

Проводится последовательность экспериментов.

В каждом эксперименте получается слово фиксированной длиной 4096 символов.

Каждый символ может принимать одно из 256 значений с заданной вероятностью $P_{x,y}$

где:

x — позиция данного символа в слове и принимает значения от 1 до 4096

y — значение данного символа и принимает значение от 1 до 256

Для любого x верно равенство:

$$\sum_{y=1}^{256} P_{x,y} = 1$$

Вероятности $P_{x,y}$ являются заданными табличными значениями и не описаны какими либо стандартными функциями распределения.

Эксперименты повторяются до тех пор, пока не получится заданное количество N **уникальных** слов.

Если в результате очередного эксперимента получается неуникальное слово, то полученный результат этого эксперимента отбрасывается, и эксперимент повторяется до получения уникального слова.

Итоговое количество уникальных слов N может быть достаточно большим и принимает значения от миллиона и более.

Каждое из получившихся N слов имеют минимальный набор из R_i уникальных **первых** символов, по которым можно однозначно идентифицировать это слово из всего набора слов, полученных в результате экспериментов, где i — номер эксперимента и может принимать значения от 1 до N .

R_i может принимать значения от 1 (слово имеет уникальную среди всех экспериментов **первую** букву) до 4096 (слово отличается от какого либо другого слова только **последней** буквой).

Требуется определить:

1. Для каждого x от 1 до 4096 количество Q_x слов, которое можно однозначно идентифицировать по первым x символам, т. е.:

Q_1 — количество слов из всех N экспериментов, которые можно однозначно идентифицировать по первому символу (при обозначенных значениях N ожидаемое значение Q_1 близко к нулю)

Q_2 — количество слов, которые можно однозначно идентифицировать по первым двум символам, и т.д.

2. Математическое ожидание (среднее) значение $M[R_i]$ для заданных N и $P_{k,l}$.

Упрощённый пример-иллюстрация:

Имеем слова-результаты экспериментов:

Номер эксперимента	Слово										R_i
1	м	а	м	а							4
2	м	а	м	о	н	т	и	х	а		7
3	м	а	с	к	а						4
4	м	а	г	л							3
5	м	а	м	о	н	т	ё	н	о	к	7
6	м	а	м	б	а						4
7	в	о	д	а							1
8	м	а	с	я	н	я					4

Красным цветом для каждого слова выделены первые R_i символов, по которым можно однозначно идентифицировать каждое из этих слов.

Искомые величины:

Количества Q_x слов, идентифицируемых по первым x буквам:

$Q_1 = 1, Q_2 = 0, Q_3 = 1, Q_4 = 4, Q_5 = 0, Q_6 = 0, Q_7 = 2$

Искомая величина $M[R_i]$ равняется:

$$M[R_i] = \frac{1}{N} \sum_{i=1}^N R_i = \frac{4+7+4+3+7+4+1+4}{8} = 4,25$$