Лабораторная работа 4

Определить понятие «количества информации» сложно. В решении этой проблемы существуют два основных подхода, которые исторически возникли одновременно. В конце 40-х годов XX века один из основоположников кибернетики американский математик Клод Шеннон развил вероятностный подход к измерению количества информации, а работы по созданию ЭВМ привели к «объемному» подходу.

Рассмотрим вероятностный подход более подробно.

В качестве примера разберем опыт, связанный с бросанием правильной игральной кости, имеющей N граней. Результаты данного опыта могут быть следующие: выпадение грани с одним из следующих знаков: 1, 2, . . . N.

Введем в рассмотрение численную величину, измеряющую неопределенность — **энтропию** (обозначим ее H).Согласно развитой теории, в случае равновероятного выпадения каждой из граней величины N и *H* связаны между собой **формулой Хартли**

$$H = log_2 N. \qquad (1)$$

Важным при введения какой-либо величины является вопрос о том, что принимать за единицу ее измерения. Очевидно, H будет равно единице при N=2. То есть, в качестве единицы принимается количество информации, связанное с проведением опыта, состоящего в получении одного из двух равновероятных исходов (примером такого опыта может служить бросание монеты, при котором возможны два исхода- «орел», «решка»). Такая единица количества информации называется «бит».

В случае, когда вероятности P_i, результатов опыта (в примере, приведенном выше — бросания игральной кости) неодинаковы, имеет место формула Шеннона

$$H = -\sum_{i=1}^{N} P_i \times log_2 P_i. \qquad (2)$$

В случае равной вероятности событий $P_i = \frac{1}{N}$, формула Шеннона переходит в формулу Хартли (1).

В качестве примера определим количество информации, связанное с появлением каждого символа в сообщениях, записанные на русском языке. Будем считать, что русский алфавит состоит из 33 букв и знака «пробел» для разделения слов. По формуле Хартли $H=\mathrm{Log}_2 34 \sim 5{,}09$ бит.

Однако в словах русского языка (равно как и в словах других языков) различные буквы встречаются неодинаково часто. В табл. 1приведенавероятностьчастотыупотребленияразличных знаков русского алфавита, полученная на основе анализа очень больших по объему текстов.

Частотность букв русского языка

i	Символ	P(i)	I	Символ	P(i)	I	Символ	P(i)
1	_	0.175	12	Л	0.035	23	Б	0.014
2	О	0.090	13	К	0.028	24	Г	0.012
3	Е	0.072	14	М	0.026	25	Ч	0.012
4	Ё	0.072	15	Д	0.025	26	Й	0.010
5	А	0.062	16	П	0.023	27	Х	0.009
6	И	0.062	17	У	0.021	28	Ж	0.007
7	Т	0.053	18	Я	0.018	29	Ю	0.006
8	Н	0.053	19	Ы	0.016	30	Ш	0.006
9	С	0.045	20	З	0.016	31	Ц	0.004
10	Р	0.040	21	Ь	0.014	32	Щ	0.003
11	В	0.038	22	Ъ	0.014	33	Э	0.003
						34	Ф	0.002

Воспользуемся для подсчета *H* формулой Шеннона: *H*~ 4.72 бит. Получение значение Н как и можно предположить, меньше вычисленного ранее. Величина Н, вычисляемая по формуле Хартли, является: максимальным количеством информации, которое могло бы проходиться на один знак.

Аналогичные подсчеты Н можно провести и для других языков, например, использующих латинский алфавит— немецкий, французский др. (26 различных букв и «пробел»). По формуле Хартли получим Н= $\log_2 27 \sim 4{,}76$ бит.

Пример №1. Подсчитать количество информации, приходящейся на один символ, в следующем тексте экономического содержания:

Организационно-правовые формы предприятий в своей основе определяют форму их собственности, то есть, кому принадлежит предприятие, его основные фонды, оборотные средства, материальные и денежные ресурсы. В за-

висимости от формы собственности в России в настоящее время различают три основные формы предпринимательской деятельности: частную, коллективную и контрактную.

Решение. Подсчет всех символов текста выполним с помощью статистики подсчета числа знаков в документе. Для этого перейдем на вкладке **Рецензирование** в разделе **Правописание,** выделим набранный текст и нажмем кнопку «статистика» 

В результате появиться окно, представленное на рис. 1.

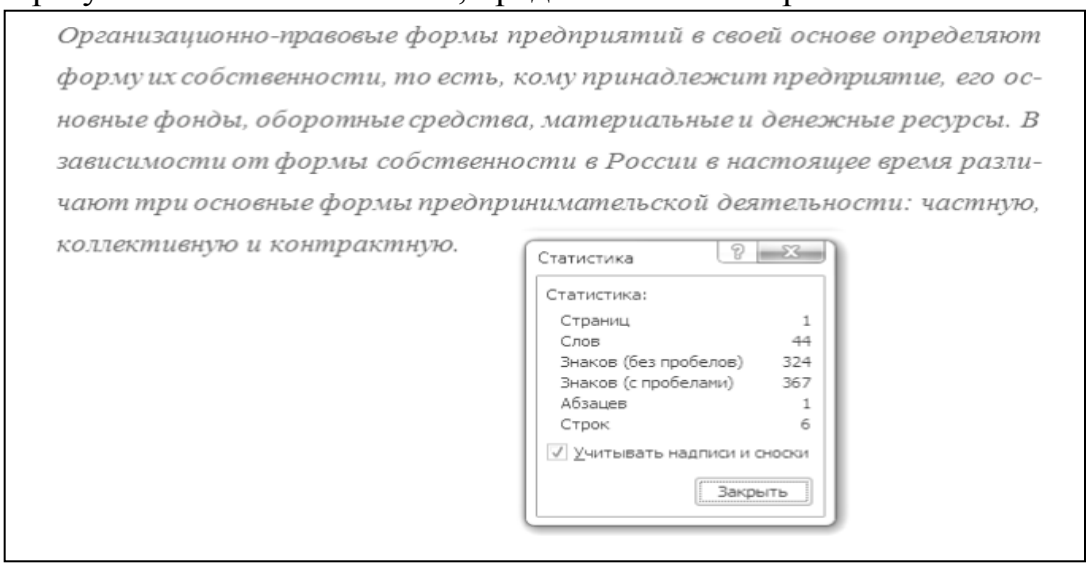
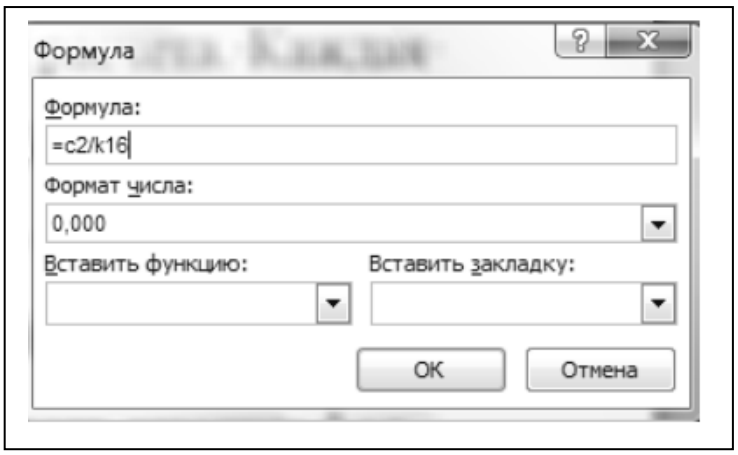


Рис. 1. Статистика подсчета числа знаков в документе

Таким образом, количество знаков в тексте вместе с пробелами – 367, число пробелов – 43 (367-324=43). Подсчитаем количество каждого символа в тексте и занесем в табл. 5. Определим вероятность P_iкаждого символа в тексте как отношение количества одинаковых символов каждого значения ко всему числу символов в тексте. Вычисления в таблице будем производить в текстовом редакторе Word с помощью добавления в ячейку таблицы формулы выполнения простого расчета. Каждая ячейка таблицы Wood имеет адрес, состоящий из номера столбца (обозначаемого буквами латинского алфавита) и номера строки (обозначаемого арабскими цифрами). Для вычисления P_i необходимо поместить курсор в ячейку D2, где будет помещен результат. На вкладке **Макет** в разделе **Данные** выбрать кнопку формула. Окно **Формула** заполнить как показано на рис. 2.



Аналогично вычислим остальные значения P_i каждого символа. Результаты вычислений в табл. 2.

Таблица 2

Частотность букв в тексте

i	Сим-вол	Кол-во симво-лов	P_i	i	Символ	Кол-во сим-волов	P_i	i	Символ	Кол-во сим-волов	P_i
1	_	43	0,117	14	Р	24	0,065	27	Б	3	0,008
2	,	6	0,016	15	В	16	0,044	28	Г	2	0,005
3	.	2	0,005	16	Л	8	0,022	29	Ч	2	0,005
4	-	1	0,003	17	К	6	0,016	30	Й	3	0,008
5	:	1	0,003	18	М	9	0,025	31	Х	1	0,003
6	О	35	0,095	19	Д	9	0,025	32	Ж	2	0,005
7	Е	32	0,087	20	П	9	0,025	33	Ю	5	0,014
8	Ё	0	0,000	21	У	6	0,016	34	Ш	0	0,000
9	А	14	0,038	22	Я	5	0,014	35	Ц	1	0,003
10	И	25	0,068	23	Ы	11	0,030	36	Щ	1	0,003
11	Т	25	0,068	24	З	3	0,008	37	Э	0	0,000
12	Н	25	0,068	25	Ь	4	0,011	38	Ф	5	0,014
13	С	23	0,063	26	Ъ	0	0,000	39			
Σ		232		Σ		110		Σ		25	
								Σ		367	

В предпоследней строке табл. 2 стоит суммарное количество знаков, а в последней строке общая сумма всех знаков текста. Далее по формуле Шеннона подсчитаем количество информации, приходящейся на один символ. Для этого выполним предварительные вычисления в табличном процессоре Excel.

Откроем лист Excel и заполним его как показано на рис. 3.

	A	B	C	D	E	F	G	H	I	J	K	L
1	i	P(i)	$log_2 P_i$	$P_i \times log_2 P_i$	i	P(i)	$log_2 P_i$	$P_i \times log_2 P_i$	i	P(i)	$log_2 P_i$	$P_i \times log_2 P_i$
2	1	0,117	=LOG(B2;2)	=B2*C2	14	0,065	=LOG(F2;2)	=F2*G2	27	0,008	=LOG(J2;2)	=J2*K2
3	2	0,016	=LOG(B3;2)	=B3*C3	15	0,044	=LOG(F3;2)	=F3*G3	28	0,005	=LOG(J3;2)	=J3*K3
4	3	0,005	=LOG(B4;2)	=B4*C4	16	0,022	=LOG(F4;2)	=F4*G4	29	0,005	=LOG(J4;2)	=J4*K4
5	4	0,003	=LOG(B5;2)	=B5*C5	17	0,016	=LOG(F5;2)	=F5*G5	30	0,008	=LOG(J5;2)	=J5*K5
6	5	0,003	=LOG(B6;2)	=B6*C6	18	0,025	=LOG(F6;2)	=F6*G6	31	0,003	=LOG(J6;2)	=J6*K6
7	6	0,095	=LOG(B7;2)	=B7*C7	19	0,025	=LOG(F7;2)	=F7*G7	32	0,005	=LOG(J7;2)	=J7*K7
8	7	0,087	=LOG(B8;2)	=B8*C8	20	0,025	=LOG(F8;2)	=F8*G8	33	0,014	=LOG(J8;2)	=J8*K8
9	8	0		0	21	0,016	=LOG(F9;2)	=F9*G9	34	0		0
10	9	0,038	=LOG(B10;2)	=B10*C10	22	0,014	=LOG(F10;2)	=F10*G10	35	0,003	=LOG(J10;2)	=J10*K10
11	10	0,068	=LOG(B11;2)	=B11*C11	23	0,03	=LOG(F11;2)	=F11*G11	36	0,003	=LOG(J11;2)	=J11*K11
12	11	0,068	=LOG(B12;2)	=B12*C12	24	0,008	=LOG(F12;2)	=F12*G12	37	0		0
13	12	0,068	=LOG(B13;2)	=B13*C13	25	0,011	=LOG(F13;2)	=F13*G13	38	0,014	=LOG(J13;2)	=J13*K13
14	13	0,063	=LOG(B14;2)	=B14*C14	26	0		0	39			=J14*K14
15	Σ	=СУММ(B2:B14)		=СУММ(D2:D14)	Σ	=СУММ(F2:F14)		=СУММ(H2:H14)	Σ	=СУММ(J2:J14)		=СУММ(L2:L14)
16									Σ	=B15+F15+J15		
17												
18												
19					H=	=-(D15+H15+L15)						
20												

Рис. 3.

При заполнении листа пользуемся встроенной функцией логарифма числа по основанию 2. Для этого на вкладке **Формулы** выбираем **Математические** и щелкаем на этой кнопке. В появившемся выпадающем меню выбираем функцию LOG. Появится окно, представленное на рис. 4. В поле число записываем адрес ячейки, в которой находится число логарифм, которого вычисляется, а в поле основание записываем 2,так как нам необходимо вычислить логарифм по основанию 2.

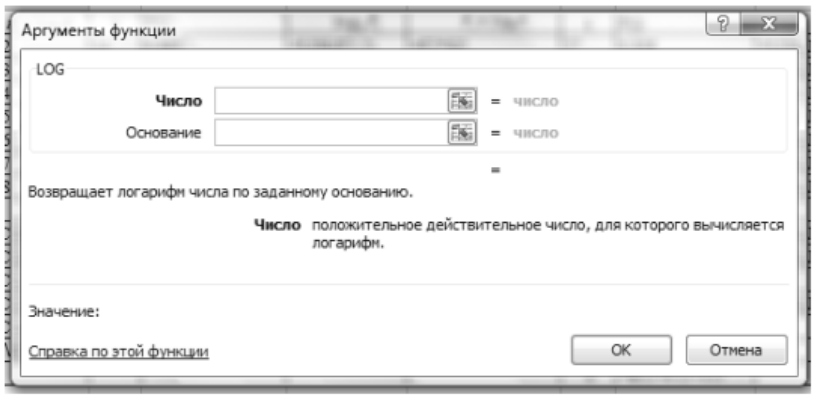


Рис. 4.

На рак. 4 в ячейках *C9*, G14, *K9 и*K12значения логарифма не вычисляются, так как количество соответствующих символов равно нулю, а следовательно и их частота в тексте равна нулю. На рис. 4 в строке с номером I5 в столбцах B, F, J вычислены суммарные вероятности по

столбцу. Это необходимо сделать для контроля. В строке с номером 16 в столбце J вычислено контрольное значение вероятности. Это значение должно быть равно единицы. В ячейке F19 вычисляется количество информации согласно формуле Шеннона.

Числовые значения. Полученные в Excel, приведены в таб. 3.

Таблица 3

Предварительные вычисления для формулы Шеннона

I	P_i	$log_2 P_i$	$P_i \times log_2 P_i$	i	P_i	$log_2 P_i$	$P_i \times log_2 P_i$	i	P_i	$log_2 P_i$	$P_i \times log_2 P_i$
1	0,117	-3,0954	-0,3622	14	0,065	-3,9434	-0,2563	27	0,008	-6,9658	-0,0557
2	0,016	-5,9658	-0,0955	15	0,044	-4,5064	-0,1983	28	0,005	-7,6439	-0,0382
3	0,005	-7,6439	-0,0382	16	0,022	-5,5064	-0,1211	29	0,005	-7,6439	-0,0382
4	0,003	-8,3808	-0,0251	17	0,016	-5,9658	-0,0955	30	0,008	-6,9658	-0,0557
5	0,003	-8,3808	-0,0251	18	0,025	-5,3219	-0,1330	31	0,003	-8,3808	-0,0251
6	0,095	-3,3959	-0,3226	19	0,025	-5,3219	-0,1330	32	0,005	-7,6439	-0,0382
7	0,087	-3,5228	-0,3065	20	0,025	-5,3219	-0,1330	33	0,014	-6,1584	-0,0862
8	0		0,0000	21	0,016	-5,9658	-0,0955	34	0		0,0000
9	0,038	-4,7179	-0,1793	22	0,014	-6,1584	-0,0862	35	0,003	-8,3808	-0,0251
10	0,068	-3,8783	-0,2637	23	0,03	-5,0589	-0,1518	36	0,003	-8,3808	-0,0251
11	0,068	-3,8783	-0,2637	24	0,008	-6,9658	-0,0557	37	0		0,0000
12	0,068	-3,8783	-0,2637	25	0,011	-6,5064	-0,0716	38	0,014	-6,1584	-0,0862
13	0,063	-3,9885	-0,2513	26	0		0,0000	39			0,0000
Σ	0,631		-2,3970	Σ	0,301		-1,5311	Σ	0,068		-0,4737
								Σ	1		

Таким образом, количество информации согласно формуле Шеннона, приходящейся на один символ, в данном тексте H=4.40199~4,40 бита.

Максимальное количество информации, которое могло бы приходиться на один знак в данном тексте, вычисляемое по формуле Хартли, H=log2 367 ~ 8,5196 бит.