BRAIN SKILLS

ПРОГРАММА КУРСА

«АНАЛИЗ ДАННЫХ»

ЦЕЛЬ: знакомство с основными средствами и инструментами анализа данных

6	12 + 30	47
НЕДЕЛЬ	ЧАСОВ ТЕОРИИ	ЧАСОВ ПРАКТИКИ

ВЫПУСКНОЙ ПРОЕКТ: «Прогнозирование стоимости недвижимости на рынке -2019»

ЭКСПЕРТ: Басалаев Макар Николаевич

1 НЕДЕЛЯ Введение в Data Science. Основы Python

Введение. Цели и задачи программы. Машинное обучение и анализ данных. Возможности использования Python для решения задач машинного обучения и анализа данных. Основы языка программирования Python: типы данных, условия, циклы.

Вебинар №1 [2 часа] Основы Python: введение, условные конструкции, циклы.

Задача.	Наибольший общий делитель
Задача.	Стандартное отклонение
Статья.	Почему Python?
Видео.	Как выучить Python с 0 до Middle
Видео-лекция.	Установка интерпретатора и текстового редактора
Презентация.	Основные конструкции языка Python
Статья.	Типы данных в Python

Вебинар №2 [2 часа] Основы Python: строковый тип данных, списки, кортежи и словари.

Задача.	Уникальные элементы
Задача.	Частотный анализ текста
Видео-лекция.	Списки (list)
Видео-лекция.	Словари (dict), а также их методы
Презентация.	Работа с основными типами данных
Конспект.	Списки в Python
Конспект.	Словари в Python

ТЕСТ [по итогам недели]. *Основные элементы синтаксиса Python*

ПРАКТИКА: 6 часов ТЕОРИЯ: 5 часов ЗАДАЧИ: 4

2 НЕДЕЛЯ Вычисления: библиотеки Numpy и Scipy

Основы Python: продолжение. Функции. Работа с файлами. Основы использования среды Jupyter. Библиотеки Numpy и Scipy: вопросы их применения.

Вебинар №3 [2 часа]	Основы Python: функции и работа с файлами.	
	Задача.	Обработка и мониторинг каталогов файлов
	Видео-лекция.	Функции в Python
	Презентация.	Python: функции, файлы
	Статья.	Работа с файлами
	Конспект.	Функции Python
	Конспект.	Работа с текстовыми файлами
Вебинар №4 [2 часа]	Jupyter. Библиотеки Numpy и Scipy.	
	Задача.	Сравнение скорости работы массивов Numpy и списков Python
	Задача.	Оптимизация функции
	Jupyter-ноутбук.	Модули Numpy и Scipy в Python
	Статья.	Преимущества Numpy
	Статья.	Python Numpy Tutorial
	Видео.	Learn NUMPY in 5 minutes!
	Видео.	Improving Python programs with NumPy and SciPy

ТЕСТ. *Структура программы на Python. Первые библиотеки.*

ПРАКТИКА: 6 часов ТЕОРИЯ: 6 часов ЗАДАЧИ: 3

3 НЕДЕЛЯ Обработка и анализ данных с Pandas. Визуализация данных

Основные элементы библиотеки Pandas. Методы визуализации данных Matplotlib. Примеры хороших визуализаций. Диаграммы, графики.

Вебинар №5 [2 часа] Библиотека Pandas: введение.

Задача.	Извлечение информации из таблиц
Jupyter-ноутбук.	Модуль Pandas
Статья.	Первичный анализ данных с Pandas
Видео.	Python: Pandas Tutorial
Книга.	Изучаем Pandas
Статья.	Анализ данных с Pandas

Вебинар №6 [2 часа] Визуализация данных с Python.

Задача.	Построение графиков разных типов
Jupyter-ноутбук.	Визуализация данных: модули matplotlib, seaborn
Статья.	Визуализация данных с Python
Видео.	Intro to Visualization with Python
Видео.	Data Visualization and Exploration with Python
Статья.	Intro to Data Visualization with Matplotlib
Статья.	Примеры визуализаций Python

ТЕСТ. *Предварительный анализ данных.*

ПРАКТИКА: 8 часов ТЕОРИЯ: 7 часов ЗАДАЧИ: 2

4 НЕДЕЛЯ Библиотека Pandas. Web-scraping

Работа с данными различных частично-структурированных форматов: .csv, .xls, .xml, .json. Извлечение данных из веб-страниц.

Вебинар №7 [2 часа]	Библиотека Pandas. Работа с данными различных форматов: .csv, .xls, .xml, .json	
	Задача.	Извлечение информации
	Jupyter-ноутбук.	Форматы .csv, .xls, .xml, .json и Pandas
	Статья.	Чтение данных из .csv
	Статья.	Чтение данных из .json
	Статья.	Чтение данных из .xml
	Видео.	Python Pandas: Excel and CSV file formats
	Видео.	Pandas: JSON file format
Вебинар №8 [2 часа]	Извлечение данных с веб-страниц (скрапинг)	
	Задача.	Извлечение данных с Википедии
	Jupyter-ноутбук.	Web-scraping
	Статья.	Веб-скрапинг — что это и как он работает
	Статья.	Введение в web-scraping
	Статья.	Web-scraping с помощью Python
	Видео-лекция.	Intro to Web Scraping with Python and Beautiful Soup
	Видео.	Добываем данные из интернета

ТЕСТ. *Структура HTML, сбор данных.*

ПРАКТИКА: 9 часов ТЕОРИЯ: 6 часов ЗАДАЧИ: 2

5 НЕДЕЛЯ Основы машинного обучения

Введение в машинное обучение. Алгоритм k - ближайших соседей (k-NN). Линейные модели машинного обучения. Ансамблевые модели (деревья решений, случайный лес, бустинг).

Вебинар №9 [2 часа]

Введение в машинное обучение. Линейные модели. Алгоритм k - ближайших соседей.

Задача.	Реализация линейной модели
Задача.	Реализация алгоритма k-NN
Jupyter-ноутбук.	Введение в машинное обучение
Видео-лекция.	How k-NN algorithm works?
Видео.	Linear and Polynomial Regression in Python
Статья.	Анализ данных — основы и терминология
Статья.	Линейная регрессия

Вебинар №10 [2 часа]

Введение в машинное обучение. Ансамблевые модели

Задача.	Предсказание выживания на Титанике
Jupyter-ноутбук.	Ансамблевые модели машинного обучения
Статья.	Решающие деревья
Видео-лекция.	Классификация. Деревья решений
Видео.	Decision Tree: how it works
Видео.	Random Forest Algorithm

ТЕСТ. *Алгоритмы машинного обучения*

ПРАКТИКА: 9 часов ТЕОРИЯ: 9 часов ЗАДАЧИ: 3

6 НЕДЕЛЯ Нейронные сети. Глубокое обучение

Модель нейронной сети. Глубокое обучение. Метод обратного распространения ошибки. Виды нейронов и слоёв. Свёрточные и рекуррентные нейронные сети.

Вебинар №11 [2 часа]	Введение в нейронные сети	
	Задача.	Обучение простейшей нейронной сети
	Jupyter-ноутбук.	Глубокое обучение. Введение
	***Статья*.**	Нейронные сети для начинающих
	Статья.	Создание нейронных сетей
	Статья.	How to Build a Neural Network
	Видео.	Нейронные сети за 30 минут: от теории до практики
	Видео.	What is a Neural Network?
Вебинар №12 [2 часа]	Введение в нейронные сети: виды нейронов и сетей	
	Задача.	Распознавание изображений датасета CIFAR10
	Jupyter-ноутбук.	Виды слоев и нейронных сетей в глубоком обучении
	Статья.	Свёрточные нейронные сети
	Статья.	Рекуррентные сети
	Видео.	CNN
	Видео-лекция.	Слои глубоких свёрточных сетей
	Видео.	Введение в RNN

ТЕСТ. *Модели глубокого обучения*

ПРАКТИКА: 9 часов ТЕОРИЯ: 9 часов ЗАДАЧИ: 2

ВЫПУСКНОЙ ПРОЕКТ

«Прогнозирование стоимости недвижимости на рынке -2019»

- выполняется в течении 6 недель
- используются полученные знания
- современные инструменты

ЭТАП 1. СБОР

Автоматический сбор информации с известного сайта продажи недвижимости.

Например, для домов извлекается местоположение, площадь, цена и т.д.

Средства: Requests и BeautifulSoup.

ЭТАП 2. ОБРАБОТКА ДАННЫХ

Выполнение разведочного анализа, предварительной и пост-обработки данных.

Средства: Pandas.

ЭТАП 3. НАСТРОЙКА МОДЕЛИ И ОЦЕНКА КАЧЕСТВА

Обучение модели прогнозирования цены дома по его различным характеристикам.

Оценка качества модели. В качестве модели используются либо традиционные алгоритмы машинного обучения или нейронные сети.

Средства: Scikit-learn, Pytorch.

РЕЗУЛЬТАТЫ ОБУЧЕНИЯ

ВЫ НАУЧИТЕСЬ

- Разрабатывать программы на Python в среде Jupyter
- Применять алгоритмы машинного обучения к задачам регрессии и классификации с использованием библиотеки Scikit-learn
- Создавать и обучать нейронные сети с Pytorch
- Извлекать данные из веб-страниц
- Представлять данные в удобном виде с помощью таблиц и графиков
- Оценивать качество алгоритмов и проводить эксперименты

ВЫ ПОЛУЧИТЕ

- Материалы курса: статьи, видео, конспекты, лекции, ссылки на полезные сервисы
- Готовые процедуры анализа данных в виде ноутбуков Jupyter
- Опыт обработки больших массивов данных

ОТВЕТЫ НА ВОПРОСЫ

ЧТО НУЖНО ЗНАТЬ, ЧТОБЫ УСПЕШНО ПРОЙТИ КУРС?

Необходимо быть уверенным пользователем персонального компьютера: уметь устанавливать программы и библиотеки, самостоятельно искать информацию в интернете. Обладать логическим мышлением и способностью делать выводы и приходить к правильным умозаключениям, используя имеющуюся информацию. Знать математику на школьном уровне и понимать математические понятия: функция, минимизация, производная, вектор, матрица, понимать математические обозначения.

Сильно проще будет тем, кто знает университетскую базовую математику (математический анализ, линейную алгебру, оптимизацию, теорию вероятностей и статистику), тем не менее все используемые понятия будут объяснены.

ЧТО ТАКОЕ АНАЛИЗ ДАННЫХ?

Это область между математикой и информатикой, которая занимается построением и исследованием различных методов извлечения полезной информации из данных.

ЗАЧЕМ ЭТО НУЖНО?

Стек возможностей применения анализа данных огромен: в медицине для диагностики, в беспилотных автомобилях для создания алгоритмов движения, задачи машинного перевода, извлечения сущностей из текста, построения рекомендательных систем в интернете для рекламы, в банках для решения о выдаче кредита и многое-многое другое. Разве не хватает аналитических алгоритмов для решения этих задач? Существуют задачи, в которых зависимости между входными и выходными данными нельзя описать какой-либо аналитической функцией. Тут помогает машинное обучение.

КАКИМИ НАВЫКАМИ НУЖНО ОБЛАДАТЬ, ЧТОБЫ РАБОТАТЬ В ЭТОЙ СФЕРЕ?

Нужно знать программирование, уметь анализировать данные и интерпретировать результаты. Знать существующие технологии и возможности их применения.

Программа обновлена 28.03.2019

© BrainSkills, Басалаев М.Н.