УДК 004.8+004.93

Віталій Мокін, д.т.н., проф., Микола Гораш, асп., Дмитро Пасічнюк, студ., Олександр Радецький, студ.

КОНЦЕПЦІЯ ІНТЕЛЕКТУАЛЬНОЇ NLP ТЕХНОЛОГІЇ ДЛЯ ГЕОПРИВ'ЯЗКИ ТА КЛАСИФІКАЦІЇ ВІДКРИТОЇ ТЕКСТОВОЇ ІНФОРМАЦІЇ ПРО МАСИВИ ВОД

Головна мета сучасної водної політики в Європі в цілому та в Україні зокрема – це забезпечення доброго екологічного стану вод в кожному водному тілі (в Україні законодавчо затверджено термін «масив вод»), тобто водному об'єкті чи ділянці річки з водозбірною площею, не меншою 10 км^2. Для кожного масиву вод слід визначити основні проблеми, які заважають досягти добрий екологічний стан або його стабілізувати у найближчій перспективі [1, 2].

Постановка задачі. Метою даного дослідження є розроблення концепції інформаційної інтелектуальної технології, яка забезпечить максимальну автоматизацію геоприв'язки та класифікації відкритої текстової інформації про масиви вод, а також створення пілотної версії для окремих етапів цієї технології.

Опис задачі

Процес, який слід автоматизувати, складається з ряду етапів: зібрати наявну інформацію про масиви вод в Україні і перевести її у формат відкритих текстових даних (ВТІ); здійснити її геоприв'язку, тобто визначити яка інформація стосується якого масиву вод; виявити серед неї усе, що стосується різних екологічних проблем, та класифікувати їх; зберегти результат у геоінформаційній системі, яка забезпечить можливість виведення і візуалізації виявлених у різний час екологічних проблем по кожному масиву вод. Причому, якщо для масивів вод, розташованих у межах обласних центрів, інформації у вигляді різноманітних звітів та публікацій чи новин надто багато, то для витоків малих річок біля лише декількох сіл, її вкрай мало і вона розпорошена по багатьох різних відомствах і веб-ресурсах. Зазвичай, цю задачу розв'язують вручну, але ми пропонуємо використати сучасні технології обробки природної мови (NLP) [3].

Розв'язання задачі

Розв'язувати поставлену задачу пропонується у ряд етапів. Етап 1. Формування опису кожного масиву вод у вигляді набору географічних назв чи понять та формування з них єдиного геопросторового словника. Етап 2. Пошук ВТІ та її прив'язка у часі та геоприв'язка кожного речення (чи блоку тексту) у ній до сутностей геопросторового словника з етапу 1 за допомогою технології NER (Named-Entity Recognition). Етап 3. Виявлення і класифікація екологічних проблем у ВТІ, яку вдалось прив'язати, передусім геопросторово, на етапі 2. Етап 4. Формування єдиної бази даних з усієї успішно класифікованою ВТІ на цих етапах, прив'язаної до геоінформаційної системи, по якій потім можна швидко робити пошук інформації та її аналіз у просторі та часі.

Ми провели попереднє дослідження усіх цих етапів і відібрали оптимальні моделі, бібліотеки та підходи для їх автоматизації. Одна з ключових проблем полягає в тому, що NLP є найбільш розвиненими лише для англійської мови. Для української мови є мультилінгвістичні моделі, наприклад bert-multilingual-cased та їх модифікації, але вони є менш точними. Ми переконались в цьому під час створення пілотної версії. Для експерименту була використана монографія одного з авторів дослідження, видана і англійською [1], і українською [2] мовами. На основі неї авторами було створено датасет з реченнями, які характеризували і не характеризували певні екологічні проблеми. На основі BERT-моделей з післяобробленням логістичною регресією з використанням бібліотек Torch та Transformers було написано програму на Python, яка показала для англомовного датасету значно вищу точність (best_score) 93.6%, ніж для україномовного.

Висновки.Розроблено концепцію інформаційної інтелектуальної технології, яка забезпечить максимальну автоматизацію геоприв'язки та класифікації відкритої текстової інформації про масиви вод. Вибрано оптимальні технології та на за їх допомогою створено пілотні версії програм, випробувані на авторському україно-англомовному датасеті для басейну р. Південний Буг.

Література

1. Pivdenny Bug River Basin Management Plan: River Basin Analysis and Measures (Summary) / Afanasiev S., Peters A., Iarochevitch O., Mokin V. etc., K.: Interservice publishing house, 2014, 188.
2. План управління річковим басейном Південного Бугу: аналіз стану та першочергові заходи / Афанасьєв С., Петерс А., Ярошевич О., Мокін В. та ін., К.: ТОВ «НВП «Інтерсервіс», 2014, 188.
3. Steven B. Natural Language Processing with Python / B. Steven, K. Ewan.. 2009 – 504 с.