**Расчетно-графическая работа**

# Обработка текстовых данных

**R1 [#60]**

Разработайте программу Count Words, которая подсчитывает число слов в текстовом файле и собирает статистику встречаемости для каждого слова.

Размер анализируемых текстовых файлов не ограничен (в программе должно действовать ограничение по количеству слов, одновременно хранящихся в оперативной памяти).

Словом считается последовательность букв русского или латинского алфавитов, все остальные символы интерпретируются как пробелы. Различные словоформы одного слова считаются разными словами.

Статистика накапливается в текстовом файле, который отсортирован в алфавитном порядке и содержит строки формата:

<слово><пробел><количество вхождений>

**R2**  
**a) [#20]**

Оптимизируйте программу Count Words таким образом, чтобы минимизировать количество обращений к файлу статистики при анализе входнного файла.

**b) [#20]**

Выполните анализ трудоемкости исходного и оптимизированного алгоритмов в терминах количества операций ввода-вывода и времени работы на контрольном примере.

**R3 [#60]**

Разработайте программу Gather Words, которая анализирует файл статистики, подготовленный программой Count Words и подсчитывает количество слов, объединяя статистику для однокоренных слов, модифицированных с помощью окончаний. Придется реализовать алгоритм, убирающий окончания слов, например *стеммер Портера*.

Например, должны считаться одним словом следующие варианты:

диск, диска, диски

стоял, стояла, стояли

красивый, красивая, красивые

Формат файла статистики

<слово>, <слово>, ..., <слово> : <количество вхождений>

**R4 [#60]**

Разработать программу, которая выполняет автоматический анализ и категоризацию русскоязычных текстов из различных предметных областей. Программа использует набор опорных текстов, для которых категоризация была выполнена ранее с помощью Count Words и GatherWords.

Опорные тексты размещены в соответствующих каталогах на диске, например: компьютеры, здравоохранение, сельское хозяйство, спорт. В каждом каталоге присутствует файл статистики, созданный программами CountWords и GatherWords. Используя эту статистику, программа анализирует документ и делает вывод о принадлежности его к одной из известных категорий, либо помещает документ в категорию «прочее».

Требования:

корректное определение категории текста не менеее, чем в 75% случаев