



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

Лабораторна робота №1

Аналіз даних

Підготував:

студент 4 курсу
групи ФІ-84

Коломієць Андрій Юрійович

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

Київ – 2021

Лабораторна робота №1

Аналіз даних

Завдання на самостійну роботу

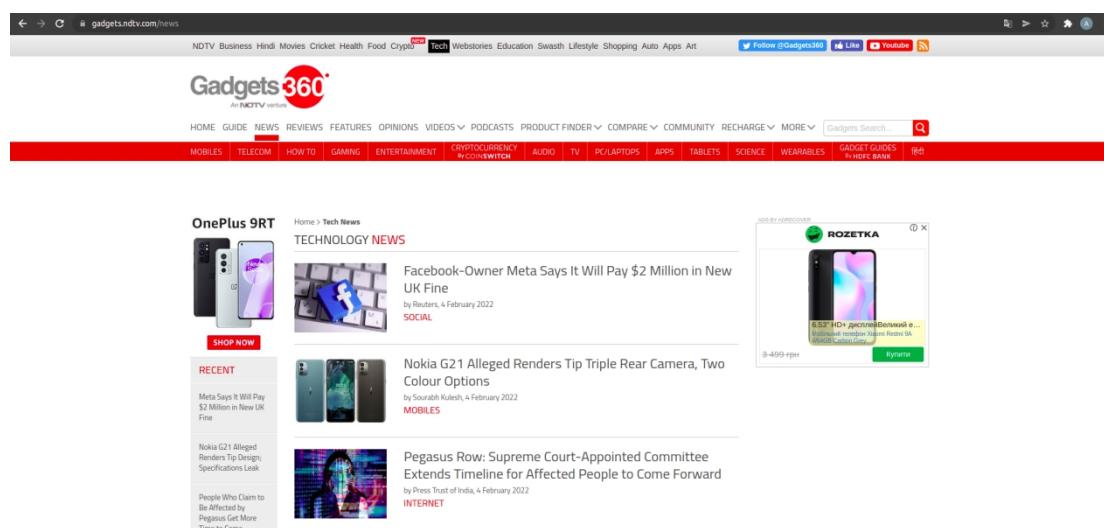
1. Розширити список телекомунікаційної RSS-фідів каналами комп'ютерної і телекомунікаційної спрямованості (але не торговельними майданчиками).
2. Створити процедуру (скрипт) для періодичного скачування інформації із створеного переліку RSS-фідів.
3. Реалізувати процедуру створення файлу, в якому об'єднуються всі скачані RSS-фіди, і підключити її до скрипту скачування.

Виконання завдання

1. Розширити список телекомунікаційної RSS-фідів каналами комп'ютерної і телекомунікаційної спрямованості (але не торговельними майданчиками).

Шукаємо вебсайт з RSS-фідами:

<https://gadgets.ndtv.com/news>



Для того щоб знайти RSS-фіди, натискаємо правою кнопкою миші на вебсайті та вибираємо переглянути джерельний код сторінки.

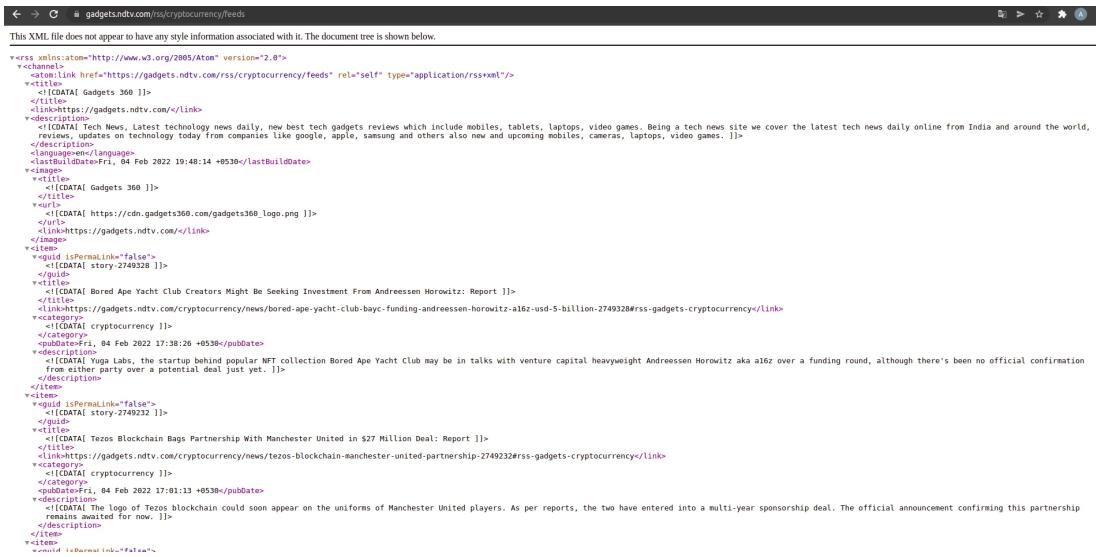
Натикаємо комбінації клавіш Ctrl+ F та шукаємо стрічку RSS.

Настикаємо на посилання табачимо список RSS-фідів на різну тематику згідно даного вебсайту. На мій вибір було обрано RSS-фіди пов'язані з криптовалютами:

<https://gadgets.ndtv.com/rss/cryptocurrency/feeds>

RSS Feeds	
The feeds are available for the following sections:	
All Stories	Reviews
Opinion	Features
Android-Hub	News
Telecom	Photos
Internet	Cryptocurrency
Home Entertainment	India-Hub
Audio	Apple-Hub
Science	Gaming
Home Appliances	Wearables
360 Daily	Transportation
Culture	Tablets
Contests	Xiaomi
	Sony-Hub
	Samsung-Hub
	Others
	How to
	Best Buys
	Breaking News
	Reviews
	Videos
	PC/Laptops
	Apps
	Transportation
	Tablets
	Xiaomi
	Sony-Hub
	Samsung-Hub
	Others
	How to
	Best Buys
	Reviews
	Videos
	PC/Laptops
	Apps
	Transportation
	Tablets
	Xiaomi
	Sony-Hub
	Samsung-Hub
	Others
	How to
	Best Buys

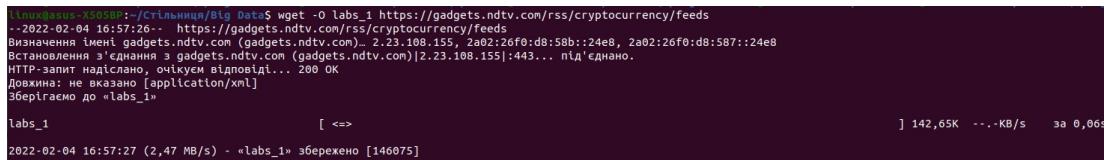
Маємо відповідну сторінку з RSS-фідів.



This XML file does not appear to have any style information associated with it. The document tree is shown below.

```
<?xml version='1.0' encoding='utf-8'?>
<rss version="2.0">
  <channel>
    <atom:link href="https://gadgets.ndtv.com/rss/cryptocurrency/feeds" rel="self" type="application/rss+xml"/>
    <title><![CDATA[ Gadgets 360 ]]>
    <link>https://gadgets.ndtv.com/</link>
    <description><![CDATA[Tech News, Latest technology news daily, new best tech gadgets reviews which include mobiles, tablets, laptops, video games. Being a tech news site we cover the latest tech news daily online from India and around the world, reviews, updates on technology today from companies like google, apple, samsung and others also new and upcoming mobiles, cameras, laptops, video games. ]]>
    <language>en</language>
    <lastBuildDate>Fri, 04 Feb 2022 19:48:14 +0530</lastBuildDate>
    <image>
      <title><![CDATA[ Gadgets 360 ]]>
      <url><![CDATA[ https://cdn.gadgets360.com/gadgets360_logo.png ]]>
      <link>https://gadgets.ndtv.com/</link>
    </image>
    <item>
      <guid isPermaLink="false"><![CDATA[ story-2749328 ]]>
      </guid>
      <title><![CDATA[ Bored Ape Yacht Club Creators Might Be Seeking Investment From Andreessen Horowitz: Report ]]>
      </title>
      <link>https://gadgets.ndtv.com/cryptocurrency/news/bored-ape-yacht-club-bayc-funding-andreessen-horowitz-a16z-usd-5-billion-2749328#rss-gadgets-cryptocurrency/</link>
      <category><![CDATA[ cryptocurrency ]]>
      <pubDate>Fri, 04 Feb 2022 17:38:26 +0530</pubDate>
      <description><![CDATA[ Yuga Labs, the startup behind popular NFT collection Bored Ape Yacht Club may be in talks with venture capital heavyweight Andreessen Horowitz aka a16z over a funding round, although there's been no official confirmation from either party over a potential deal just yet. ]]>
      </description>
    </item>
    <item>
      <guid isPermaLink="false"><![CDATA[ story-2749322 ]]>
      </guid>
      <title><![CDATA[ Tezos Blockchain Bags Partnership With Manchester United in $27 Million Deal: Report ]]>
      </title>
      <link>https://gadgets.ndtv.com/cryptocurrency/news/tezos-blockchain-manchester-united-partnership-2749322#rss-gadgets-cryptocurrency/</link>
      <category><![CDATA[ cryptocurrency ]]>
      <pubDate>Fri, 04 Feb 2022 17:01:13 +0530</pubDate>
      <description><![CDATA[ The logo of Tezos blockchain could soon appear on the uniforms of Manchester United players. As per reports, the two have entered into a multi-year sponsorship deal. The official announcement confirming this partnership remains awaited for now. ]]>
      </description>
    </item>
  </channel>
</rss>
```

RSS-фіди мають версію більш сучасну.

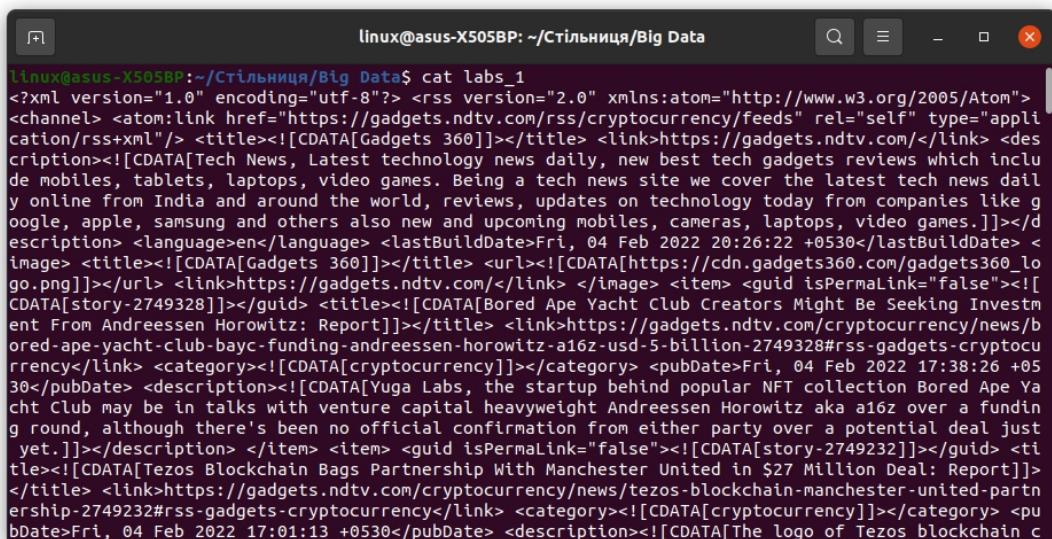


```
linux@asus-X505BP:~/Стільниця/Big Data$ wget -O labs_1 https://gadgets.ndtv.com/rss/cryptocurrency/feeds
--2022-02-04 16:57:26- https://gadgets.ndtv.com/rss/cryptocurrency/feeds
Визначення імені gadgets.ndtv.com (gadgets.ndtv.com)... 2.23.108.155, 2a02:26f0:d8:58b::24e8, 2a02:26f0:d8:587::24e8
Встановлення з'єднання з gadgets.ndtv.com (gadgets.ndtv.com)|2.23.108.155|:443... під'єднано.
НТТР-запит надіслано, очікує відповіді... 200 OK
Довжина: не вказано [application/xml]
Зберігаємо до «labs_1»

labs_1                                         [ =>                               ] 142,65K  ---.KB/s   за 0,065
2022-02-04 16:57:27 (2,47 MB/s) - «labs_1» збережено [146075]
```



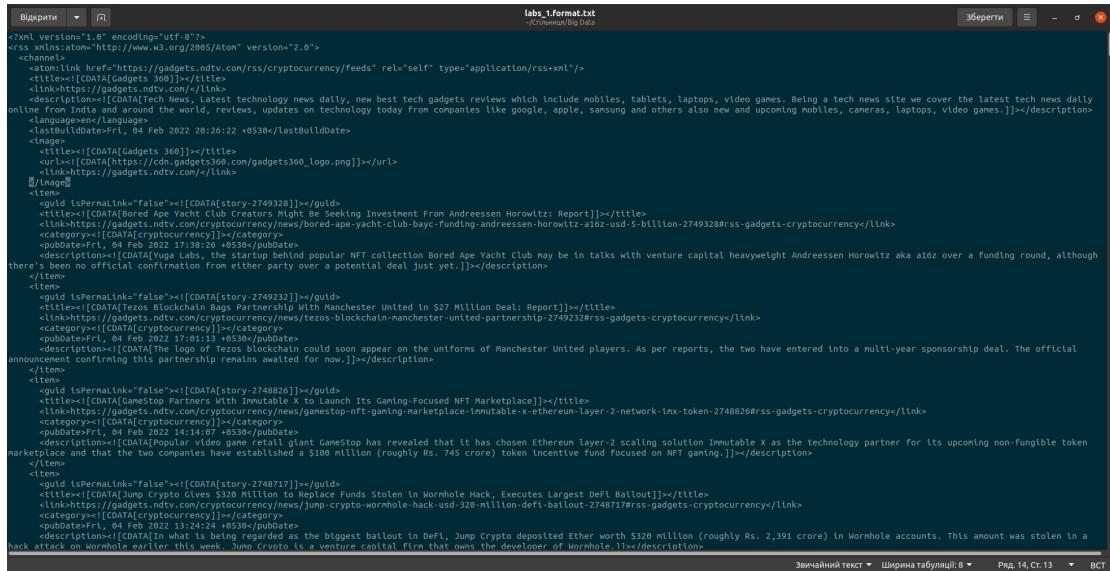
Завантажимо наші данні через утиліти консолі Linux.



```
linux@asus-X505BP:~/Стільниця/Big Data$ cat labs_1
<?xml version="1.0" encoding="utf-8"?> <rss version="2.0" xmlns:atom="http://www.w3.org/2005/Atom">
  <channel> <atom:link href="https://gadgets.ndtv.com/rss/cryptocurrency/feeds" rel="self" type="application/rss+xml"/> <title><![CDATA[Gadgets 360]]></title> <link>https://gadgets.ndtv.com/</link> <description><![CDATA[Tech News, Latest technology news daily, new best tech gadgets reviews which include mobiles, tablets, laptops, video games. Being a tech news site we cover the latest tech news daily online from India and around the world, reviews, updates on technology today from companies like google, apple, samsung and others also new and upcoming mobiles, cameras, laptops, video games.]]></description> <language>en</language> <lastBuildDate>Fri, 04 Feb 2022 20:26:22 +0530</lastBuildDate> <image> <title><![CDATA[Gadgets 360]]></title> <url><![CDATA[https://cdn.gadgets360.com/gadgets360_logo.png]]></url> <link>https://gadgets.ndtv.com/</link> </image> <item> <guid isPermaLink="false"><![CDATA[story-2749328]]></guid> <title><![CDATA[Bored Ape Yacht Club Creators Might Be Seeking Investment From Andreessen Horowitz: Report]]></title> <link>https://gadgets.ndtv.com/cryptocurrency/news/bored-ape-yacht-club-bayc-funding-andreessen-horowitz-a16z-usd-5-billion-2749328#rss-gadgets-cryptocurrency/</link> <category><![CDATA[cryptocurrency]]></category> <pubDate>Fri, 04 Feb 2022 17:38:26 +0530</pubDate> <description><![CDATA[Yuga Labs, the startup behind popular NFT collection Bored Ape Yacht Club may be in talks with venture capital heavyweight Andreessen Horowitz aka a16z over a funding round, although there's been no official confirmation from either party over a potential deal just yet.]]></description> </item> <item> <guid isPermaLink="false"><![CDATA[story-2749322]]></guid> <title><![CDATA[Tezos Blockchain Bags Partnership With Manchester United in $27 Million Deal: Report]]></title> <link>https://gadgets.ndtv.com/cryptocurrency/news/tezos-blockchain-manchester-united-partnership-2749322#rss-gadgets-cryptocurrency/</link> <category><![CDATA[cryptocurrency]]></category> <pubDate>Fri, 04 Feb 2022 17:01:13 +0530</pubDate> <description><![CDATA[The logo of Tezos blockchain could soon appear on the uniforms of Manchester United players. As per reports, the two have entered into a multi-year sponsorship deal. The official announcement confirming this partnership remains awaited for now.]]></description> </item> </channel>
</rss>
```

Формат **RSS-фідів** має неструктурований вигляд. Зробимо адекватну візуалізацію завантаженим даним:

```
linuX@asus-X505BP:~/Стільниця/Big Data$ xmllint --format labs_1 >labs_1.format.txt
```



```
<?xml version="1.0" encoding="utf-8"?>
<rss xmlns:atom="http://www.w3.org/2005/Atom" version="2.0">
  <channel>
    <title><![CDATA[Gadgets 360]]></title>
    <link>https://gadgets.ndtv.com/</link>
    <description><![CDATA[Tech News, Latest technology news daily, new best tech gadgets reviews which include mobiles, tablets, laptops, video games. Being a tech news site we cover the latest tech news daily online from India and around the world, reviews, updates on technology today from companies like google, apple, samsung and others also new and upcoming mobiles, cameras, laptops, video games.]]></description>
    <language>en</language>
    <lastBuildDate>Fri, 04 Feb 2022 20:26:22 +0530</lastBuildDate>
    <image>
      <title><![CDATA[Gadgets 360]]></title>
      <url><![CDATA[https://cdn.gadgets360.com/gadgets360_logo.png]]></url>
      <link>https://gadgets.ndtv.com/</link>
    </image>
    <item>
      <guid isPermaLink="false"><![CDATA[story-2749328]]></guid>
      <title><![CDATA[Bored Ape Yacht Club Creators Might Be Seeking Investment From Andreessen Horowitz: Report]]></title>
      <link>https://gadgets.ndtv.com/cryptocurrency/news/boredape-yacht-club-bayc-funding-andreessen-horowitz-alice-usd-2749328#rss-gadgets-cryptocurrency</link>
      <category><![CDATA[cryptocurrency]]></category>
      <pubDate>Fri, 04 Feb 2022 17:38:26 +0530</pubDate>
      <description><![CDATA[Vuga Labs, the startup behind popular NFT collection Bored Ape Yacht Club may be in talks with venture capital heavyweight Andreessen Horowitz aka alice over a funding round, although there is no official confirmation from either party over a potential deal just yet.]]></description>
    </item>
    <item>
      <guid isPermaLink="false"><![CDATA[story-2749322]]></guid>
      <title><![CDATA[Tezos Blockchain Lags Partnership With Manchester United in $27 Million Deal: Report]]></title>
      <link>https://gadgets.ndtv.com/cryptocurrency/news/tezos-blockchain-manchester-united-partnership-2749232#rss-gadgets-cryptocurrency</link>
      <category><![CDATA[cryptocurrency]]></category>
      <pubDate>Fri, 04 Feb 2022 17:01:18 +0530</pubDate>
      <description><![CDATA[Popular video game retail giant GameStop has revealed that it has chosen Ethereum layer-2 scaling solution Immutable X as the technology partner for its upcoming non-fungible token marketplace and that the two companies have established a $100 million (roughly Rs. 745 crore) token incentive fund focused on NFT gaming.]]></description>
    </item>
    <item>
      <guid isPermaLink="false"><![CDATA[story-2748820]]></guid>
      <title><![CDATA[GameStop Partners With Immutable X to Launch Its Gaming-Focused NFT Marketplace]]></title>
      <link>https://gadgets.ndtv.com/cryptocurrency/news/gamestop-nft-gaming-marketplace-immutable-x-ethereum-layer-2-network-lmx-token-2748820#rss-gadgets-cryptocurrency</link>
      <category><![CDATA[cryptocurrency]]></category>
      <pubDate>Fri, 04 Feb 2022 16:14:08 +0530</pubDate>
      <description><![CDATA[Popular video game retail giant GameStop has revealed that it has chosen Ethereum layer-2 scaling solution Immutable X as the technology partner for its upcoming non-fungible token marketplace and that the two companies have established a $100 million (roughly Rs. 745 crore) token incentive fund focused on NFT gaming.]]></description>
    </item>
    <item>
      <guid isPermaLink="false"><![CDATA[story-2748717]]></guid>
      <title><![CDATA[Jump Crypto Gives $320 Million to Replace Funds Stolen in Wormhole Hack, Executes Largest DeFi Bailout]]></title>
      <link>https://gadgets.ndtv.com/cryptocurrency/news/jump-crypto-wormhole-hack-usd-320-million-defi-bailout-2748717#rss-gadgets-cryptocurrency</link>
      <category><![CDATA[cryptocurrency]]></category>
      <pubDate>Fri, 04 Feb 2022 13:24:24 +0530</pubDate>
      <description><![CDATA[In what was being regarded as the biggest bailout in DeFi, Jump Crypto deposited Ether worth $320 million (roughly Rs. 2,391 crore) in Wormhole accounts. This amount was stolen in a hack attack on Wormhole earlier this week. Jump Crypto is a venture capital firm that owns the developer of Wormhole.]]></description>
    </item>
  </channel>
</rss>
```

Аналогічно так можна виконати процедуру для багатьох вебсайтів котрі мають **RSS-фіди**.

У зазначеному прикладі погано вибрано тематику, оскільки вона про валюти. З цього ж самого сайту можна вибрати іншу тематику пов'язану з технологіями чи новинами:

<https://gadgets.ndtv.com/rss/laptops/feeds>

<https://gadgets.ndtv.com/rss/mobiles/feeds>

також візьмемо ще котрийсь вебсайт з **RSS-фідами** наприклад:

<https://www.euronews.com/news/international>

посиланя на **RSS-фід**:

<https://www.euronews.com/rss?level=theme&name=news>

2. Створити процедуру (скрипт) для періодичного скачування інформації із створеного переліку RSS-фідів. Реалізувати процедуру створення файлу, в якому об'єднуються всі скачані RSS-фіди, і підключити її до скрипту скачування.

Скрипт

```
#!/bin/bash

rm -R file

mkdir file

COUNTER=0

for i in <list website with RSS-feeds>

do
    echo '-----'
    echo $COUNTER'. ' ${i}
    echo '-----'

    wget -O file/file_$COUNTER.xml ${i}
    grep -Eo "<item>.*</item>" file/file_$COUNTER.xml > file/xml_$COUNTER.xml
    rm file/file_$COUNTER.xml

    let COUNTER++

done

tidy -xml -i file/xml* >> rss.xml

rm -R file
```

Під <list website with RSS-feeds> розуміється:

```
'http://economictimes.indiatimes.com/tech/software/rssfeeds/13357555.cms'
'http://nypost.com/tech/feed'
'http://www.smh.com.au/rssheadlines/technology-news/article/rss.xml'
'http://zeenews.india.com/rss/science-technology-news.xml'
'http://www.washingtontimes.com/rss/headlines/culture/technology/'
'https://www.thehindu.com/sci-tech/technology/?service=rss'
'https://www.economist.com/science-and-technology/rss.xml'
'https://feeds.skynews.com/feeds/rss/technology.xml'
'http://www.innertemplibrary.com/category/secure-hospitals/feed/'
'http://fakty.ua/rss_feed/science'
'https://gadgets.ndtv.com/rss/cryptocurrency/feeds'
'https://gadgets.ndtv.com/rss/laptops/feeds'
'https://gadgets.ndtv.com/rss/mobiles/feeds'
'https://www.euronews.com/rss?level=theme&name=news'
```

Код

```
Відкрити Зберегти
labs.sh
~/Стільниця/Big Data/Lab-1
#!/bin/bash

rm -R file

mkdir file

COUNTER=0

for i in 'http://economictimes.indiatimes.com/tech/software/rssfeeds/13357555.cms' 'http://nypost.com/-tech/feed' 'http://www.smh.com.au/rssheadlines/technology-news/article/rss.xml' 'http://zeenews.india.com/rss/science-technology-news.xml' 'http://www.washingtontimes.com/rss/headlines/culture/technology/' 'https://www.thehindu.com/sci-tech/technology/?service=rss' 'https://www.economist.com/science-and-technology/rss.xml' 'https://feeds.skynews.com/feeds/rss/technology.xml' 'http://www.innertemplelibrary.com/category/secure-hospitals/feed/' 'http://fakty.ua/rss_feed/science' 'https://gadgets.ndtv.com/rss/crypto/currency/feeds' 'https://gadgets.ndtv.com/rss/laptops/feeds' 'https://gadgets.ndtv.com/rss/mobiles/feeds' 'https://www.euronews.com/rss?level=theme&name=news'

do
    echo '-----'
    echo $COUNTER'. '$i
    echo '-----'

    wget -O file/file_${COUNTER}.xml ${i}
    grep -Eo "<item>.*</item>" file/file_${COUNTER}.xml > file/xml_${COUNTER}.xml
    rm file/file_${COUNTER}.xml

    let COUNTER++

done

tidy -xml -i file/xml* >> rss.xml

rm -R file

sh └ Ширина табуляції: 8 └ Ряд. 26, Ст. 11 └ ВСТ
```

Маємо результати

```
Відкрити Зберегти
rss.xml
~/Стільниця/Big Data/Lab-1
<items>
    <item>
        <title>FAUC launches on Play Store worldwide</title>
        <description><a href="https://economictimes.indiatimes.com/tech/software/fauc-launches-on-play-store-worldwide/slideshow/80747391.cms">&gt;</a><br/>&gt;</a></description>
        <link>
            https://economictimes.indiatimes.com/tech/software/fauc-launches-on-play-store-worldwide/slideshow/80747391.cms
        </link>
        <image>
            https://img.etimg.com/thumb/width-1200,lnsize-58833,reslzenode-4,nsid-80747295/fauc-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><a href="https://economictimes.indiatimes.com/tech/software/covid-19-major-it-companies-to-extend-wfh-till-march-2021/videoshow/79640742.cms">&gt;</a><br/>&gt;</a></description>
        <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,lnsize-8302,reslzenode-4,nsid-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 Capot Software for $6.34 million</title>
        <description><a href="https://economictimes.indiatimes.com/tech/software/persistent-to-acquire-us-based-capot-software-for-6-34-million/articleshow/78681554.cms">&gt;</a><br/>&gt;</a></description>
        <link>
            https://economictimes.indiatimes.com/tech/software/persistent-to-acquire-us-based-capot-software-for-6-34-million/articleshow/78681554.cms
        </link>
        <image>
            https://img.etimg.com/thumb/width-1200,lnsize-147954,reslzenode-4,nsid-78681554/persistent-to-acquire-us-based-capot-software-for-6-34-million.jpg
        </image>
        <guid>Article at EconomicTimes.com with article Id : 78681554</guid>
        <pubDate>2020-10-15T17:33:06+05:30</pubDate>
    </item>
</items>
```

Питання до практичної роботи

1. Що таке RSS і як він використовується?

RSS — це родина XML-форматів, що використовується для публікації та постачання інформації, що часто змінюється, наприклад, нових записів в блозі, заголовків новин, анонсів статей, зображень, аудіо і відео матеріалів (в стандартизованому форматі).

2. Що таке RSS канал?

RSS-канали (Rely Simple Syndication або Rich Site Summary) дозволяють користувачам збирати оновлення з веб-сайтів, не вводячи їх. Користувачам потрібен інструмент, який називається "RSS reader" чи "RSS aggregator", який перевіряє, чи оновили вміст веб-сайтів на підписку.

3. З чого складається RSS?

RSS базується на стандарті XML. Перший тег в RSS-документі обов'язково вказує на формат XML, що застосовується.

*Після нього йде тег <rss> з обов'язковим атрибутом **version**, який вказує на версію документа.*

У обов'язковому порядку RSS-документ містить 2 теги: <channel> i <item>.

Основну інформацію про цей RSS-канал містить тег <channel>. Він зустрічається лише 1 раз.

У обов'язковому порядку він містить такі три теги:

<title> – ім'я каналу. Може збігатися із ім'ям сайту.

<link> – посилання на сайт, з яким пов'язаний канал.

<description> – опис каналу.

Опціональні теги:

<language> – мова стрічки.

<copyright> – авторські права.

<managingEditor> – електронна пошта редактора вмісту каналу

<webMaster> – електронна пошта веб-майстра каналу

<pubDate> – дата публікації каналу.

<lastBuildDate> – дата останньої зміни вмісту каналу.

<category> – категорії контенту каналу.

<generator> – програма, за допомогою якої було згенеровано канал.

<docs> – посилання на документацію використовуваного формату RSS.

<ttl> – час актуальності каналу в хвилинах.

<image> – зображення, яке відображається з каналом.

У свою чергу, тег має такі параметри:

<Title> – заголовок.

<Description> – опис (аналог тега ALT в HTML).

<Link> – посилання на сайт, з яким пов'язаний канал.

<URL> – адреса зображення.

<Width> – ширина зображення.

<Height> – висота зображення.

<skipHours> – скільки годин не вимагати оновлення з каналу

<skipDays> – скільки днів не вимагати оновлення з каналу

Тег *<item>* містить інформацію про публікацію. Обов'язкові вкладені теги:

<title> – назва публікації.

<link> – посилання на сторінку з повним текстом публікації.

<description> – короткий текст публікації.

Необов'язкові вкладені теги:

<Author> – електронна пошта автора

<Category> – категорія повідомлення

<Comments> – посилання на сторінку з коментарями до повідомлення

<Enclosure> – приєднаний мультимедійний об'єкт.

Його параметри:

- <URL> – адреса об‘єкта
- <Length> – розмір об‘єкта в байтах
- <Type> – MIME-тип файлу
- <Guid> – ідентифікатор повідомлення
- <PubDate> – дата публікації.

4. У чому різниця між веб-стрічками та RSS?

У Всесвітній мережі веб - канал (або стрічка новин) — це формат даних, який використовується для надання користувачам вмісту, який часто оновлюється. Тобто може бути, як посиланням на веб-сторінку ресурсу, в форматі відомому від RSS.

Типи вмісту, що надається через веб-канал, як правило, це HTML (вміст веб-сторінки) або посилання на веб-сторінки та інші види цифрових медіа. Часто, коли веб- сайти надають веб-канали для сповіщення користувачів про оновлення вмісту, вони включають лише підсумки у веб-канал, а не повний вміст. Багато новинних веб- сайтів , веб- журналів , шкіл і подкастерів керують веб-каналами. Оскільки веб-канали призначені для машиночитання , а не для зчитування людиною , їх також можна використовувати для автоматичної передачі інформації з одного веб- сайту на інший без будь- якого втручання людини.