МАТЕРІАЛИ XXVII МІЖНАРОДНОГО МОЛОДІЖНОГО ФОРУМУ

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

ХАРКІВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ РАДІОЕЛЕКТРОНІКИ

РАДІОЕЛЕКТРОНІКА ТА МОЛОДЬ У XXI СТОЛІТТІ

2023

ТОМ 4

ХАРКІВ

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ХАРКІВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ РАДІОЛЕКТРОНІКИ

МАТЕРІАЛИ 27-го МІЖНАРОДНОГО МОЛОДІЖНОГО ФОРУМУ
«РАДІОЕЛЕТРОНІКА І МОЛОДЬ У XXI СТОЛІТТІ»

10 – 12 травня 2023 р.

Том 4

КОНФЕРЕНЦІЯ

«ПЕРСПЕКТИВИ РОЗВИТКУ ІНФОКОМУНІКАЦІЙ ТА ІНФОРМАЦІЙНО-ВИМІРЮВАЛЬНИХ ТЕХНОЛОГІЙ»

Харків 2023

УДК 004:[621.317+621.391](06)

27-й Міжнародний молодіжний форум «Радіоелектроніка та молодь у XXI столітті». Зб. Матеріалів форуму. Т.4. – Харків: ХНУРЕ. 2023. – 192 с.

В збірник включені матеріали 27-го Міжнародного молодіжного форуму «Радіоелектроніка і молодь у XXI столітті».

Видання підготовлено факультетом інфокомунікацій
Харківського національного університету радіоелектроніки

61166 Україна, Харків, просп. Науки, 14
тел./факс.: (057) 7021397

E-mail: mref21@nure.ua

Програмний комітет конференції

Снігуров А.В. к.т.н., декан факультету ІК
Безрук В.М. д.т.н, зав. каф. ІМІ
Лемешко О.В. д.т.н., зав. каф. ІКІ
Захаров І.П. д.т.н., зав. каф. ІВТ

УДК 004.91

ВИЛУЧЕННЯ ТЕКСТУ З ІНТЕРНЕТУ НА ОСНОВІ НАВЧАННЯ МАШИН

Шалатов В.О.
Науковий керівник – к.т.н., доц. Кривенко С.А.
Харківський національній університет радіоелектроніки, каф. ІМІ
м. Харків, Україна
тел. +38(012) 345-67-89, e-mail vasyl.shalatov@nure.ua,

In this work, Beautiful Soup is used to extract the titles, authors, summaries, published data, and hyperlinks from blog posts. The extracted text could then be used in a downstream NLP task, such as topic extraction, sentiment analysis, text-to-speech, or translation.

Першим етапом програми NLP є завантаження та обробка тексту. Цей етап можна розглядати як такий, що складається з трьох під етапів. По-перше, необхідно отримати дані з джерел даних. Наприклад, можна отримати текст із веб-сайтів або інших веб-ресурсів. Цей процес відомий як веб-збирання. Якщо виконується завантаження даних з документів, необхідно перетворити їх у форму, яку вимагає застосований компонент завантаження. Для більшості реальних програм є бажання автоматизувати процес вилучення. Витягнувши текст, його можна завантажити в конвеєр перетворення. Цей процес можна виконати за допомогою бібліотек Python, але також можна автоматизувати процес за допомогою Amazon Texttract. Нарешті, можна перетворити текст у числове представлення для використання обраної моделі навчання машин (ML). Метою даної роботи є розробка моделі використання Beautiful Soup [1], щоб видобувати заголовки, авторів, резюме, опубліковані дані та гіперпосилання з публікацій блогу. Щоб потім витягнутий текст можна було б використати в подальших завданнях NLP, таких як виділення теми, аналіз настроїв, перетворення тексту в мовлення або переклад. Повідомлення в блозі, яке аналізувалося, є блогом навчання машин AWS [2].

За допомогою веб-браузера була відкрита сторінка AWS Machine Learning. Використовувався режим інспектора браузера, щоб дізнатися структуру сторінки. У Mozilla FireFox і Google Chrome можна відкрити інспектор, натиснувши CTRL+SHIFT+C. Якщо використовується інший браузер, необхідно звертатися до документації браузера.

Були переглянуті різні елементи веб-сторінки, переміщаючи вказівник на сторінку. Переміщенням вказівника на наступні елементи було визначено, чи можна знайти теги, які використовуються для ідентифікації інформації: заголовок публікації в блозі; автор; дата публікації; короткий текст; гіперпосилання на публікацію в блозі.

Покрокова методика пошуку тегів наведена нижче.

Код статусу HTTP дорівнює 200, це передумова виконання наступних

кроків.

Вміст зі сторінки **content** був завантажений в об'єкт **soup**.

Всю сторінка доступна для перегляду за допомогою функції *soup.prettify()*.

Примітка. Вміст зі сторінки блогів AWS може бути довгим. Щоб перейти до наступного завдання, необхідно прокручувати блокнот **JupyterLab** вниз.

До всіх елементів сторінки можна отримати доступ за допомогою крапкової нотації (.). Таким чином, щоб переглянути заголовок, можна використовувати **soup.title**. Якщо потрібен лише текст, можна використовувати текстовий елемент **soup.h2.text**. **Best Egg** досяг утричі швидшого навчання моделі ML за допомогою автоматичного налаштування моделі **Amazon SageMaker.** Коли використовувався інспектор для пошуку тегів на сторінці блогів AWS, було виявлено, що вміст публікації в блозі впорядковано **organized/categorized/marked** позначено тегами <article>, які вказують на окрему одиницю вмісту.

Заголовок можна знайти на **soup.article.h2.span**. Щоб відобразити лише текст, використалась властивість *text*. Дата публікації статті знайдена за допомогою: *soup.article.time.text*. Далі короткий зміст статті витягнутий за допомогою: *soup.article.section.p.text*. Прізвище автора вказано у нижньому колонтитулі. Допис у блозі може мати кількох авторів. Однак спочатку було отримано лише першого автора: *soup.article.footer.span.prettify().*

Гіперпосилання на повний текст статті є останньою інформацією, яка була знайдена: *soup.article.a['href'].* Тепер коли були визначили всі відповідні елементи. Можна знайти всі статті за допомогою функції find_all(). Визначивши формат даних, можна додати результати до масиву. Далі був завантажений масив у фрейм даних **pandas**. Стовпець **published** тобто значення дати й часу були перетворені за допомогою метода *to_datetime().*

Ширину стовпця було налаштовано для **pandas** і відображені перші п'ять рядків фрейму даних.

Тепер, коли дані знаходяться у фреймі даних **pandas**, їх можна використовувати у наступних завданнях NLP.

Список використаних джерел

1. Beautiful Soup Documentation [Online]. Available: https://www.crummy.com/software/BeautifulSoup/bs4/doc/. [Accessed 24 03 2023].

2. AWS Machine Learning Blog [Online]. Available: https://aws.amazon.com/blogs/machine-learning/. [Accessed 24 03 2023].

АЛФАВІТНИЙ ПЕРЕЛІК