Задача № 6

Прогнозирование временных рядов с помощью методов сжатия данных

То, что в теории информации называется "сообщением", в прогнозировании называется "временным рядом" (ударение на букву 'ы'). Задача ставится следующим образом. Заданы алфавит $A=\{a_1,\ldots,a_N\}$ и ряд $x_1,\ldots,x_t$, в котором все $x_i \in A$. Требуется найти условное распределение вероятностей появления следующего символа $P(X_{t+1}\,|\,x_t,\ldots,x_1)$ или нескольких следующих символов $P(X_{t+k},\ldots,X_{t+1}\,|\,x_t,\ldots,x_1)$ (для всех комбинаций букв алфавита).

По формуле умножения вероятностей

$$P(a\,|\,x_t,\ldots x_1)=\frac{P(x_1,\ldots,x_t,a)}{P(x_1,\ldots,x_t)}$$

для всех букв $a \in A$. Оценить вероятности, стоящие в знаменателе и числителе, можно через длины кодов для соответствующих сообщений. Предположим, что мы можем построить совершенные коды (имеющие нулевую избыточность). Обозначим через l длину такого кода для $x_1,\ldots,x_t$, а через l_a – длину кода для того же сообщения с присоединенной к нему буквой a (длина кода выражается в битах и для совершенного кода может быть нецелым числом). Тогда по теореме Шеннона о кодировании источника

$$P(x_1,\ldots,x_t)=2^{-l},\quad P(x_1,\ldots,x_t,a)=2^{-l_a},$$

откуда

$$P(a\,|\,x_t,\ldots,x_1)=2^{-(l_a-l)}.$$

На практике все коды имеют избыточность, поэтому через длины кодов мы получаем только некоторую *оценку* вероятности. Полученные таким образом оценки не обязаны суммироваться в единицу. Если для дальнейшего использования оценок вероятностей желательно привести их к виду какого-то распределения, то можно каждую оценку поделить на сумму всех оценок.

Задание. Для нескольких неоконченных текстов разного типа спрогнозировать

1) распределение вероятностей следующего символа;

2) оценить вероятности нескольких возможных и невозможных продолжений.

В качестве методов сжатия использовать стандартные архиваторы и собственную программу, разработанную на предыдущих лабораторных работах.