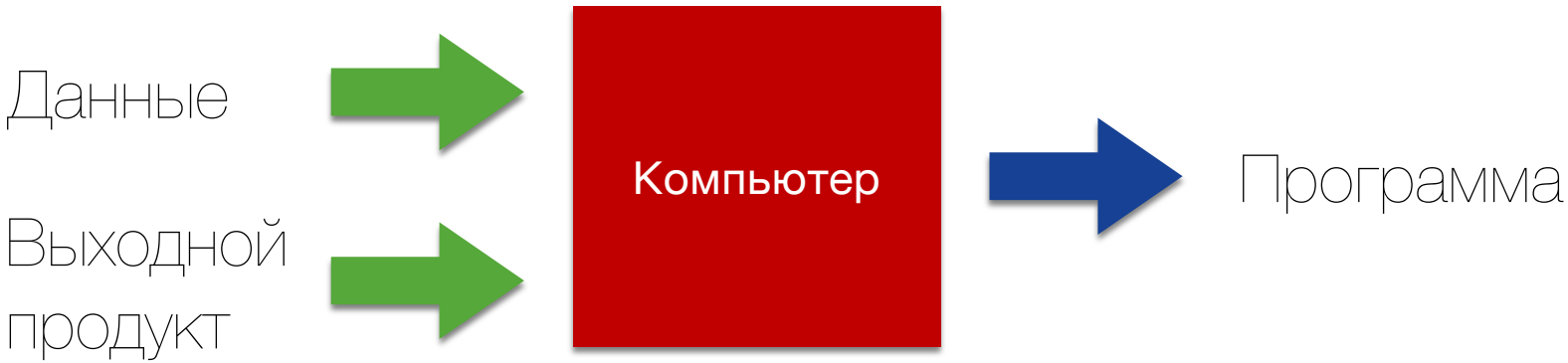
Темы

1. **Что такое машинное обучение?**
2. Databricks (введение)
3. Создание конвейеров обработки данных
    1. Как извлечь данные онлайн-вакансий при помощи скрейпинга и как построить наш конвейер обработки данных на основе Spark
    2. Акцент на категоризации профессии

Машинное обучение

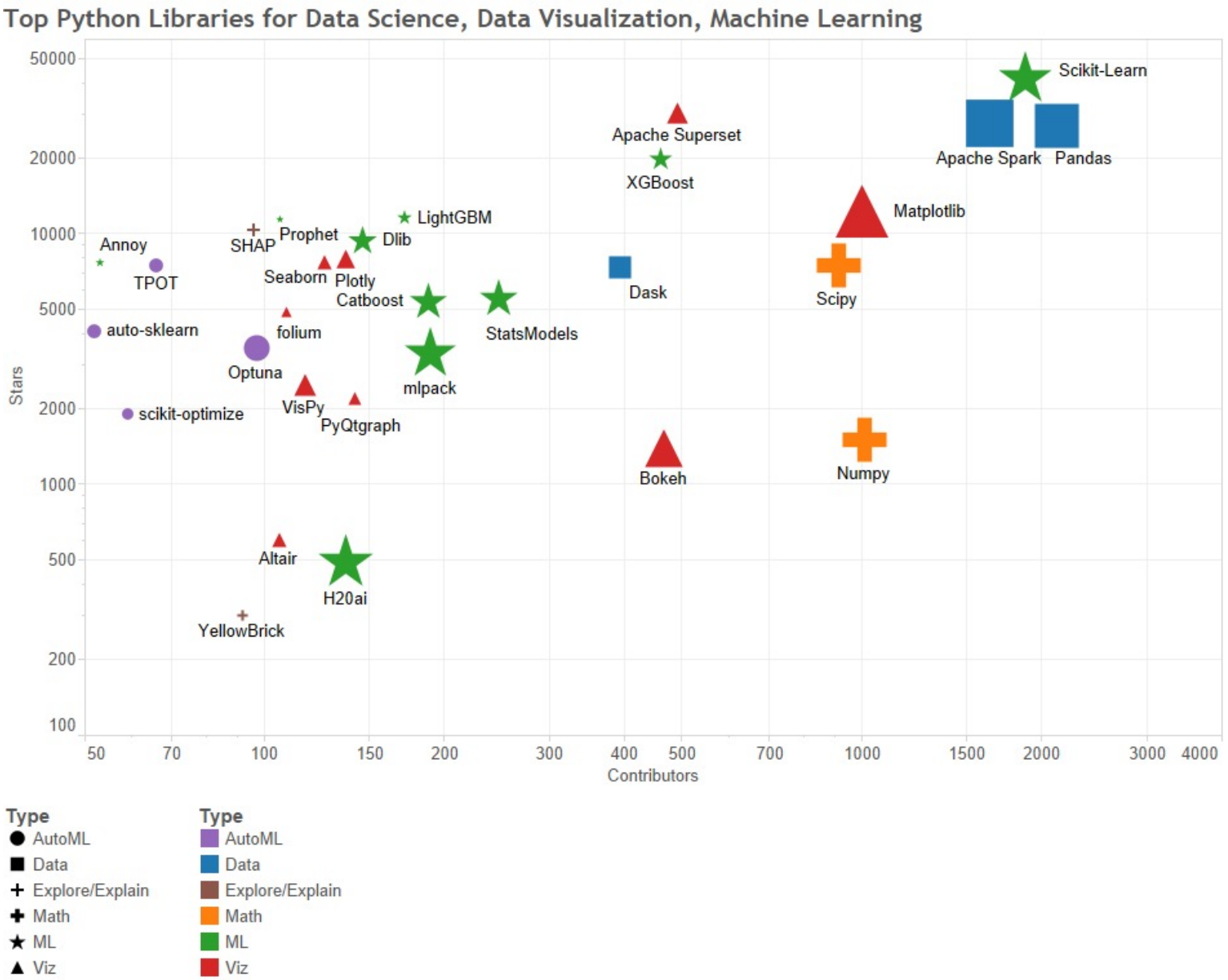
Обучение — это любой процесс, при помощи
которого система повышает свою
эффективность на основании опыта.

Герберт Саймон



Определение машинного обучения

Считается, что компьютерная программа обучается на опыте (E) в отношении определенного класса задач (T) и меры эффективности (P), если ее эффективность при выполнении задач T, при измерении мерой P, улучшается с опытом E.

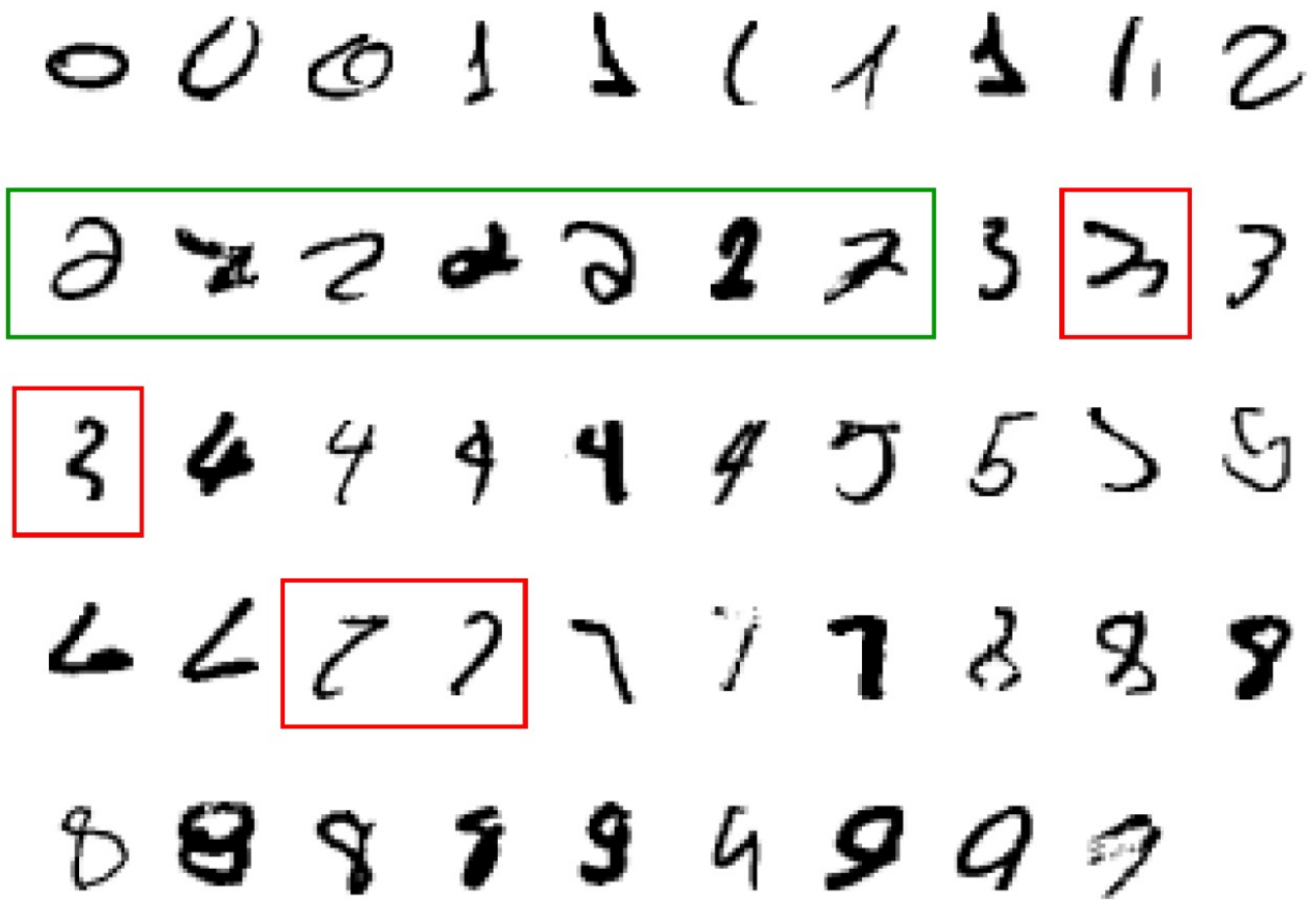


https://www.kdnuggets.com/2020/11/top-python-libraries-data-science-data-visualization-machine-learning.html

Машинное обучение

Тренировка — это процесс развития у системы способности учиться.

Классический пример задачи, требующей машинного обучения:
Очень сложно сказать, какие знаки представляют собой 2



Slide credit: Geoffrey Hinton

Тип обучения

Обучение с учителем (индуктивное)

- Дано: тренировочные данные + желаемые результаты (метки)

Обучение без учителя

- Дано: тренировочные данные (без желаемых результатов)

Обучение с частичным привлечением учителя

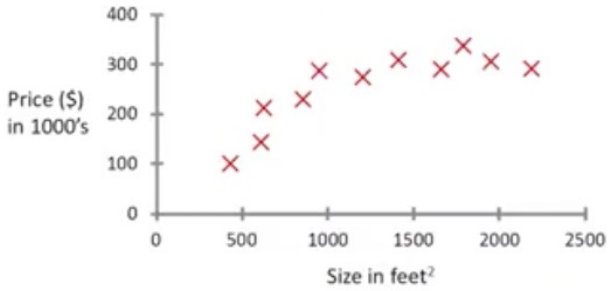
- Дано: тренировочные данные + несколько желаемых результатов

Обучение с подкреплением

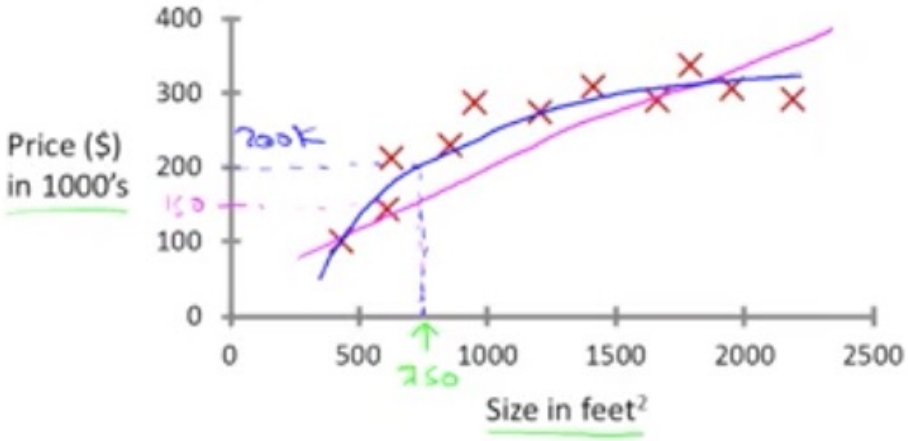
- Награды за выполнение последовательности действий

Обучение с учителем

Прогноз цен на жилье



Прогноз цен на жилье



Мы знаем, к какому классу принадлежат наблюдения

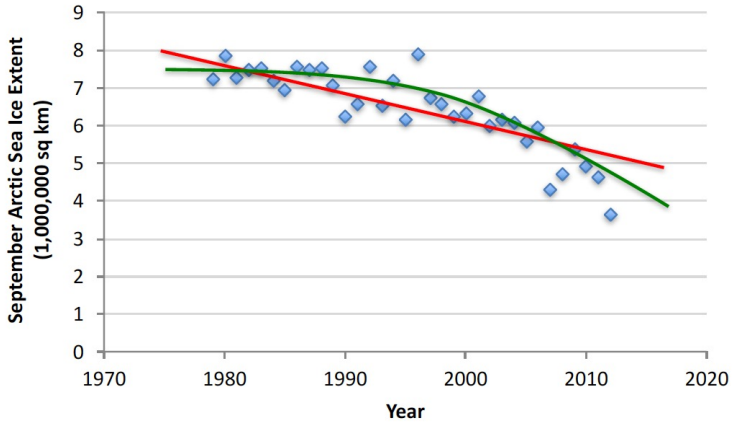
Проблема классификации: к какому классу относится новое наблюдение?

Площадь в футах^2	Количество комнат	Год постройки	Цена ($)
500	3	1983	100 000
1000	4	2005	165 000
1000	3	2016	230 000

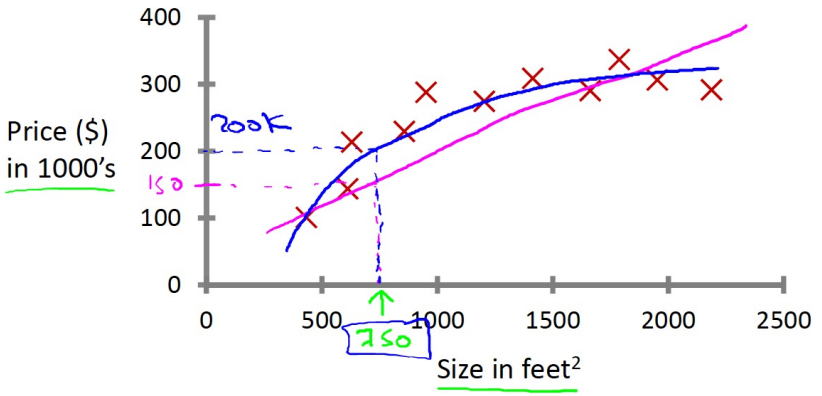
Обучение с учителем
Регрессия/прогнозирование

Дано $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$
Узнать функцию f(x), чтобы спрогнозировать y, когда известно x
y — **действительнозначно** —> **регрессия/прогнозирование**

Оледенение Северного Ледовитого Океана в сентябре (млн. кв. км)

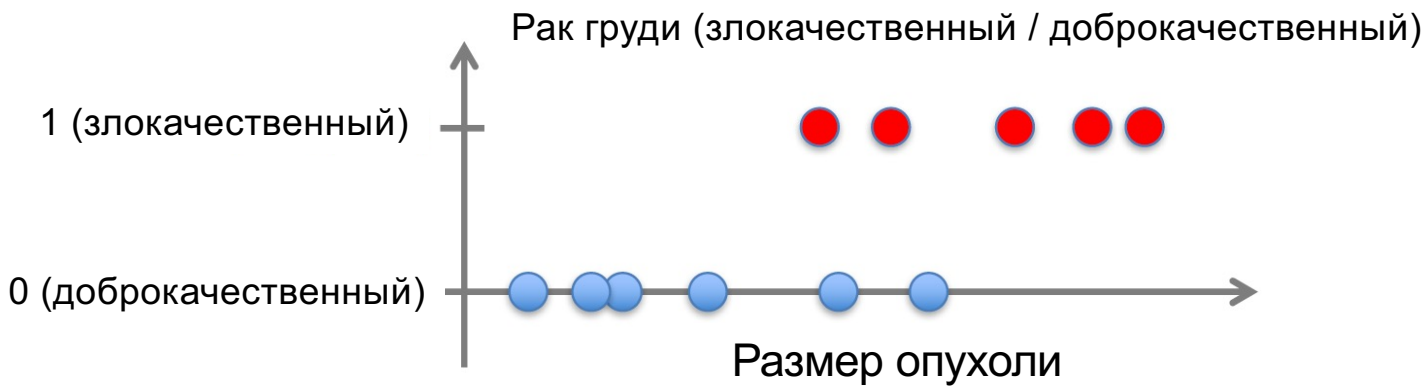


Прогноз цен на жилье



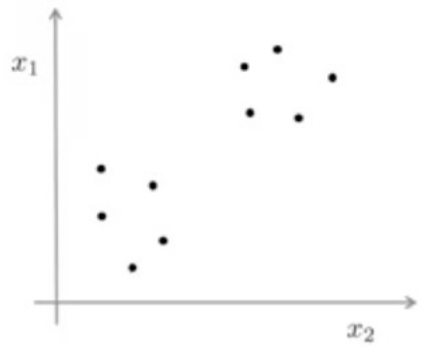
Обучение с учителем
Классификация

Дано $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$
Узнать функцию f(x), чтобы спрогнозировать y, когда известно x
y — **категорийно** —> **классификация**



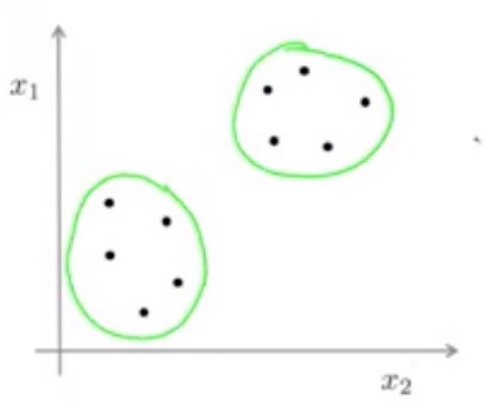
Обучение без учителя

Обучение без учителя



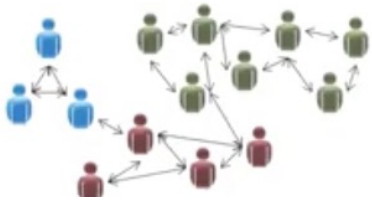
У нас нет информации о классе, к которому относятся наблюдения.
Мы ищем новые признаки, скрытые в наших данных, и пытаемся их интерпретировать.

Обучение без учителя



Применение кластеризации

Сегментация рынка



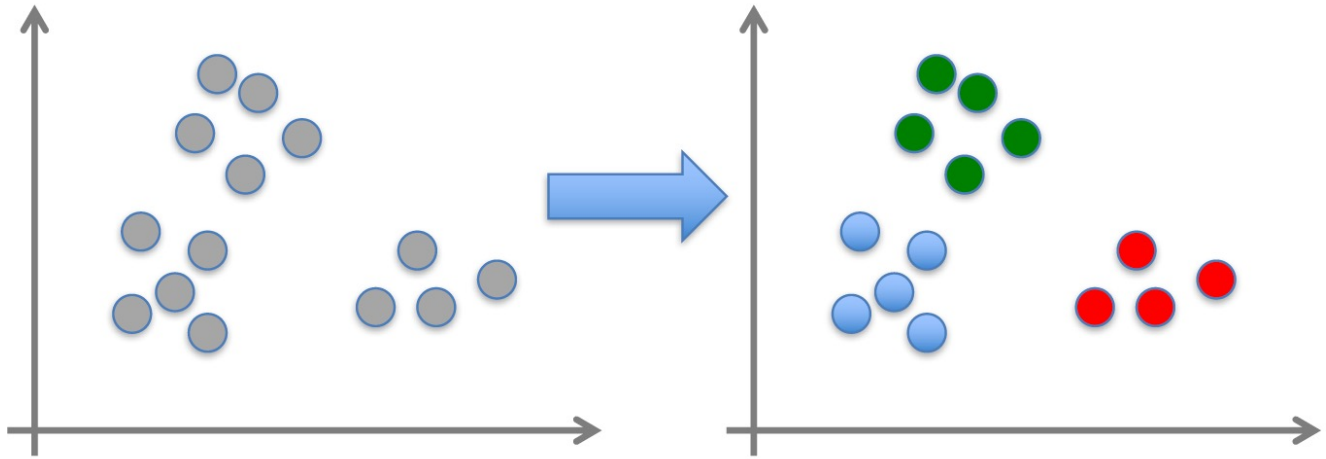
Анализ социальной сети

Организация компьютерных кластеров

Анализ астрономических данных

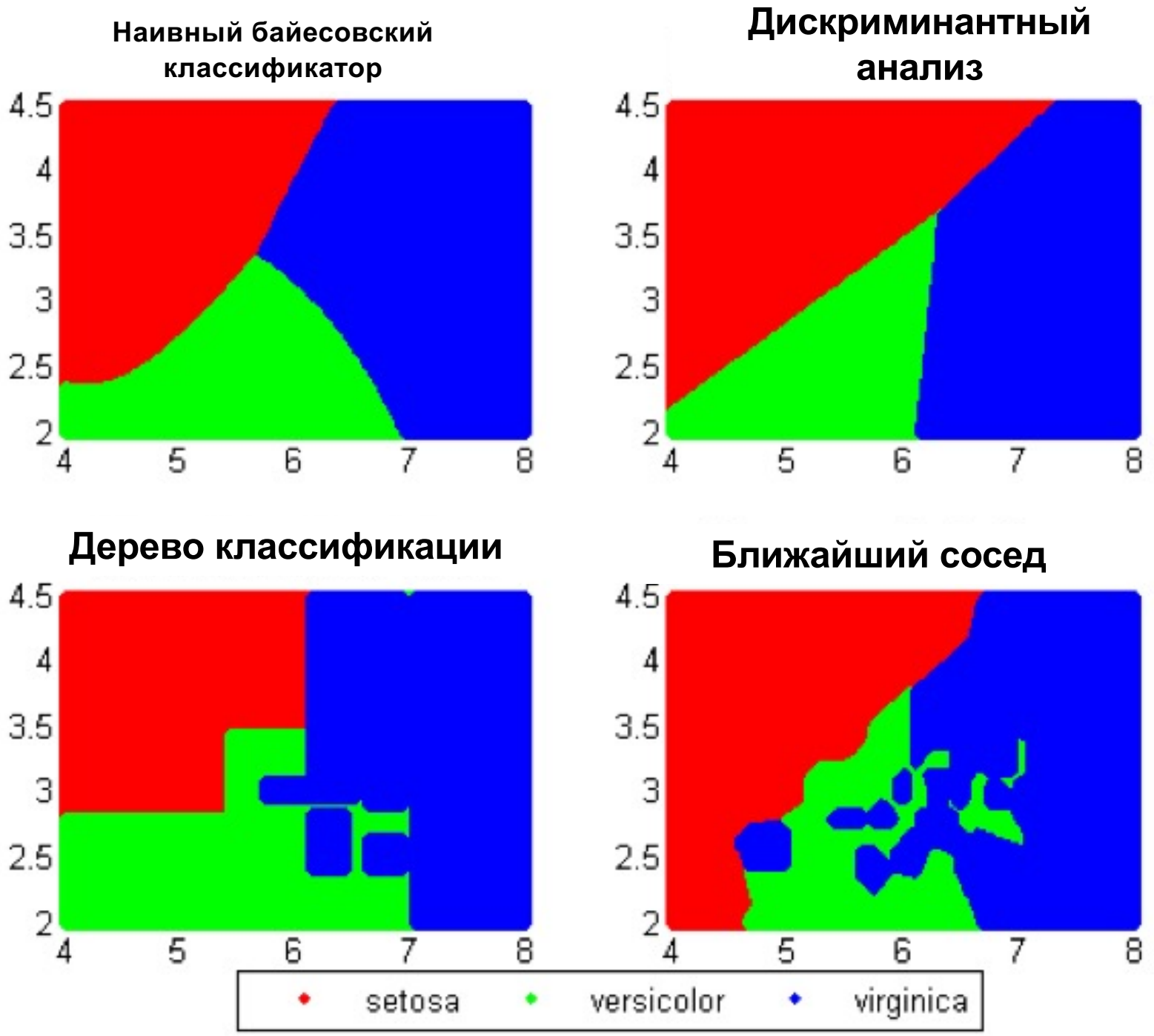
Обучение без учителя
Классификация

Дано (x_1, y_1), (x_2, y_2), ..., (x_n, y_n) (без меток)
Скрытая структура результата, которую обозначает x
Например —> **кластеризация, оценка распределения вероятностей, поиск ассоциаций (в признаках), снижение размерности**



Четыре разных этапа

1. Построение **моделей** регрессии или классификации на основе выборки или набора данных, для которых наблюдаются/известны значения как объясняющих переменных (признаков) Xi, так и зависимой переменной Y.
2. **Оценка эффективности** разных моделей на наборе независимых данных (**проверочный набор**), который не использовался для построения моделей.
3. **Измерение эффективности лучшей модели** на наборе независимых данных.
4. **Применение лучшей модели к новым примерам** (баллы)


Наивный байесовский классификатор
Дискриминантный анализ
Дерево классификации
Ближайший сосед
4.5
4
3.5
3
2.5
2
4
5
6
7
8
setosa
versicolor
virginica

Показатели эффективности

Прогнозируемый класс

Фактический класс	P	N
P	Истинно-положительные (TP)	Ложно отрицательные (FN)
N	Ложно положительные (FP)	Истинно-отрицательные (TN)

$$\text{Правильность} = \frac{tp + tn}{tp + tn + fp + fn}$$

$$\text{Точность} = \frac{tp}{tp + fp}$$

какое количество выданных документов верно (> 0,6)

$$\text{Повторение} = \frac{tp}{tp + fn}$$

сколько положительных результатов выдает модель (> 0,6)

Параметр F

$$F = 2 \cdot \frac{\text{точность} \cdot \text{повторение}}{\text{точность} + \text{повторение}}$$

Показатель эффективности
Точность

Важно **получить ошибочные результаты в виде отдельных числовых значений**.
Иначе будет трудно оценить эффективность вашего алгоритма.

$$\text{Точность} = \frac{tp}{tp + fp}$$

Точность: какое количество классифицированных документов верно
(высокая точность = без «мусора»)

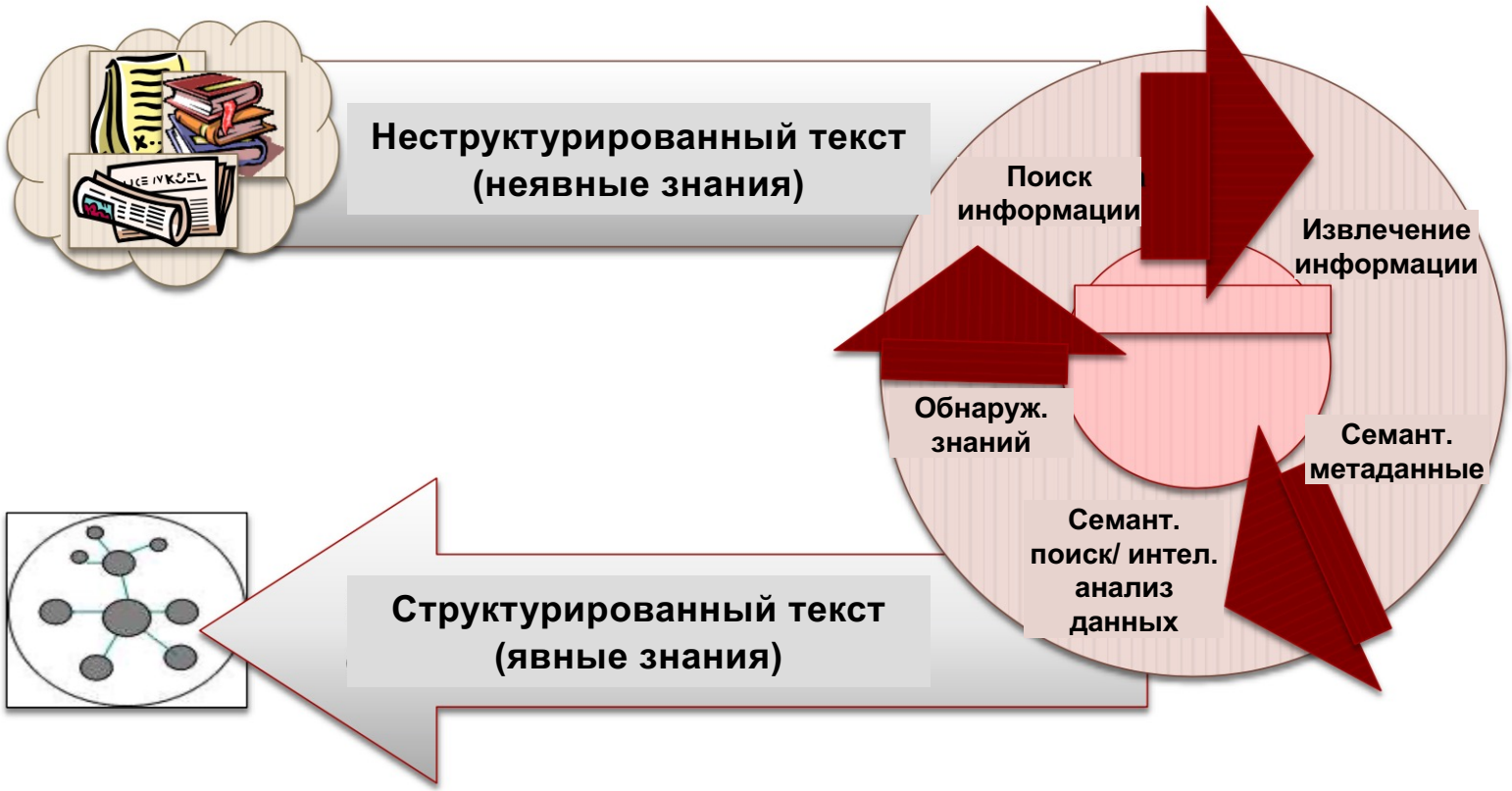
Из всех спрогнозированных нами случаев, где y = разработчик ПО,
какая часть действительно относится к разработчикам ПО?

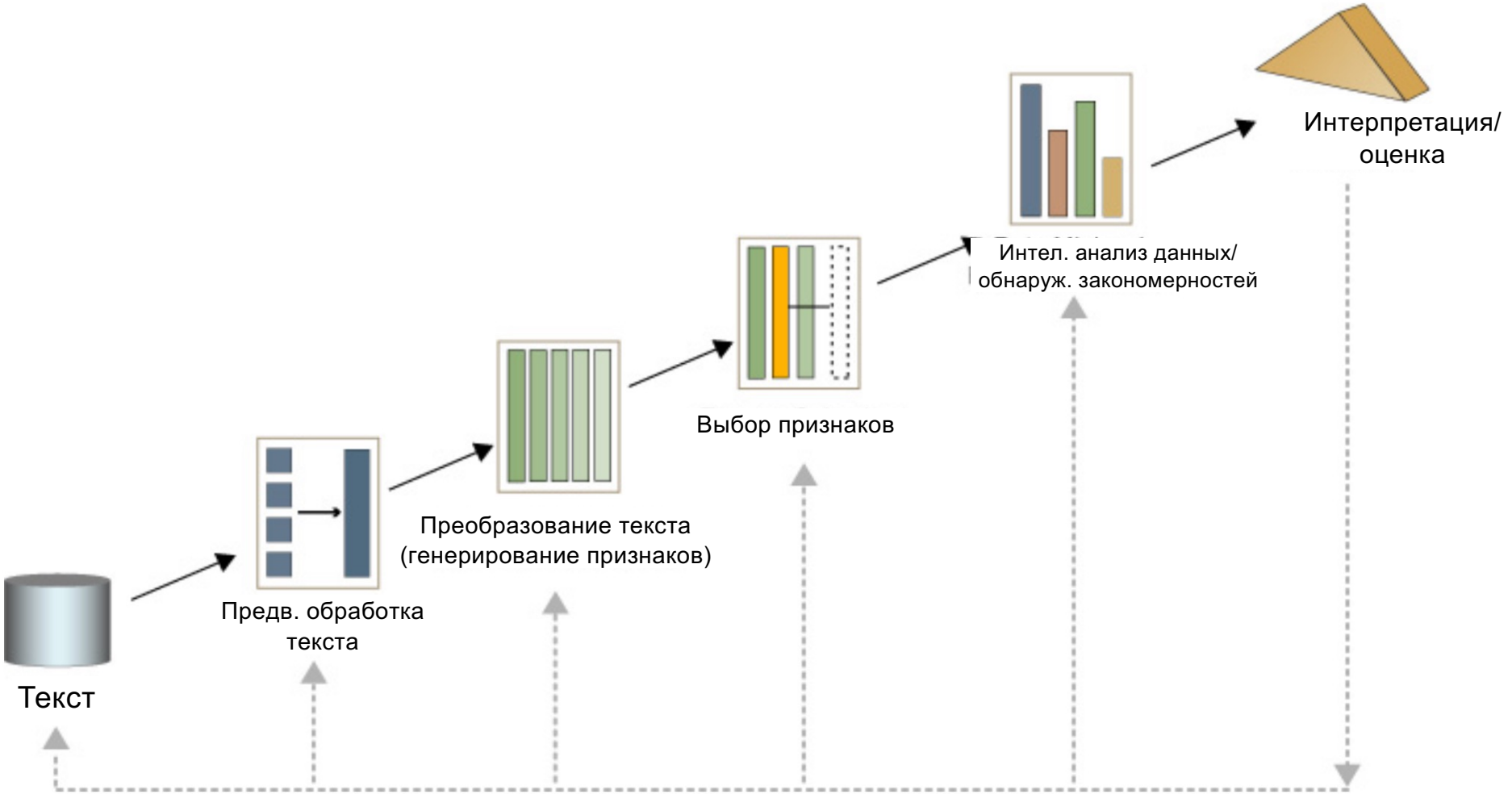
Определение

- Интеллектуальным анализом текста обычно называют процесс извлечения интересующей информации и знаний из неструктурированного текста.

- Интеллектуальный анализ текста можно определить как наукоемкий процесс, в котором пользователь в течение времени взаимодействует с массивом документов при помощи набора аналитических инструментов.

- Интеллектуальный анализ текста предназначен для извлечения полезной информации из источников данных (массивов документов) путем выявления и изучения интересующих закономерностей.


Неструктурированный текст
(неявные знания)
Поиск
информации
Извлечение
информации
Обнаруж.
знаний
Семант.
метаданные
Семант.
поиск/ интел.
анализ
данных
Структурированный текст
(явные знания)


Интерпретация/
оценка
Интел. анализ данных/
обнаруж. закономерностей
Выбор признаков
Преобразование текста
(генерирование признаков)
Предв. обработка
текста
Текст

Структурирование текстовой информации

Разработано много методов анализа структурированных данных.

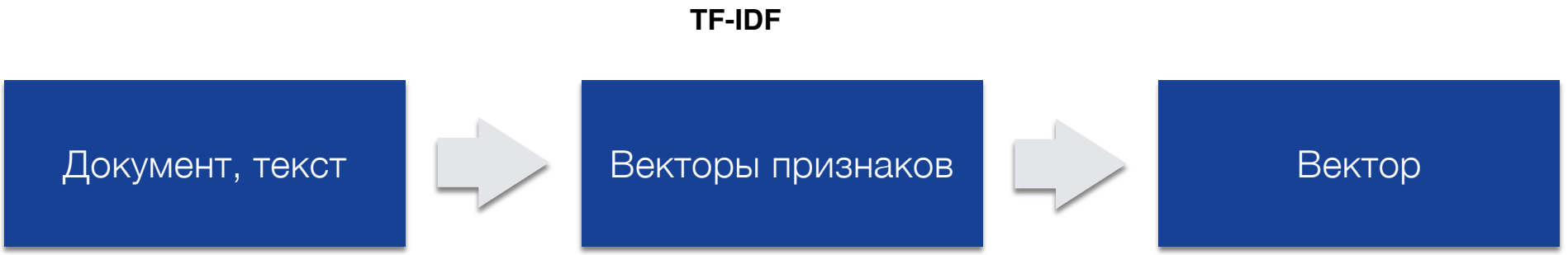
Если нам удастся представить документы в виде набора атрибутов, мы сможем использовать существующие методы интеллектуального анализа.

Использование статистики, чтобы добавить к неструктурированному тексту числовой аспект.

Как представить документ?

- Векторное представление —> набор слов
- Частотность термина (TF)
- Частотность документа (DF)
- TF-IDF
- Объем документа

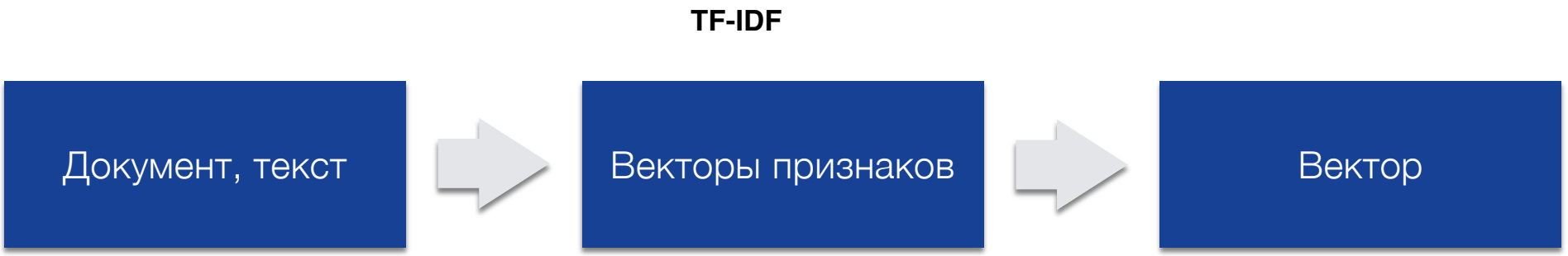
Схема взвешивания для частотности терминов



$$\text{TF-IDF}(w, d) = \text{TermFreq}(w, d) \cdot \log\left(N / \text{DocFreq}(w)\right)$$

TermFreq(w, d): частотность слова (w) в документе (d)
N: количество документов в массиве
DocFreq(w): количество содержащихся в массиве документов, в которых есть слово (w)

Схема взвешивания для частотности терминов

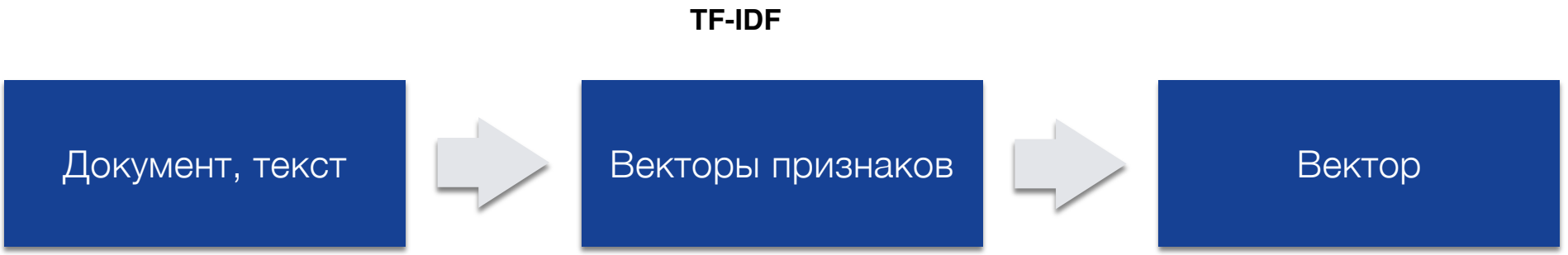


$$TF\text{-}IDF(w, d) = TermFreq(w, d) \cdot \log (N / DocFreq(w))$$

TermFreq(w, d): частотность слова (w) в документе (d)
N: количество документов в массиве
DocFreq(w): количество содержащихся в массиве документов, в которых есть слово (w)

Термину, который встречается в документе много раз, присваивается высокое значение TF-IDF, если он не является обычным для всего массива документов:
это РЕДКИЕ и ВАЖНЫЕ термины

Схема взвешивания для частотности терминов



$$TF\text{-}IDF(w, d) = TermFreq(w, d) \cdot \log (N / DocFreq(w))$$

TermFreq(w, d): частотность слова (w) в документе (d)
N: количество документов в массиве
DocFreq(w): количество содержащихся в массиве документов, в которых есть слово (w)

Термины с низким значением TF-IDF либо встречаются в документах редко, либо очень распространенные во всем массиве.
Распространенные в массиве — ОБЫЧНЫЕ СЛОВА И ШУМ

Темы

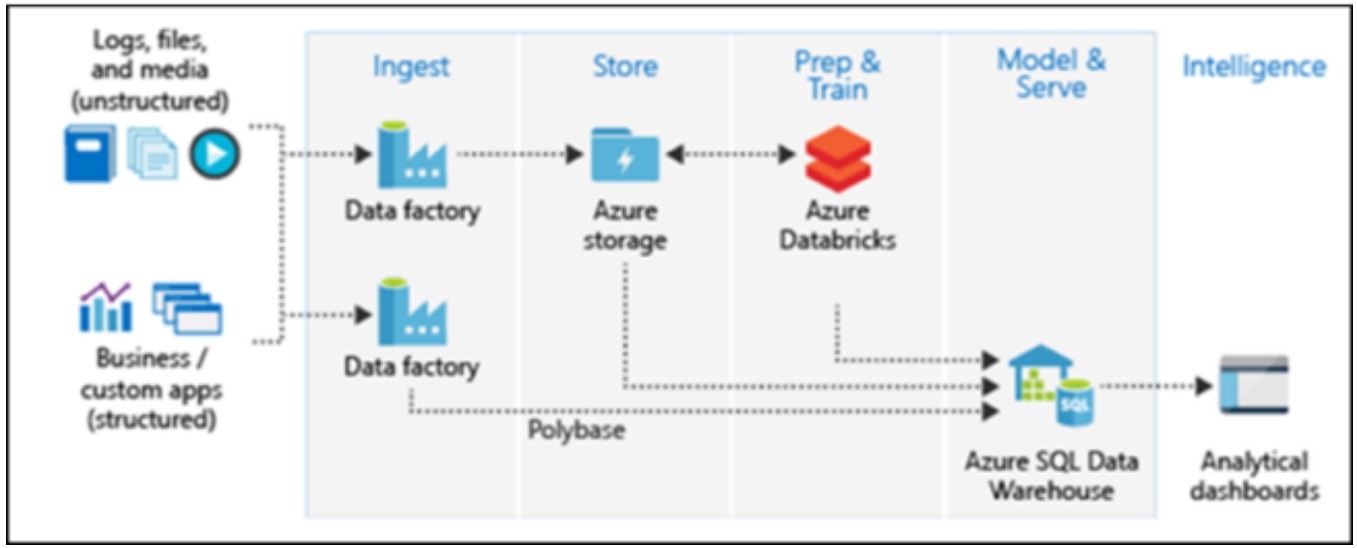
1. Что такое машинное обучение?
2. **Databricks (введение)**
3. Создание конвейеров обработки данных
    1. Как извлечь данные онлайн-вакансий при помощи скрейпинга и как построить наш конвейер обработки данных на основе Spark
    2. Акцент на категоризации профессии

Databricks

- Это простая аналитическая платформа совместного пользования на базе Apache Spark.
- В Databricks представлены все модули Spark (SparkSQL, Streaming, ML, GraphX).
- «Основная цель: устранить все трудности и сложности получения и управления кластером Spark».
- Установка одним нажатием и управление параметрами.
- Предлагает упрощенные рабочие процессы и интерактивную рабочую область для содействия сотрудничеству между исследователями данных, разработчиками и бизнес-аналитиками.
- Интеграция с ведущими поставщиками облачных услуг, такими как Amazon AWS и Microsoft Azure.

Databricks


User Views / Clicks
Impressions
Conversions
User Conversions
Amazon Kinesis / Kafka
databricks
ETL, Aggregation, Attribution
Real-Time Reporting in BI Tools


Logs, files, and media (unstructured)
Business / custom apps (structured)
Ingest
Store
Prep & Train
Model & Serve
Intelligence
Data factory
Azure storage
Azure Databricks
Data factory
Polybase
Azure SQL Data Warehouse
Analytical dashboards

Databricks

Кластеры

Блокноты

Работы

Данные

Databricks
Блокнот

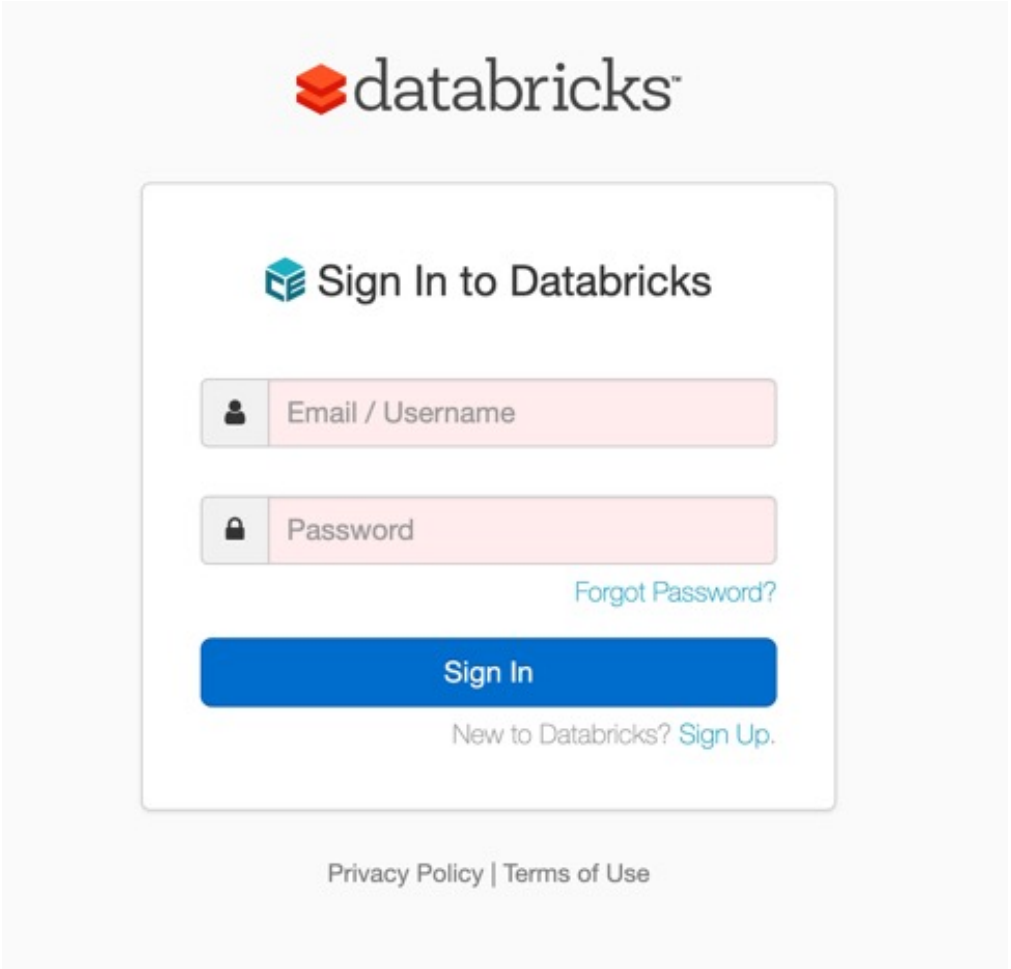
- Подобен блокнотам Jupyter или Zeppelin
- Поддерживаемые языки
  - Python, Scala и SQL (также R…)
  - все они могут быть использованы в рамках одного блокнота
- Сессия Spark уже определена для каждого блокнота в качестве глобальной переменной Spark
- Как только блокнот создан, он должен быть подключен к активному кластеру

Databricks
Версия и совместная работа

- Databricks — аналитическая платформа для совместной работы, где пользователи могут предоставлять друг другу доступ к рабочим областям, кластерам и работам в рамках одного интерфейса.

- Существует возможность создания совместных моделей в таком же блокноте в режиме реального времени, повторного использования активов данных, библиотек в том же кластере, либо повторного использования/мониторинга запланированных работ.

- Databricks поддерживает интеграцию с Github, Bitbucket Cloud и Azure DevOps Services.



https://community.cloud.databricks.com/login.html

Темы

1. Что такое машинное обучение?
2. Databricks (введение)
3. **Создание конвейеров обработки данных**
    1. Как извлечь данные онлайн-вакансий при помощи скрейпинга и как построить наш конвейер обработки данных на основе Spark
    2. Акцент на категоризации профессии

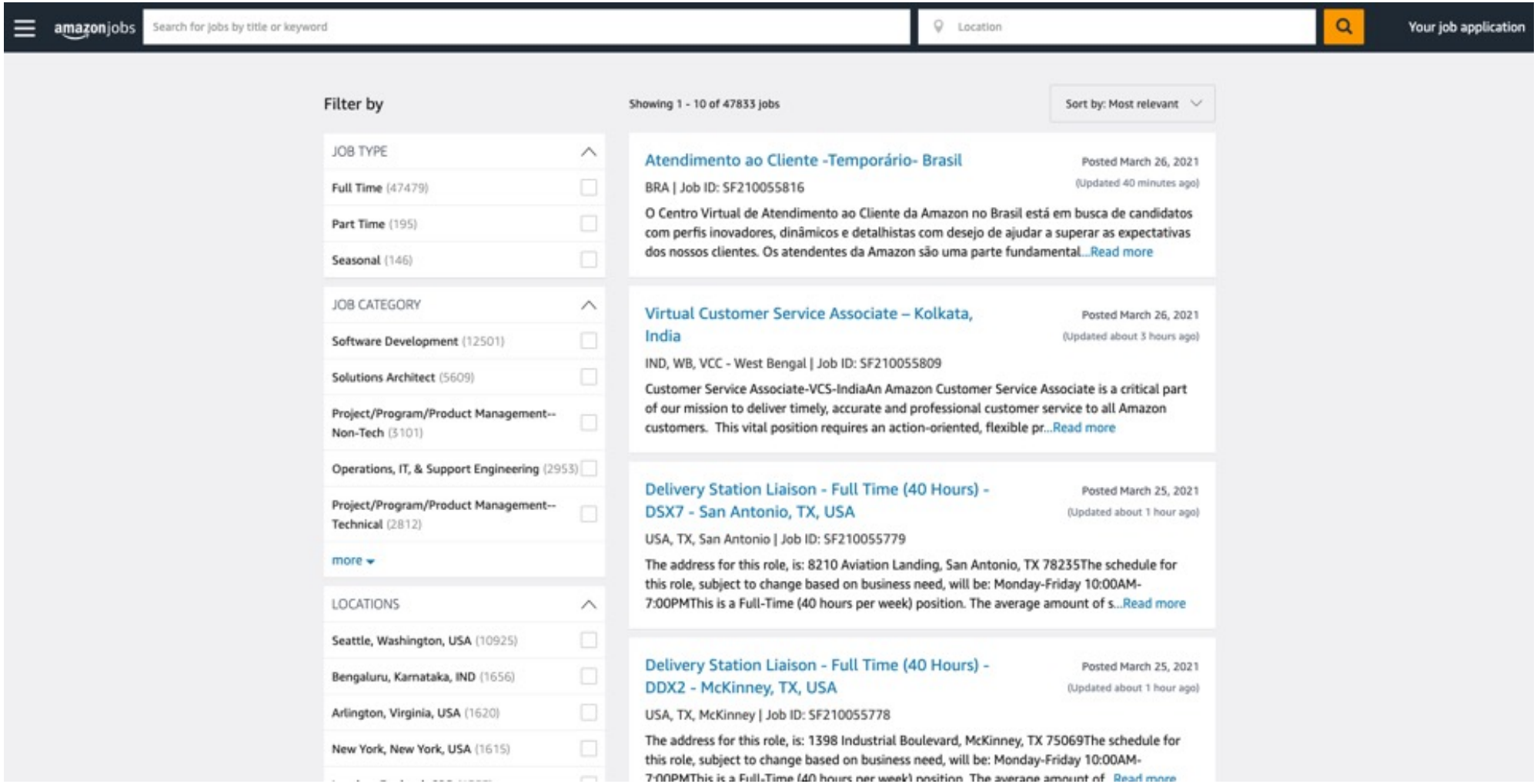
Что нам следует сделать?

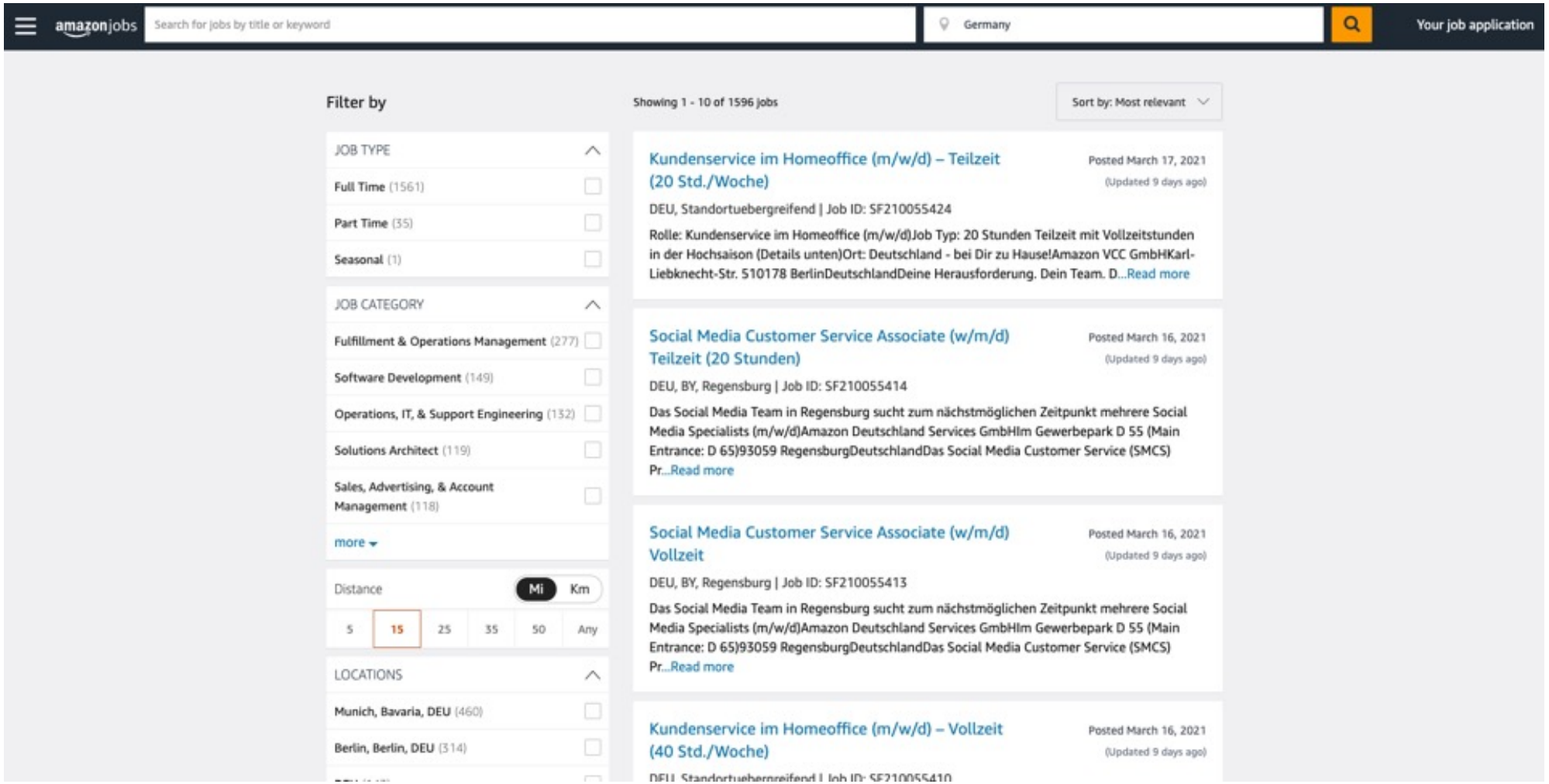
- Собрать определенное количество объявлений о работе
- Создать кластер
- Натренировать модель машинного обучения классифицировать профессии

Темы

1. Что такое машинное обучение?
2. Databricks (введение)
3. Создание конвейеров обработки данных
    1. **Как извлечь данные онлайн-вакансий при помощи скрейпинга и как построить наш конвейер обработки данных на основе Spark**
    2. Акцент на категоризации профессии

https://www.amazon.jobs/it/search?base_query=&loc_query=




amazonjobs
Search for jobs by title or keyword
Germany
Your job application
Filter by
Showing 1 - 10 of 1596 jobs
Sort by: Most relevant
JOB TYPE
Full Time (1561)
Part Time (35)
Seasonal (1)
JOB CATEGORY
Fulfillment & Operations Management (277)
Software Development (149)
Operations, IT, & Support Engineering (132)
Solutions Architect (119)
Sales, Advertising, & Account Management (118)
more
Distance
Mi
Km
5
15
25
35
50
Any
LOCATIONS
Munich, Bavaria, DEU (460)
Berlin, Berlin, DEU (314)
Kundenservice im Homeoffice (m/w/d) – Teilzeit (20 Std./Woche)
Posted March 17, 2021
(Updated 9 days ago)
DEU, Standortuebergreifend | Job ID: SF210055424
Rolle: Kundenservice im Homeoffice (m/w/d)Job Typ: 20 Stunden Teilzeit mit Vollzeitstunden in der Hochsaison (Details unten)Ort: Deutschland - bei Dir zu Hause!Amazon VCC GmbHKarl-Liebknecht-Str. 510178 BerlinDeutschlandDeine Herausforderung. Dein Team. D...Read more
Social Media Customer Service Associate (w/m/d) Teilzeit (20 Stunden)
Posted March 16, 2021
(Updated 9 days ago)
DEU, BY, Regensburg | Job ID: SF210055414
Das Social Media Team in Regensburg sucht zum nächstmöglichen Zeitpunkt mehrere Social Media Specialists (m/w/d)Amazon Deutschland Services GmbHIm Gewerbepark D 55 (Main Entrance: D 65)93059 RegensburgDeutschlandDas Social Media Customer Service (SMCS) Pr...Read more
Social Media Customer Service Associate (w/m/d) Vollzeit
Posted March 16, 2021
(Updated 9 days ago)
DEU, BY, Regensburg | Job ID: SF210055413
Das Social Media Team in Regensburg sucht zum nächstmöglichen Zeitpunkt mehrere Social Media Specialists (m/w/d)Amazon Deutschland Services GmbHIm Gewerbepark D 55 (Main Entrance: D 65)93059 RegensburgDeutschlandDas Social Media Customer Service (SMCS) Pr...Read more
Kundenservice im Homeoffice (m/w/d) – Vollzeit (40 Std./Woche)
Posted March 16, 2021
(Updated 9 days ago)

Набор данных

- ~100 объявлений о работе в Интернете
  - С сайта amazon.jobs
  - Германия

https://colab.research.google.com/notebooks/intro.ipynb#recent=true

https://pandas.pydata.org/getting_started.html

```
[ ] import pandas as pd
    ds_items = pd.DataFrame(items_details)
    ds_items.set_index("job_id")
    ds_items.head()
```

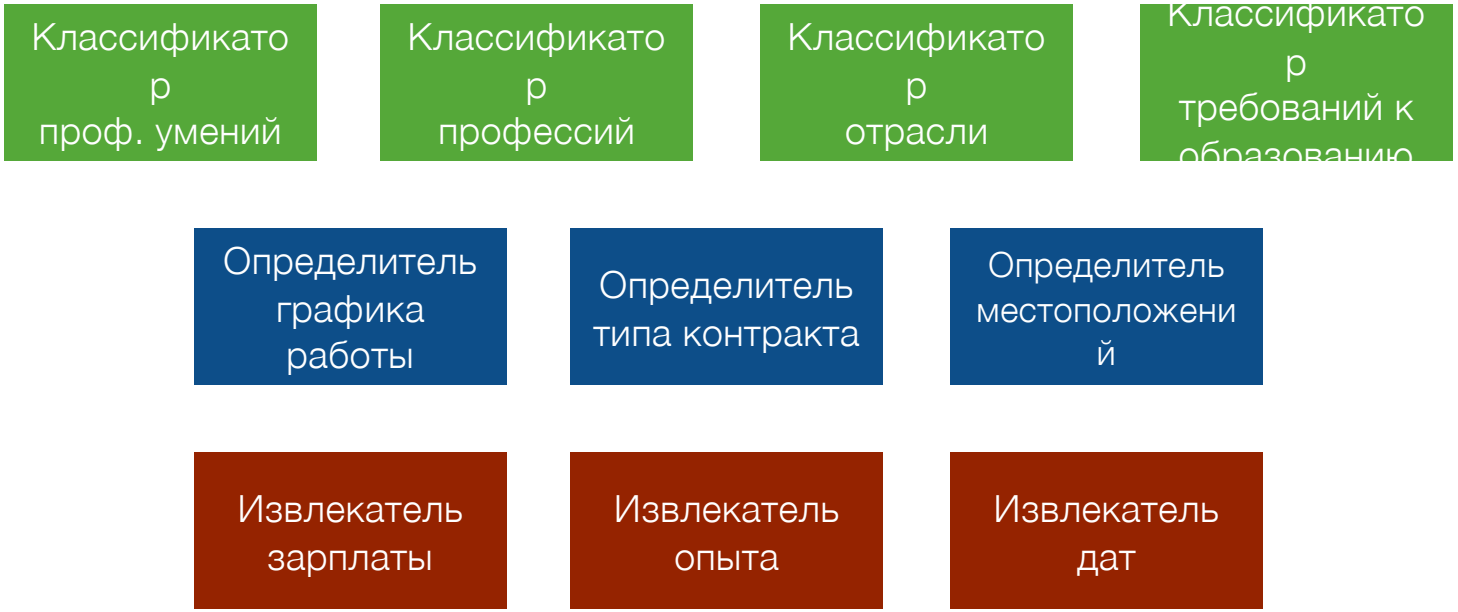
	title	url	location	description	job_id
0	Kundenservice im Homeoffice (m/w/d) – Teilzeit...	https://www.amazon.jobs/en/jobs/SF210055424/ku...	[location]	DESCRIPTIONvnRolle: Kundenservice im Homeoffic...	SF210055424
1	Social Media Customer Service Associate (w/m/d...	https://www.amazon.jobs/en/jobs/SF210055414/so...	[location]	DESCRIPTIONvnDas Social Media Team in Regensbu...	SF210055414
2	Social Media Customer Service Associate (w/m/d...	https://www.amazon.jobs/en/jobs/SF210055413/so...	[location]	DESCRIPTIONvnDas Social Media Team in Regensbu...	SF210055413
3	Kundenservice im Homeoffice (m/w/d) – Vollzeit...	https://www.amazon.jobs/en/jobs/SF210055410/ku...	[location]	DESCRIPTIONvnKundenservicemitarbeiter*innen im...	SF210055410
4	Kundenservice (m/w/d) – Berlin – deutsprachige...	https://www.amazon.jobs/en/jobs/SF210054326/ku...	[location]	DESCRIPTIONvnKundenservicemitarbeiter*innen / ...	SF210054326

```
ds_items.to_csv('ds_items.csv')
```

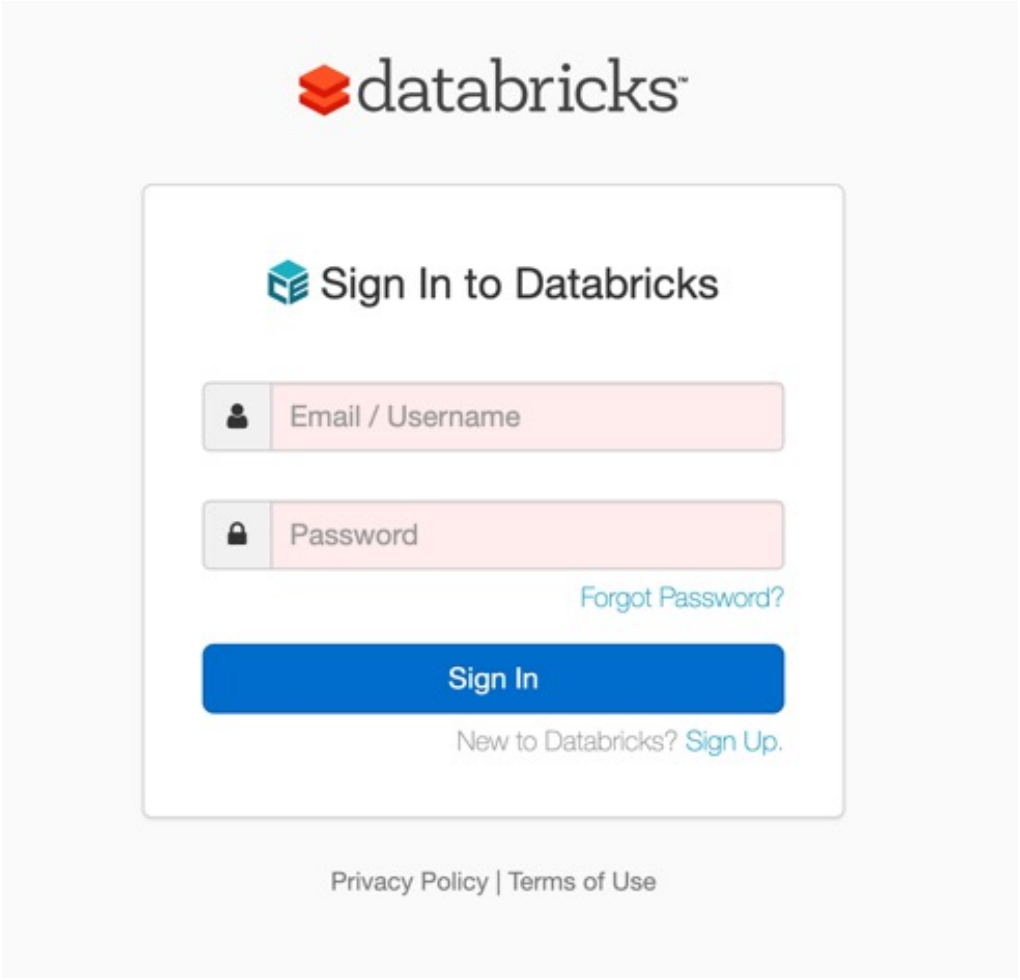
Темы

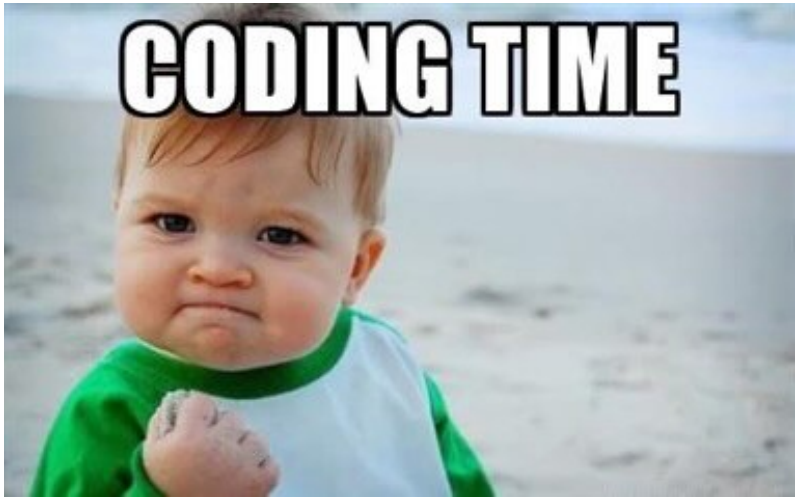
1. Что такое машинное обучение?
2. Databricks (введение)
3. Создание конвейеров обработки данных
    1. Как извлечь данные онлайн-вакансий при помощи скрейпинга и как построить наш конвейер обработки данных на основе Spark
    2. **Акцент на категоризации профессии**

Микросервисы по классификации

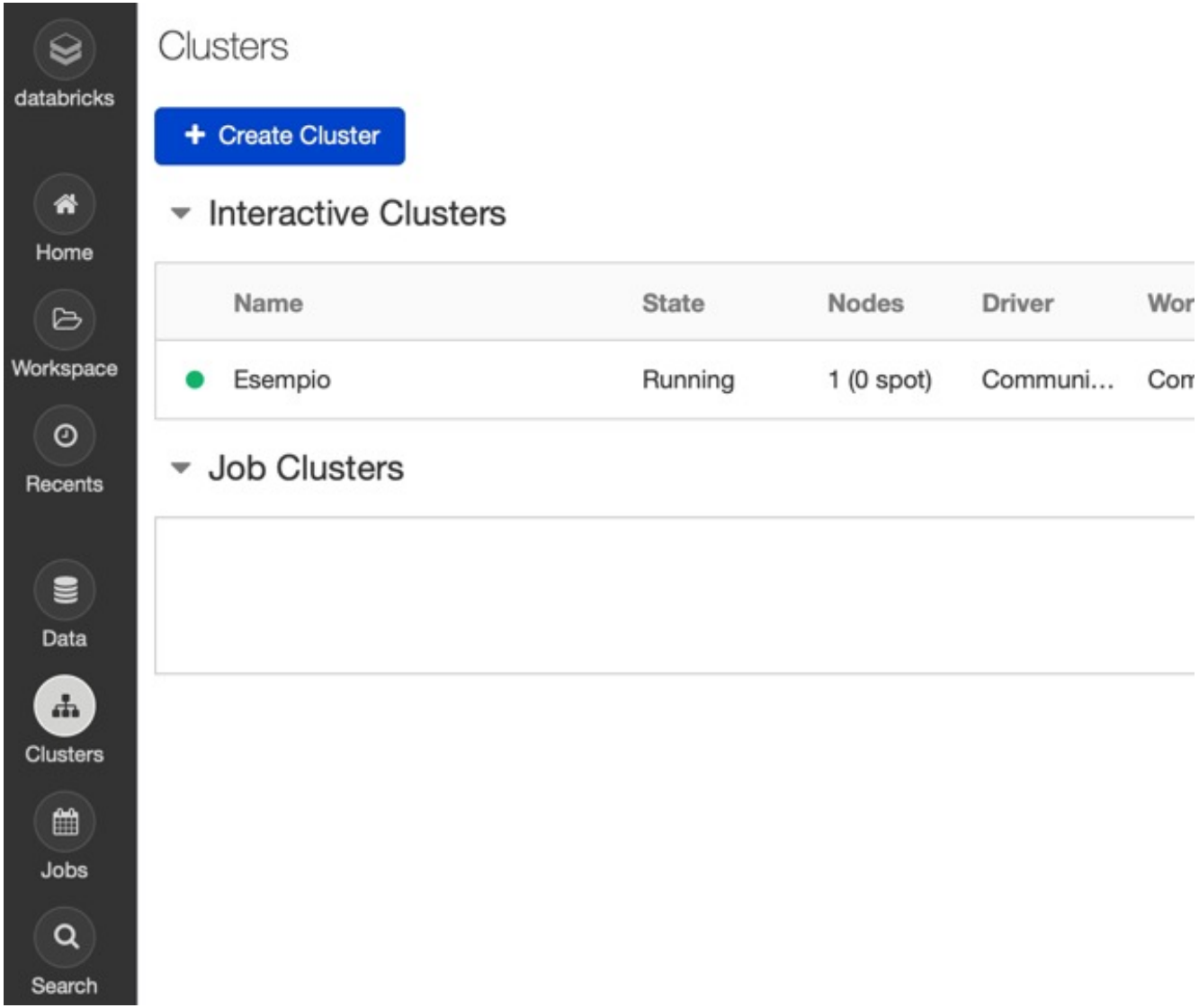


https://community.cloud.databricks.com/login.html

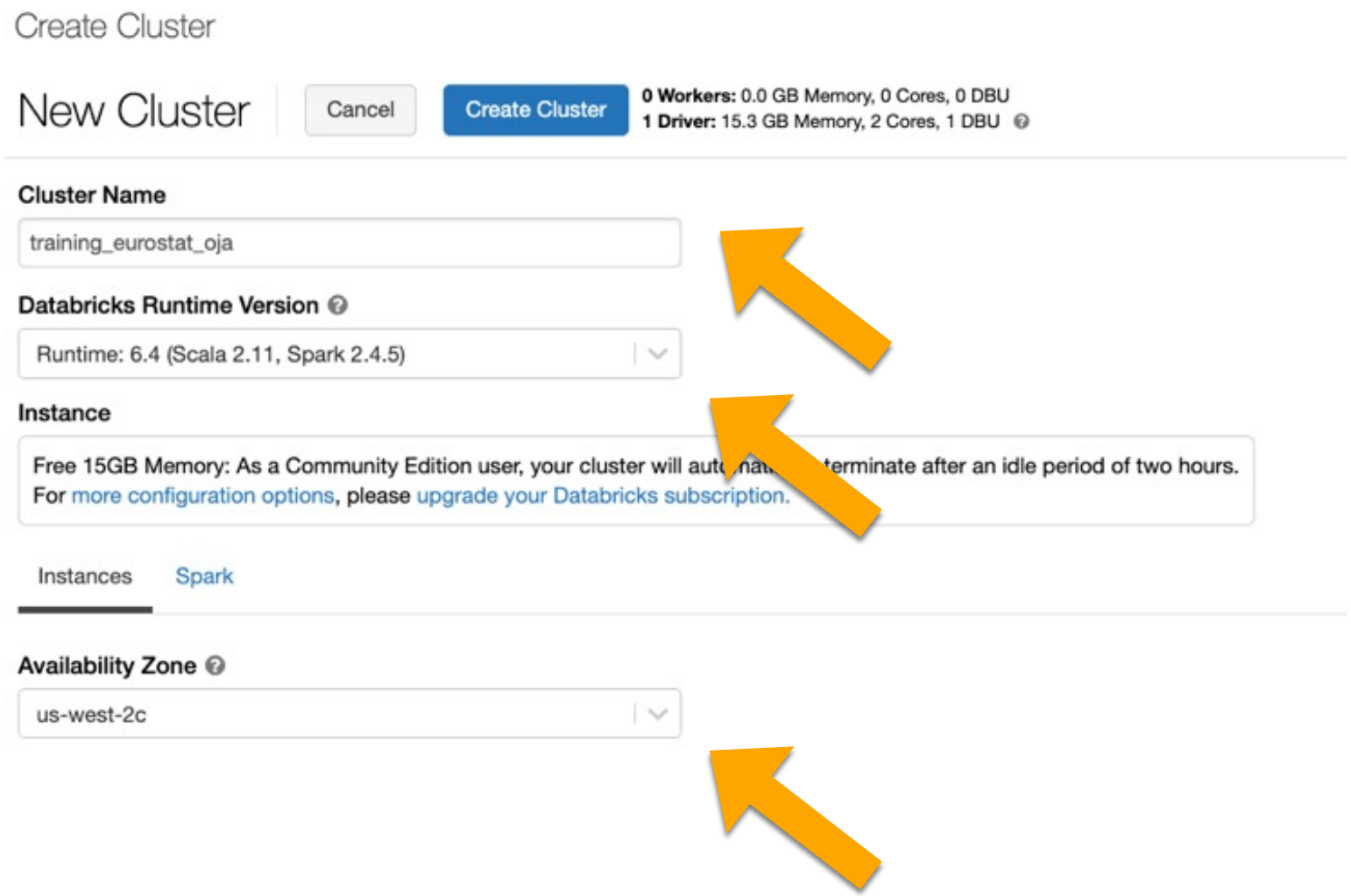




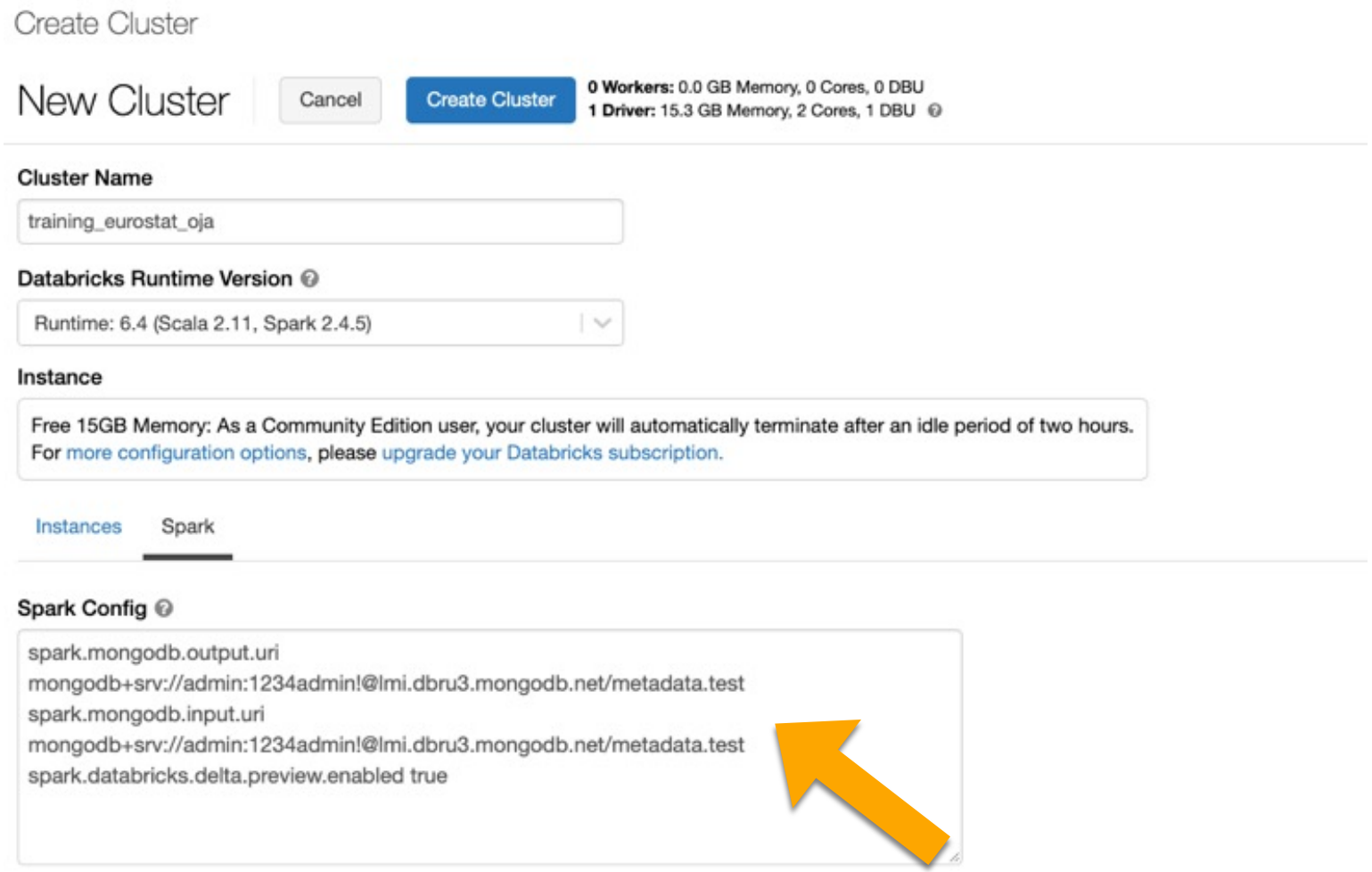
Создать новый кластер



Создать новый кластер



Создать новый кластер

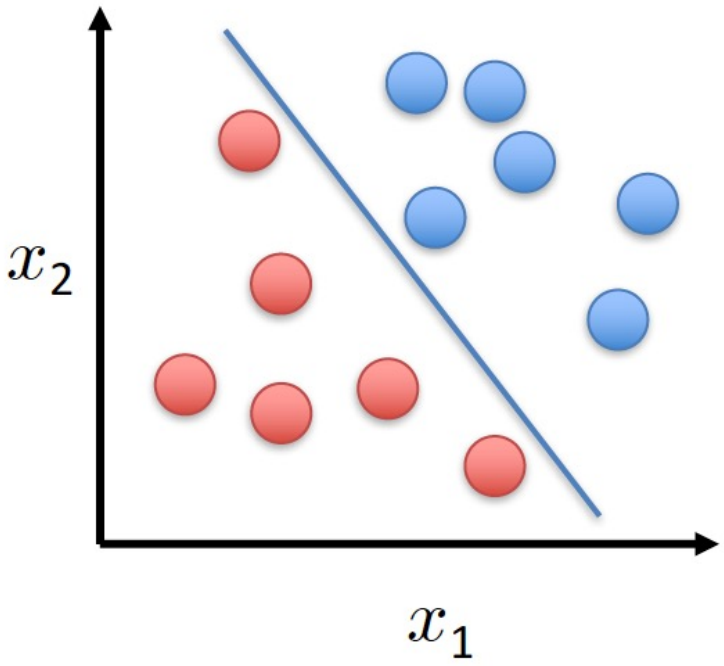


Цели

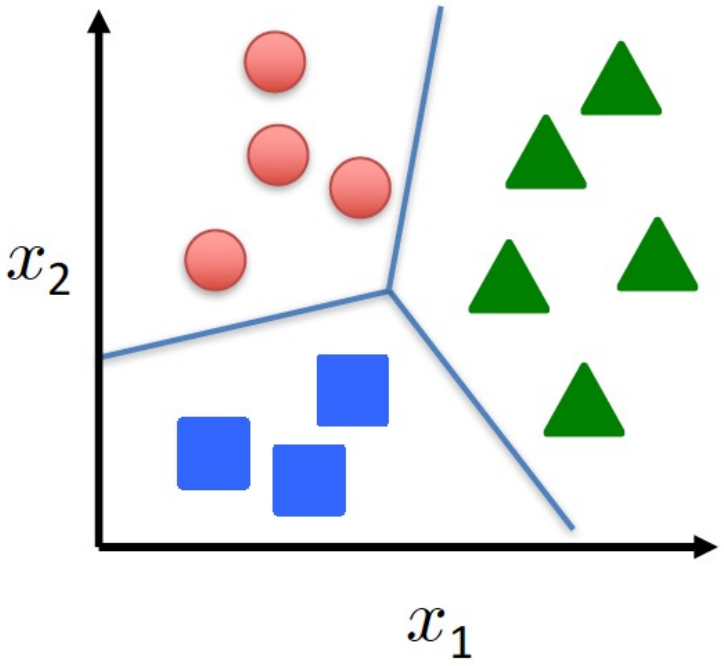
- Классифицировать профессии, исходя из заголовков объявлений
- Разработать общий подход, который будет работать для всех 25 официальных языков Европейского союза (а также дополнительных государственных языков)
- Сократить использование «золотых» наборов данных, чтобы минимизировать влияние человеческих ошибок и неоднозначностей
    - ~100 000 вручную размеченных наблюдений для каждого языка
- Разработать систему, которой будет легко управлять и будет возможно улучшить в случае неверной классификации: важность объяснимых результатов

Многоклассовая классификация

Двоичная классификация:

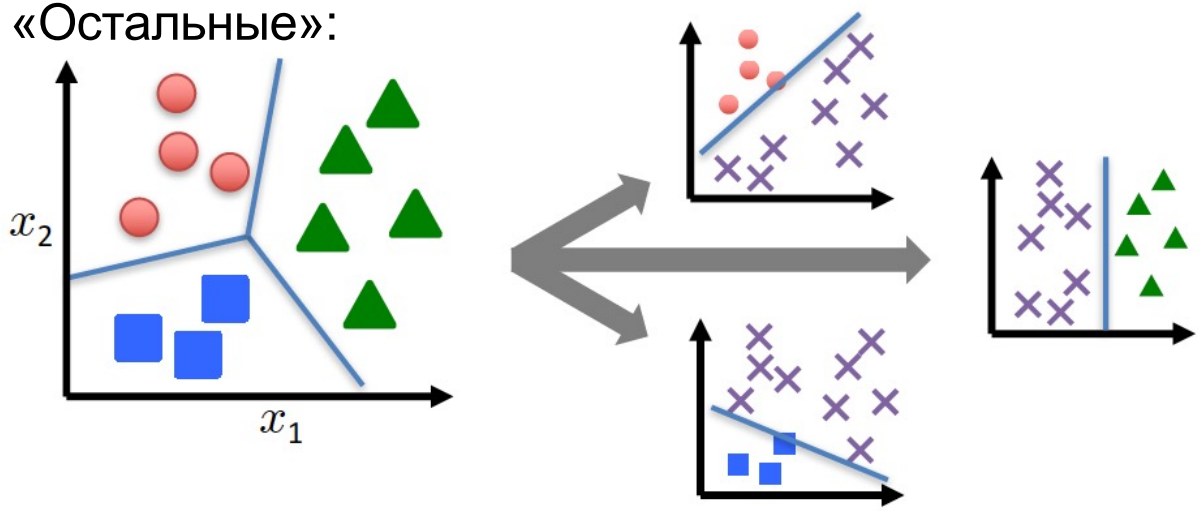


Многоклассовая классификация:



Многоклассовая логистическая регрессия

Разбивка на «Одни» и «Остальные»:



- Натренировать классификатор логистической регрессии для каждого класса *i*, чтобы спрогнозировать вероятность того, что *y* = *i* при помощи формулы:

$$h_c(\boldsymbol{x}) = \frac{\exp(\boldsymbol{\theta}_c^\mathsf{T}\boldsymbol{x})}{\sum_{c=1}^{C}\exp(\boldsymbol{\theta}_c^\mathsf{T}\boldsymbol{x})}$$

Выходной продукт модели Многоклассовая модель

[
1330 0,12
2511 0,11
2512 0,98
2513 0,97
....
Спрогнозировать вероятности
(для каждого класса)]

Класс работы: 2512 — Разработчик ПО
Иерархия работы: (251 — Профессия в ИТ, 25 —, ….)
Макс. вероятность: 0,98

Путь прогнозирования

{
Прогнозирование на основе текста: ложно
Прогнозирование на основе метаданных: верно
Тип модели Онтология / машинное обучение
Язык модели: EN
Расстояние (если применяется): Равенства / сходство Жаккара / Джаро — Винклера
Расстояние (показатели): 0..100
Путь онтологии: разработчик ПО
}

Набор данных

- 20 тыс. онлайн-вакансий
- 4 класса (5 тыс. для каждого класса)
  - Специалисты в области рекламы и маркетинга
  - Разработчики программного обеспечения
  - Математики, актуарии и статистики
  - Промышленные и производственные инженеры

Table: esco_en_dataset_csv

esco_en_dataset_csv

Refresh

test (clone)

Schema:

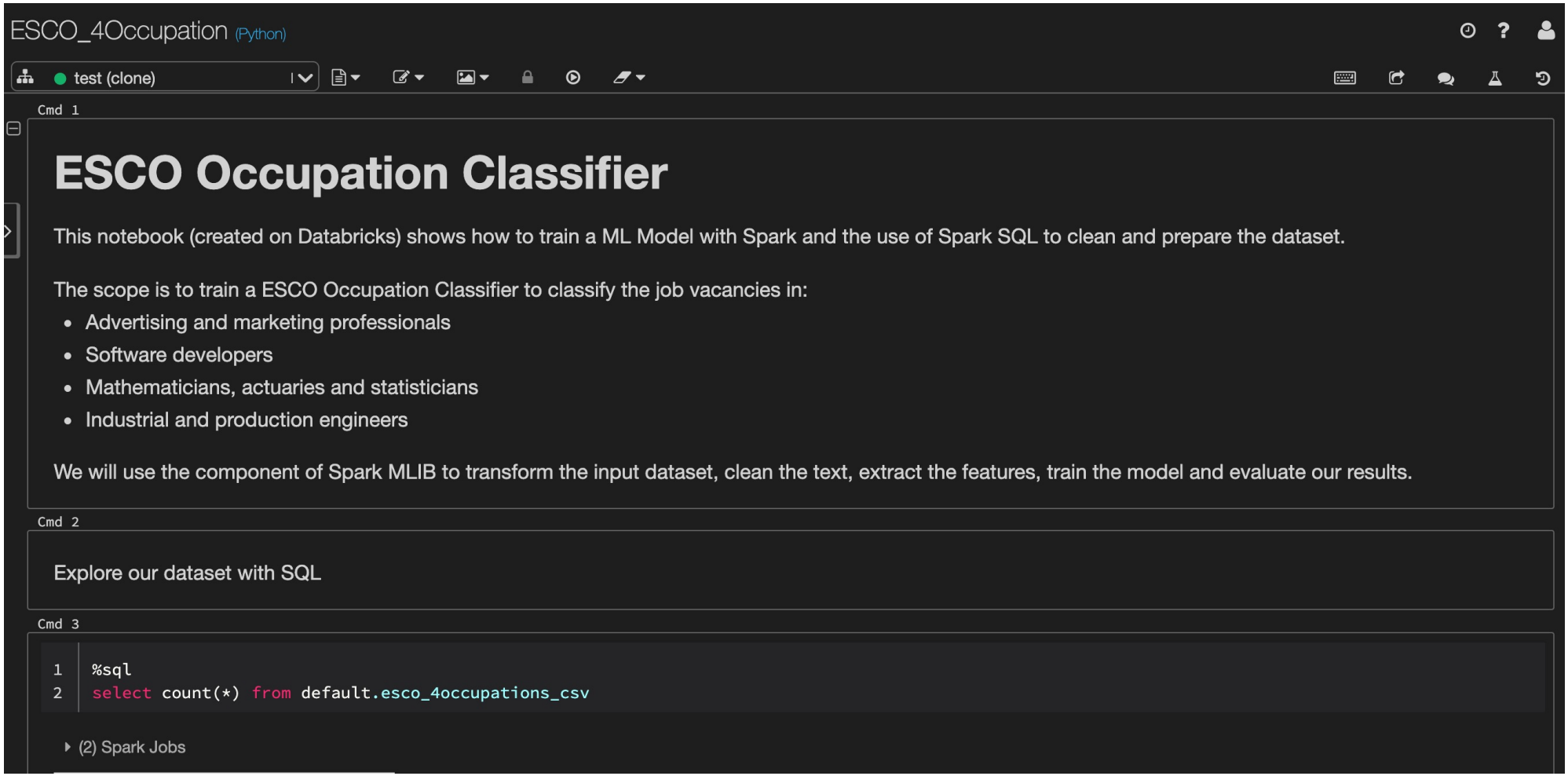
	col_name	data_type	comment
1	title	string	null
2	idesco_level_4	int	null
3	esco_level_4	string	null

Showing all 3 rows.

Sample Data:

	title	idesco_level_4	esco_level_4
1	B93-C04 Softwareentwickler C++ und C#/.NET (m/w)	2512	Software developers
2	Gezocht: Oracle Developer #Freelance #PandS #Jobs #Vacatures (Req:9096–Loc:Bxl)	2512	Software developers
3	Senior (GXP Process Excellence) Engineer	2512	Software developers
4	Software-Entwickler (m/w/d) Buildsystem / Integration	2512	Software developers
5	Business Intelligence Developer	2512	Software developers
6	Microsoft Dynamics NAV Functional Consultant	2512	Software developers

Showing all 20 rows.


ESCO_4Occupation (Python)
test (clone)
Cmd 1
ESCO Occupation Classifier
This notebook (created on Databricks) shows how to train a ML Model with Spark and the use of Spark SQL to clean and prepare the dataset.
The scope is to train a ESCO Occupation Classifier to classify the job vacancies in:
Advertising and marketing professionals
Software developers
Mathematicians, actuaries and statisticians
Industrial and production engineers
We will use the component of Spark MLIB to transform the input dataset, clean the text, extract the features, train the model and evaluate our results.
Cmd 2
Explore our dataset with SQL
Cmd 3
1 %sql
2 select count(*) from default.esco_4occupations_csv
(2) Spark Jobs

Резюме и ключевые слова

- Жизненный цикл проекта в области анализа и обработки данных
- Spark и SparkMlib
- Процесс интеллектуального анализа текста в Spark
- Как оценить модель?

Вопросы?