

Марія Талах
Валентина Дворжак

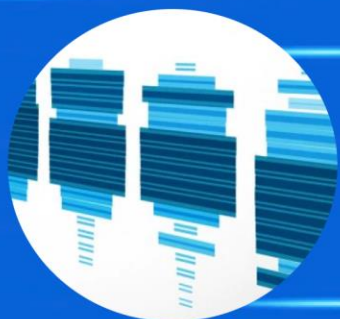
ІНТЕЛЕКТУАЛЬНИЙ АНАЛІЗ ДАНИХ

Частина I

ІНТЕЛЕКТУАЛЬНИЙ АНАЛІЗ ДАНИХ. Частина I



КАФЕДРА КОМП'ЮТЕРНИХ НАУК



МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ЧЕРНІВЕЦЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
ІМЕНІ ЮРІЯ ФЕДЬКОВИЧА

М.В. Талах, В.В. Дворжак

ІНТЕЛЕКТУАЛЬНИЙ АНАЛІЗ
ДАНИХ
ЧАСТИНА 1

Навчальний посібник

Технодрук

2022

УДК 681.3

T-16

*Рекомендовано до друку вченою радою навчально-наукового інституту фізико-технічних та комп'ютерних наук Чернівецького національного університету імені Юрія Федьковича
(протокол № 10 від 24 листопада 2022 р.)*

Рецензент:

Виклюк Я.І., д.т.н., проректор з наукової роботи, професор кафедри комп'ютерних систем та технологій приватного вищого навчального закладу «Буковинський університет»

Талах М.В., Дворжак В.В.

T-16 Інтелектуальний аналіз даних. Частина 1 /

М.В. Талах, В.В. Дворжак – Чернівці: Технодрук, 2022. – 367 с.

Навчальний посібник призначений для студентів, що навчаються за спеціальністю 122 «Комп'ютерні науки», освітньо-професійна програма «Інтелектуальний аналіз даних в комп'ютерних інформаційних системах» за першим (бакалаврським) освітнім рівнем та всіх бажаючих.

У навчальному посібнику викладено методи вирішення задач, що постають у аналізі даних з використанням базових алгоритмів машинного навчання. У першій частині посібнику розглядаються базові питання теорії і практики інтелектуального аналізу даних: задачі класифікації та прогнозу, кластерного аналізу та редукції даних та ін. Друга частина присвячена підходам до аналізу тексту, як специфічного типу даних. Для реалізації прикладних задач Data Mining запропоновано використовувати мову програмування Python.

УДК 681.3

©Чернів. нац. ун-т, 2023

©ПВКФ «Технодрук», 2023

©Талах М.В., 2022

Зміст

ВСТУП	12
ЧАСТИНА I. ОСНОВИ МАШИННОГО НАВЧАННЯ	13
РОЗДІЛ 1. ТЕОРІЯ МАШИННОГО НАВЧАННЯ	13
1.1. Для чого використовують машинне навчання?.....	13
1.2. Основні види машинного навчання	18
1.3. Приклади алгоритмів машинного навчання для вирішення простих задач.....	20
1.4. Коротка характеристика інших алгоритмів машинного навчання	29
1.4.1. Виявлення аномалій та новизни	29
1.4.2. Асоціативні правила	30
1.4.3. Напівконтрольоване навчання (Semi-supervised learning)...	31
1.4.4. Навчання з підкріпленням	32
1.4.5. Пакедне і динамічне навчання	34
1.4.5.1. Пакедне навчання	34
1.4.5.2. Динамічне навчання	35
1.4.6. Навчання на основі зразків або на основі моделей	37
1.4.6.1. Навчання на основі зразків	38
1.4.6.2. Навчання на основі моделей	39
1.5. Основні проблеми машинного навчання	43
1.5.1. Недостатній розмір навчальних даних.....	43
1.5.2. Нерепрезентативні навчальні дані	43
1.5.3. Приклади зміщення вибірки.....	45
1.5.4. Дані поганої якості	45
1.5.5. Конструювання ознак та проблема перенавчання моделі ..	46

1.6. Крок назад.....	49
1.7. Випробування і перевірка	50
1.8. Налаштування гіперпараметра(ів) і підбір моделі	51
РОЗДІЛ 2. ОСНОВИ РОБОТИ З PYTHON	54
2.1. Створення робочого середовища	54
2.2. Створення ізольованого середовища	56
2.3. Вектори, матриці, масиви.....	58
2.3.1. Створення вектору.....	59
2.3.2. Створення матриці	59
2.3.3. Створення розрідженої матриці.....	60
2.3.4. Вибір елементів.....	62
2.3.5. Опис матриці	63
2.3.6. Проведення операцій з елементами	64
2.3.7. Максимальное і мінімально значення.....	65
2.3.8. Обчислення описових статистик (середнього значення, дисперсії і стандартного відхилення).....	65
2.3.9. Операції з масивами	67
2.3.10. Транспонування вектору в матрицю.....	68
2.3.11. Додавання і віднімання матриць	69
2.3.12. Множення матриць.....	69
2.3.13. Генерування випадкових значень.....	71
2.4. Завантаження даних	72
2.4.1. Завантаження вбудованого набору даних	73
2.4.2. Створення потрібного набору даних	74

2.4.3. Завантаження файлу CSV	77
2.4.4. Завантаження файлу Excel	78
2.4.5. Завантаження файлу JSON	79
2.4.6. Завантаження файлу SQL	79
РОЗДІЛ 3. ВПОРЯДКУВАННЯ ДАНИХ	81
3.1. Створення фрейму даних.....	82
3.2. Опис даних	83
3.3. Навігація фреймами даних	85
3.4. Вибір рядків на основі умовних операторів і конструкцій ...	87
3.5. Заміна значень.....	88
3.6. Перейменування стовпчиків	89
3.7. Знаходження мінімуму, максимуму, суми, середнього арифметичного і кількості	90
3.8. Знаходження унікальних значень.....	91
3.9. Відбір пропущених значень	92
3.10. Вилучення стовпчика	94
3.11. Видалення рядка	95
3.12. Видалення повторюваних рядків.....	96
3.13. Групування рядків за значеннями	98
3.14. Групування рядків за часом	99
3.15. Циклічні операції з даними	101
3.16. Застосування функції до всіх елементів у стовпчику	102
3.17. Застосування функції до груп	102
3.18. Конкатенація фреймів даних.....	103

3.19. Злиття фреймів даних	104
РОЗДІЛ 4. РОБОТА З РІЗНИМИ ТИПАМИ ДАНИХ.....	108
4.1. Робота з числовими даними	109
4.1.1. Шкалювання, стандартизація і нормалізація ознак	109
4.1.1.1. Шкалювання.....	110
4.1.1.2. Стандартизація ознак.....	111
4.1.1.3. Нормалізація даних	113
4.1.2. Виявлення і робота з викидами	116
4.1.3. Дискретизація ознак	121
4.1.4. Видалення спостережень з пропущеними значеннями..	123
4.1.5. Заповнення пропущених значень	125
4.2. Робота з категоріальними даними	126
4.2.1. Кодування номінальних категоріальних ознак.....	127
4.2.2. Кодування порядкових категоріальних ознак	130
4.2.3. Кодування словників ознак	132
4.2.4. Видалення пропущених класів	134
4.2.5. Робота з незбалансованими класами.....	136
РОЗДІЛ 5. SCIKIT-LEARN, ЯК ОСНОВНА БІБЛІОТЕКА МАШИННОГО НАВЧАННЯ У PYTHON: ОГЛЯД ОСНОВНИХ АЛГОРИТМІВ.....	141
5.1. Представлення даних у Scikit-Learn	141
5.2. API оцінювача (estimator) Scikit-Learn	144
5.2.1. Основи API	145
5.2.2. Приклад навчання з вчителем: проста лінійна регресія (задача прогнозу).....	146
5.2.3. Приклад навчання з вчителем (задача класифікації)	150

5.2.4. Навчання без вчителя: зниження розмірності.....	151
5.3. Гіперпараметри та перевірка моделі.....	152
5.3.1. Перевірка моделі.....	153
5.3.2. Вибір найкращої моделі.....	156
5.3.3. Криві навчання.....	164
5.3.4. Налаштування на практиці: сітковий пошук.....	168
5.4. Дерева рішень.....	170
5.4.1. Створення дерева рішень.....	171
5.5. Метод опорних векторів.....	174
5.5.1. Вибір найкращої моделі.....	176
5.5.2. Реалізація алгоритму метод опорних векторів.....	177
5.5.2.1. Поза лінійними межами: ядро SVM.....	180
5.5.2.2. Налаштування SVM: запобігання перенавчанню моделі ..	183
5.5.2.3. Зведена інформація про.....	185
5.6. Регресія.....	186
5.6.1. Проста лінійна регресія.....	186
5.6.2. Нелінійна регресія.....	188
5.6.3. Регуляризація.....	191
5.7. Кластеризація k-середніх.....	197
5.7.1. Алгоритм k-середніх.....	198
5.7.2. Підбір оптимальної кількості кластерів.....	200
5.7.3. Приклади використання підходів кластеризації для роботи з зображеннями.....	205
РОЗДІЛ 6. ПРИКЛАД ПОВНОГО ПРОЕКТУ МАШИННОГО НАВЧАННЯ.....	212
6.1. Робота з реальними даними.....	212

6.2. З'ясування загальної картини	214
6.3. Постановка задачі	214
6.4. Конвеєри	215
6.5. Вибір критеріїв якості роботи	217
6.6. Перевірка припущень.....	218
6.6.1. Побіжний погляд на структуру даних	218
6.6.2. Створення навчального набору	222
6.6.3. Візуалізація даних для розуміння їх сутності	228
6.6.4. Візуалізація географічних даних	228
6.6.5. Пошук зв'язків.....	231
6.6.6. Експериментування з комбінаціями ознак.....	234
6.6.7. Підготовка даних	236
6.6.8. Очищення даних	237
6.6.9. Обробка текстових і категоріальних ознак	239
6.7 Спеціальні трансформатори.....	242
6.7.1. Масштабування ознак	243
6.7.2. Конвеєри трансформації	244
ЧАСТИНА II АНАЛІЗ ТЕКСТУ, ЯК ОСОБЛИВОГО ТИПУ ВХІДНИХ ДАНИХ	247
РОЗДІЛ 7. ПРИРОДНІ МОВИ ТА ЇХ ОБРОБКА	247
7.1. Додатки для обробки даних, засновані на аналізі природної мови	248
7.2. Конвеєр додатків для обробки даних.....	249
7.3. Мова як дані	251
7.3.1 Комп'ютерна модель мови.....	252

7.3.2. Лінгвістичні ознаки	254
7.3.3. Контекстні ознаки	258
7.3.4. Структурні ознаки	260
РОЗДІЛ 8. СТВОРЕННЯ ВЛАСНОГО КОРПУСУ	264
8.1. Що таке корпус?	264
8.2. Предметні корпуси	265
8.3. Рушій збору даних Valeen.....	266
8.4. Управління корпусом даних	268
8.5. Структура корпусу на диску	270
8.6. Структура каталогів на диску для Valeen.....	272
8.7. Об'єкти читання корпусів	273
8.8. Поточковий доступ до даних за допомогою NLTK.....	276
8.9. Читання корпусу HTML	279
8.10. Моніторинг корпусу.....	282
8.11. Читання корпусу з бази даних	283
РОЗДІЛ 9. ПОПЕРЕДНЯ ОБРОБКА ТА ПЕРЕТВОРЕННЯ КОРПУСУ	286
9.1. Розбивка документів	287
9.2. Виявлення і витяг основного контенту.....	287
9.3. Сегментація: виділення пропозицій.....	291
9.4. Попередня обробка даних із використанням NLTK:.....	293
9.4.1. Токенізація	294
9.4.2. Частотний розподіл слів	295
9.4.3. Фільтрування і видалення стоп-слів.....	296
9.4.4. Стемінг	297

9.4.5. Лемматизація	298
9.5. Лексемізація: виділення лексем.....	298
9.6. Маркування частинами мови	300
9.7. Проміжний аналіз корпусу	301
9.8. Трансформація корпусу	303
9.9. Попередня обробка і збереження проміжного результату	304
9.10. Запис в стислий архів.....	307
9.11. Читання попередньо обробленого корпусу.....	308
РОЗДІЛ 10. ПІДГОТОВКА ТЕКСТОВИХ ДАНИХ.....	311
10.1. Простір ознак основи текстової інформації.....	312
10.2. Частотні вектори.....	315
10.2.1. Застосування NLTK.....	315
10.2.2. Застосування Scikit-Learn.....	316
10.2.3. Застосування Gensim.....	316
10.3. Пряме кодування.....	317
10.3.1. Застосування NLTK.....	318
10.3.2. Застосування Scikit-Learn.....	318
10.3.3. Застосування Gensim	319
10.4. Частота слова – зворотна частота документа	320
10.4.1. Застосування NLTK.....	323
10.4.2. Застосування Scikit-Learn.....	324
10.4.3. Застосування Gensim	324
10.5. Розподілене представлення	326
10.5.1. Застосування Gensim	327
10.6. Scikit-Learn API	330

10.6.1. Інтерфейс BaseEstimator	330
10.6.2. Розширення TransformerMixin.....	332
10.7. Створення свого перетворювача для векторизації на основі Gensim	333
10.8. Розробка метода для нормалізації тексту.....	335
10.9. Конвеєри	337
10.9.1 Основи конвеєрів	338
10.9.2. Сітковий пошук оптимальних параметрів	340
10.9.3. Удосконалення вилучення ознак за допомогою об'єктів FeatureUnion.....	341
РОЗДІЛ 11. КЛАСИФІКАЦІЯ В ТЕКСТОВОМУ АНАЛІЗІ .	346
11.1. Класифікація тексту	347
11.1.1. Ідентифікація завдань класифікації.....	347
11.1.2. Моделі класифікації.....	349
11.2. Створення додатків класифікації тексту	351
11.2.1 Застосування перехресної перевірки: поточний доступ до к- блоків	352
11.2.2. Конструювання моделі.....	354
11.2.3. Оцінка моделі.....	356
11.3. Експлуатація моделі.....	360
ПЕРЕЛІК ДЖЕРЕЛ	362

Навчальне видання

**Талах Марія Віталіївна
Дворжак Валентина Володимирівна**

**Інтелектуальний аналіз даних
Частина 1**

Навчально-методичний посібник

Літературний редактор: О. В. Лупул

Папір офсетний. Формат 60x84/16.
Умов. друк. арк.. 21,33. Обл.- вид. арк. 22,93. Тираж – 50.

Видавець та виготівник: ПВКФ «Технодрук»
Свідоцтво суб'єкта видавничої справи ДК №1841 від 10.06.2004 р.
58000, м. Чернівці, вул. І. Франка, 20, оф.18, тел. (0372) 55-05-85