Практическая работа №1

Вычисление энтропии Шеннона

Цель работы: Экспериментальное вычисление оценок энтропии Шеннона текстов. Изучение свойств энтропии Шеннона.

Язык программирования: C, C++, C#, Python

Результат: программа, тестовые примеры, отчет.

Задание:
1. Для выполнения работы потребуются три текстовых файла с различными свойствами. Объем файлов больше 10 Кб, формат txt.
В первом файле содержится последовательность символов, количество различных символов больше 2 (3,4 или 5). Символы **последовательно и независимо** с равными вероятностями генерируются с помощью датчика псевдослучайных чисел и записываются в файл.

Для генерации второго файла необходимо сначала задать набор вероятностей символов (количество символов такое же, как и в первом файле), а затем **последовательно и независимо** генерировать символы с соответствующей вероятностью и записывать их в файл, вероятности в процессе записи файла не меняются.

В качестве третьего файла необходимо выбрать художественный текст на русском (английском) языке. Для алфавита текста предполагается, что строчные и заглавные символы не отличаются, знаки препинания опущены, к алфавиту добавлен пробел, для русских текстов буквы «е» и «ё», «ь» и «ъ» совпадают.

2. Составить программу, определяющую несколько оценок энтропии созданных текстовых файлов. Вычисление значения по формуле Шеннона **настоятельно рекомендуется** оформить в виде отдельной функции, на вход которой подается массив (список) вероятностей символов, выходной параметр – значение, вычисленное по формуле Шеннона.

Вычислить три оценки энтропии Шеннона для каждого из файлов. Рекомендуется вычисление оценки оформить в виде отдельной функции с параметром имя файла:
Первая оценка H_1. Сначала определить частоты отдельных символов файла, т.е. отношения количества отдельного символа к общему количеству символов в файле. Далее используя полученные частоты как оценки вероятностей, рассчитать оценку энтропии по формуле Шеннона.

Вторая оценка H_2. Определить частоты всех последовательных пар символов в файле. Для того правильной оценки энтропии H_2 пары символов нужно рассматривать с перехлестом.
Пример. Пусть имеется такая последовательность ФЫВАФПРО
Под парами понимаются пары соседних символов, т.е.

ФЫ ЫВ ВА АФ ФП ПР РО

Для подсчета оценки энтропии H_2 необходимо подсчитать частоту каждой пары символов и подставить в формулу Шеннона. Полученное значение оценки энтропии следует разделить на 2.

Третья оценка H_3. Определить частоты всех последовательных троек символов в файле. Для того правильной оценки энтропии H_3 **тройки** символов нужно рассматривать с перехлестом.
Для подсчета оценки энтропии H_3 необходимо подсчитать частоту каждой тройки символов и подставить в формулу Шеннона. Полученное значение оценки энтропии следует разделить на 3.
По желанию можно продолжить процесс вычисления оценок с использованием частот четверок, пятерок символов и т.д.

3. После тестирования программы необходимо заполнить таблицу для отчета и в отчете **проанализировать** полученные результаты, объяснить замеченные эффекты. Для получения теоретических значений энтропии использовать наборы вероятностей, которые использовались при генерации файлов, для файла с текстом на естественном языке не заполнять.

Название файла	H_1	H_2	H_3	Максимально возможное значение энтропии	Теоретическое значение энтропии
файл 1					
файл 2					
файл 3					

4. Оформить отчет, должен содержать заполненную таблицу и анализ полученных результатов, по желанию в отчет можно включить описание программной реализации. **В отчет не нужно включать содержимое этого файла.** Загрузить отчет в электронную среду.

5. Анализ полученных результатов можно оформить в виде ответов на вопросы

1. Каким образом реализована генерация файлов?
2. Как соотносятся между собой полученные оценки для каждого файла? Равны? не равны? Поясните в каких случаях получаются примерно равные оценки, а в каких – нет.
3. Если продолжить процесс вычисления оценок энтропии, используя более длинные последовательности символов, то можно получить последовательность значений. Спрогнозируйте поведение такой последовательности – последовательность будет возрастать или убывать? имеет ли предел такая последовательность оценок?
4. Какие значения энтропии Шеннона по вашему мнению имеют тексты из файлов?