"Автоматизированные методы обработки текстовой информации"

Лабораторная работа №1
Тема "*Работа со списками Python*"

1. Реализуйте процедуру, которая получает как входной параметр ссылку на числовой массив, а возвращает кортеж из трех элементов: число отрицательных, число нулевых и число положительных элементов массива.

2. Реализуйте процедуру, которая получает на вход две ссылки на два массива и возвращает ссылку на тот из них, который имеет большую сумму своих элементов.

3. Реализуйте процедуру, которая для входного текста создает словарь, ключами которого будут символы, а значениями -- количество раз, сколько встречается символ в строке.

4. Реализуйте процедуру, которая вычисляет сумму всех элементов входного числового массива.

5. Реализуйте процедуру, которая разбивает входную строку на подстроки заданной длины и возвращает массив этих частей строки.

Лабораторная работа №2

Тема "*Чтение и запись данных в файл*"

1. По заданному целочисленному параметру n командной строки программы сгенерировать n строк треугольника Паскаля.

2. В строках текстового файла через пробел перечислены целые числа. В выходной файл записать в каждой строке записать значение суммы чисел строк входного файла.

3. То же что и в задании 2, но суммирование производить столбцов. Соответственно в выходном файле должна быть строка состоящая из значений сумм чисел в столбцах входного файла.

4. Во входном файле в каждой строке записано одинаковое количество чисел. В выходной файл записать те же числа, но строки и столбцы поменять местами.

5. В каждой строке файла записаны строки вида
ФамилияСтудента число баллов.
При этом фамилии студентов могут повторяться. В выходной файл записать строки вида ФамилияСтудента суммарное число баллов.

6. То же что и задание 5, но создать для каждого студента выходной файл по его фамилии и записать строку состоящую из баллов этого студента.

7. В каждой строке входного файла записаны числа -- коэффициенты многочлена по возрастающим степеням.
При запуске программы через командную строку передается значение x и программа в выходной файл записывает в каждой строке значение соответствующего многочлена в точке x.

Лабораторная работа №3

Тема "*Чтение и запись данных в файл"*

1. Необходимо из слов файла организовать словарь. Ключами словаря являются буквы русского алфавита, а значениями списки, содержащие слова, начинающиеся на букву, указанную в ключе словаря.

Пример такого словаря:
{
'а':['арбуз','апельсин'],
'б':['барабан','бочка','билет']}

2. Для заданного текстового файла построить словарь,
ключами которого будут слова, а значениями списки слов длина которых равна значению ключа.

3. Из заданного текстового файла выбрать слова заданной длины.

4. Из заданного текстового файла выбрать слова, заканчивающиеся на -ый,-ая,-ое.

5. В файле, в каждой строке записаны данные в формате:

Иванов 12 21 14 15 22
Петров 5 23 31
Сидоров 7 18 25 11

Требуется в отдельный файл записать данные

Иванов 84 зачтено
Петров 59 не зачтено
Сидоров 61 зачтено.

Здесь числа равны суммам чисел в соответствующих строках исходного файла.

6. В файле записаны слова через пробел. Записать в другой файл слова первого файла в алфавитном порядке.

Лабораторная работа №4

Тема "*Регулярные выражения*"

1. Из заданного текста выделить все числа и вывести их на экран.

Использовать регулярное выражение r'[0-9]+'.

2. Из заданного текста найти слова, начинающиеся на -пре, -про, -при. Вывести их на экран.

Использовать регулярное выражение r' ((пре|про|при)[^]*)'

3. Из заданного текста найти слова, оканчивающиеся на -ое, -ая, -ый, -ые.

Использовать регулярное выражение r'([^]+(ое|ая|ый|ые)) '

4. В заданном текстовом файле посчитать количество предложений. Учесть, что предложения могут заканчиваться точкой, вопросительным, восклицательным знаком и многоточием.

Использовать регулярное выражение r'(\.{3}|[.!?])'.

5. Слова в тексте могут быть разделены пробелами и знаками препинания. Из заданного текста извлечь все слова.

Использовать регулярное выражение r'[^ .?!,]+'

6. В тексте, в каждой строке записано пронумерованное слово:

10. слово
23. слово
35. слово

Составить словарь, в котором ключ будет номер, значением соответствующее слово.

Использовать регулярное выражение r'\n(\d+)\. ([^ 0-9\n]+)'.

Лабораторная работа №5

Тема "*Регулярные выражения*"

1. Найти в данном тексте предложения, состоящие из не более 4-х слов.

2. Найти в данном тексте все слова, содержащие от 2-х до 4-х гласных букв.

3. Задана строка для записи многочленов с целыми коэффициентами в виде

5x^4-21x^2+x-7.

Извлечь из этой строки все коэффициенты в виде словаря, в котором ключем является показатель степени, а значением -- соответствующий коэффициент.

4. В файле записаны строки вида
Текст: число Текст: число
Текст: число Текст: число Текст: число
Текст: число Текст: число
Текст: число
...
Найти сумму этих чисел.

5. Найти в текстовом файле первые слова всех предложений.

6. Из текста выбрать слова, начинающиеся на из-, ис-.

7. В каждой строке файла записан номер телефона. Средствами регулярных выражений привести все такие записи к одному виду: +79033176536.

Лабораторная работа №6

Тема "*Задачи кластеризации текстов*"

1. Имеется набор текстов в виде текстовых файлов размещенных в указанном каталоге. Используя модуль sklearn выполните подсчет слов в этих текстах.

2. Для указанных выше текстов вычислите меру их сходства на основе модуля sklearn.

3. Произвести кластеризацию сообщений с помощью метода K средних.