Задание 1

Оценить вероятность появления символов и вычислить величину энтропии символов $H_i = \log_2 \frac{1}{p(x_i)}$ и текста $H = -\sum p(x_i) H_i$.

In [2]:
```
def H(ps):
    ps = list(ps.values())
    return -np.sum(ps * np.log2(ps))
```

Задание 2

Вычислить значение энтропии $H*$ для пар символов (x_i, x_j) как $-\sum p(x_i/x_j) * p(x_j) * \log p(x_i/x_j)$, где $p(x_i/x_j)$ — вероятность встречи пары.

In [3]:
```
def H_star(ps, pair_ps):
    return -sum(p_ij * ps[pair[1]] * np.log2(p_ij) for pair, p_ij in pair_ps.items())
```

Таблица к заданию 1

In [4]:
```
def analyze_letters(file):
    # letter_probabilities is a native function, see entropy/src/lib.rs
    probabilities = entropy.letter_probabilities(file)
    t = pd.DataFrame()
    t['Символ'] = probabilities.keys()
    t['Вероятность символа'] = probabilities.values()
    t['Энтропия символа'] = np.log2(1 / t['Вероятность символа'])
    display(Latex(f'Значение энтропии $H$ = {H(probabilities):.4f}'))
    display(t)

print('Введите путь к файлу:')
file = input()
analyze_letters(file)
```

```
Введите путь к файлу:
machine_50k
```

Значение энтропии H = 4.1529

	Символ	Вероятность символа	Энтропия символа
0		0.1745	2.5183
1	.	0.0300	5.0604
2	a	0.0638	3.9709
3	b	0.0125	6.3219
4	c	0.0190	5.7170
5	d	0.0368	4.7661
6	e	0.1020	3.2938
7	f	0.0148	6.0766
8	g	0.0139	6.1667
9	h	0.0557	4.1651
10	i	0.0562	4.1544
11	j	0.0006	10.5941
12	k	0.0057	7.4502
13	l	0.0315	4.9870
14	m	0.0201	5.6339
15	n	0.0522	4.2611
16	o	0.0577	4.1147
17	p	0.0139	6.1732
18	q	0.0007	10.4608
19	r	0.0439	4.5087
20	s	0.0508	4.2980
21	t	0.0781	3.6783
22	u	0.0235	5.4126
23	v	0.0078	6.9937
24	w	0.0184	5.7620
25	x	0.0013	9.6175
26	y	0.0140	6.1560
27	z	0.0004	11.1560

Таблица к заданию 2

In [5]:
```
def analyze_letter_pairs(files):
    file_ps = [(entropy.letter_probabilities(f),
                entropy.letter_pair_probabilities(f)) for f in files]
    t = pd.DataFrame()
    t['Файл'] = files
    t['Энтропия H'] = [H(ps) for ps, _ in file_ps]
    t['Энтропия H*'] = [H_star(ps, pps) for ps, pps in file_ps]
    display(t.T)

print('Введите путь к файлам через запятую:')
files = [f.strip() for f in input().split(",")]
analyze_letter_pairs(files)
```

```
Введите путь к файлам через запятую:
machine_50k, rama_60k, brave_30k
```

	0	1	2
Файл	machine_50k	rama_60k	brave_30k
Энтропия H	4.1529	4.1909	4.2036
Энтропия H*	0.3860	0.3901	0.3782