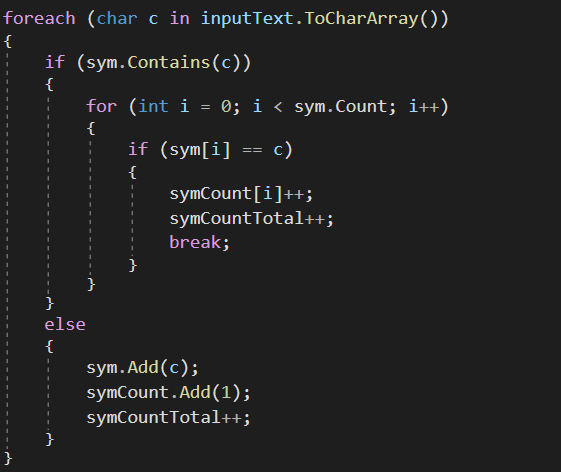
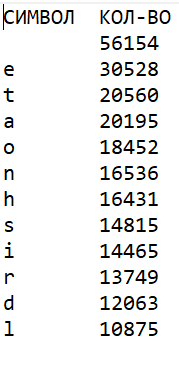
##### 1. Частотность букв русского и английского языков

1.1. Создайте в произвольной программной оболочке программу со следующим функционалом:

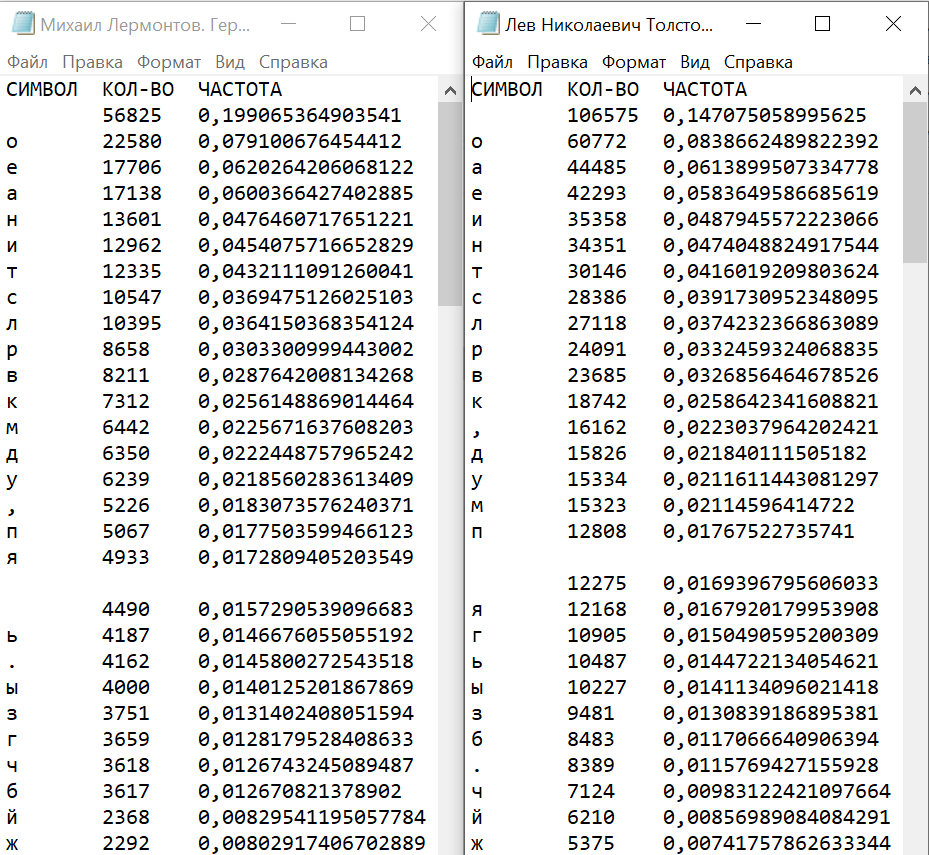
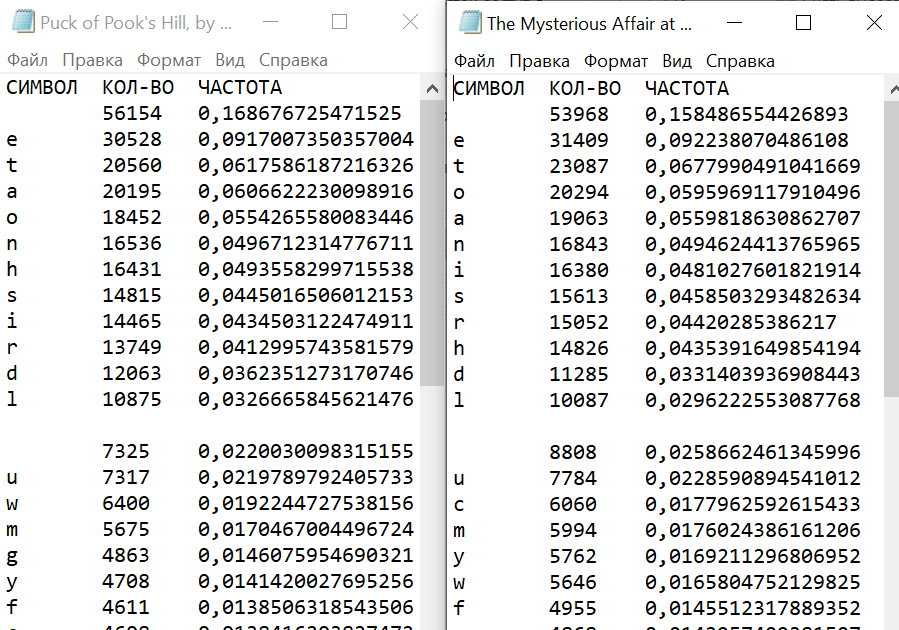
— на вход программы поступает текстовый файл;

— в результате обработки программа формирует таблицу из двух столбцов: символ или его код и количество появлений данного символа в тексте.



1.2 (необязательное задание, +2 балла). Выберите два различных документа с осмысленным текстом на английском языке и два различных документа с осмысленным текстом на русском языке. Примеры документов находятся в подпапке «Тексты».

Для каждого документа рассчитайте таблицу частот появления печатных символов и отсортируйте её по убыванию частоты. Одинаков ли порядок символов для разных текстов на одном языке? НЕТ



1.3. Выберите любой из русскоязычных файлов подпапки «Тексты» (далее X). Рассчитайте, считая алфавитом печатные символы:

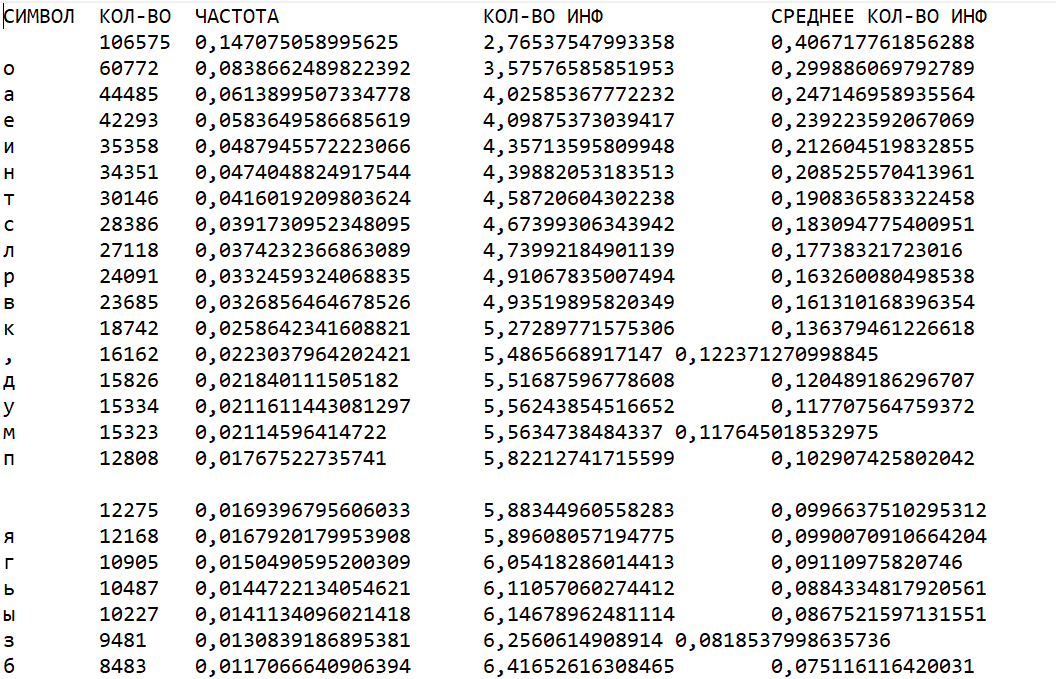
— относительные частоты символов X (количество появлений данного символа в тексте, нормированное к 1, то есть делённое на общее количество символов);

— количество информации для каждого символа X;

— среднее количество информации на символ X;

— общее количество информации в X.

ВОЙНА И МИР



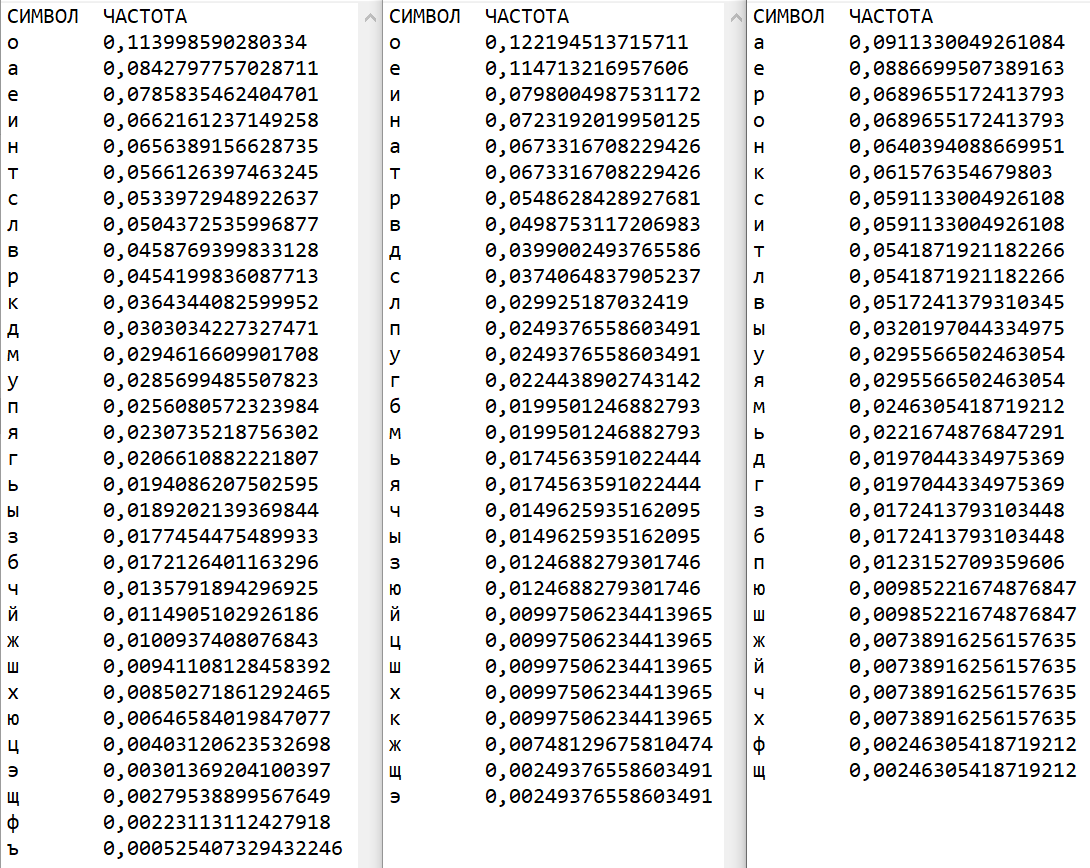


1.4. Выделите из текста файла X два различных небольших (длиной около 500 символов) фрагмента.

Рассчитайте для каждого фрагмента относительные частоты употребления букв (заглавные и строчные не различать, для русских текстов учитывать только русский алфавит, нормировать к 1, то есть делить на сумму вхождений всех рассматриваемых букв) и сравните полученные частоты с рассчитанными по всему файлу. Одинаковы ли частоты символов в целом файле X и его фрагментах? НЕТ

Отсортируйте рассчитанные частоты символов в целом файле X и двух фрагментах и по убыванию частоты. Одинаков ли порядок символов для X и его фрагментов? НЕТ

Сравните полученные частоты с таблицей частот раздела «Справочные данные». На сколько процентов различаются частоты самой частой буквы в полном файле X, фрагментах и корпусе русского/английского языка (таблице справочных данных)?



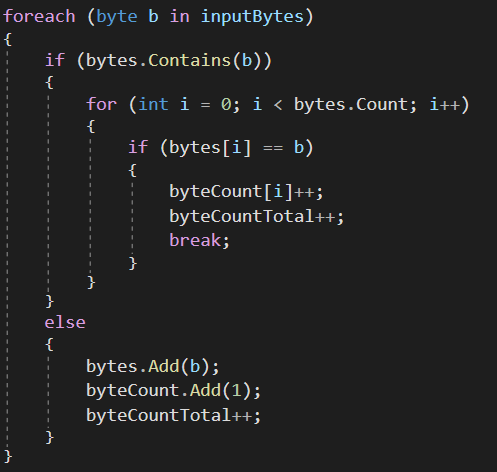
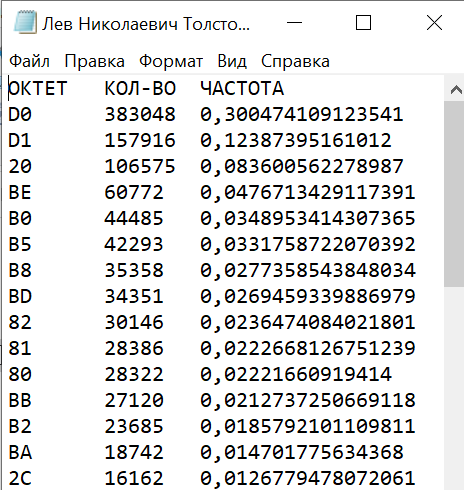
4% 11% 13%

##### 2. Частотность октетов в различных кодировках русского языка

2.1. В произвольной программной оболочке создайте программу со следующим функционалом:

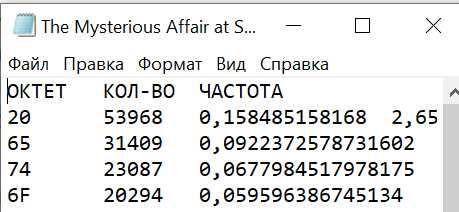
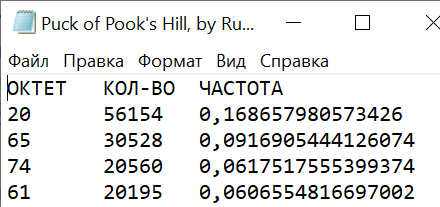
— на вход программы поступает произвольный файл;

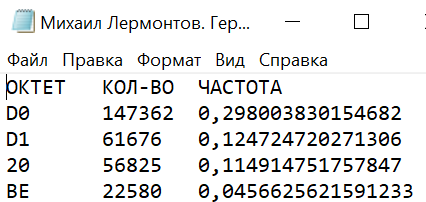
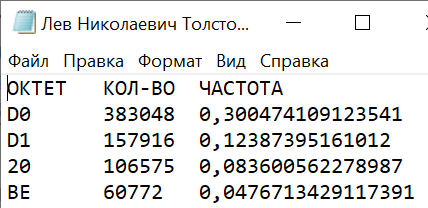
— в результате обработки программа формирует таблицу из двух столбцов: октет (для текстовых файлов — код символа в однобайтовых кодировках либо его часть в многобайтовых) и количество появлений этого октета в файле. Таблица должна быть отсортирована по убыванию частоты.



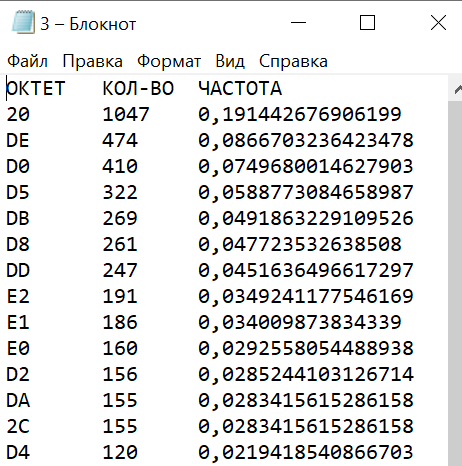
2.2. Рассчитайте частоты появления октетов в текстах, представленных в различных кодировках (папка «Тексты»).

Отсортируйте таблицы частот появления октетов по убыванию частоты и определите 4 наиболее частых октета среди всех используемых и 4 наиболее частых октета, не являющихся кодами печатных символов ASCII. Обратите внимание на распределение октетов многобайтовых кодировок.





2.3. Рассчитайте частоты появления октетов в тексте, соответствующем варианту (Z). Определите, является ли Z русскоязычным текстом в одной из стандартных кодировок (один из вариантов представляет собой нерусскоязычный текст); если да — определите кодировку и декодируйте в UTF-8.



2.4. Для файла X рассчитайте, считая алфавитом октеты:

— относительные частоты символов X;

— количество информации для каждого символа X;

— среднее количество информации на символ X;

— общее количество информации в X.

Перекодируйте файл X из UTF-8 в однобайтовую кодировку (например, KOI8-R) и выполните для полученного файла те же расчёты.

Сравните полученные величины с результатами 1.3.

