УДК 004.912

М. М. Стовбчатий

ЗАСТОСУВАННЯ МЕТОДІВ МАШИННОГО НАВЧАННЯ ДЛЯ ВИЗНАЧЕННЯ АВТОРСТВА УКРАЇНОМОВНОГО ТЕКСТУ

Вінницький національний технічний університет

Анотація

Запропоновано підхід до аналізу текстової інформації з метою визначення авторства україномовного тексту. Параметри для характеристики тексту являють собою опис дерев залежностей речень і отримується через україномовний парсер. Розвиток підходу передбачає застосування методів машинного навчання за допомогою лінгвістичного пакету NLTK, що дозволить автоматизувати процес ідентифікації автора.

Ключові слова: україномовний текст; визначення автора; парсер; граф; дерево залежностей; NLTK; машинне навчання.

Abstract

An approach to the analysis of textual information is proposed in order to determine the authorship of the Ukrainian-language text. Parameters for the description of the text are a description of dependency trees of the sentences and are obtained through the Ukrainian-language parser. The development of the approach involves the use of methods of machine learning using the linguistic package NLTK, which will automate the process of identification of the author.

Key words: Ukrainian text; definition of the author; parser; graph; dependency tree; NLTK; machine learning

Вступ

Стиль письма людини являє собою яскравий приклад поведінкової біометрії. Словниковий запас, який активно використовують люди, і те, як вони структурують свої повідомлення та речення тексту, характеризуються індивідуальними (якісними та кількісними) ознаками, тому їх часто можна використовувати для ідентифікації автора конкретного твору [1]. Задача визначення авторства тексту є актуальною у багатьох сферах діяльності людини, зокрема при виявленні плагіату, у розшукових заходах правоохоронних органів, для проведення лінгвістичної експертизи і літературознавчих досліджень тощо.

В сучасних умовах експерти-лінгвісти все частіше застосовують формальні методи і моделі, а також засоби ІТ для визначення авторства. Зокрема популярними є методи машинного навчання [2], особливо для англійської мови, оскільки вільно доступні англомовні лінгвістичні ресурси існують для більшості програмних платформ [3]. Проте надзвичайно актуальним можна вважати задачу ідентифікації автора україномовного тексту в зв’язку з відсутністю відповідних лінгвістичних пакетів вільного доступу.

Результати дослідження

На основі творчої співпраці з лабораторією комп'ютерної лінгвістики КНУ ім. Тараса Шевченка через вільний ресурс [4] було отримано csv файл, у якому представлені графи речень україномовного тексту у вигляді рядків. Загальний формат файлу, що пропонується – всі слова в тексті послідовно пронумеровані, між ними встановлені бінарні зв'язки: номер головного слова, номер підлеглого слова, номер типу зв'язку, а також номер тексту і номер речення. Нижче представлено приклади поширених типів зв'язку:

{ "ПМ","b75c03","сполука з цифрою"},
{ "ІС","ba00ff","іменникова сполука"},
{ "СУ","22d626","сурядна сполука"},
{ "ПП","0012ff","прийменникова сполука"},
{ "ДС","0090ff","дієслівна сполука"},
{ "КЗ","ff0000","координаційний зв'язок, сполука підмета і присудка"},
{ "ЗС","ffcc00","займенникова сполука"},
{ "ДЯ","d5ee48","займенник+дієслово (залежне слово у препозиції)"},
{ "АС","ff7800","прикметникова сполука"},
{ "ЧС","22d6d4","числівникова сполука"},
{ "РС","a60800","прислівникова сполука"},
{ "ФЗ","d48484","фразеологічна сполука".

За допомогою отриманої інформації такого типу пропонується виконати математичний опис дерев залежностей (графів) кожного з речень тексту. При цьому використовують такі діагностичні параметри:

- кількість слів в реченні;
- кількість рівнів в дереві;
- кількість гілок в дереві;
- кількість коренів в дереві;
- кількість простих речень в складному.

Вихідний файл розробленого програмного забезпечення узагальнює дані параметри за всіма реченнями для тексту в цілому. В даній роботі було проведено дослідження трьох авторів – Миколи Вінграновського, Івана Франко, Івана Нечуй-Левицького.

При аналізу твору Івана Нечуй-Левицького «Два брата» були отримані такі параметри (рис.1):

```
Кількість слів в реченні: 11.253061224489796
Кількість рівнів в дереві: 5.1204081632653065
Кількість гілок в дереві: 7.983673469387755
Кількість коренів в дереві: 5.020408163265306
Кількість простих речень в складному: 8.983673469387755
```

Рис. 1 – Математичні параметри Іван Нечуй-Левицького

При аналізу твору Миколи Вінграновського «Кінь на горі» були отримані такі чисельні параметри (рис.2):

```
Кількість слів в реченні: 18.73846153846154
Кількість рівнів в дереві: 4.1692307692307695
Кількість гілок в дереві: 8.461538461538462
Кількість коренів в дереві: 9.015384615384615
Кількість простих речень в складному: 9.461538461538462
```

Рис. 2 – Математичні параметри Миколи Вінграновського

При аналізу твору Івана Франка «Батьківщина» були отримані такі параметри (рис.3):

```
Кількість слів в реченні: 9.869700103412617
Кількість рівнів в дереві: 4.463288521199586
Кількість гілок в дереві: 7.334022750775595
Кількість коренів в дереві: 4.618407445708376
Кількість простих речень в складному: 8.334022750775594
```

Рис. 3. – Математичні параметри Івана Франка

Тепер задача визначення авторства твору зводиться до вибору платформи машинного навчання та проведення відповідних експериментів з метою автоматизації цього процесу. Пропонується обрати пакет NLTK, який є однією з провідних платформ для роботи з природно-мовною інформацією на основі мови програмування Python [3, 5]. Даний програмний пакет забезпечує зручні у використанні інтерфейси для більш ніж 50 корпусів і лексичних ресурсів, таких як WordNet, одночасно з набором бібліотек обробки тексту для класифікації, токенізації, лемізації, стемізації та інших інструментів машинного навчання і автоматизованого розв’язання задач комп’ютерної лінгвістики.

Висновки

Встановлено, що запропонований підхід дозволяє формально розв’язати задачу ідентифікації автора україномовного тексту за 5 чисельними параметрами, що характеризують речення тексту як граф залежностей. Для цього використовуються результати синтаксичного аналізу (парсерингу) тексту, отримані завдяки творчій співпраці з лабораторією комп'ютерної лінгвістики КНУ ім. Тараса Шевченка.

З метою автоматизації процесу визначення авторства пропонується застосувати популярну технологію, що підтримує сучасні методи машинного навчання: вільно доступний лінгвістичний пакет NLTK + мова програмування Python. При проведенні відповідних експериментів з пакетом NLTK,

порівняно невелике число параметрів (5 для кожного речення), ймовірно, доведеться збільшувати загальнотекстовими параметрами, що характеризують застосування автором властивих для нього типів синтаксичних зв'язків.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Ramyaa, Congzhou He, Khaled Rasheed. Using Machine Learning Techniques for Stylometry [Електронний ресурс]. – Режим доступу: https://www.cs.nmt.edu/~ramyaa/publications/ml_techniques_Stylometry.pdf.
2. Луис Педро Коэльо, Вилли Ричарт. Построение систем машинного обучения на языке Python / М., ДМК Пресс. – 2016. – 302 с.
3. Steven Bird Natural Language Processing with Python Analyzing Text with the Natural Language Toolkit / Steven B., Ewan K., Edward L // Sebastopol: O`REILLY. – 2010. – P. 504 – 512.
4. Mova.info [Електронний ресурс] лінгвістичний портал. – Режим доступу: http://www.mova.info/.
5. Natural Language Toolkit [Електронний ресурс]. – Режим доступу: https://www.nltk.org/.

Стовбчатий Максим Михайлович, студент групи 1АКІТ-18, факультет комп'ютерних систем та автоматики ВНТУ, м. Вінниця

Науковий керівник: ***Бісікало Олег Володимирович*** — д-р техн. наук, професор, декан факультету КСА, Вінницький національний технічний університет, м. Вінниця, e-mail: obisikalo@gmail.com