ХЕРСОНСЬКИЙ НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ
(повне найменування вищого навчального закладу)
ФАКУЛЬТЕТ ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ ТА ДИЗАЙНУ
(повне найменування інституту, назва факультету (відділення))
КАФЕДРА ПРОГРАМНИХ ЗАСОБІВ І ТЕХНОЛОГІЙ
(повна назва кафедри (предметної, циклової комісії))

Пояснювальна записка

до кваліфікаційної роботи
магістра
(освітній рівень)

на тему: «Дослідження методів класифікації новин за достовірністю з використанням методів веб-краулінгу та машинного навчання»

Виконав: студент групи 6ПР
спеціальності
121 - «Інженерія програмного забезпечення»
(шифр і назва спеціальності)

Серебрянський Владислав Вадимович
(прізвище та ініціали)
Керівник к.т.н., доцент Огнєва О.Є.
(прізвище та ініціали)
Рецензент _____
(прізвище та ініціали)

Херсон - 2021

РЕФЕРАТ

Загальний обсяг роботи 86 с., 15 рис., 4 табл., 1 додаток та 19 джерел.

Об’єкт дослідження – процес екстракції текстових даних з подальшою обробкою методами машинного навчання.

Предмет дослідження – методи та засоби екстракції та аналізу структурованих текстових даних.

Мета роботи – створення програмного інструментарію екстракції структурованих даних з веб-сторінок новинних ресурсів для подальшої класифікації за достовірністю.

Наукова новизна роботи полягає в тому, що створено простий жадібний алгоритм у якому суміщено процеси пошуку посилань та видобування інформації, доведено доцільність використання простих алгоритмів для збору даних з ресурсів у мережі Інтернет з ціллю використання у тренуванні алгоритмів машинного навчання.

Практичний результат роботи полягає в тому, що доведено здатність класичних алгоритмів навчання досягати результатів, співставних з такими у нейронних мережах, таких як мережі ДКЧП, та показано, що такі моделі здатні працювати на двомовних наборах даних.

Ключові слова: ВЕБ-СКРАПІНГ, ЕКСТРАКЦІЯ ДАНИХ З ВЕБ-СТОРІНОК, КРАУЛІНГ, ПОШУК ПОСИЛАНЬ, МАШИННЕ НАВЧАННЯ, КЛАСИФІКАЦІЯ НОВИН, ДОВГА КОРОТКОЧАСНА ПАМ’ЯТЬ, НЕЙРОННІ МЕРЕЖІ.

ЗМІСТ

ВСТУП 8

1. ЗАГАЛЬНІ ПОЛОЖЕННЯ 11

1.1. Опис предметного середовища 11

1.2. Веб-скрапінг 11

1.2.1. Загальні положення про веб-скрапінг 11

1.2.2. Концепція веб-скрапінгу 14

1.2.3. Існуючі рішення у сфері веб-скрапінгу 16

1.3. Класифікація дезінформативних новин 17

1.3.1. Загальні положення про класифікацію дезінформативних новин 17

1.3.2. Концепція класифікації дезінформативних новин 20

1.3.3. Існуючі рішення в сфері класифікації дезінформативних новин 22

1.4. Постановка задачі 25

1.4.1. Призначення розробки 25

1.4.2. Цілі та задачі розробки 26

1.5. Висновок до розділу 26

2. МЕТОДИ КРАУЛІНГУ ТА МАШИННОГО АНАЛІЗУ ОТРИМАНИХ ТЕКСТІВ 28

2.1. Алгоритм жадібного краулінгу 28

2.2. Алгоритм підготовки даних 31

2.3. Алгоритм TF-IDF-векторизації 32

2.4. Алгоритм токенізації 34

2.5. Пасивно-агресивний класифікатор 36

2.6. Двустороння довга короткочасна пам’ять 37

2.6.1. Архітектура ДКЧП-мереж 37
2.6.2. Обчислення рекурентних мереж 39
2.6.3. Обчислення ДКЧП-мереж 40
2.6.4. Обчислення двусторонніх ДКЧП-мереж 41
2.6.5. Ембедінги та їх обчислення 42
2.6.6. Щільна нейронна мережа 42
2.7. Висновок до розділу 44
3. ПРОГРАМНЕ ЗАБЕЗПЕЧЕННЯ СКРАПІНГУ ТА МАШИННОГО АНАЛІЗУ ОТРИМАНИХ ТЕКСТІВ 45
3.1. Екстрактор 45
3.2. Скрипт підготовки даних 46
3.3. Класифікатор на основі пасивно-агресивного класифікатора 47
3.4. Класифікатор на основі ДКЧП-мережі 49
3.5. Альтернативний скрапер 53
3.6. Порівняння результатів 55
3.7. Висновки до розділу 58
4. Розробка додатку для браузера Chrome 60
4.1. Опис функціональності побудованого браузерного додатку 60
4.2. Опис архітектури побудованого браузерного додатку 62
4.2.1. Контент-сценарій 62
4.2.2. Фоновий сценарій 63
4.3. Інструкція з установки браузерного додатку 66
4.4. Інструкція з використання скрапера 68
4.5. Інші застосування 71

4.6. Висновки до розділу 72

ВИСНОВКИ 73

ПЕРЕЛІК ПОСИЛАНЬ 75

ДОДАТОК А. ТЕКСТИ ПРОГРАМНОГО КОДУ 77

ВСТУП

Стрімкий зріст інтернет-технологій, а також здешевлення та покращення доступності як і користувацьких пристроїв з доступом у інтернет, так і серверних потужностей для створення власного контенту у мережі.

Мільярди тільки проіндексованих пошуковими сервісами сторінок містять дані на будь-яку тематику, і як і об’єм, так і різномаїття цих даних з кожним роком тільки збільшується.

Завдяки цьому інтернет став важливим, потужним та всеосяжним джерелом абсолютного різноманіття неструктурованих даних.

Дані (а особливо – видобута з даних істотна інформація) є ключовими у прийнятті рішень, здійснення досліджень, тощо.

Коректні, повні дані у достатньому обсязі дають можливість розуміння досліджуваного предмету, явища, проблеми. Саме тому видобування даних з мережі Інтернет є розвиненою та поширеною практикою як у веденні бізнесу, так і у здійсненні досліджень.

Якісний веб-скрапінг є поширеною та популярною послугою, однак для отримання якісних масових даних з багатьох ресурсів і досі потрібно розробляти спеціалізовані екстрактори даних з конкретних ресурсів або закладати істотні фінансові витрати на розробку таких екстракторів підрядниками які на цьому спеціалізуються або закладати співставні витрати на придбання ліцензій промислових сервісів, які надають API для використання високотехнологічних екстракторів, побудованих на алгоритмах машинного зору та застосовують засекречені алгоритми очистки та перевірки видобутих даних.

Однак, для багатьох досліджень та бізнесів такий підхід є фінансово недосяжним, отже у простих аналітичних алгоритмів є місце для розвитку – як з точки зору дешевизни та простоти розробки та підтримки, так і у невибагливості до умов експлуатації (наявних обчислювальних потужностей.

Дана робота присвячена поліпшенню простих алгоритмів та доведенню доцільності їх використання. завдань, які потрібно розробити. Для цього необхідно було провести аналіз існуючих рішень у сфері екстракції структурованої інформації з множини веб-сторінок, розробити інструмент екстракції, виконати порівняння інструменту екстракції з існуючими аналогами у вигляді постачання даних для систем машинного навчання.

Метою дослідження є створення програмного інструментарію екстракції структурованих даних з веб-сторінок новинних ресурсів для подальшої класифікації за достовірністю.

Для досягнення поставленої мети було окреслено та виконано наступні завдання:

- провести огляд існуючих підходів та програмних аналогів у областях екстракції даних з веб-ресурсів та оцінки якості новин;
- позробити та реалізувати алгоритми екстракції, підготовки та класифікації даних;
- порівняти результати, отримані розробленим алгоритмом та результатами тренування алгоритмів машинного навчання на даних, видобутих ним з існуючим аналогом та результатами тренування на даних аналогу.

Відповідно до завдань дослідження, в рамках даної роботи було проведено аналіз існуючих методів екстракції даних з ресурсів у мережі інтернет, та розроблено алгоритм видобування новинних матеріалів.

Також було розглянуто методи класифікації новинних текстів.

Розроблений алгоритм було імплементовано та порівняно з наявним у індустрії програмним забезпеченням за допомогою тренування алгоритмів машинного навчання на зібраних відповідними екстракторами даних, та застосовано результати натренованих алгоритмів для розробки браузерного додатку попередження про неякісні новини.

Наукова новизна роботи полягає в тому, що створено простий жадібний алгоритм у якому суміщено процеси пошуку посилань та видобування інформації, доведено доцільність використання простих алгоритмів для збору даних з ресурсів у мережі Інтернет з ціллю використання у тренуванні алгоритмів машинного навчання.

Практичний результат роботи полягає в тому, що доведено здатність класичних алгоритмів навчання досягати результатів, співставних з такими у нейронних мережах, таких як мережі ДКЧП, та показано, що такі моделі здатні працювати на двомовних наборах даних.

Отримані результати програмної реалізації запропонованих алгоритмів повністю підтверджують їх працездатність.