УДК: 004.43

ПОРІВНЯННЯ МОВ ПРОГРАМУВАННЯ PYTHON ТА R В DATA SCIENCE

Андрусенко Б.Г., Мамчуровський В.С., Трунова О.В.
Чернігівський національний технологічний університет

У сучасному світі майже будь-яку задачу можна розв'язати за допомогою програмування. Вказати точну кількість мов програмування майже неможливо. На сьогодні існує близько семи сотень мов програмування, серед них 245 відомих [1]. Серед усього цього різноманіття є як універсальні (Java, C++, Python та ін.), так і вузькоспеціалізовані (1C, R, Swift та ін.). До того ж, популярність мов програмування змінюється кожного місяця, що досить просто відстежувати за індексом TIOBE. Станом на червень 2020 року трійку найпопулярніших мов програмування складають C, Java і Python (див. рис. 1).

Jun 2020	Jun 2019	Change	Programming Language	Ratings	Change
1	2	^	C	17.19%	+3.89%
2	1	v	Java	16.10%	+1.10%
3	3		Python	8.36%	-0.16%
4	4		C++	5.95%	-1.43%
5	6	^	C#	4.73%	+0.24%
6	5	v	Visual Basic	4.69%	+0.07%
7	7		JavaScript	2.27%	-0.44%
8	8		PHP	2.26%	-0.30%
9	22	^^	R	2.19%	+1.27%
10	9	v	SQL	1.73%	-0.50%

Рис. 1. Індекс TIOBE станом на червень 2020 [2]

Не можна не помітити різкого стрибка популярності мови R, що зумовлено зростанням потреби в обробці різноманітних статистичних даних, якою займається напрям програмування – Data Science. Його завдання – аналіз, обробка та представлення у цифровому вигляді великих обсягів даних. Для роботи у даному напрямку перевага надається наступним мовам програмування: R і Python.

Проведемо порівняння даних мов програмування на основі загальної та технічної інформації. Мова Python є значно універсальнішою, оскільки може використовуватися у різних напрямках, на відміну від R, яка в більшій мірі, орієнтована на статистичну обробку даних. Обидві мови є: імперативними, об'єктно-орієнтованими, функціональними, процедурними та рефлексивними (див. табл.1).

Таблиця 1. Порівняння мов програмування Python і R [3]

General and technical information	Python	R
Intended Use	Applications, general, web, scripting, AI, scientific computing	Application, statistics
Imperative	Yes	Yes
Object-oriented	Yes	Yes
Functional	Yes	Yes
Procedural	Yes	Yes
Generic	Yes	
Reflective	Yes	Yes
Event-driven	Yes	
Other paradigm(s)	aspect-oriented	
Standardized?	"De facto" standart via PEPs	No

Python підтримує узагальнене програмування, яке дозволяє записувати алгоритми, що приймають різні типи даних та є подійно-орієнтованою мовою, тобто виконання програми визначається подіями (дії користувача, повідомлення інших програм або потоків, події операційної системи).

Метою даної роботи є порівняння мов програмування Python та R у процесі обробки експериментальних даних.

Для нашого дослідження були використані результати кросплатформенного тесту Cinebench R15. Дані про процесор було отримано з довідки: Name, Mark, TDP, L3, Frequency, Turbo, Cores, Threads, Technology, RamSpeed) [4].

Дослідження на двох мовах відбувалося за такими напрямками: парсинг даних з сайту; побудова кореляційної матриці та регресійних полів, проведення регресійного аналізу; прогнозування значень на основі побудованих моделей.

За результатами дослідження була складена таблиця (див. табл. 2)., яка демонструє кількість часу (мс), за який виконується відповідна обробка даних на кожній з мов. Порівняння часу виконання певних завдань не дає нам остаточної відповіді, яку з мов краще використовувати при обробці статистичних даних. Оскільки, збереження та виведення даних у R займає більше часу, ніж у Python. Але в той же час парсинг веб-сторінки та розрахунки статистичних даних на R відбуваються швидше.

Таблиця 2. Час (мс) виконання певних операцій

Напрямок дослідження	Python	R
Парсинг веб-сторінки та запис до файлу з розширенням .xlsx	784982	681660
Зчитування даних з файлу .xlsx та виведення їх на екран	1136	5838
Побудова таблиці кореляції	991	1405
Побудова регресійного поля	2648	1750
Створення регресійних моделей і побудова ліній регресії	9259	10478
Створення регресійної моделі та виконання передбачення по заданим значенням	1388	2395
Створення мультирегресійної моделі	1921	3645
Створення мультирегресійної моделі та передбачення по заданим значенням	1772	812
Всього	804097	707983

Порівняння якості та кількості інформації при побудові лінійної мультирегресійної моделі, яку пропонують Python (рис. 2) та R (рис. 3), показало, що, у цілому, виведення інформації є достатньо інформативним. На обох мовах програмування: вказуються значення коефіцієнтів, коефіцієнт кореляції, середньоквадратичні відхилення, величини статистики Стьюдента для перевірки гіпотези щодо значущості відповідних параметрів. Проте Python ще пропонує значення критерію Дарбіна-Уотсона, який використовується для тестування автокореляції.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                   Mark   R-squared:                       0.956
Model:                            OLS   Adj. R-squared:                  0.956
Method:                 Least Squares   F-statistic:                     3476.
Date:                Wed, 10 Jun 2020   Prob (F-statistic):               0.00
Time:                        18:06:39   Log-Likelihood:                -6483.8
No. Observations:                 964   AIC:                         1.298e+04
Df Residuals:                     957   BIC:                         1.302e+04
Df Model:                           6
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept   -547.2389     55.301     -9.896      0.000    -655.763    -438.714
TDP            1.5617      0.267      5.843      0.000       1.037       2.086
L3             8.4982      0.895      9.490      0.000       6.741      10.256
Turbo        182.0977     13.043     13.962      0.000     156.502     207.693
Threads       60.1787      1.431     42.042      0.000      57.370      62.988
Technology    -8.4436      1.285     -6.569      0.000     -10.966      -5.921
RamSpeed       0.1056      0.020      5.210      0.000       0.066       0.145
==============================================================================
Omnibus:                      224.585   Durbin-Watson:                   1.491
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             2803.048
Skew:                           0.689   Prob(JB):                         0.00
Kurtosis:                      11.239   Cond. No.                     1.83e+04
==============================================================================
```

Рис. 2. Опис лінійної мультирегресійної моделі на Python

```
Call:
lm(formula = Mark ~ TDP + L3 + Turbo + Threads + Technology +
    RamSpeed, data = dataset)

Residuals:
    Min      1Q  Median      3Q     Max
-918.87  -86.98  -12.46   91.82 1597.96

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -547.23886   55.30059  -9.896  < 2e-16 ***
TDP            1.56168    0.26728   5.843 7.03e-09 ***
L3             8.49822    0.89548   9.490  < 2e-16 ***
Turbo        182.09767   13.04257  13.962  < 2e-16 ***
Threads       60.17874    1.43139  42.042  < 2e-16 ***
Technology    -8.44362    1.28529  -6.569 8.28e-11 ***
RamSpeed       0.10561    0.02027   5.210 2.32e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 202.5 on 957 degrees of freedom
Multiple R-squared:  0.9561,    Adjusted R-squared:  0.9559
F-statistic:  3476 on 6 and 957 DF,  p-value: < 2.2e-16
```

Рис. 3. Опис лінійної мультирегресійної моделі на *R*

R і Python є свого роду конкурентами за звання «найкращого» інструменту для роботи з даними, маючи власні переваги та недоліки. На питання: «Яка мова краще?» однозначної відповіді дати неможливо, оскільки все залежить від конкретної ситуації, необхідних інструментів, рівня знань. Зазвичай R використовується тоді, коли для аналізу даних необхідні окремі сервери або ж виділені обчислювальні потужності. Python же допоможе у тих випадках, коли задачі, пов’язані з аналізом даних, вплітаються в роботу веб-додатків, а також прекрасно підходить для реалізації алгоритмів з їх подальшим практичним використанням [5].

Література:

1. How Many Computer Languages Are There? URL: https://careerkarma.com/blog/how-many-coding-languages-are-there/
2. TIOBE Index for June 2020. URL: https://www.tiobe.com/tiobe-index/
3. Comparison of programming languages. URL: https://en.wikipedia.org/wiki/Comparison_of_programming_languages
4. Cinebench R15 (Multi-Core). URL: https://www.cpu-monkey.com/en/cpu_benchmark-cinebench_r15_multi_core-8(
5. R и Python – достойные соперники? URL: https://habr.com/ru/company/piter/blog/263457/