**Лабораторная работа №6**

Средства анализа данных

по курсу "Методы и средства защиты информации"

Рабочее задание

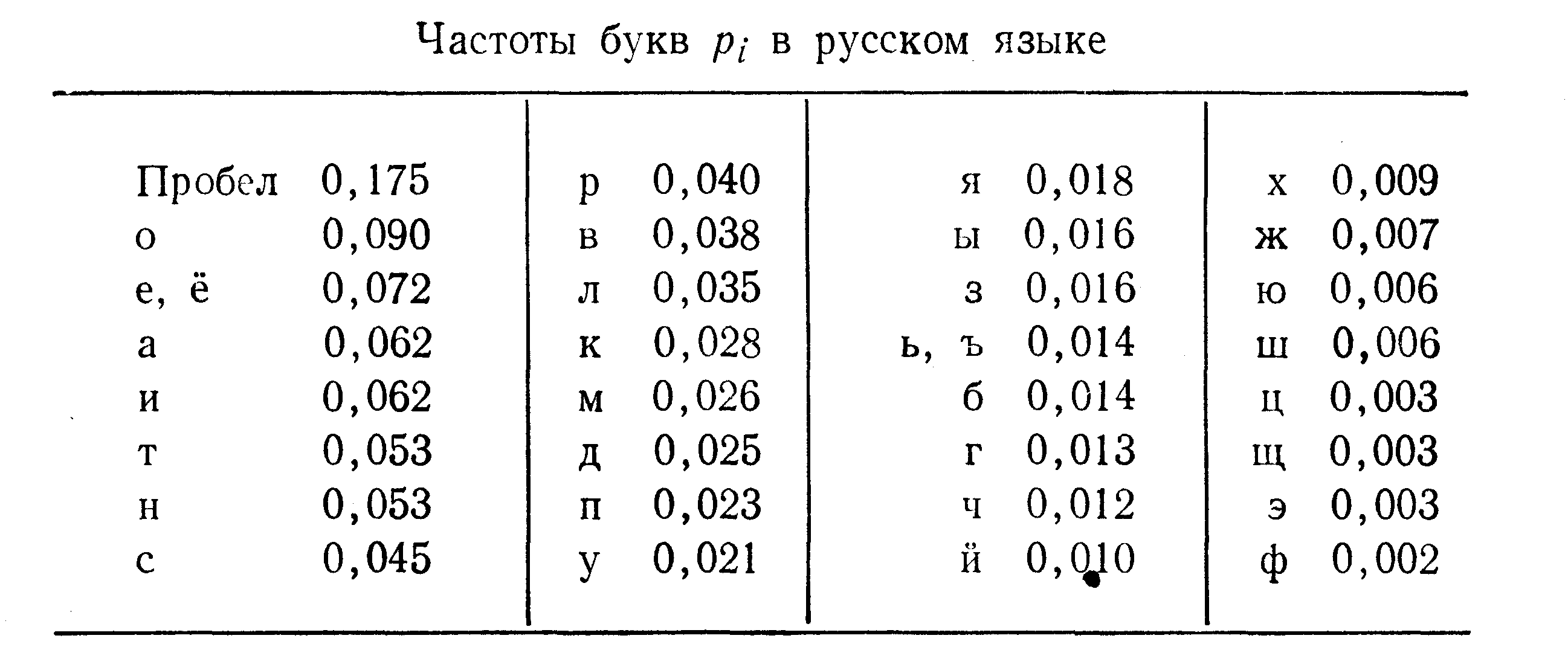
1. Подготовить два текстовых файла (\*\*\*.txt) размером 100 и 2500 знаков из plaintext.doc, начиная со страницы с номером N (N- номер по списку в группе) и сохранить их в рабочей папке в кодировке ANSI.
2. Импортировать каждый текстовый файл в свой ноутбук (документ) Mathematica, следующим образом: plntext=FromCharacterCode[Import["File Path...","Byte"],"WindowsCyrillic"],

указав путь к выбранному файлу в меню Insert //File Path//…

1. Провести преобразования в импортированных файлах: поменять прописные буквы на строчные , оставить только строчные буквы и пробелы , получить две строки длиной 30 и 1500 символов. Сохранить ноутбуки в рабочей папке.
2. Сформировать имена файлов (строк) в следующем виде: fname30=FileNameJoin[{NotebookDirectory[],"text30.txt"}] и fname1500=FileNameJoin[{NotebookDirectory[],"text1500.txt"}].
3. Провести экспрот (запись) соответствующих строк в текстовые файлы: Export[fname\*\*\*,str\*\*\*,CharacterEncoding-> "UTF8"]. Проверить наличие файлов в рабочей папке. Проверить корректность импорта Import[fname\*\*\*].

Для работы с редактором WinHex создать копию файла text30.txt в кодировке "WindowsCyrillic".

1. Запусить редактор WinHex, открыть созданный файл text30.txt, и с помощью программы PrintKеy зафиксировать полученный результат (полный экран).
2. В меню Инструменты выбрать опцию Analise File и получить на экране распределение символов, содержащихся в выбранном файле. Зафиксировать (на произвольном носителе) те буквы и их количество , вероятность появления которых в выбранном тексте выше 0.05 (5%), например : E0h а 0.0795 96 (информация появляется при наведении курсора на элемент гистограммы).
3. В пакете Mathematica создать список полученных в предыдущем пункте букв и их частот появления в виде: ch0 = {{"а", 0.0795}, {"е", 0.0902}, {"и", 0.0637}, {"н", 0.0604}, {"о", 0.0762}, {"р", 0.0513}, {"т", 0.613}, {" ", 0.1358}}
4. Подключить блок статистической обработки Needs["StatisticalPlots`"] и построить диаграмму Парето для введенного списка ParetoPlot[ch0].
5. Вычислить длину списка (вектора) Length[ch0].
6. Проверить сумму вероятностей элементов списка используя две операции: сначала создать список состоящий только из значений вероятности p0=ch0[[All,2]], а затем подсчитать сумму вероятностей summch0=Sum[p0[[i]],{i, Length[ch0]}] .
7. Ввести список наиболее вероятных частот букв русского алфавита в виде alfru={{"а",0.062},……..}. Возможно использование файла alfru.doc.



1. Сформировать вектор частот pa=alfru[[All,2]], определить длину списка na=Length[pa] и проверить сумму вероятностей summp=Sum[pa[[i]],{i,na}].
2. Рассчитать величину информационной энтропии H (энтропию языка):



1. Построить диаграмму Парето для наиболее вероятного распределения букв русского языка в тексте.
2. С помощью команды v1= ReadList["file",Byte,30] считать первые 30 байт из файла text1500.txt, где file - это полный путь к файлу, который можно ввести с помощью команд меню Insert\ File Path, 30 –число вводимых символов на первом этапе.
3. Построить гистограммы распределения букв (символов),используя команды Histogram[v1], Histogram[v1,{200,255,1}].
4. Установить параметр n2=60 и вести новый список данных v2=ReadList["file1500",Byte,n2], где file1500 –условное имя файла из п.13.
5. Определить число символов, соответствующих буквам русского языка в векторе v2: freq2=Tally[v2].
6. Создать список частот для введенных n2=60 символов текста:

p2=N[freq2[[All,2]]/Length[v2]],

где N[ ] – преобразование к действительной форме представления чисел.

1. Определить длину списка p2, сумму вероятностей, и информационную энтропию.
2. Подготовить список для записи энтропии 120 последовательно увеличивающихся сегментов подготовленного текста entropytextout=Range[120].
3. Используя оператор Do[*expr*, {j,jmax}], построить вектор значений энтропии сегментов текста, увеличивающихся каждый раз на 10 символов:

Do[{n2=10\*j;

v2=ReadList["file",Byte,n2]; freq2=Tally[v2];

p2=N[freq2[[All,2]]/Length[v2]];np2=Length[p2]};

entropytextout[[j]]=N[-Sum[p2[[i]]\*Log[2,p2[[i]]],{i,np2}]],{j,120}]

1. Построить точечный график зависимости энтропии сообщения от его длины: ListPlot[entropytextout]
2. Построить гистограмму распределения рассчитанных значений энтропии.
3. Определить среднее значение Mean[entropytextout] и дисперсию Variance[entropytextout] для распределения энтропии.
4. Найти энтропию сообщения из 1000 первых байт файла text1500.txt с точностью 7 знаков после запятой.