УДК 004.9+556

К.О. Бондалєтов[1]
В.Б. Мокін[1]
М.В. Григорчук[1]
С.В. Джура[1]
М.О. Кищук[1]
О.В. Нереуцький[1]
С.Д. Неволя[1]
А.М. Фурман[1]
В.В. Гіжевський[1]

ПОБУДОВА ДАТАСЕТУ ДЛЯ ТРЕНУВАННЯ ІНТЕЛЕКТУАЛЬНИХ МОДЕЛЕЙ ВЕБ-СИСТЕМИ З ІНФОРМАЦІЄЮ ПРО ЕКОЛОГІЧНІ ПРОБЛЕМИ ТА ЗАХОДИ У МАСИВАХ ВОД БАСЕЙНУ Р. ПІВДЕННИЙ БУГ WISEST-SBB

[1]Вінницький національний технічний університет, Україна

Abstract

Робота присвячена побудові датасету для тренування інтелектуальних моделей веб-системи з інформацією про екологічні проблеми та заходи у масивах вод басейну р. Південний Буг «WISEST-SBB». Наведено розроблений шаблон, який наповнювали автори з використанням автоматичних та напівавтоматичних способів веб-скрапингу, парсингу та класифікації даних. Охарактеризовано результати випробування на цього датасеті ряду інтелектуальних засобів для забезпечення автоматичності цих процесів у подальшого.

Ключові слова: NLP, Python, геоприв'язування даних, україномовний природномовний текст, штучний інтелект, масив вод, р. Південний Буг.

Abstract

The work is devoted to the construction of a dataset for training intelligent models of the web system with information about environmental problems and measures in the water bodies of the Southern Bug River basin "WISEST-SBB". The developed template, which was filled by the authors using automatic and semi-automatic methods of web scraping, parsing and data classification, is given. The results of the test on this dataset of a number of intelligent tools to ensure the automaticity of these processes in the future are characterized.

Keywords: NLP, Python, data geobinding, Ukrainian natural language text, artificial intelligence, water body, Southern Bug River.

Вступ

За рішенням басейнової ради Південного Бугу (протокол № 12 від 7 грудня 2022 року) на кафедрі системного аналізу та інформаційних технологій Вінницького національного технічного університету командою під керівництвом професора Мокіна В.Б. ведеться створення інтелектуальної веб-системи з інформацією про екологічні проблеми, природоохоронні заходи тощо про масиви вод басейну р. Південний Буг «WISEST Southern Bug Basin» (скорочено – «WISEST-SBB» або «WISESTR-SBB» – англ.: Water Information SystEm with Spatial and Temporal references for the Southern Bug Basin – «Водна інформаційна система з просторовою і часовою прив'язкою для басейну Південного Бугу») (http://wisestr.ai/). Така система дозволить здійснювати підтримку прийняття рішень під час розроблення та супроводу плану управління річковим басейном Південного Бугу [1, 2].

Наповнення системи здійснюється в напівавтоматичний спосіб, коли частина інформації знаходиться вручну, потім по ній тренуються інтелектуальні засоби для її пошуку та розмітки. Далі вручну перевіряється результат та удосконалюється. Потім – знов автоматичний етап і так – в циклі. Запорукою успіху є створення базового датасету для тренування цих інтелектуальних засобів.

Метою даного дослідження було створення датасету для тренування інтелектуальних моделей веб-системи з інформацією про екологічні проблеми та заходи у масивах вод басейну р. Південний Буг WISEST-SBB.

Систематизація інформації

Для розв'язання поставленої задачі Бондалєтовим К.О. та Мокіним В.Б. було знайдено 20 перших фактів про стан масивів вод і на них відпрацьовано, по-перше, шаблон для їх систематизації, а по-друге, інтелектуальні засоби для їх оброблення з використанням зібраного раніше досвіду [3-5].

Було прийнято рішення робити шаблон на базі MS Excel (рис. 1).

id	text	link	source	source_type	geonames_text	geodata_text_other	date_doc	dates_text	dates	dates_text_other	status	issues	measures	comments
3	Екологи припустили чому в Південному Бузі гине риба. Хмельницькі екологи перевірили рівень кисню та поспілкувались з місцевими рибалками. Кажуть, риба точно гине не через брак кисню. Дізнавайтеся яку причину мору назвали працівники управління екології та що кажуть хмельничани. Велику рибину пузом догори дніми знову опублікували в соцмережах. Нагадаємо, перед цим в Південному Бузі відібрали проби води на дослідження. В обласному центрі контролю та профілактики хвороб повідомили про результати лабораторних досліджень. В управлінні з питань екології та контролю за благоустроєм, кажуть також провели всі можливі перевірки стану водойми та прибережної смуги. Відтак, стверджують, що кисню в Південному Бузі вистачає. — Після останньої нічної перевірки кисню у річці, який був на рівні 6-7 мгО2/дм3 в усіх місцях проб (відповідно до вимог вміст розчиненого кисню повинен становити не менше 4,0 мгО2/дм3), можемо констатувати, що риба	https://vsim.ua/Podi/ekologi-pripustili-chomu-v-pivdennomu-buzi-gine-riba-11638925.html	Веб-сторінка новин Хмельниччини	сайт	Південний Буг, м. Хмельницький		28.07.2022		28.07.2022		1	1	1	
4	У Миколаївській області відбувається масовий мор риби: зникла частина русла Південного Бугу. Місцева влада та жителі вже почали вирішувати проблему. У місті Первомайськ Миколаївської області на річці Південний Буг розгортається екологічна катастрофа. Відбувається масовий мор риби. За її словами, через ремонт місцевої дамби за кілька днів буквально зникла частина русла річки. "Що з рибою добре видно на фото, люди залишаться без води (з якої в місті і район і без того самі проблеми)", — написала Мікульська. Як додав інший місцевий житель Валерій Чекодонов, ситуація склалася дуже критична. "Зникнула велика кількість риби, раків, різних мікроорганізмів в річці тощо, що є екологічною катастрофою, додалась ще проблема з тим, що якщо рівень води впаде ще на 20 см, то пів міста може просто залишитись без води. Єдиним рішенням цього питання є підняття рівня води, але на греблях вище по течії не можуть надати нам потрібної кількості води і, навіть, якщо вони відкриють нам воду, то в кращому випадку вона буде йти декілька діб" — написав ви-	https://tsn.ua/ukrayina/u-mikolayevskiy-oblasti-vidbuvayetsya-masoviy-mor-ribi-znikla-chastina-rusla-pivdennogo-buga-1635808.html	Веб-сторінка служби ТСН	сайт	Південний Буг, м. Первомайськ, Миколаївська область		30.09.2020		30.09.2020		0	1	0	

Рисунок 1. Структура шаблону з розміткою датасету

Як видно на рис. 1, шаблон передбачає інформацію про сам текст, його джерело, прив'язку у просторі (до карти) та у часі та класифікацію за рядом ознак. Ще можливими є примітки, коли розробник датасету не впевнений в деяких своїх оцінках цих даних.

Для розв'язання поставленої задачі Бондалєтовим К.О. та Мокіним В.Б. було знайдено 20 перших фактів про стан масивів вод і на них відпрацьовано, по-перше, шаблон для їх систематизації, а по-друге, інтелектуальні засоби для їх оброблення з використанням зібраного раніше досвіду [3-5].

Тестування інтелектуальних засобів

Далі було перевірено яким чином розроблені Мокіним В.Б. та Бондалєтовим К.О. інтелектуальні засоби зможуть здійснити геоприв'язку з використанням бібліотеки SpaCy та класифікацію знайденої інформації з використанням мовних моделей BERT та різних моделей машинного навчання для класифікації даних (рис. 2).

```
Score = 0.333 for WB code UA_M5.4_0002 and maybe: UA_M5.4_0002 UA_M5.4_0003 UA_M5.4_0004 UA_M5.4_
0011 UA_M5.4_0013 UA_M5.4_0019 UA_M5.4_0028 UA_M5.4_0969 UA_M5.4_0970
['Південний Буг', 'Олександрівське водосховище'] 0.16666666666666663
['Південний Буг', 'Олександрівське водосховище'] 0.21428571428571422
['Південний Буг', 'Олександрівське водосховище'] 0.11111111111111109
['Південний Буг', 'Олександрівське водосховище'] 0.09999999999999996
['Південний Буг', 'Олександрівське водосховище'] 0.03448275862068965
['Південний Буг', 'Олександрівське водосховище'] 0.0
['Південний Буг', 'Олександрівське водосховище'] 0.046874999999999986
['Південний Буг', 'Олександрівське водосховище'] 0.0
['Південний Буг', 'Олександрівське водосховище'] 0.06730769230769229
['Південний Буг', 'Олександрівське водосховище'] 0.08535533905932736
['Південний Буг', 'Олександрівське водосховище'] 0.07142857142857141
```

Рисунок 2. Приклад роботи NLP-програми на Python з автоматичним прив'язуванням інформації до масивів вод

Аналіз показав, що потрібно удосконалювати інтелектуальні засоби, враховуючи певні особливості української мови та певну стилістику подання новин, щоб підвищити точність автоматичної геоприв'язки та класифікації новин та іншої інформації.

Висновки

Робота присвячена побудові датасету для тренування інтелектуальних моделей веб-системи з інформацією про екологічні проблеми та заходи у масивах вод басейну р. Південний Буг «WISEST-SBB». Наведено розроблений шаблон, який наповнювали автори з використанням автоматичних та напівавтоматичних способів веб-скрапингу, парсингу та класифікації даних. Охарактеризовано результати випробування на цього датасеті ряду інтелектуальних засобів для забезпечення автоматичності цих процесів у подальшого.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Водний кодекс України : Кодекс України від 06.06.1995 р. № 213/95-ВР : станом на 19 серп. 2022 р. URL: https://zakon.rada.gov.ua/laws/show/213/95-вр#Text (дата звернення: 07.06.2023).
2. Про затвердження Порядку розроблення плану управління річковим басейном : Постанова Каб. міністрів України від 18.05.2017 р. № 336. URL: https://www.kmu.gov.ua/npas/249999756 (дата звернення: 04.06.2023).
3. Мокін В. Б. Інформаційна інтелектуальна технологія автоматизованої геоприв'язки екологічної текстової природно-мовної інформації / В. Б. Мокін, М. А. Гораш, Є. М. Крижановський, Т. Є. Вуж // Наукові праці ВНТУ [Електронний ресурс]. — 2020. — № 4. — Режим доступу: https://praci.vntu.edu.ua/index.php/praci/article/view/624
4. Vitalii Mokin. NLP for UA : BERT CLS & 10 Classifiers. Kaggle: Your Machine Learning and Data Science Community. URL: https://www.kaggle.com/code/vbmokin/nlp-for-ua-bert-cls-10-classifiers (date of access: 07.06.2023).
5. Vitalii Mokin. Kostiantyn Bondaletov. SpaCy for ukrainian text similarity. Kaggle: Your Machine Learning and Data Science Community. https://www.kaggle.com/code/bondaletov/spacy-for-ukrainian-text-similarity (date of access: 07.06.2023).

Бондалєтов Костянтин Олегович — аспірант кафедри системного аналізу та інформаційних технологій; e-mail: bondaletov.k@gmail.com;

Мокін Віталій Борисович — д-р техн. наук, професор, завідувач кафедри системного аналізу та інформаційних технологій; e-mail: vbmokin@vntu.edu.ua;

Марина Василівна Григорчук — студент кафедри системного аналізу та інформаційних технологій; hruhorchuk08@gmail.com;

Сергій Вікторович Джура — студент кафедри системного аналізу та інформаційних технологій; dzurasergij4@gmail.com;

Максим Олександрович Кищук — студент кафедри системного аналізу та інформаційних технологій; makskishuk.78@gmail.com;

Олег Віталійович Нереуцький — студент кафедри системного аналізу та інформаційних технологій; onereutsk@gmail.com;

Сергій Дмитрович Неволя — студент кафедри системного аналізу та інформаційних технологій; nevolya2003@gmail.com;

Анна Михайлівна Фурман — студент кафедри системного аналізу та інформаційних технологій; furmanania1@gmail.com;

Владислав Віталійович Гіжевський - студент кафедри системного аналізу та інформаційних технологій; vladgiz2000@gmail.com.

Bondalietov Kostiantyn O. — Post-graduate student of the Chair of System Analysis and Information Technology, e-mail: bondaletov.k@gmail.com;

Mokin Vitalii B. — Dr. Sc. (Eng.), Professor, Head of the Chair of System Analysis and Information Technology, e-mail: vbmokin@vntu.edu.ua;

Hryhorchuk Maryna V. — student of Vinnytsia National Technical University, e-mail: hruhorchuk08@gmail.com;

Dzhura Serhii V. — student of Vinnytsia National Technical University, e-mail: dzurasergij4@gmail.com;

Kyshchuk Maksym O. — student of Vinnytsia National Technical University, e-mail: makskishuk.78@gmail.com;

Nereutskyi Oleh V. — student of Vinnytsia National Technical University, e-mail: onereutsk@gmail.com;

Nevolia Serhii D. — student of Vinnytsia National Technical University, e-mail: nevolya2003@gmail.com;

Furman Anna M. — student of Vinnytsia National Technical University, e-mail: furmanania1@gmail.com.

Hizhevskyi Vladyslav V. — student of Vinnytsia National Technical University, e-mail: vladgiz2000@gmail.com.