Оптимальное кодирование.

Построить коды Фано и Хаффмана, найти цены кодирования, сравнить с энтропией и длиной равномерного кода:

1. $A = (a_1, \dots, a_{12})$, $B = \{0, 1\}$, $p = (0,18; 0,139; 0,128; 0,113; 0,103; 0,098; 0,097; 0,059; 0,044; 0,017; 0,013; 0,009)$.
2. $A = (a_1, \dots, a_{12})$, $B = \{0, 1\}$ $p = (0,094; 0,066; 0,052; 0,138; 0,148; 0,033; 0,009; 0,002; 0,101; 0,117; 0,103; 0,137)$.
3. $A = \{0, 1\}$, $p = \{2/3, 1/3\}$. Построить код Хаффмана для множества A^3, состоящего из всех слов длины 3. Найти среднее количество бит на символ, сравнить со средней ценой оптимального побуквенного кодирования и энтропией.

4.* Рассмотреть кодирование русского и английского алфавитов.

5.* Случайная величина X задана распределением $P(X = 2^n) = 1/2^n$, $n = 1, 2, ...$ Найти энтропию X. Придумать оптимальное кодирование для X, вычислить его среднюю длину.

Алогоритм построения кода Фано (или Шеннона–Фано).

Для алфавита $A = \{a_1, \dots, a_m\}$ известны вероятности появления букв в сообщении $p_1 \geqslant \dots \geqslant p_m > 0$. Алфавит A разбивается на две группы $A_1 = \{a_1, \dots, a_k\}$ и $A_2 = \{a_{k+1}, \dots, a_m\}$ так, чтобы суммарные вероятности в каждой группе были как можно ближе друг другу, т.е. определяется номер k такой, что

$|(p_1 + p_2 + \dots + p_k) - (p_{k+1} + p_{k+2} ... + p_m)| \leqslant |(p_1 + p_2 + \dots + p_l) - (p_{l+1} + p_{l+2} ... + p_m)| \quad \forall l = 1, \dots, m.$

Буквам первой группы приписывается код 0, второй — 1. После этого каждая из полученных групп A_1, A_2 снова разбивается на две с наиболее близкими суммарными вероятностями. К коду букв первой из полученных подгрупп приписывается 0, второй 1. И т.д. Каждая группа разбивается до тех пор, пока в не останется одна буква. Коды всех букв построены. По построению схема является префиксной, а значит и разделимой.

Пример. $p = (0.4; 0.2; 0.2; 0.1; 0.05; 0.05)$. Получим таблицу (жирным шрифтом выделены элементарные коды букв).

0,4	0	**00**		
0,2	0	**01**		
0,2	1	**10**		
0,1	1	11	**110**	
0,05	1	11	111	**1110**
0,05	1	11	111	**1111**

Таблица частот букв для английского текста

$A - 0,082$	$B - 0,014$	$C - 0,028$	$D - 0,038$	$E - 0,131$	$F - 0,029$	$G - 0,020$	$H - 0,053$	$I - 0,063$
$J - 0,001$	$K - 0,004$	$L - 0,034$	$M - 0,025$	$N - 0,071$	$O - 0,080$	$P - 0,020$	$Q - 0,001$	$R - 0,068$
$S - 0,061$	$T - 0,105$	$U - 0,025$	$V - 0,009$	$W - 0,015$	$X - 0,002$	$Y - 0,020$	$Z - 0,001$	

Таблица частот букв для русского текста

а $- 0,0802$	б $- 0,0159$	в $- 0,0435$	г $- 0,0178$	д $- 0,0320$	е $- 0,0848$	ё $- 0,0037$	ж $- 0,0091$	з $- 0,0167$
и $- 0,0723$	й $- 0,0100$	к $- 0,0353$	л $- 0,0470$	м $- 0,0371$	н $- 0,0673$	о $- 0,1110$	п $- 0,0275$	р $- 0,0458$
с $- 0,0536$	т $- 0,0597$	у $- 0,0234$	ф $- 0,0027$	х $- 0,0074$	ц $- 0,0048$	ч $- 0,0149$	ш $- 0,0062$	щ $- 0,0032$
ъ $- 0,0003$	ы $- 0,0192$	ь $- 0,0179$	э $- 0,0038$	ю $- 0,0062$	я $- 0,0197$			

Оптимальное кодирование.

Построить коды Фано и Хаффмана, найти цены кодирования, сравнить с энтропией и длиной равномерного кода:

1. $A = (a_1, \dots, a_{12})$, $B = \{0, 1\}$, $p = (0,18; 0,139; 0,128; 0,113; 0,103; 0,098; 0,097; 0,059; 0,044; 0,017; 0,013; 0,009)$.
2. $A = (a_1, \dots, a_{12})$, $B = \{0, 1\}$ $p = (0,094; 0,066; 0,052; 0,138; 0,148; 0,033; 0,009; 0,002; 0,101; 0,117; 0,103; 0,137)$.
3. $A = \{0, 1\}$, $p = \{2/3, 1/3\}$. Построить код Хаффмана для множества A^3, состоящего из всех слов длины 3. Найти среднее количество бит на символ, сравнить со средней ценой оптимального побуквенного кодирования и энтропией.

4.* Рассмотреть кодирование русского и английского алфавитов.

5.* Случайная величина X задана распределением $P(X = 2^n) = 1/2^n$, $n = 1, 2, ...$ Найти энтропию X. Придумать оптимальное кодирование для X, вычислить его среднюю длину.

Код Фано (или код Шеннона–Фано).

Для алфавита $A = \{a_1, \dots, a_m\}$ известны вероятности появления букв в сообщении $p_1 \geqslant \dots \geqslant p_m > 0$. Алфавит A разбивается на две группы $A_1 = \{a_1, \dots, a_k\}$ и $A_2 = \{a_{k+1}, \dots, a_m\}$ так, чтобы суммарные вероятности в каждой группе были как можно ближе друг другу, т.е. определяется номер k такой, что

$|(p_1 + p_2 + \dots + p_k) - (p_{k+1} + p_{k+2} ... + p_m)| \leqslant |(p_1 + p_2 + \dots + p_l) - (p_{l+1} + p_{l+2} ... + p_m)| \quad \forall l = 1, \dots, m.$

Буквам первой группы приписывается код 0, второй — 1. После этого каждая из полученных групп A_1, A_2 снова разбивается на две с наиболее близкими суммарными вероятностями. К коду букв первой из полученных подгрупп приписывается 0, второй 1. И т.д. Каждая группа разбивается до тех пор, пока в не останется одна буква. Коды всех букв построены. По построению схема является префиксной, а значит и разделимой.

Пример. $p = (0.4; 0.2; 0.2; 0.1; 0.05; 0.05)$. Получим таблицу (жирным шрифтом выделены элементарные коды букв).

0,4	0	**00**		
0,2	0	**01**		
0,2	1	**10**		
0,1	1	11	**110**	
0,05	1	11	111	**1110**
0,05	1	11	111	**1111**

Таблица частот букв для русского текста

а $- 0,0802$	б $- 0,0159$	в $- 0,0435$	г $- 0,0178$	д $- 0,0320$	е $- 0,0848$	ё $- 0,0037$	ж $- 0,0091$	з $- 0,0167$
и $- 0,0723$	й $- 0,0100$	к $- 0,0353$	л $- 0,0470$	м $- 0,0371$	н $- 0,0673$	о $- 0,1110$	п $- 0,0275$	р $- 0,0458$
с $- 0,0536$	т $- 0,0597$	у $- 0,0234$	ф $- 0,0027$	х $- 0,0074$	ц $- 0,0048$	ч $- 0,0149$	ш $- 0,0062$	щ $- 0,0032$
ъ $- 0,0003$	ы $- 0,0192$	ь $- 0,0179$	э $- 0,0038$	ю $- 0,0062$	я $- 0,0197$			