



Стандартизация и очистка НСИ

Программа электронного курса

Модулей: 2

Стандартизация и очистка НСИ

Тип: лекции

Лекция 1. Общие сведения

Введение: нормативно-справочная информация (НСИ), объемы данных, понятие качества данных. Причины загрязнения данных: ручной ввод данных, не заполнение/некорректное заполнение поля, потеря информации, отсутствие ограничений на вводимые значения, отсутствие стандартов. Виды ошибок в данных: опечатки, пропуски, противоречия, дубликаты, несоответствие форматов, неправдоподобие данных. Понятие очистки данных. Понятие стандартизации данных. Причины необходимости очистки. Показатели качества данных.

Лекция 2. Методы очистки данных

Основные методы очистки. Использование словарей и таблиц замены: виды словарей, способы использования, понятие классификатора. Общероссийские классификаторы. Классификаторы адресов: КЛАДР, ФИАС. Анализ строк: расстояние Левенштейна, расстояние Дамерау-Левенштейна. Регулярные выражения: понятие, решаемые задачи. Частотный анализ: частота появления значений, комбинации значений. Контрольные числа: возможности использования, проверка числа, восстановление числа, алгоритм формирования контрольного числа СНИЛС.

Лекция 3. Основы регулярных выражений

Понятие регулярного выражения. Понятие метасимволов. История появления регулярных выражений: истоки и использование. Диалекты регулярных выражений. Регулярные выражения Perl. Регулярные выражения в стиле Perl. Возможности регулярных выражений: поиск по шаблону, проверка соответствия формату, извлечение подстроки, поиск с заменой. Инструменты для проверки регулярных выражений.

Лекция 4. Синтаксис регулярных выражений Perl

Понятие метасимволов. Основные метасимволы. Экранирование символов. Диапазоны символов. Метпоследовательности. Мнимые метасимволы. Понятие квантификаторов. Основные квантификаторы. «Жадные» квантификаторы. Дополнительные опции. Поиск любого числа. Поиск числа определенного размера: ограничение размера, использование символов-разделителей, использование символа границы слова.

Лекция 5. Объединение методов

Применимость методов. Использование одного метода. Комбинирование методов. Общий алгоритм очистки данных.

Самопроверочный тренажер

16 вопросов с выбором вариантов ответа.

Очистка и стандартизация информации в Loginom

Тип: пошаговая симуляция

Занятие 1. Регулярные выражения

Компонент **Калькулятор**. Функции по работе с регулярными выражениями. Описание и примеры использования. Прочие функции компонента, используемые при очистке и стандартизации.

Занятие 2. Очистка регионов

Решение задачи по очистке и стандартизации списка регионов. Описание наборов данных. Проведение очистки в четыре этапа: поиск по коду региона, сверка со справочником, использование регулярных выражений, расчет расстояния Дамерау-Левенштейна. Оценка результатов.

Контрольное задание

Две задачи (вписать вариант ответа)

Темы: очистка и стандартизация списка организаций. Очистка записей с номерами телефонов.