Національний технічний університет України
«Київський політехнічний інститут»
Фізико-Технічний Інститут
Кафедра математичних методів захисту інформації

Аналітика Big Data

Практичне заняття №2

Формування JSON-файлу, придатного для завантаження в систему ElasticSearch. Реалізувати процедуру конвертування RSS-фідів в формат JSON мовами Python або R

Виконав:
студент групи ФІ-73
Чіхладзе Вахтанг

Перевірив:
Ланде Д.В.

Київ 2021

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.

1),2),3)В першій лабораторній роботі ми отримали xml файл, що містить item елементи. Була реалізована програма на мові python3.8.3, що використовує регулярні вирази для обробки та пошуку шаблонних текстів, які пететворюють item з формату xml у формат json.

```
import re
import os
import json


def prepare_xml(filename):
    file=open(filename,'r')
    xml=file.read()
    xml=xml.replace('\n','')
    xml=xml.replace('<items>','')
    xml=xml.replace('</items>','')
    xml=xml.replace('</item>','</item>\n')
    xml=xml.split('\n')
    return xml


def parse_component(item,component):
    regex_result=re.findall(r'<'+component+'>(.*)</'+component+'>',item)
    if len(regex_result)>0:
        return regex_result[0]
    else:
        return ""
def parse_source(link):
    regex_result=re.findall(r'https?://(.*?)/',link)
    if len(regex_result)>0:
        return regex_result[0]
    else:
        return ""

def parse_items(filename):
    items=prepare_xml(filename)
    file=open("rss_feeds.json",'a')
    for item in items[1:]:
        title=parse_component(item,'title')
        description=parse_component(item,'description')
        pubDate=parse_component(item,'pubDate')
```

```
            link=parse_component(item,'link')
            source=parse_source(link)
            item_dictionary={
            "title":title,
            "textBody":description,
            "source": source,
            "pubDate": pubDate,
            "URL": link
            }
            item_json=json.dumps(item_dictionary,indent = len(item_dictionary),
ensure_ascii=False)
            file.write(item_json+",\n")
            file.flush()

if __name__ == "__main__":
    parse_items("rss_feeds.xml")
```

Висновок: отже, я поглибив знання з обробки xml файлів та отримав навички роботи з json.