Облікова картка дисертації

I. Загальні відомості

Державний обліковий номер: 0823U100710

Особливі позначки: відкрита

Дата реєстрації: 27-09-2023

Статус: Захищена

Реквізити наказу МОН / наказу закладу:

II. Відомості про здобувача

Власне Прізвище Ім'я По-батькові:

1. Шаптала Роман Віталійович

2. Roman V. Shaptala

Ідентифікатор ORCHID ID: 0000-0002-4367-5775

Вид дисертації: доктор філософії

Шифр наукової спеціальності: 122

Назва наукової спеціальності: Комп'ютерні науки

Галузь / галузі знань: інформаційні технології

Освітньо-наукова програма зі спеціальності: Комп'ютерні науки

Дата захисту: 11-09-2023

Спеціальність за освітою: Комп'ютерні науки та інформаційні технології

Місце роботи здобувача:

Код за ЄДРПОУ:

Місцезнаходження:

Форма власності:

Сфера управління:

Ідентифікатор ROR: Не застосовується

Сектор науки: Не застосовується

III. Відомості про дисертацію

Шифр спеціалізованої вченої ради (разової спеціалізованої вченої ради): ДФ 26.002.31; ID 2014

Повне найменування юридичної особи: Національний технічний університет України "Київський політехнічний інститут імені Ігоря Сікорського"

Код за ЄДРПОУ: 02070921

Місцезнаходження: проспект Берестейський, буд. 37, Київ, 03056, Україна

Форма власності: Державна

Сфера управління: Міністерство освіти і науки України

Ідентифікатор ROR:

Сектор науки: Університетський

IV. Відомості про підприємство, установу, організацію, в якій було виконано дисертацію

Повне найменування юридичної особи: Національний технічний університет України "Київський політехнічний інститут імені Ігоря Сікорського"

Код за ЄДРПОУ: 02070921

Місцезнаходження: проспект Берестейський, буд. 37, Київ, 03056, Україна

Форма власності: Державна

Сфера управління: Міністерство освіти і науки України

Ідентифікатор ROR:

Сектор науки: Університетський

V. Відомості про дисертацію

Мова дисертації: Українська

Коди тематичних рубрик: 28.23, 28.23.37

Тема дисертації:

1. Класифікація документів на основі векторних представлень словників при обробці природної мови у малоресурсному середовищі

2. Dictionary embeddings for document classification in low-resource natural language processing

Реферат:

1. Метою дисертаційного дослідження є розробка та вдосконалення методів класифікації документів, написаних природною мовою, у малоресурсному середовищі за допомогою побудови векторних графових представлень словників природної мови. Зважаючи на те, що 63% контенту Інтернету написано англійською мовою, і більшість мов представлена менш ніж 1% веб-сторінок, величезна кількість мов є малоресурсними та, відповідно, менш дослідженими з точки зору підходів до обробки природних мов. Це призводить до того,

що інформаційні системи, які вимушені працювати на основі малопредставлених мов, часто потерпають від низької якості, порівняно з їх англомовними аналогами. Тому, покращення вже існуючих та розробка нових методів обробки природної мови у малоресурсному середовищі є актуальною задачею. Результати проведених досліджень показали, що векторні представлення словників на основі методів кодування вершин графів можна поєднувати з типовими векторними представленнями документів для покращення якості класифікації документів за допомогою підходів машинного навчання. Кожен крок запропонованого методу має набір параметрів та гіперпараметрів, від яких залежить результат та ефективність фінального рішення. Тому додатково наведено аналіз даних опцій, а також порівняння різних підходів до побудови представлень вершин графів у контексті словників. Для досягнення найкращих результатів пропонується використання методу на основі випадкових блукань - Node2Vec, який перетворює елементи словника у вектори за прийнятний час, не вимагає багато ресурсів та отримує вищі оцінки при подальшій класифікації документів. Для наступного кроку, а саме злиття векторних представлень документів та словникової інформації оптимальним виявився метод зваженої суми. Додатково наводяться практичні рекомендації по роботі з подібними даними, а саме особливості отримання, збереження та передобробки документів, побудови словників для кожного з методів класифікації документів, збереження та обробки словника синонімів, а також аналіз статистичної значущості результатів. Наукова новизна одержаних результатів полягає у наступному: Вперше запропоновано метод класифікації документів на основі векторних представлень словників при обробці природної мови у малоресурсному середовищі, який відрізняється від методів доповнення даних, що базуються на словниках, тим що у ньому поєднуються векторні представлення документів з векторними представленнями елементів лінгвістичних словників, що дозволяє збільшити F1-міру якості класифікації документів у малоресурсному середовищі; Запропоновано векторну модель слів зі словника синонімів, яка на відміну від інших будується на основі векторних представлень вузлів графу словника, що надає можливість її повторного використання в різних задачах обробки природної мови через трансферне навчання; Модифіковано методи конкатенації та зваженої суми при злитті векторних представлень слів додаванням етапу пошуку відповідності слів з документу словам з словника синонімів, що дозволяє покрити відсутні у словнику словоформи без побудови моделей визначення частини мови та пошуку словоформ, що суттєво ускладнено у малоресурсних середовищах. Практичне значення одержаних результатів полягає у тому, що: Розроблений метод дозволяє значно підвищити F1-міру якості систем класифікації документів у малоресурсних середовищах. Таким чином розробники даних систем можуть зменшити час та витрати на розробку, адже вища якість системи досягатиметься з меншою кількістю розмітки, розширення якої може бути не доступним, або вимагати додаткових часових чи фінансових інвестицій; Розроблено векторні представлення слів у словнику синонімів української мови, які можна перевикористовувати за допомогою трансферного навчання при створенні програмних систем у інших прикладних областях; Представлено набір даних для класифікації тем петицій, націлений на тестування методів обробки природної мови у малоресурсному середовищі. Документи написані українською мовою та мають вузьку урбаністичну спеціалізацію, що робить набір даних відмінним від корпусів загального призначення; Запропоновано застосування розробленого методу до класифікації петицій до Київської міської ради за темами, яка дозволяє автоматично пропонувати тему петиції при ручній розмітці, що може суттєво скоротити час на їх аналіз.

2. The objective of this research is to develop and improve document classification methods in low-resource natural language processing through graph embeddings of linguistic dictionaries. Considering that 63% of the Internet is written in English, and most of natural languages are represented in less than 1% of all web pages, a lot of natural languages are considered low-resource, and are less researched in the field of natural language processing. This leads to information systems built to work with low-resource languages having lower quality than their English counterparts. Consequently, improving existing low-resource natural language processing methods and the development of new ones is a relevant research problem. The results of the research showed that vector representations of dictionaries based on graph node embedding methods can be combined with common vector representations of documents to improve the quality of document classification using machine learning

approaches. Each step of the proposed method has a set of parameters and hyperparameters, which the result and effectiveness of the final solution depend on. Therefore, an analysis of these options is additionally given, as well as a comparison of different approaches to the construction of graph node embeddings in the context of dictionaries. To achieve the best results, it is suggested to use random-walk based method - Node2Vec, which converts dictionary elements into vectors in an acceptable time, does not require a lot of resources, and receives higher F1-scores further down the pipeline – during document classification. For the next step, namely the fusion of vector representations of documents and dictionary information, the weighted sum method turned out to be better than concatination. In addition, practical recommendations for working with such data are provided, namely, the process of obtaining, saving and preprocessing documents for each of the proposed methods, saving and processing of a synonyms dictionary, as well as the analysis of statistical significance of the results. Scientific novelty of the results includes: For the first time, a method of document classification based on dictionary embeddings during low-resource natural language processing is proposed, which differs from dictionary-based methods of data augmentation in that it fuses vector representations of documents with vector representations of elements of linguistic dictionaries, which allows to increase F1-score of document classification in a low-resource environment; A vector model of words from the dictionary of synonyms is proposed, which, unlike others, is built on the basis of vector representations of the nodes of the dictionary graph, which makes it possible to reuse it in various tasks of natural language processing through transfer learning; The methods of concatenation and weighted sum during vector representations of words fusion have been modified by adding a stage of matching words from the document to words from the dictionary of synonyms, which allows for covering word forms missing from the dictionary without building models for part of speech tagging and word form generation, which is significantly complicated in low-resource environments. The practical significance of the results includes: The proposed method makes it possible to significantly increase the F1-score of document classification systems in low-resource environments. This way, developers of these systems can reduce development time and costs, because higher system quality will be achieved with less labeling, the process which may not be available or require additional time or financial investment; Vector representations of words in the dictionary of synonyms of the Ukrainian language were developed, which can be reused with the help of transfer learning when creating software systems in other applied areas; A data set for the classification of petition topics is presented, aimed at testing low-resource natural language processing methods. The documents are written in Ukrainian and have a narrow urban specialization, which makes the data set different from general-purpose corpora; It is proposed to apply the developed method to the topic classification of petitions to the Kyiv City Council, which allows for automatic suggestions of topic for the petition during manual labeling. This can significantly reduce the time for their analysis.

Державний реєстраційний номер ДіР:

Пріоритетний напрям розвитку науки і техніки: Інформаційні та комунікаційні технології

Стратегічний пріоритетний напрям інноваційної діяльності: Розвиток сучасних інформаційних, комунікаційних технологій, робототехніки

Підсумки дослідження: Нове вирішення актуального наукового завдання

Публікації:

- R. Shaptala and G. Kyselov, “Enhancing document representations with synonyms graph node embeddings,” J. Theor. Appl. Inf. Technol., vol. 100, no. 1, pp. 70–80, Jan. 2022.
- Р. Шаптала і Г. Кисельов, «Метод злиття багатомодальних векторних представлень слів у малоресурсному середовищі», ВОТТП, вип. 1, с. 174–179, Бер. 2023.
- Р. Шаптала і Г. Кисельов, “Класифікація текстових документів з використанням доповнення векторних представлень документів графовими представленнями елементів словника синонімів,” Інформаційні технології та суспільство, вип. 3 (5), с. 49–55, Січ. 2023.

- Р. Шаптала і Г. Кисельов, “Огляд методів злиття векторних представлень,” Телекомунікаційні та інформаційні технології, вип. 4 (77), с. 84–89, 2022.
- R. V. Shaptala and G. D. Kyselev, “Using graph embeddings for Wikipedia link prediction,” Bull. Natl. Tech. Univ. “KhPI”. Ser. Syst. Anal. Control Inf. Technol., vol. 0, no. 1 SE-INFORMATION TECHNOLOGY, pp. 48–52, Jul. 2019.
- Shaptala R.V. and Kyselov G.D., “Vector space models of Kyiv city petitions,” Sci. notes Taurida Natl. V.I. Vernadsky Univ. Ser. Tech. Sci., vol. 32, no. 1, pp. 169–177, 2021.
- A. Samvelyan, R. Shaptala, and G. Kyselov, “Exploratory data analysis of Kyiv city petitions,” in 2020 IEEE 2nd International Conference on System Analysis Intelligent Computing (SAIC), 2020, pp. 1–4.

Наукова (науково-технічна) продукція:

Соціально-економічна спрямованість:

Охоронні документи на ОПІВ:

Впровадження результатів дисертації: Впроваджено

Зв'язок з науковими темами: 0120U103046 0121U110624 0122U002655

VI. Відомості про наукового керівника/керівників (консультанта)

Власне Прізвище Ім'я По-батькові:

1. Кисельов Геннадій Дмитрович

2. Hennadii D. Kyselov

Ідентифікатор ORCHID ID: Не застосовується

Додаткова інформація:

Повне найменування юридичної особи: Національний технічний університет України "Київський політехнічний інститут імені Ігоря Сікорського"

Код за ЄДРПОУ: 02070921

Місцезнаходження: проспект Берестейський, буд. 37, Київ, 03056, Україна

Форма власності: Державна

Сфера управління: Міністерство освіти і науки України

Ідентифікатор ROR:

Сектор науки: Університетський

VII. Відомості про офіційних опонентів та рецензентів

Власне Прізвище Ім'я По-батькові:

1. Завгородній Валерій Вікторович

2. Valerii Zavgorodnii

Ідентифікатор ORCHID ID: Не застосовується

Додаткова інформація:

Повне найменування юридичної особи: Державний університет інфраструктури та технологій

Код за ЄДРПОУ: 41330257

Місцезнаходження: вул. Кирилівська, буд. 9, Київ, 04071, Україна

Форма власності: Державна

Сфера управління: Міністерство освіти і науки України

Ідентифікатор ROR:

Сектор науки: Університетський

Власне Прізвище Ім'я По-батькові:

1. Жебка Вікторія Вікторівна

2. Victoria V. Zhebka

Ідентифікатор ORCHID ID: 0000-0003-4051-1190

Додаткова інформація:

Повне найменування юридичної особи: Державний університет інформаційно-комунікаційних технологій

Код за ЄДРПОУ: 38855349

Місцезнаходження: вул. Солом'янська, буд. 7, Київ, 03680, Україна

Форма власності: Державна

Сфера управління: Міністерство освіти і науки України

Ідентифікатор ROR:

Сектор науки: Університетський

Власне Прізвище Ім'я По-батькові:

1. Шаповалова Світлана Ігорівна

2. Shapovalova I. Svetlana

Ідентифікатор ORCHID ID: Не застосовується

Додаткова інформація:

Повне найменування юридичної особи: Національний технічний університет України "Київський політехнічний інститут імені Ігоря Сікорського"

Код за ЄДРПОУ: 02070921

Місцезнаходження: проспект Берестейський, буд. 37, Київ, 03056, Україна

Форма власності: Державна

Сфера управління: Міністерство освіти і науки України

Ідентифікатор ROR:

Сектор науки: Університетський

Власне Прізвище Ім'я По-батькові:

1. Бідюк Петро Іванович

2. Petro I. Bidyuk

Ідентифікатор ORCHID ID: Не застосовується

Додаткова інформація:

Повне найменування юридичної особи: Національний технічний університет України "Київський політехнічний інститут імені Ігоря Сікорського"

Код за ЄДРПОУ: 02070921

Місцезнаходження: проспект Берестейський, буд. 37, Київ, 03056, Україна

Форма власності: Державна

Сфера управління: Міністерство освіти і науки України

Ідентифікатор ROR:

Сектор науки: Університетський

VIII. Заключні відомості

Власне Прізвище Ім'я По-батькові голови ради: Аушева Наталія Миколаївна

Власне Прізвище Ім'я По-батькові головуючого на засіданні: Аушева Наталія Миколаївна

Відповідальний за підготовку облікових документів: Шаптала Роман Віталійович

Реєстратор: УкрІНТЕІ

Керівник відділу УкрІНТЕІ, що є відповідальним за реєстрацію наукової діяльності

Юрченко Тетяна Анатоліївна