УДК 004.424.5

**Р. О. Добера**

# Пакети Python для Data Science

Вінницький національний технічний університет

**Анотація**

*Дана стаття містить інформацію про кращі пакети мови програмування Python для Data Science.*

**Ключові слова:** пакет, Python, NumPy, Data Science, Matplotlib, Pandas, Seaborn, Theano, SciPy, SAGE.

**Abstract**

*This article provides information about the best Python programming language packages for Data Science.*

**Keywords:** package, Python, NumPy, Data Science, Matplotlib, Pandas, Seaborn, Theano, SciPy, SAGE.

## Вступ

На практиці в багатьох випадках знайти точне рішення виниклої математичної задачі не вдається. Це відбувається не тому, що ми не вміємо цього зробити, так як дані рішення зазвичай не виражаються в звичних для нас елементарних або інших відомих функціях. Тому важливе значення набули чисельні методи, особливо у зв'язку зі зростанням ролі математичних методів в різних галузях науки і техніки і з появою високопродуктивних ЕОМ.

## Результати дослідження

Останнім часом Python став затребуваним в області Data Science. Це стало можливим завдяки появі бібліотек, здатних обробляти та візуалізувати великий обсяг даних на рівні MATLAB, Mathematica та R.

Python – це сучасна потужна високорівнева кросплатформна мова програмування, яка може використовуватись практично у будь-якій області розробки (автономній, клієнт-серверній, Web-проектування). Головною причиною її успіху є прозорий і логічний синтаксис, який дозволяє максимально прискорити процес освоєння мови чи створення проектів. У середовище виконання Python входить тільки інтерпретатор, який одночасно є і компілятором, який компілює початковий код Python безпосередньо в машинний код цільової платформи[1].

Data Science — це міждисциплінарна галузь про наукові методи, процеси і системи, які стосуються добування знань із даних у різних формах, як структурованих так і неструктурованих. Наука про дані є продовженням деяких галузей аналізу даних, таких як статистика, класифікація, кластеризація, машинне навчання, добування даних і передбачувальна аналітика.

Інша складова успіху Python – це її кросплатформні пакети розширення. Для роботи в галузі Data Science існує ряд високоефективних пакетів.

Python SciPy Stack - набір бібліотек, спеціально призначених для наукових обчислень. Кожен, хто зібрався використовувати Python в науці, повинен познайомитися з цим стеком.

Найбільш фундаментальний пакет - NumPy. Він додає Python підтримку великих багатовимірних масивів і матриць, разом з великою бібліотекою високорівневих математичних функцій для операцій з цими масивами. Основні пакети, які доповнюють NumPy, це: SciPy і Matplotlib[1].

SciPy є відкритою бібліотекою високоякісних наукових інструментів і містить модулі для оптимізації, інтеграції, спеціальних функцій, обробки сигналів, обробки зображень, генетичних алгоритмів, розв’язку звичайних диференціальних рівнянь і інших завдань, що зазвичай

вирішуються в наукових дослідженнях. Для візуалізації під час використання SciPy часто застосовують бібліотеки Matplotlib та Dislin[2].

Matplotlib – це бібліотека для побудови графіків і візуалізації даних. Особливістю Matplotlib є те, що за його допомогою можна виводити формули у вигляді TeX, однак існують проблеми з відображенням кириличних букв. Графіки, побудовані за допомогою Matplotlib можна масштабувати для перегляду області, що цікавить, причому як програмно із скрипта, так і через інтерфейс за допомогою миші. Бібліотека Dislin для побудови дво- і тримірних графіків працює дуже швидко, але зовнішній вигляд налаштовується тільки програмно. Навіть тривимірні графіки не можна обертати за допомогою миші[2].

Pandas - це пакет, призначений для простий і інтуїтивно зрозумілою роботи з «поміченими» і «реляційними» даними. Теж працює в зв'язці з NumPy, і крім математичних обчислень забезпечує їх агрегацію і візуалізацію.

Seaborn базується на Matplotlib, але оптимізований для візуалізації статистичних моделей: теплові карти, розподілу, результати математичних операцій. Незважаючи на зазначені можливості, більшість розробників використовує бібліотеку для відображення простих тимчасових розподілів.

Theano - одна з найпотужніших бібліотек з усього переліку. Ось кілька причин:

- тісна інтеграція з NumPy;
- використання CPU та GPU для покращення продуктивності;
- вбудовані механізми оптимізації коду;
- розширення для юніт-тестування й самоперевірки.

Theano використовується там, де необхідно зробити обчислення з великою точністю максимально швидко, наприклад в нейронних мережах і машинному навчанні.

Завдяки перерахованим пакетам Python перетворюється на ефективну мову високого рівня, здатну замінити Matlab в інтерактивній обробці даних і будувати повнофункціональний, призначений для користувача інтерфейс для контролю над експериментами. Не дивно, що Python для наукових обчислень використовують NASA, Los Alamos, JPL і Fermilab.

Для повноцінної безкоштовної заміни математичних середовищ, зокрема Magma, Maple, Mathematica, і MATLAB, розроблена і активно розвивається система комп'ютерної алгебри SageMath, яка покриває багато областей математики, включно з алгеброю, комбінаторикою, обчислювальною математикою і матаналізом. SAGE - це безкоштовне і вільно поширюване математичне програмне забезпечення з відкритими початковими кодами для дослідницької роботи і навчання в найрізноманітніших областях, включно з алгеброю, геометрією, теорією чисел, криптографією, числовими обчисленнями тощо. Серед можливостей системи підтримка паралельних обчислень з використанням як багатоядерних процесорів, так і багатопроцесорних систем і систем розподілених обчислень. Матаналіз реалізований на основі систем Maxima і SymPy. Лінійна алгебра, бібліотеки елементарних і спеціальних математичних функцій, статистичні бібліотеки функцій реалізовані на основі систем GSL, R, SciPy і NumPy. Дво- і тривимірні графіки для функцій і даних реалізовані за допомогою Matplotlib та Dislin. Розвинуті засоби для обробки зображень, візуалізації і аналізу теорії графів з використанням PyLab[3].

## Висновок

Мова програмування Python, завдяки функціональності пакетів NumPy, Matplotlib, SciPy тощо та розробленої на її основі системи комп’ютерної математики SAGE є потужною основою для наукових обчислень і, зокрема, для Data Science.

## СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Python [Електронний ресурс] — Режим доступу: URL : https://uk.wikipedia.org/wiki/Python
2. Python-библиотеки для Data Science [Електронний ресурс] — Режим доступу: URL : https://geekbrains.ru/posts/python_data_science
3. Welcome to Python.org [Електронний ресурс] — Режим доступу: URL : https://www.python.org/

***Добера Роман Олександрович*** − студент групи 2КН-19м, факультет інформаційних технологій та комп’ютерної інженерії, Вінницький національний технічний університет, Вінниця, e-mail: 2kn15b.dobera@gmail.com

Науковий керівник: ***Володимир Володимирович Колодний*** – кандидат техн. наук, доцент, доцент кафедри комп'ютерних наук, Вінницький національний технічний університет, м. Вінниця.

***Roman O. Dobera*** − Student of Information Technologies and Computer Engineering Department, Vinnytsia National Technical University, Vinnytsia, e-mail: 2kn15b.dobera@gmail.com

Supervisor: ***Kolodnyi Volodymyr V.*** **–** Ph.D., Docent, Docent of the Chair of Computer Science, Vinnytsia National Technical University, Vinnytsia.