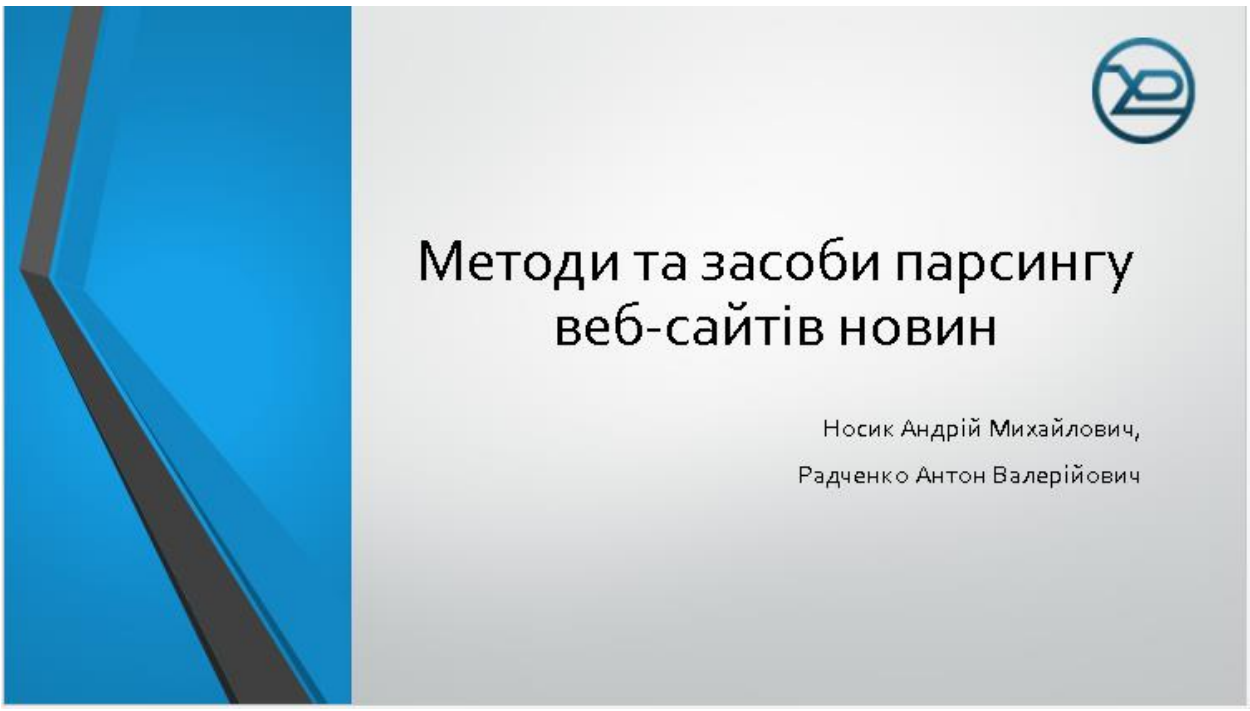
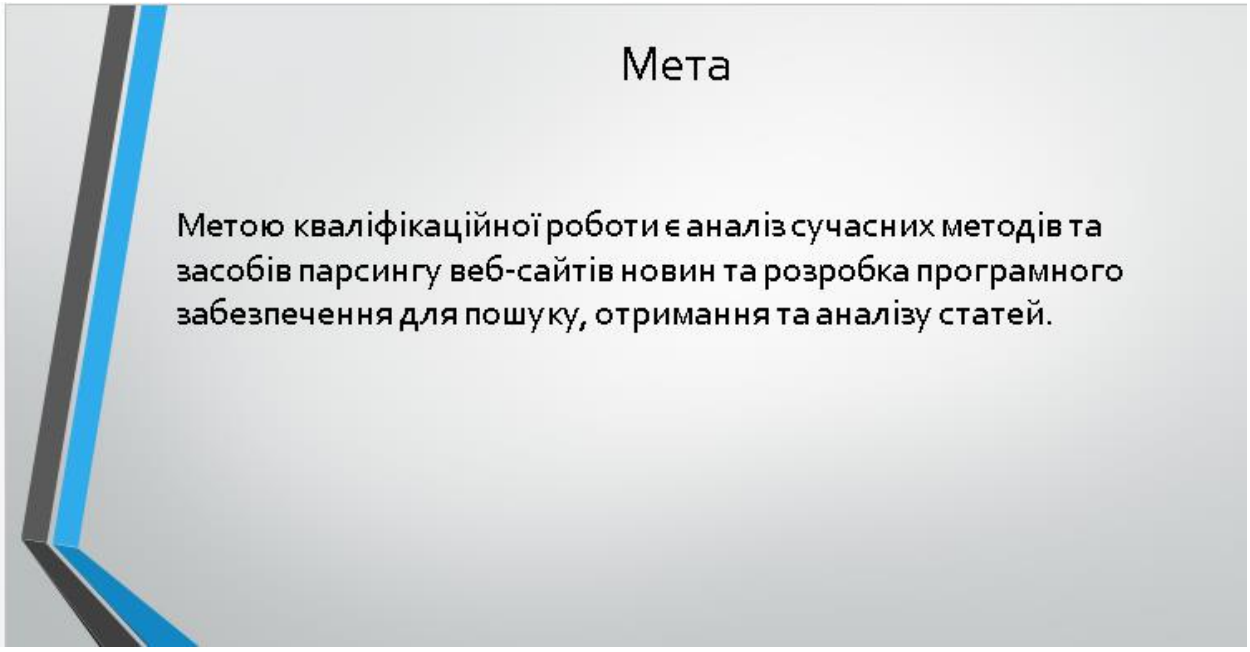
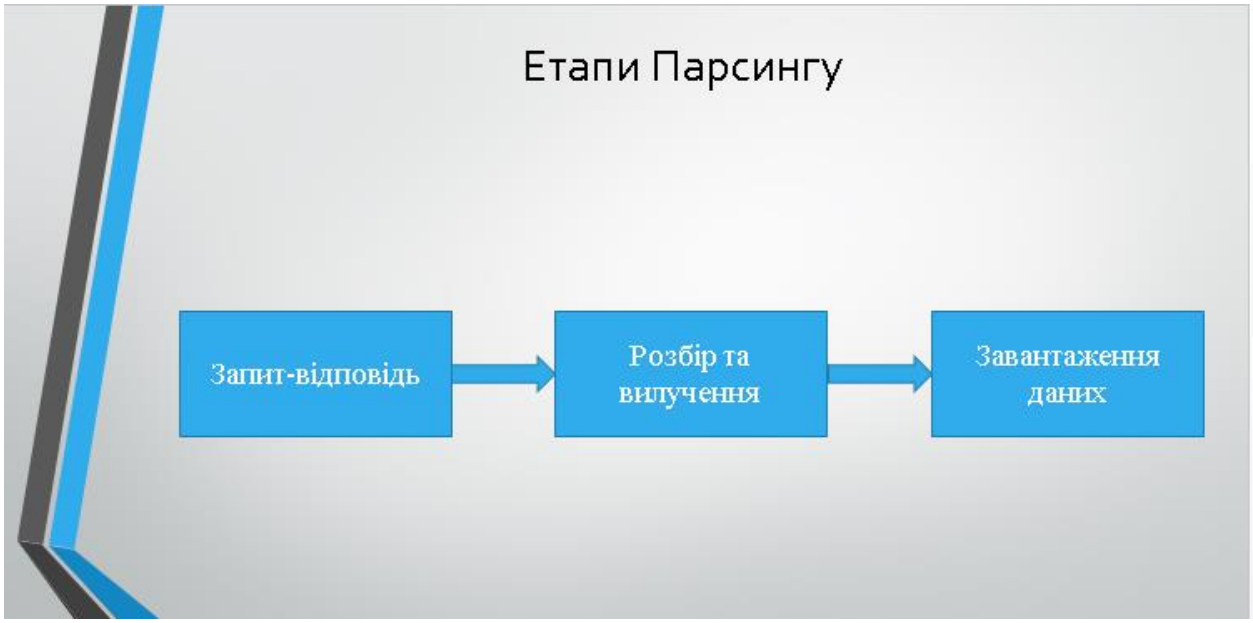
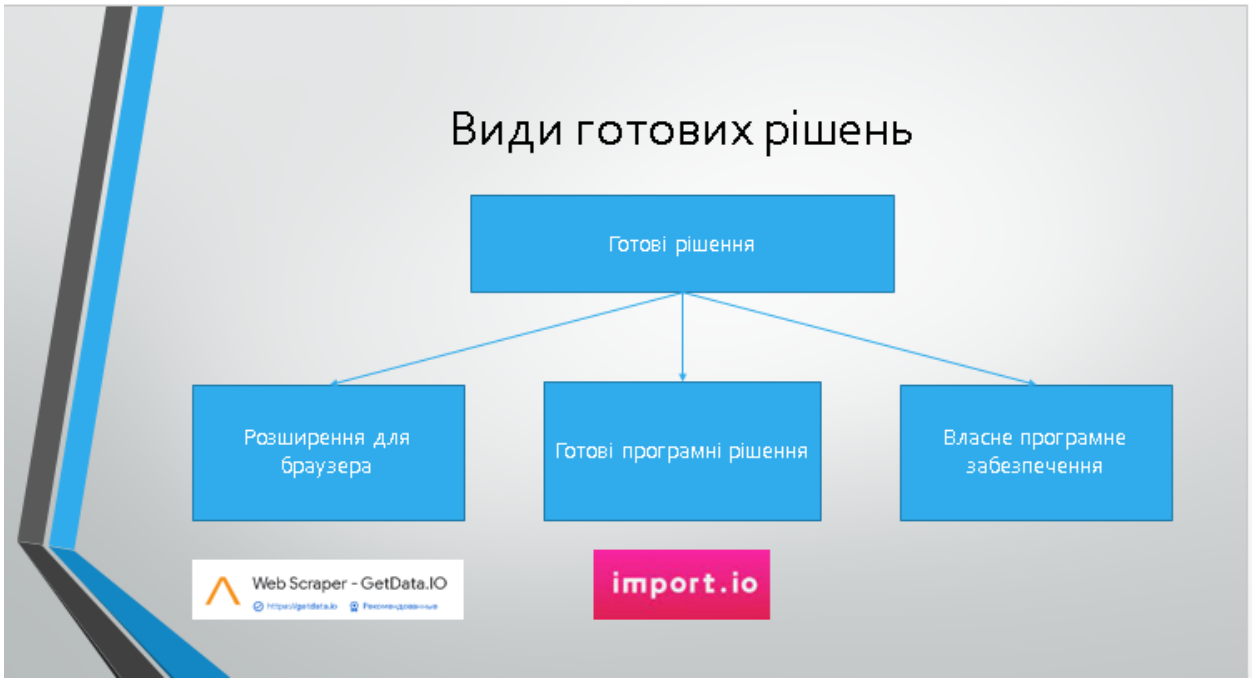
ДОДАТОК А
ГРАФІЧНИЙ МАТЕРІАЛ КВАЛІФІКАЦІЙНОЇ РОБОТИ


Методи та засоби парсингу веб-сайтів новин
Носик Андрій Михайлович,
Радченко Антон Валерійович


Мета
Метою кваліфікаційної роботи є аналіз сучасних методів та засобів парсингу веб-сайтів новин та розробка програмного забезпечення для пошуку, отримання та аналізу статей.


Етапи Парсингу
Запит-відповідь
Розбір та вилучення
Завантаження даних


Методи парсингу
Методи веб-парсингу
Ручне копіювання та вставка
Аналіз HTML, DOM
Аналіз за допомогою комп'ютерного зору



Переваги та недоліки додатків

Переваги розширень для браузера:

- Легко встановити та налаштувати
- Ефективний під час збору невеликої к-сті даних

Недоліки розширень для браузера:

- Не ефективний під час збору великої к-сті інформації
- Привязаність до конкретного браузера
- Малий обсяг інформації про дані

Переваги готових програмних рішень:

- Великий обсяг інформації про дані
- Ефективний під час збору великої к-сті інформації
- Недоліки готових програмних рішень:
- В основному всі ПО є платними
- Не ефективні для парсингу новин

Інструменти які були використані

- мова програмування Python;
- середовище розробки Visual Studio Code;
- мова розмітки HTML та CSS;
- структура данних JSON,CSV,TXT;
- машинне навчання.



Власне рішення

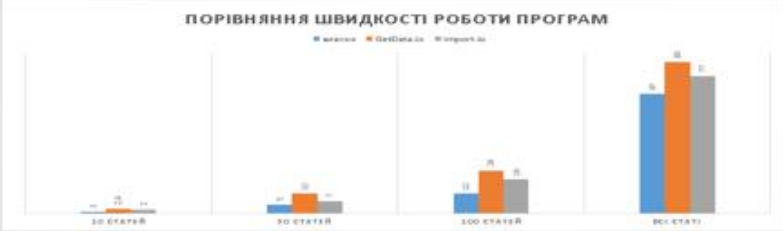

```
"headline": "Facebook is spending $5.7 billion to capitalize on internet boom",
"author": ["Sherisse Pham", "Cnn Business"],
"date_publish": "2022-04-22 04:33:39",
"date_modify": "None",
"image_url": "None",
"filename": "https://www.cnn.com/2020/04/22/tech/facebook-reliance-jio/index.html.json",
"description": "Facebook is investing $5.7 billion into Jio Platforms, the digital techno
"publication": "CNN",
"category": "tech",
"source_domain": "www.cnn.com",
"article": "Hong Kong (CNN Business) Facebook (FB) is spending billions of dollars for a
"summary": "Hong Kong (CNN Business) Facebook (FB) is spending billions of dollars for a
"keyword": [
  "mobile","billion", "platforms","57", "partnership", "boom","indias", "stake", "users",
],
"title_page": "None",
"title_rss": "None",
"url": "https://www.cnn.com/2020/04/22/tech/facebook-reliance-jio/index.html"
```

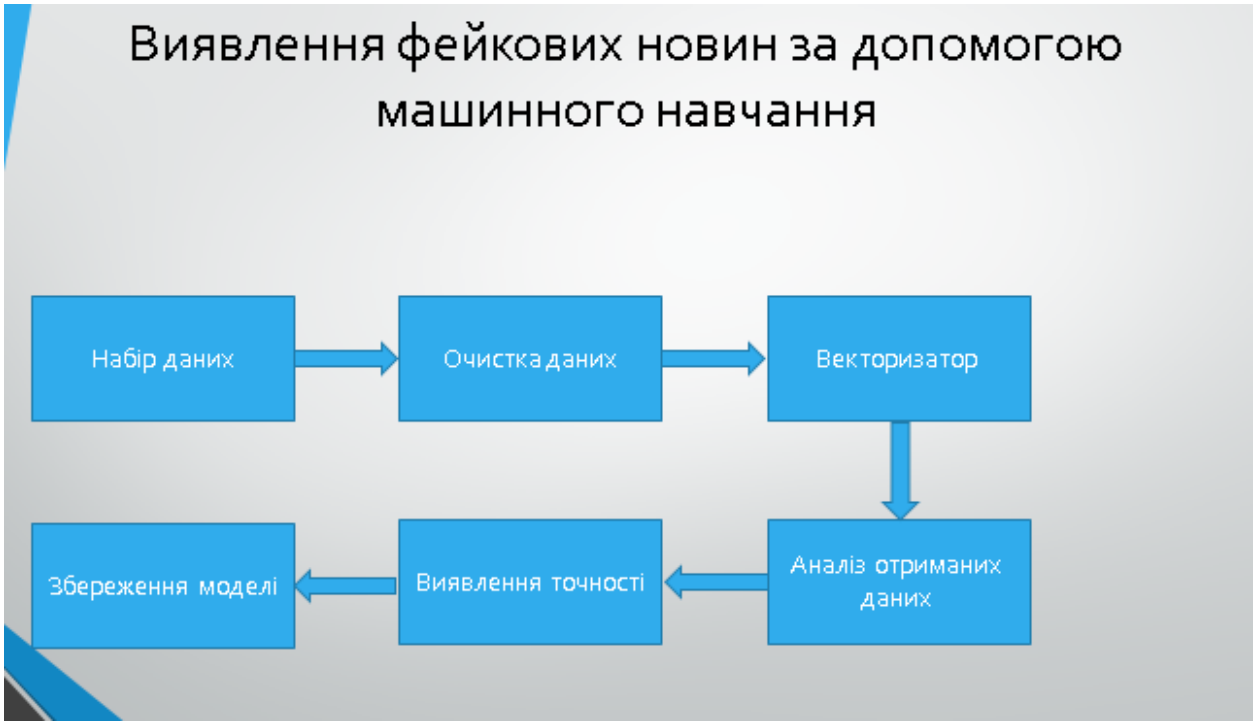
Рішення аналогів


```
"data":[
{
"Headline": "Facebook is spending $5.7 billion to capitalize on internet boom",
"Author": [ "Cnn Business"],
"Datepublish": "2022-04-22 04:33:39",
"Datemodify": "None",
"Imageurl": "None",
"Content": "Hong Kong (CNN Business) Facebook (FB) is spending billions of dollars for

"Url": "https://www.cnn.com/2020/04/22/tech/facebook-reliance-jio/index.html"
}
]
```


Порівняння результатів





Результати роботи моделі

	precision	recall	f1-score	support
0	0.96	0.95	0.96	13385
1	0.96	0.96	0.96	14615
accuracy			0.96	28000
macro avg	0.96	0.96	0.96	28000
weighted avg	0.96	0.96	0.96	28000

	precision	recall	f1-score	support
0	0.86	0.89	0.88	3213
1	0.90	0.87	0.88	3522
accuracy			0.88	6735
macro avg	0.88	0.88	0.88	6735
weighted avg	0.88	0.88	0.88	6735

Висновки

- Для досягнення поставленої мети кваліфікаційної роботи було проаналізовано методи парсингу новин, існуючі програмні забезпечення, виявлені їх недоліки та переваги, змодельовано системута вибрані оптимальні інструменти для розробки. Був забезпечений експорт даних, розроблена модель для вилучення необхідної інформації, розроблений та протестований модуль для аналізу отриманих даних на їх правдивість.
- Результатом цих заходів стало створення програмного забезпечення пасрингу веб-сайтів новин. Розроблене програмне забезпечення у повній мірі використовує основні можливості новітніх технологій для вирішення прикладної задачі.
- На підставі проведеного дослідження була запропонована та розроблена програма, парсингу та аналізу веб-сайтів новин з можливістю чіткого виявлення статей, отримання всієї доступної та необхідної інформації про неї, можливістю зберігання інформації у різних форматах для експорту в інші програми. Також було розроблено аналіз статей на їх правдивість за допомогою машинного навчання.
- В ході дослідження була проведена оцінка роботи програми, а саме, її час роботи, точність, та обе'м отриманої інформації, яка показала свою високу ефективність та покращила результати своїх аналогів у півтора рази.
- Функціональні можливості спроектованого та розробленого веб-застосунку можуть бути доопрацьовані та розширені у відповідності до вимог тієї сфери, де буде застосовано розроблену систему.