

ОБЗОР ЯЗЫКОВ ПРОГРАММИРОВАНИЯ ДЛЯ МЕДИКО-БИОЛОГИЧЕСКИХ ИССЛЕДОВАНИЙ

Щербаков Д.О.

Белорусский государственный университет информатики и радиоэлектроники

г. Минск, Республика Беларусь

Чураков А.В. - д-т кафедры ЭТТ, к.м.н.

Данный доклад предназначен для разработчиков в сфере медицинских исследований, а также анализа данных. Доклад содержит краткий обзор основных технологий и языков программирования с основными преимуществами и недостатками. Основной идеей было донести начинающим разработчикам, что для анализа данных, несмотря на разнообразие различных языков программирования, уже много время лидирующее место занимает Python, за счёт простоты изучения, а также количества приложений с которыми Python имеет глубокую интеграцию.

ПО нашло своё применение во всех сферах нашей жизни, и медицина не стала исключением. Компьютерная хирургия - это быстро развивающаяся область медицины, которая сочетает в себе медицинскую экспертизу с компьютерным интеллектом, чтобы дать более быстрые и более точные результаты в хирургических процедурах и таких примеров множество.

Тема доклада очень тесно связана с обработкой и анализом большого количества данных, а, следовательно, язык программирования должен иметь высокую скорость обработки данных, а также простоту синтаксиса, чтобы код был читабельным. На данный момент самыми популярными языками для медико-биологических исследований являются: R, Python, MatLab. Далее мы рассмотрим основные преимущества данных языков программирования при работе с данными.

Итак, R- язык программирования, который является прямым потомком старшего языка программирования S, был выпущен в далеком 1995 году и с тех пор становится все совершеннее. Распространяется - бесплатно.

Преимущества:

1) отличный набор высококачественных предметно-ориентированных пакетов с открытым исходным кодом. R имеет в своем распоряжении пакеты практически для любого количественного и статистического применения, которое можно только себе представить. Сюда входят нейронные сети, нелинейная регрессия, филогенетика, построение сложных диаграмм, графиков и многое-многое другое;

2) вместе с базовой установкой в довесок нам предоставляется возможность установки обширных встроенных функций и методов. Кроме того, R прекрасно обрабатывает данные матричной алгебры;

3) возможность визуализации данных является немаловажным преимуществом наряду с возможностью использования различных библиотек, например, ggplot2.

Недостатки:

1) низкая производительность. Здесь нечего сказать: R не является быстрым ЯП;

2) специфичность. R прекрасно подходит для статистических исследований и науки о данных, но он не так хорош, когда дело доходит до программирования для общих целей.

Другие особенности. R имеет несколько необычных особенностей, которые могут сбить с толку программистов, привыкших работать с другими ЯП: индексирование начинается с 1, использование нескольких операторов присваивания, нетрадиционные структуры данных.

MATLAB – это признанный язык для численных расчетов, используемый как в научных целях, так и в индустрии. Он был разработан и лицензирован MathWorks, компанией, созданной в 1984 году, основной целью которой являлось коммерциализация программного обеспечения. Лицензия - условно бесплатная.

Преимущества:

1) MATLAB, предназначенный для численных вычислений, хорошо подходит для использования количественного анализа со сложными математическими требованиями, такими как обработка сигналов, преобразования Фурье, матричная алгебра и обработка изображений;

2) MATLAB имеет ряд встроенных возможностей построения графиков и диаграмм;

3) MATLAB часто можно встретить во многих курсах бакалавриата по точным наукам, таким как физика, инженерия и прикладная математика. Таким образом, он широко используется в этих областях.

Недостатки:

1) платная лицензия. Вне зависимости от выбранного вами варианта (для научных, личных целей или целей компании) вам придется раскошелиться на дорогостоящую лицензию. Наш совет: обратите внимание на бесплатную альтернативу – Octave;

2) MATLAB – это не самый подходящий язык программирования для общего назначения.

Благодаря своему широкому использованию в различных количественных вычислениях как для научных целей, так и для целей индустрии, MATLAB стал достойным вариантом для применения в области науки о данных. Он придется как нельзя кстати, если для ваших ежедневных целей необходима интенсивная, продвинутая математическая функциональность, собственно, для чего MATLAB и был разработан.

В 1991 году Гвидо ван Россум представил язык программирования Python. С тех пор этот язык стал чрезвычайно популярным ЯП общего назначения и широко используется в сообществе специалистов по данным. В настоящее время основными версиями являются Python 3,7 и Python 2,7. Лицензирование происходит бесплатно.

Преимущества:

- 1) Python – это очень популярный, широко используемый язык программирования общего назначения. Он имеет обширный набор специально разработанных модулей и широко используется разработчиками. Многие онлайн-сервисы предоставляют API для Python;
- 2) Python очень прост в изучении. Низкий порог вхождения делает его идеальным первым языком для тех, кто занимается программированием;
- 3) такие программные пакеты как pandas, scikit-learn и Tensorflow, делают Python надежным вариантом для современных приложений в области машинного обучения;
- 4) очень высокий уровень интеграции с различными сторонними платформами, например, sim4life.

Недостатки:

- 1) типобезопасность. Python – это динамически типизированный язык, а это значит, что вы должны быть осторожными при работе с ним. Ошибки несоответствия типов (например, передача строки (string) в качестве аргумента методу, который ожидает целое число (integer)) могут время от времени случаться;
- 2) у Python существует реализация PyPy, которая по скорости близка к Java, но в ней есть не все возможности оригинального языка. Python не подходит для задач, требующих большого объема памяти, — их лучше решать вставками на C или C++;
- 3) другим недостатком является сильная зависимость языка от системных библиотек, из-за чего затрудняется перенос на другие системы. Для этих целей существует инструмент Virtualenv, но и он с недостатками: избыточность полных методов изоляции, костыли, дублирование системных библиотек;
- 4) еще одна проблема — в том, что Global Interpreter Lock (GIL) не позволяет выполнять несколько потоков Python одновременно в реализации CPython. Однако GIL можно отключить на какое-то время, как это сделано в математическом пакете NumPy;
- 5) например, в случае если имеются конкретные цели статистического анализа и анализа данных, то обширный набор пакетов языка R дает ему преимущество перед Python. Кроме того, существуют более быстрые и безопасные альтернативы Python среди языков программирования.

В Python есть несколько библиотек, которые можно использовать для проведения исследований и вычислений: SciPy — библиотека с научными инструментами, NumPy — расширение, которое добавляет поддержку матриц и многомерных массивов, а также математические функции для работы с ними, Matplotlib — библиотека для работы с 2D- и 3D-графикой.

Python является хорошим вариантом для целей науки о данных (data science), и это утверждение справедливо как для начального, так и для продвинутого уровней работы в данной области. Большая часть науки о данных сосредоточена вокруг процесса ETL (извлечение-преобразование-загрузка). Эта особенность делает Python идеально подходящим для таких целей языком программирования. Библиотеки, такие как Tensorflow от Google, делают Python очень интересным языком для работы в области машинного обучения.

Благодаря библиотекам и простоте освоения языка многие ученые выбирают Python — особенно он популярен у математиков и физиков.

Python — один из самых используемых в Data Science языков. На нем пишут алгоритмы программ с машинным обучением и аналитические приложения. С помощью него обслуживают хранилища данных и облачные сервисы.