Работа с разными форматами данных

Николай Свиридов

Николай Свиридов

О спикере:

- Backend-разработчик, IT-блогер

Проверка связи

Отправьте, пожалуйста, смайлик в чат

Если у вас все отлично со связью **:-)**
Если у вас есть какие-то проблемы со связью **:-(**

Поставьте **“+”**, если меня видно и слышно

Если у вас нет звука:

- убедитесь, что на вашем устройстве и на колонках включен звук
- обновите страницу вебинара (или закройте страницу и заново присоединитесь к вебинару)
- откройте вебинар в другом браузере
- перезагрузите компьютер (ноутбук) и заново попытайтесь зайти

План занятия

1. Введение
2. Формат CSV
3. Формат JSON
4. Формат YAML
5. Формат XML
6. Проблема кодировок

ВВЕДЕНИЕ

CSV	JSON	YAML	XML
плоский	дерево	дерево	дерево

Сериализация

Сериализация — процесс преобразования объекта в поток байтов для сохранения или передачи в память, базу данных или файл.

Предназначена для того, чтобы сохранить состояния объекта для последующего воссоздания при необходимости.

Обратный процесс называется **десериализацией**.

ФОРМАТ CSV

Формат CSV

```
1   title,link,description,pubDate
2   Как не править Python тесты,https://habr.com/ru/post/502278/?utm_campaign=502278&
    utm_source=habrahabr&utm_medium=rss,"<p>И вынести тестируемые результаты вне кода. Это
    статья об автоматизации и увеличения удобства тестирования на Python.</p><p>У меня был
    проект, который разрабатывался уже несколько лет. В проекте отсутствовали тесты. А также
    у него были активные зависимости от других команд, которые также влияли на
    результат.</p><p>Регрессионное тестирование было одним из шагов для более уверенной
    разработки. Его суть в сравнении вычисленных данных с последним канонизированным
    результатом работы программы.</p><p>Результаты выполнения можно проверять в python коде
    тестов. Это близко к контексту выполнения и зачастую удобно.</p>","Wed, 20 May 2020
    11:25:18 +0300"
3   Как построить диаграмму Венна с 50 кругами? Визуализация множеств и история моего
    Python-проекта с открытым кодом,https://habr.com/ru/post/501924/?utm_campaign=501924&
    utm_source=habrahabr&utm_medium=rss,"<p>Сегодня хочу рассказать вам про задачу
```

	title	link	description	pubDate
2	Как не править Python тесты	https://habr.com/ru/post/502278	<p>И вынести тестируемые результ	20.05.2020 11:25
3	Как построить диаграмму Венна с 50 кругами? Визуализация мно	https://habr.com/ru/post/501924	<p>Сегодня хочу рассказать вам п	20.05.2020 16:13
4	jupyter-book - делаем интерактивную книгу из Jupyter Notebooks	http://github.com/executablebook		19.05.2020 11:14
5	Исключаем дефекты с изображения с помощью OpenCV	https://www.pyimagesearch.com/		20.05.2020 1:38

Формат CSV

Основное применение — **выгрузки данных**

- Для хранения больших объемов (до нескольких Гб) единообразных данных
- Самый компактный формат из всех
- Самый популярный формат для обмена данными у аналитиков
- Поддерживается MS Excel (эквивалентен плоской таблице)
- Не подходит для иерархических данных

https://docs.python.org/3/library/csv.html

ФОРМАТ JSON

Формат JSON

```
{
  "channel": {
    "title": "Дайджест новостей о python",
    "link": "https://pythondigest.ru/",
    "description": "Рускоязычные анонсы свежих новостей о python и близлежащих технологиях.",
    "lastBuildDate": "Wed, 20 May 2020 16:13:18 +0300",
    "items": [
      {
        "title": "Как не править Python тесты",
        "link": "https://habr.com/ru/post/502278/?utm_campaign=502278&utm_source=habrahabr&
        utm_medium=rss",
        "description": "<p>И вынести тестируемые результаты вне кода. Это статья об автоматизации и
        увеличения удобства тестирования на Python.</p>",
        "pubDate": "Wed, 20 May 2020 11:25:18 +0300"
      }
    ]
  }
}
```

Формат JSON

Основное применение — **базы данных, выгрузки данных**

- Для импорта/экспорта данных в базы данных (в т.ч. bson)
- Для сохранения вложенных структур данных
- При передаче данных клиент <-> сервер для сериализации иерархических объектов
- Самый популярный и простой в использовании формат для Python и Java программистов
- Является подмножеством формата YAML

https://docs.python.org/3/library/json.html документация JSON
https://jsoneditoronline.org/ онлайн редактор JSON

ФОРМАТ YAML

Формат YAML

```yaml
channel:
  description: Рускоязычные анонсы свежих новостей о python и близлежащих технологиях.
  items:
  - description: '<p>И вынести тестируемые результаты вне кода. Это статья об автоматизации
      и увеличения удобства тестирования на Python.</p>'
    link: https://habr.com/ru/post/502278/?utm_campaign=502278&utm_source=habrahabr&
    utm_medium=rss
    pubDate: Wed, 20 May 2020 11:25:18 +0300
    title: Как не править Python тесты
  lastBuildDate: Wed, 20 May 2020 16:13:18 +0300
  link: https://pythondigest.ru/
  title: Дайджест новостей о python
```

Формат YAML

Основное применение — **файлы конфигурации**

- Самый компактный язык разметки
- Для создания файлов настроек
- Для описания классов, ресурсов и манифестов в API

https://yaml.org/
https://github.com/yaml/pyyaml

ФОРМАТ XML

XML vs JSON

```xml
<?xml version="1.0" ?>
<rss xmlns:ns0="http://www.w3.org/2005/Atom" version="2.0">
  <channel>
    <title>Дайджест новостей о python</title>
    <link>https://pythondigest.ru/</link>
    <description>Рускоязычные анонсы свежих новостей о python и близлежащих технологиях.</description>
    <ns0:link href="https://pythondigest.ru/rss/" rel="self"/>
    <language>ru-ru</language>
    <lastBuildDate>Wed, 20 May 2020 16:13:18 +0300</lastBuildDate>
    <item>
      <title>Как не править Python тесты</title>
      <link>https://habr.com/ru/post/502278/?utm_campaign=502278&amp;utm_source=habrahabr&amp;utm_medium=rss</link>
      <description>&lt;p&gt;И вынести тестируемые результаты вне кода. Это статья об автоматизации и увеличения удобства тестирования на Python.&lt;/p&gt;</description>
      <pubDate>Wed, 20 May 2020 11:25:18 +0300</pubDate>
      <guid>https://habr.com/ru/post/502278/?utm_campaign=502278&amp;utm_source=habrahabr&amp;utm_medium=rss</guid>
    </item>
    <item>
      <title>Как построить диаграмму Венна с 50 кругами? Визуализация множеств и история моего Python-проекта с открытым кодом</title>
      <link>https://habr.com/ru/post/501924/?utm_campaign=501924&amp;utm_source=habrahabr&amp;utm_medium=rss</link>
      <description>&lt;p&gt;Сегодня хочу рассказать вам про задачу визуализации пересекающихся множеств. Поехали!&lt;/p&gt;</description>
      <pubDate>Wed, 20 May 2020 16:13:18 +0300</pubDate>
      <guid>https://habr.com/ru/post/501924/?utm_campaign=501924&amp;utm_source=habrahabr&amp;utm_medium=rss</guid>
    </item>
```

```json
{
  "channel": {
    "title": "Дайджест новостей о python",
    "link": "https://pythondigest.ru/",
    "description": "Рускоязычные анонсы свежих новостей о ру
    технологиях.",
    "lastBuildDate": "Wed, 20 May 2020 16:13:18 +0300",
    "items": [
      {
        "title": "Как не править Python тесты",
        "link": "https://habr.com/ru/post/502278/?utm_campai
        utm_source=habrahabr&utm_medium=rss",
        "description": "<p>И вынести тестируемые результаты
        статья об автоматизации и увеличения удобства тестир
        Python.</p>",
        "pubDate": "Wed, 20 May 2020 11:25:18 +0300"
      },
      {
        "title": "Как построить диаграмму Венна с 50 кругами
        множеств и история моего Python-проекта с открытым к
        "link": "https://habr.com/ru/post/501924/?utm_campai
        utm_source=habrahabr&utm_medium=rss",
        "description": "<p>Сегодня хочу рассказать вам про з
        пересекающихся множеств. Поехали!</p>",
        "pubDate": "Wed, 20 May 2020 16:13:18 +0300"
      },
```

Формат XML

Основное применение — **сериализация объектов любой сложности**

- Применяется при передаче данных клиент<->сервер для сериализации объектов
- Стандарт обмена данными и сообщениями большинства информационных систем
- Для создания файлов конфигурации

https://docs.python.org/3/library/xml.etree.elementtree.html

Формат XML

XPath

Поиск одного элемента: `root.find(query)`

Поиск нескольких элементов: `root.findall(query)`

`query = XPath`

https://www.w3schools.com/xml/xpath_syntax.asp синтаксис XPath

Проблема кодировок

Кодировки: проблемы с кириллицей

Один и тот же текст в кодировке **Windows 1251** и **utf-8**

Windows 1251 (один байт на букву)

```
00000000: C8 F1 F2 EE F0 E8 FF 20|F3 20 EC E5 ED FF 20 F1  | История у меня с
00000010: EB E5 E4 F3 FE F9 E0 FF|3A 20 EF EE E7 ED E0 EA  | ледующая: познак
00000020: EE EC E8 EB F1 FF 20 F1|20 E4 E5 E2 F3 F8 EA EE  | омился с девушко
00000030: E9 20 E8 E7 20 F1 E2 EE|E5 E3 EE 20 E8 ED F1 F2  | й из своего инст
00000040: E8 F2 F3 F2 E0 2E 0D 0A|D1 ED E0 F7 E0 EB E0 20  | итута...Сначала
```

utf-8 (два байта на букву)

```
00000000: EF BB BF D0 98 D1 81 D1|82 D0 BE D1 80 D0 B8 D1  | п»ї.СЃС‚РѕСЂРёС
00000010: 8F 20 D1 83 20 D0 BC D0|B5 D0 BD D1 8F 20 D1 81  | Џ СѓРјРµРSСЏ СЃ
00000020: D0 BB D0 B5 D0 B4 D1 83|D1 8E D1 89 D0 B0 D1 8F  | Р»РµРґСѓСЋС‰Р°СЏ
00000030: 3A 20 D0 BF D0 BE D0 B7|D0 BD D0 B0 D0 BA D0 BE  | : РїРѕР·РЅР°РєРѕ
00000040: D0 BC D0 B8 D0 BB D1 81|D1 8F 20 D1 81 20 D0 B4  | РјРёР»СЃСЏ СЃ Рґ
```

Домашнее задание

Пройдите квиз по теме «Работа с разными форматами данных»

Ваши вопросы?