УДК 004 **Технические науки**

Разин Сергей Александрович, *студент 3 курс, кафедра «Информационные системы и технологии» Поволжский государственный университет телекоммуникаций и информатики Россия, г. Самара*

ЧТО ДОЛЖЕН ЗНАТЬ РАЗРАБОТЧИК НА ЯЗЫКЕ PYTHON, РАБОТАЯ В СФЕРЕ DATA SCIENCE

Аннотация: Из статьи читатель узнает о знаниях, которые должен приобрести учащийся языку программирования Python. Будут представлены инструменты для работы в сфере Data Science и прописаны лучшие базы данных. Читатель узнает, что такое машинной обучение и глубокое обучение, сможет составить план обучения, благодаря данной статье, который поможет ему правильно прокачать «скилы» в этом направлении.

Ключевые слова: Python, Machine Learning, Deep Learning, Data Science, Алгоритмы обучения.

Annotation: From the article, the reader will learn about the knowledge that a student of the Python programming language should acquire. Tools for working in the field of Data Science will be presented and the best databases will be registered. The reader will learn what machine learning and deep learning are, and will be able to make a training plan, thanks to this article, which will help him correctly pump "skills" in this direction.

Keywords: Python, Machine Learning, Deep Learning, Data Science, learning Algorithms.

Язык программирования Python простой для изучения и легкий в составлении кода. Понадобиться написать всего несколько строк, чтобы создать искусственный интеллект, который бы самообучался. Этот язык был выбран

разделом информатики, изучающим проблемы анализа, обработки и представления их в цифрах, иначе Data Science [1].

«Даталогия» обрабатывает большие объемы информации. Разработчики, работающие в этой сфере, умеют:

- собирать большие объемы информации и конвертировать в их единый удобный для пользователей формат;
- решать задачи бизнеса с использованием цифровой информации;
- писать коды на языке Python, R и других подобных;
- систематизировать информацию и вести статистику;
- тестовый анализ, Machine Learning и Deep Learning;
- изучают современные технологии для успешных разработок в компании, где эти специалисты и работают.

Так что нужно уметь и знать разработчику Python, чтобы стать экспертом по аналитическим данным в этой сфере или Data Scientist [2]?

Линейные уравнения и математический анализ

Математика играет главенствующую роль в разделе Data Science. Например, линейная алгебра используется для работы с громадным объемом информации. Она включает в себя:

- матрицы;
- векторы;
- линейные уравнения;
- алгоритмы классификации и кластеризации.

А оптимизация собранной информации осуществляется посредством математического анализа [3].

Статистика

Разработчик на языке Питон, работая с Data Scientist должен знать, как сделать определенные выводы из полученных результатов, которые были выведены из собранной информации. А этим занимается наука – статистика, в сфере которой должен быть хорошо подкован эксперт по аналитическим данным.

Статистика позволяет делать выводы на основе использования способов и приемов по сбору, обработке и представлению анализа числовых данных. Разработчик должен уметь делать:

- выборку;
- распределять частоты;
- выявлять среднее значение;
- принимать взвешенное среднее значение;
- определять вероятность;
- распределять ее же;
- тестировать значимости.

Найти курсы по Data Science можно самостоятельно и обучиться им. Такие уроки на английском языке распространены в интернете. Они помогут начинающему разработчику быстрее освоить разделы статистики [4].

Дополнительные инструменты Python и библиотеки языка

Чтобы проводить все вышеописанные вычисления на компьютере понадобятся библиотеки языка Python.

Например, библиотека Джона Хантера под названием MatplotLib. Позволяет строить двумерные числовые построения. Она способна упростить сложные числовые вычисления. Посредством Матплотлиб разработчик сможет вести контроль за видом линий, свойствами системы координат.

Seaborn – эта библиотека построена на базе предыдущей. Обладает нормальными настройками по дефолту для работы с графиками. Она позволяет сделать картинки симпатичными.

Numpy позволяет добавить поддержку многомерных массивов и матриц. В ней заложены многоуровневые функции, которые помогают быстро решать задачи и проводить операции с многомерными массивами.

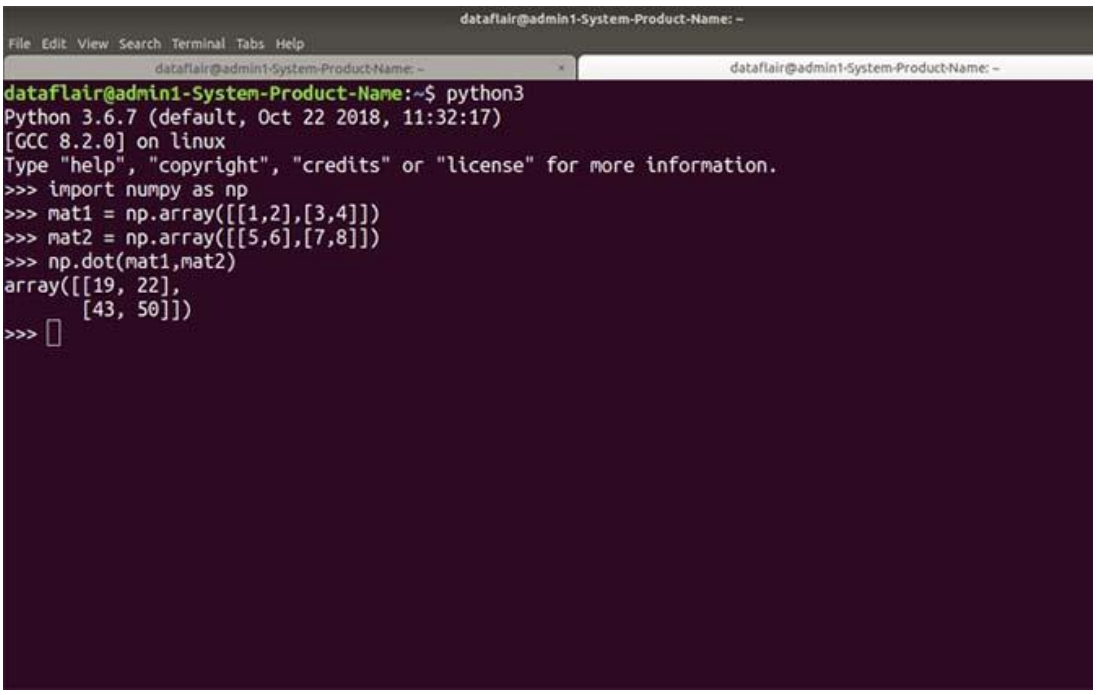


Рис. 2 Вычисления в библиотеке NumPy

SciPy имеет открытый исходный код. С помощью ее выполняют инженерные расчеты и математические вычисления. Возможности ее таковы:

- позволяет находить минимумы и максимумы функций;
- помогает решать интегралы функций;
- обрабатывает сигналы;
- обрабатывает картинки;
- позволяет работать с генетическими алгоритмами.

Pandas строится поверх Numpy и отвечает за обработку и анализ информации. Она относится к высокоуровневым библиотекам в отличие от предыдущей. В ней разработчик найдет структуры данных для манипуляций с числовыми таблицами и временными рядами. А также она позволяет проводить с ними операции.

Выборку можно производить из разных источников типа Эксель, SQL, CSV [5].

Такие библиотеки будут отличным пособником для специалиста в Python в работе с графикой и собранной информацией.

Одним из лучших дополнительных инструментов является Jupyter Notebook. С помощью него можно сохранять большое количество информации. Так как инструмент позволяет хранить код, картинки и графики, функции в одном месте.

Базы данных

Так как речь идет о громадном объеме информации, разработчик должен уметь работать с БД. Он должен знать, как вытаскивать из них информацию и обрабатывать ее, уметь выполнять сложные запросы.

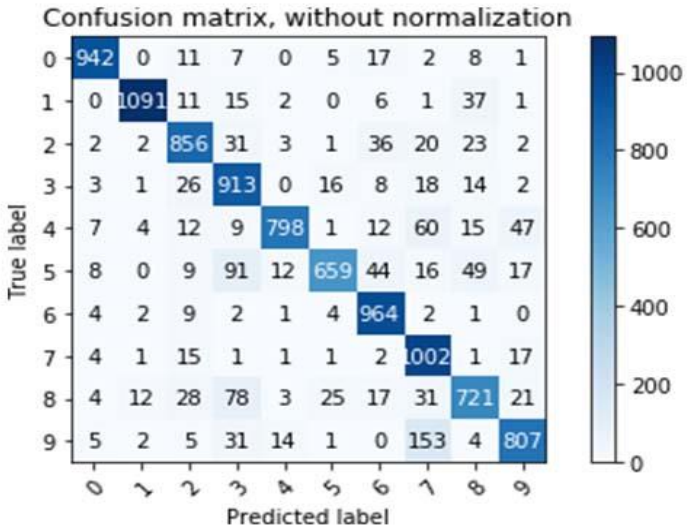
Базы данных бывают двух видов: реляционные и нереляционные:

- реляционные – это SQL. Они имеют структурированный язык запросов. К ним относятся MySQL, PostgreSQL, Microsoft SQL;
- нереляционные – это NoSQL. Это MongoDB, BigTable, Redis.

Однако с последними специалисты Data Sceince реже сталкиваются, чем с первыми. Оптимальными для работы разработчика являются: SQL Server, PostgreSQL, MySQL.

Machine Learning

Машинное обучение – это еще одна ветвь развития компьютеризации. Главной задачей его является умение компьютера обучаться опознавать лица, голоса, подбирать картинки или кино в Youtube, учитывая при этом просмотренные пользователем видео ранее. То есть он не должен работать только по заранее записанному коду.



Алгоритмы обучения

Ниже даны начальные алгоритмы, которые закладываются в интеллект машины для дальнейшего самостоятельного познания компьютерным мозгом:

- линейная регрессия;
- логистическая регрессия;
- random forest;
- кластеризация способом k-средних;
- интегрированная модель авторегрессии скользящего среднего.

Для Machine Learning используют и библиотеки.

Библиотеки Machine Learning

К подобным библиотекам относятся:

- scikit-learn – позволяет работать с классическими алгоритмами самостоятельного обучения искусственного интеллекта;
- TensorFlow и Keras. С помощью ее проводится работа с нейронными сетями.

Исходя из вышеописанного разработчик на языке Python должен понимать алгоритмы кластеризации, классификации, нейронных сетей, решать задачи посредством представленных библиотек. И только потом Data Scientist сможет работать над глубоким обучением и составлением своего портфолио.

Библиографический список:

1. Прохоренок Н.А. Самое необходимое. — СПб.: БХВ-Петербург, 2011. — 416 с.
2. Доусон М. Программируем на Python. – СПб.: Питер, 2014. – 416 с.
3. Шелудько, В. М. Основы программирования на языке высокого уровня Python: учебное пособие / В. М. Шелудько. – Ростов-на-Дону, Таганрог: Издательство Южного федерального университета, 2017. – 146 с. – ISBN 978-5-9275-2649-9. – Текст: электронный // Электронно-библиотечная система IPR

BOOKS: [сайт]. – URL: http://www.iprbookshop.ru/87461.html (дата обращения: 13.02.2020). – Режим доступа: для авторизир. Пользователей.

4. Любанович Билл Простой Python. Современный стиль программирования. – СПб.: Питер, 2016. – 480 с.: – (Серия «Бестсепперы O’Reilly»).

5. Лутц М. Программирование на Python, том II, 4-е издание. – Пер. с англ. – СПб.: Символ-Плюс, 2011. – 992 с.