МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ТЕРНОПІЛЬСЬКИЙ НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ
ІМЕНІ ІВАНА ПУЛЮЯ
ФАКУЛЬТЕТ КОМП'ЮТЕРНО-ІНФОРМАЦІЙНИХ СИСТЕМ І ПРОГРАМНОЇ ІНЖЕНЕРІЇ

ВІВЧАРИК ВОЛОДИМИР МИХАЙЛОВИЧ

УДК 004.056

ВИЗНАЧЕННЯ АВТОРСТВА ДОКУМЕНТУ З ДОПОМОГОЮ МЕТОДІВ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ТЕКСТУ

125 «Кібербезпека»

Автореферат
дипломної роботи на здобуття
освітнього рівня «магістр»

Тернопіль
2019

Роботу виконано на кафедрі кібербезпеки Тернопільського національного технічного університету імені Івана Пулюя Міністерства освіти і науки України

Керівник роботи: доктор технічних наук, професор кафедри кібербезпеки
Карпінський Микола Петрович,
Тернопільський національний технічний університет імені Івана Пулюя,

Рецензент: доктор наук із соціальних комунікацій, професор кафедри комп'ютерних наук
Кунанець Наталія Едуардівна,
Тернопільський національний технічний університет імені Івана Пулюя

Захист відбудеться 23 грудня 2019 р. о $9^{.00}$ годині на засіданні екзаменаційної комісії №32 у Тернопільському національному технічному університеті імені Івана Пулюя за адресою: 46001, м. Тернопіль, вул. Руська, 56, навчальний корпус №1, ауд. 806

ЗАГАЛЬНА ХАРАКТЕРИСТИКА РОБОТИ

Актуальність теми роботи. Як відомо, проблема аутентифікації повідомлення, тобто підтвердження автора повідомлення, є одним з наріжних каменів криптографії, яка традиційно розв'язується алгоритмами цифрового підпису. Проте існує кардинально інша проблема ідентифікації та встановлення автору документа шляхом аналізу контенту документу. Ця проблема є менш поширеною, проте не менш важливою, адже задача ідентифікації автора розв'язується для боротьби з плагіатом, встановлення авторства анонімних текстів, програмного коду, шкідливого програмного забезпечення, експертизи та встановлення особистості в криміналістиці, запобігання злочинів та багатьох інших застосувань.

Мета роботи: створення програмного забезпечення для визначення автора документу на основі обґрунтованої моделі мульти-класифікації, дослідження точності моделі для різного розміру простору векторних ознак та визначення впливу на точність прогнозування різних попередніх етапів для нормалізації вхідних даних.

Об’єкт, методи та джерела дослідження. Об'єкт дослідження – процес визначення авторства для текстових документів. Предмет дослідження – моделі та методи для ідентифікації автора тексту.Методи дослідження: загальнонаукові методи пізнання як порівняльний та системний аналіз, методи машинного навчання та математичної статистики.

Наукова новизна отриманих результатів: в роботі проведено порівняльний аналіз точності прогнозування автору документу в залежності від частки обраних ключових слів від загальної кількості слів та досліджено вплив на точність таких методів нормування тексту, як видалення стоп-слів та стемінгу.

Практичне значення отриманих результатів полягає в тому, що було досліджено на практичних даних вплив розміру простору ознак та методів нормалізації тексту на точність класифікації для визначення авторства документу.

Апробація. Окремі результати роботи доповідались на VII науково-технічній конференції «Інформаційні моделі, системи та технології», Тернопіль, ТНТУ, 11 – 12 грудня 2019 р.

Структура роботи. Робота складається з розрахунково-пояснювальної записки та графічної частини. Розрахунково-пояснювальна записка складається з вступу, 7 частин, висновків, переліку посилань та додатків. Обсяг роботи: розрахунково-пояснювальна записка – 111 арк. формату А4, ілюстративна частина– 13 слайдів.

ОСНОВНИЙ ЗМІСТ РОБОТИ

У вступі сформульовано актуальність проблеми визначення авторства текстових документів та сформульовано мету і основні завдання роботи.

У першому розділі описується, проблема визначення авторства текстового документу, її підвиди та можливі застосування для практичних задач. Наводиться також класифікація методів дослідження стилю написання текстів та короткі теоретичні відомості про базові методи класифікації, які можуть бути використані для ідентифікації автора документу.

У другому розділі розглядаються особливості побудови та оцінки якості класифікаційних моделей визначення авторства документу.

Третій розділ – практична частина. У ньому описано деталі налаштування необхідних бібліотек Python, наведено лістинги основних етапів алгоритмів підготовки, опрацювання, нормалізації даних та власне класифікації. Проведено оцінку точності для різної розмірності простору ознак та різних методів нормалізації тексту.

В четвертому розділі наведено основні типи та структури даних Python, які використовувались в практичній реалізації запропонованої моделі.

У розділі "Обґрунтування економічної ефективності" обчислено собівартість та термін окупності проекту.

У шостому розділі описано інструкції з охорони праці при роботі з комп'ютером та фактори виробничого середовища і їх вплив на життєдіяльність людини.

В розділі "Екологія" описано питання зниження енергоємності та енергозбереження та індексний метод в екології.

У загальних висновках щодо дипломної роботи описано основні результати, отримані в роботі та сформульовано висновки, отримані в результаті проведення практичного експерименту.

В додатках до пояснювальної записки приведено тези.

В ілюстративній частині приведено Основні завдання визначення авторства, Практичне застосування, Місце задачі ідентифікації автора в розрізі інших суміжних наук, Види задач машинного навчання, Класифікаційна модель наївного Байєса, Етапи очищення та нормалізації тексту, Показники оцінки якості моделі, Критерії якості моделі класифікації, Порівняння показників точності моделі для різних налаштувань моделі, Порівняльний аналіз точності моделі для різної частки кількості термінів в словнику ознак, Висновки.

ВИСНОВКИ

У дипломній роботі розроблено програмне забезпечення для визначення авторства документу.

Загалом було вирішено наступні задачі:

- Проведено огляд літературних джерел в області дослідження.
- Досліджено основні завдання та сфери практичного застосування задачі визначення авторства.
- Обґрунтувано вибір моделі наївного класифікатора Байєса для задачі мульти-класифікації визначення авторства.
- Навчальний та тестовий набір даних взято з репозитарію машинного навчання UCI.
- Розроблене програмне забезпечення для реалізації обраної моделі з використанням мови програмування Python та бібліотек Pandas, Scikit-learn та NLTK/
- Проведено порівняльний аналіз точності класифікаційної моделі для різної розмірності простору ознак. Виявилось, що збільшення розмірності простору ознак не значно впливає на точність, але в значній мірі сповільнює алгоритм.

- Досліджено вплив на якість прогнозної моделі різних методів нормалізації текстових документів. Всупереч загально-прийнятій думці, що стемінг повинен значно покращити результати текстової класифікації, у даному випадку відбулось мізерне зростання точності.

СПИСОК ОПУБЛІКОВАНИХ АВТОРОМ ПРАЦЬ ЗА ТЕМОЮ РОБОТИ

1. Вівчарик В. Особливості методів аналізу текстів для ідентифікації авторства документу [текст] / Вівчарик В. Збірник тез VII науково-технічної конференції Тернопільського національного технічного університету імені Івана Пулюя «Інформаційні моделі, системи та технології» – Тернопіль (11 – 12 грудня 2019 р.), ТНТУ, 2019. – с.29.

АНОТАЦІЯ

В роботі досліджено основні завдання та можливі сфери застосування задачі визначення авторства деякого документу, обґрунтовано вибір моделі класифікації та програмного середовища Python для практичної реалізації методу визначення автора документу. Проведено тестування імплементованої класифікаційної моделі наївного Байєса для реальних даних, здійснено порівняння основних показників точності моделі для різного розміру простору ознак, проведено аналіз впливу на точність різних методів нормування текстових документів для задач класифікації.

Ключові слова: ІДЕНТИФІКАЦІЯ, ВИЗНАЧЕННЯ АВТОРА, МАШИННЕ НАВЧАННЯ, КЛАСИФІКАЦІЯ, НАЇВНИЙ КЛАСИФІКАТОР БАЙЄСА, СТОП-СЛОВА, СТЕМІНГ, ЛЕМАТИЗАЦІЯ.

ANNOTATION

The main tasks and possible application areas of author of a document determining problem are investigated; the choice of the classification model and the Python software environment for practical implementation of the method of the document authorship identification is substantiated. Tests of the implemented naive Bayes classification model of for real data were carried out. The basic criteria of model accuracy for different size of feature vector space were compared, the influence of different methods of text documents normalization on accuracy for classification problems was analyzed.

Key words: IDENTIFICATION, AUTHOR DETERMINING PROBLEM, MACHINE LEARNING, CLASSIFICATION, NAIVE BAYES CLASSIFIER, STOP-WORDS, STEMMING, LEMMATIZATION.