



**МІНІСТЕРСТВО ОСВІТИ, НАУКИ, МОЛОДІ ТА СПОРТУ УКРАЇНИ
НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ»
ФІЗИКО-ТЕХНІЧНИЙ ІНСТИТУТ**

Лабораторна робота №2

Аналіз даних

Підготував:

студент 4 курсу

групи ФІ-84

Коломієць Андрій Юрійович

E-mail: *andrew.kolomiets.work@gmail.com*

Київ – 2021

Лабораторна робота №2

Аналіз даних

Завдання на самостійну роботу

1. Написати програму формування пакетного файлу в форматі **JSON**.
2. Встановити на комп'ютері бібліотеку для роботи із регулярними виразами у середовищі мови програмування (**Python**).
3. Ознайомитися із основними можливостями мови програмування **Python** щодо роботи із строковими даними.

Виконання завдань

Інсталювання необхідних бібліотек

```
$pip3 install re  
$pip3 install datetime
```

Програма конвертації даних із формату RSS до формату JSON

```
#Підключення модулів для роботи із регулярними виразами і часом  
  
import re  
import datetime  
  
#Відкриття файлу rss.xml у режимі «читання»  
  
f = open("rss.xml", "r")  
  
#Зчитування вмісту файлу rss.xml у змінну t  
  
t = f.read()  
  
#Закриття файлу rss.xml  
  
f.close()  
  
#Розбиття файлу по рядкам і склеювання рядків через пропуск  
  
rss = t.split('\n')  
  
t=""  
  
for i in range(len(rss)):  
    t=t+" "+rss[i]
```

```

#Видалення перших пропусків

t=re.sub('^\s','',t)

#Формування масиву заголовків

title = re.findall('<title>(.*?)</title>', t)

#Перший заголовок - назва фіду, далі - його специфічна обробка

source=title[0]
source=re.sub('[\s\-\]*$','',source)
source=re.sub("'",'',source)

#Розмірність масиву заголовків

x=range(len(title))

#Формування масиву описів

text = re.findall('<description>(.*?)</description>', t)

#Формування масиву гіперпосилань

link = re.findall('<link>(.*?)</link>', t)

#Формування дати і часу в форматі "YY-MM-MMTTHH:MM:00Z"

now = datetime.datetime.now()

tim=now.strftime("%Y-%m-%dT%H:%M:00Z")

#Виведення результатів

for i in range(1, len(title)):
    print ("{"title\":"+"title[i]+\","")

    #Специфічна обробка тексту

    text[i]=re.sub('[\s\-\]*$','',text[i])
    text[i]=re.sub("'",'',text[i])
    text[i]=re.sub('\','&',text[i])

    #Подальша виведення результатів

    print ("textBody\":"+"text[i]+\","")
    print ("source\":"+"source+"\","")
    print ("PubDate\":"+"tim+"\","")
    print ("URL\":"+"link[i]+\n")

if i<len(title)-1:
    print (",")

```

Запуск програми

```
linux@asus-X505BP:~/Стіленьця/Big Data/Lab-2$ python3 RSS_to_JSON.py > result.json
```

Результати

RSS

```
<item>
  <title>FAUG launches on Play Store worldwide</title>
  <description><![CDATA[
    href="https://economictimes.indiatimes.com/tech/software/faug-launches-on-play-store-worldwide/slideshow/80747391.cms">
    width="100" height="75" border="0" hspace="10" align="left"
    src="https://img.etimg.com/photo/80747295.cms"
    />
    </description>
  </item>
  <link>
    https://economictimes.indiatimes.com/tech/software/faug-launches-on-play-store-worldwide/slideshow/80747391.cms</link>
  <image>
    https://img.etimg.com/thumb/width=1200,ingsize=58833,resizemode=4,msid=80747295/faug-launches-on-play-store-worldwide.jpg</image>
  <guid>Article at EconomicTimes.com with article id :
    80747295</guid>
  <pubDate>2021-02-08T13:57:35+05:30</pubDate>
</item>
<item>
  <title>Covid-19: Major IT companies to extend WFH till March 2021</title>
  <description><![CDATA[
    href="https://economictimes.indiatimes.com/tech/software/covid-19-major-it-companies-to-extend-wfh-till-march-2021/videoshow/79640742.cms">
    width="100" height="75" border="0" hspace="10" align="left"
    src="https://img.etimg.com/photo/79640742.cms"
    />
    </description>
  </item>
  <link>
    https://economictimes.indiatimes.com/tech/software/covid-19-major-it-companies-to-extend-wfh-till-march-2021/videoshow/79640742.cms</link>
  <image>
    https://img.etimg.com/thumb/width=1200,ingsize=83302,resizemode=4,msid=79640742/covid-19-major-it-companies-to-extend-wfh-till-march-2021.jpg</image>
  <guid>Article at EconomicTimes.com with article id :
    79640742</guid>
  <pubDate>2020-12-09T13:49:11+05:30</pubDate>
</item>
<item>
  <title>Persistent to acquire US-based Caplot Software for $6.34 million</title>
  <description><![CDATA[
    href="https://economictimes.indiatimes.com/tech/software/persistent-to-acquire-us-based-caplot-software-for-6-34-million/articleshow/78681554.cms">
    width="100" height="75" border="0" hspace="10" align="left"
    src="https://img.etimg.com/photo/78681554.cms"
    />
    </description>
  </item>
  <link>
    https://economictimes.indiatimes.com/tech/software/persistent-to-acquire-us-based-caplot-software-for-6-34-million/articleshow/78681554.cms</link>
  <image>
    https://img.etimg.com/thumb/width=1200,ingsize=147954,resizemode=4,msid=78681554/persistent-to-acquire-us-based-caplot-software-for-6-34-million.jpg</image>
  <guid>Article at EconomicTimes.com with article id :
    78681554</guid>
  <pubDate>2020-12-09T13:49:11+05:30</pubDate>
</item>
</rss>
```

Відповідну думку знайдено у рядку 1

XML Ширина таблиці: 8 Ряд. 1, Ст. 1 ВСТ

JSON

```
{
  "title": "Covid-19: Major IT companies to extend WFH till March 2021",
  "textBody": "FAUG launches on Play Store worldwide",
  "source": "FAUG launches on Play Store worldwide",
  "pubDate": "2021-02-08T13:57:35+05:30",
  "url": "https://economictimes.indiatimes.com/tech/software/covid-19-major-it-companies-to-extend-wfh-till-march-2021/videoshow/79640742.cms"
},
{
  "title": "Persistent to acquire US-based Caplot Software for $6.34 million",
  "textBody": "Persistent to acquire US-based Caplot Software for $6.34 million",
  "source": "Persistent to acquire US-based Caplot Software for $6.34 million",
  "pubDate": "2020-12-09T13:49:11+05:30",
  "url": "https://economictimes.indiatimes.com/tech/software/persistent-to-acquire-us-based-caplot-software-for-6-34-million/articleshow/78681554.cms"
},
{
  "title": "Revise royalty definition to include software payments by subsidiaries to parent: Developing countries tell UN tax committee",
  "textBody": "Many developing countries including India want to tweak royalty definition in upcoming tax treaties",
  "source": "FAUG launches on Play Store worldwide",
  "pubDate": "2022-02-10T18:49:00Z",
  "url": "https://economictimes.indiatimes.com/tech/software/revise-royalty-definition-to-include-software-payments-by-subsidiaries-to-parent-developing-countries-tell-un-tax-committee/articleshow/78602676.cms"
},
{
  "title": "Springworks joins hands with Cure.fit, Plum, YourDOST, Mindhouse, ClinKK, Nova, GoodHealth to launch Employee Wellness bundle for startups",
  "textBody": "Springworks joins hands with Cure.fit, Plum, YourDOST, Mindhouse, ClinKK, Nova, GoodHealth to launch Employee Wellness bundle for startups",
  "source": "Springworks joins hands with Cure.fit, Plum, YourDOST, Mindhouse, ClinKK, Nova, GoodHealth to launch Employee Wellness bundle for startups",
  "pubDate": "2022-02-10T18:49:00Z",
  "url": "https://economictimes.indiatimes.com/tech/software/springworks-joins-hands-with-cure-fit-plum-yourdost-mindhouse-clinck-nova-goodhealth-to-launch-employee-wellness-bundle-for-startups/articleshow/78655386.cms"
},
{
  "title": "Microsoft allows employees to work from home permanently",
  "textBody": "Microsoft allows employees to work from home permanently",
  "source": "Microsoft allows employees to work from home permanently",
  "pubDate": "2022-02-10T18:49:00Z",
  "url": "https://economictimes.indiatimes.com/tech/software/microsoft-allows-employees-to-work-from-home-permanently/videoshow/78579699.cms"
},
{
  "title": "Microsoft to let employees work from home permanently: Report",
  "textBody": "Microsoft to let employees work from home permanently: Report",
  "source": "Microsoft to let employees work from home permanently: Report",
  "pubDate": "2022-02-10T18:49:00Z",
  "url": "https://economictimes.indiatimes.com/tech/software/microsoft-to-let-employees-work-from-home-permanently-report/articleshow/78578406.cms"
}
```

Відповідну думку знайдено у рядку 7

JSON Ширина таблиці: 8 Ряд. 1, Ст. 1 ВСТ

Питання до практичної роботи

1. Що таке *JSON-формат*?

JSON (англ. **JavaScript Object Notation**, укр. запис об'єктів **JavaScript**, вимовляється джейсон) — це текстовий формат обміну даними між комп'ютерами. **JSON** базується на тексті, може бути прочитаним людиною. Формат дає змогу описувати об'єкти та інші структури даних. Цей формат використовується переважно для передачі структурованої інформації через мережу.

2. Які основні переваги *JSON-формату*?

JSON відрізняється від **XML**, оскільки:

- **JSON** не використовує кінцевий тег
- **JSON** коротший
- **JSON** швидше читається і записується
- **JSON** може використовувати масиви

Найбільша різниця:

XML потрібно проаналізувати за допомогою аналізатора **XML**. **JSON** можна проаналізувати за допомогою стандартної функції **JavaScript**.

3. Які ви знаєте правила створення структури *JSON-файлу* в об'єкті, масиві і при присвоєнні значення?

JSON будується на двох структурах:

- Набір пар назва/значення. У різних мовах програмування це реалізовано як об'єкт, запис, структура, словник, хеш-таблиця, список із ключем або асоціативним масивом.
- Впорядкований список значень. У багатьох мовах це реалізовано як масив, вектор, список або послідовність.

Синтаксис **JSON** походить від синтаксису позначення об'єктів **JavaScript**:

- Дані містяться в парах ім'я/значення
- Дані розділяються комами
- Фігурні дужки утримують об'єкти
- Квадратні дужки містять масиви

Масив записується у квадратних дужках «**[]**». Значення поділяються комами. Масив може бути порожнім, тобто не містити жодного значення. Значення не більше одного масиву можуть мати різний тип.

В якості значень **JSON** можуть бути використані як числа, так і рядки.

4. Що таке пошукова система **Elasticsearch** та її призначення?

Elasticsearch — вільне програмне забезпечення, пошуковий сервер, розроблений на базі **Lucene**. Надає розподілений, мультиарендний повнотекстовий пошуковий рушій з **HTTP** вебінтерфейсом і підтримкою безсхемних **JSON** документів. **Elasticsearch** призначена для отримання повної картини даних.

Elasticsearch може використовуватись для індексування та пошуку будь-яких типів документів. Він надає масштабний пошук, має пошук близький до реального часу і підтримку мультиарендності.

Elasticsearch має можливість розподілення, індекси можуть бути розділені на сегменти, при чому кожен сегмент може мати нуль чи більше реплік. Кожен вузол містить один чи більше сегментів і діє як координатор делегування операцій на потрібний сегмент. Балансування та маршрутизація виконується автоматично.