Лабораторная работа 0
Язык программирования Python

1. Материалы для изучения

Вариант 1: необходимый минимум

Изучение Python за час: https://www.youtube.com/watch?v=x_2QpLcRdeY. Там же есть указания по установке python. ***Для лабораторных нам подойдет любая версия python, начиная с третьей (python 3.X)***.

Также стоит пройти уроки: https://pythonworld.ru/samouchitel-python, нам достаточно уроков 1-16, 19, 20. Или можно использовать этот самоучитель как справочник, обращаясь к нему, когда необходимо.

Помимо самоучителя, на сайте есть просто статьи на различные темы с кратким и понятным объяснением.

Вариант 2: для тех, кто хочет лучше изучить Python

По желанию можно также пройти целый обучающий курс: https://stepik.org/course/67. Это потребует больше времени (курс рассчитан на 3 недели, но если вы знаете любой другой язык, то за две пройти более чем возможно), однако вы получите хорошую базу и привыкнете к языку. Далее, если вам понравится язык и вы захотите изучить его получше, можно взять курс поинтереснее, например, такой: https://stepik.org/course/512.

2. Задание

Можно использовать любую среду разработки на python, даже блокнот пока подойдет. Удобно использовать jupyter notebook (http://jupyter.org/). Если вы любите сидеть в отладчике или планируете работать с проектами, можно установить полноценную среду разработки (https://www.jetbrains.com/pycharm/, вам нужна community-версия), однако стоит учесть, что на слабых машинах она может заметно тормозить.

Очень советую, как только возникает вопрос «А как сделать в python ЭТО», искать ответ в интернете. Лучше, если на английском. Например, на stackoverflow можно найти хорошие ответы. При этом старайтесь разобраться в теме, сравнить варианты реализации; если вам предлагают какой-то встроенный

метод python — почитайте документацию, посмотрите, какие у него есть аргументы.

1.1 Общие задачки

Решите 8 задачек: https://pythonworld.ru/osnovy/tasks.html

К отчету приложите скрипт со всеми этими функциями. Если вы прошли предложенный курс по python и у вас есть хотя бы 25 баллов из 31 за «Задачи по материалам недели», то это задание можно не выполнять.

1.2 Создадим файлы

При решении задач машинного обучения часто приходится работать с большим числом данных — списками, словарями, файлами, поэтому, чтобы далее не отвлекаться на разбор таких вспомогательных задач, посвятим это и следующее задания им.

Напишите программу генерации файлов.

Сначала скачайте базу даных смс с диска курса:

datasets->Text->SpamSMS

Там находися текстовый файл с самой базой и файл-readme с описанием формата данных.

Создадите две папки для спама и обычных сообщений: «spam» и «ham».

Затем напишите цикл по содержимому текстового файла с базой так, чтобы каждая отдельная запись базы (сообщение) была сохранена в отдельный текстовый файл в соответствующую папку (spam/ham). При этом необходимо оставить только текст самого смс-сообщения, меток spam/ham и лишних пробелов в полученном текстовом файле быть не должно.

В качестве имён генерируемых файлов возьмите номер сообщения, причем нумерация должна быть отдельно по спаму и отдельно по обычным сообщениям.

1.3 Проанализируем файлы

Теперь, когда у нас есть полная папка нагенерированных файлов, давайте получим некоторые их характеристики. *Все полученные характеристики запишите в отчет по лабораторной.*

- Найдите минимальную, максимальную и среднюю длину содержимого файлов. Используйте встроенные функции python для этого.
- Посчитайте, сколько раз каждый символ (включая пробел) встречается во всех файлах и выведите эти значения, отсортированные по величине. Тут будет удобно и полезно использовать словарь. Ключом будет сам символ, значением — число, обозначающее, сколько раз этот символ встречается. При разборе каждого нового файла вам нужно будет увеличивать значения в словаре по ключам, содержащимся в файле.
- Найдите самые популярные слова. Словом будем считать любой набор букв, отделенный с двух сторон пробелами либо другими разделительными символами — точкой, запятой, восклицательным знаком, и т.д. Или отделенный с одной стороны, если этот набор находится в начале/конце сообщения.