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]
```

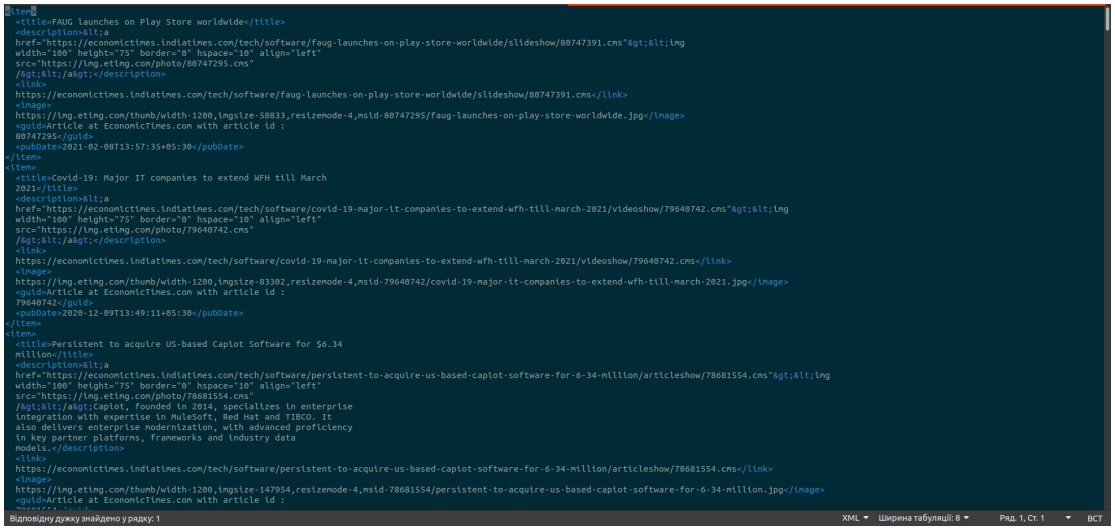
```
#Видалення перших пропусків

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-MMTHH: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 ("{\n\"title\":\""+title[i]+"\",")

     #Специфічна обробка тексту

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

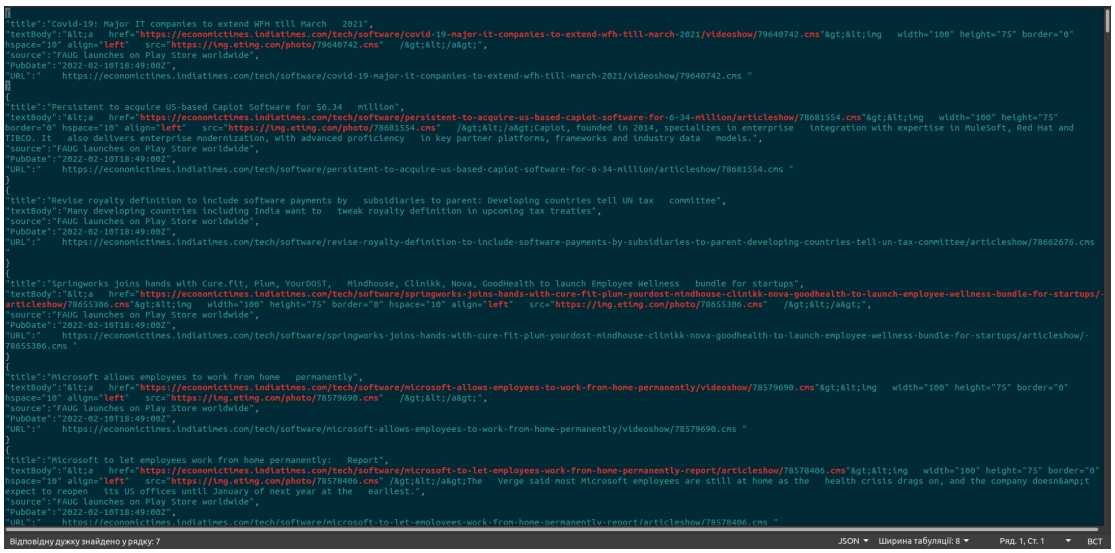
```
linux@asus-X505BP:~/Стільниця/Big Data/Lab-2$ python3 RSS_to_JSON.py > result.json
```

Результати

RSS



JSON



Питання до практичної роботи

1. *Що таке **JSON–формат**?*

JSON (англ. **JavaScript Object Notation**, укр. запис об'єктів **JavaScript**, вимовляється джéйсон) — 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 масштабовний 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.