Задание №4. Условие Фано

Теория

Кодирование/декодирование информации

Код/кодовое слово - правило, по которому *информация* превращается в *цепочку каких-либо знаков или символов*. *В случае 4 задания, мы присваиваем буквам определенную последовательность нулей и единиц.*

Кодовая последовательность - последовательность кодов/кодовых слов.

Виды кодирования:

1. При равномерном кодировании коды имеют одинаковую длину
2. При неравномерном кодировании коды имеют разные длины

Кодирование - процесс представления информации в виде, удобном для её дальнейшего хранения, передачи и обработки. *В случае 4 задания, кодирование происходит в двоичный код, то есть последовательность нулей и единиц.*

Декодирование - процесс восстановления исходной информации из последовательности кодов.

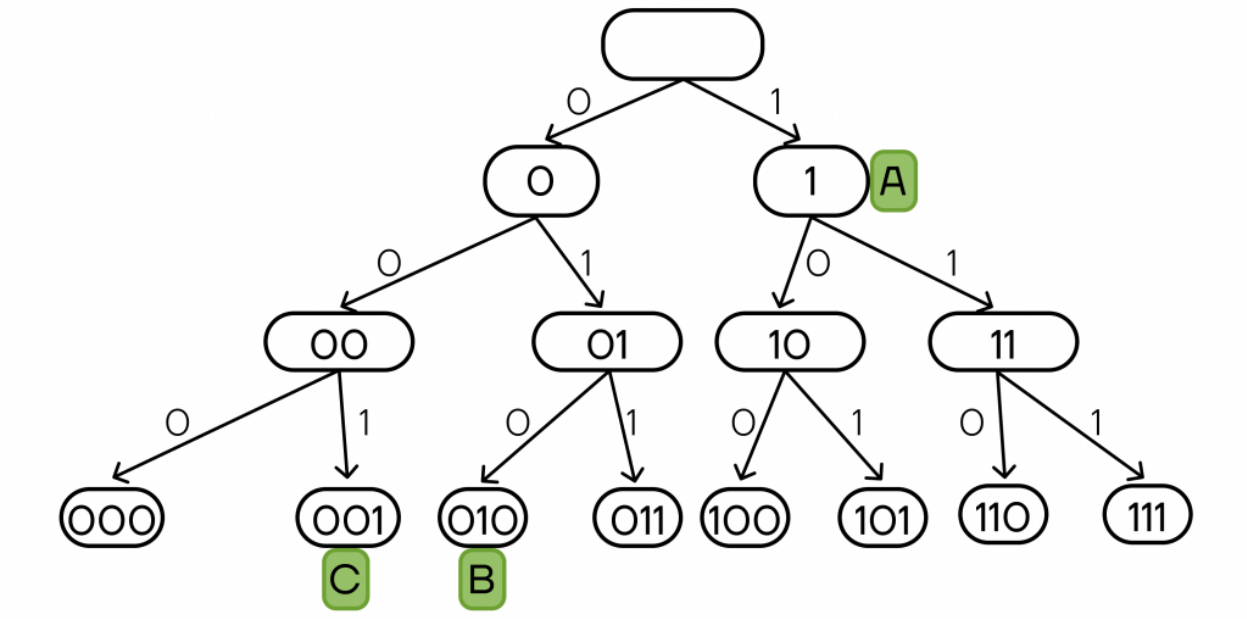
Условие Фано (условие однозначного декодирования): ни одно кодовое слово не может быть началом другого кодового слова.

Пусть букве А соответствует код 10, букве Б код 1011, букве В - 11. Из-за того, что 10 (А) - это начало 1011 (Б), мы не можем однозначно декодировать кодовую последовательность 101111: это может быть как и АВВ, так и БВ. Условие Фано не выполняется.

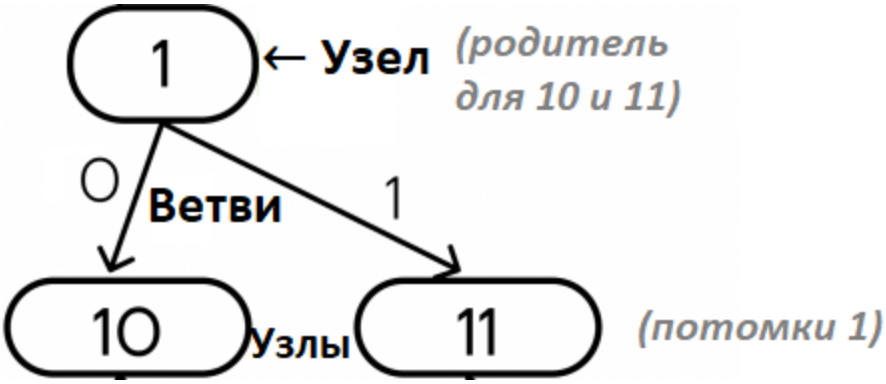
Но при этом набор Б, В не нарушает условие Фано: хоть 11 и является частью 1011, но не его началом. Поэтому, т.к. декодирование происходит слева направо, кодовая последовательность 11101111 остается однозначной: ВББВ.

Алгоритм решения

Дерево Фано - бинарное дерево (каждый узел имеет не более двух потомков, в нашем случае 1 и 0), используемое для решения задачи на определение кодовых слов, удовлетворяющих условию однозначного декодирования.



- *Элементы дерева:*



- *Для того, чтобы не путать обозначения узлов, старайтесь приучить себя всегда записывать единички по одну сторону (например, всегда справа), а нули - по другую (слева).*
- *Значения в узлах - это кодовые слова, которые вы можете присваивать символам.*
- ***Когда какому-то узлу присваивается символ, то все кодовые слова, соответствующие его потомкам, перестают удовлетворять условию Фано, поэтому их использовать нельзя!***

Задачи в основном делятся на два типа по тому, что требуется найти:

1. наименьшее по длине кодовое слово для определённой буквы из набора (нужно указать сам код или его длину),
2. наименьшую возможную длину всей кодовой последовательности для заданного слова (немного сложнее первого типа).

Пример задачи

Усложненная, разбиралась на занятии

По каналу связи передаются сообщения, содержащие только буквы из набора: Б, О, Р, Т, Ф, Я. Для передачи используется двоичный код, удовлетворяющий условию Фано. Кодовые слова для некоторых букв известны: Р – 01, Ф – 110. Для четырех оставшихся букв Я, Б, О, Т кодовые слова неизвестны. Какое количество двоичных знаков потребуется для кодирования слова ФОТОРОБОТ, если известно, что оно закодировано минимально возможным количеством двоичных знаков?

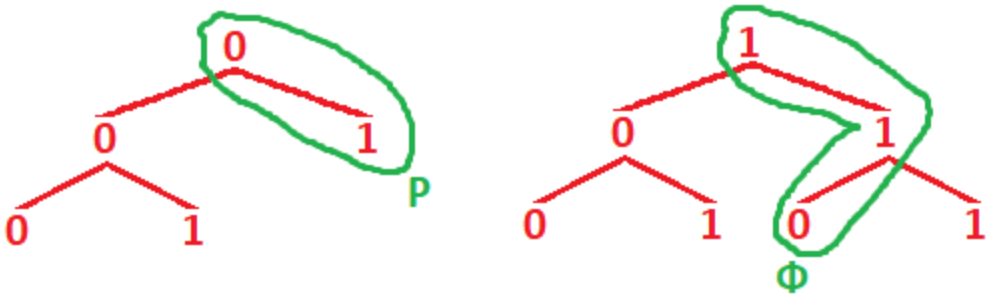
Решение

Это решение сделано достаточно подробно для обучения, но в реальной практике советую упрощать некоторые моменты в угоду скорости решения. Например, можно не расписывать всю таблицу, особенно если для решения важна лишь длина кодового слова.

1. Прежде всего смотрим на заданный набор букв: Б, О, Р, Т, Ф, Я. **Не всегда все буквы из набора могут использоваться в слове, предназначенном для кодирования, но это не значит, что их можно не учитывать!**
2. Составляем таблицу для всех букв из набора, считаем сколько раз они встречаются в слове из условия **ФОТОРОБОТ**:

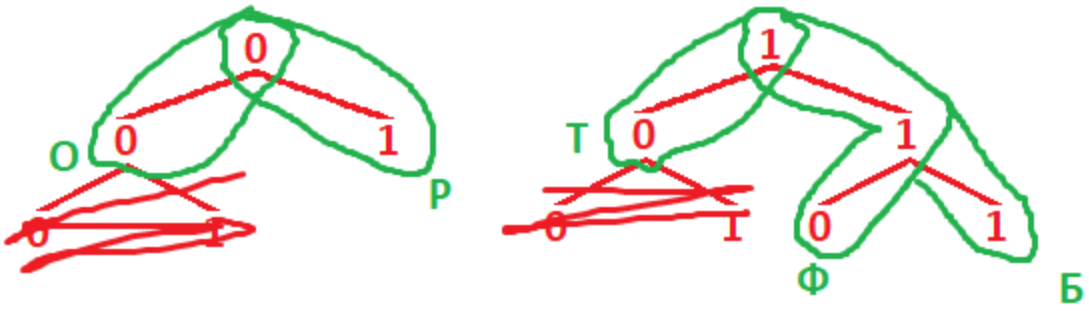
Буква	Кодовое слово	Встречается в слове	Длина кодового слова
О		4	
Т		2	
Р	01	1	2
Ф	110	1	3
Б		1	
Я		0	

   Очевидно, что кодовое слово для Я может быть бесконечно длинным, но на ответ оно не повлияет. Главное оставить для Я саму возможность назначить код, соответствующий условию Фано.
3. Т.к. нам надо закодировать слово наименьшим количеством знаков, очевидно, что кодовые слова с наименьшим количеством знаков следует отдавать наиболее часто встречающимся словам.
4. Строим дерево Фано, сразу отмечая на них Р и Ф, заданные по условию:



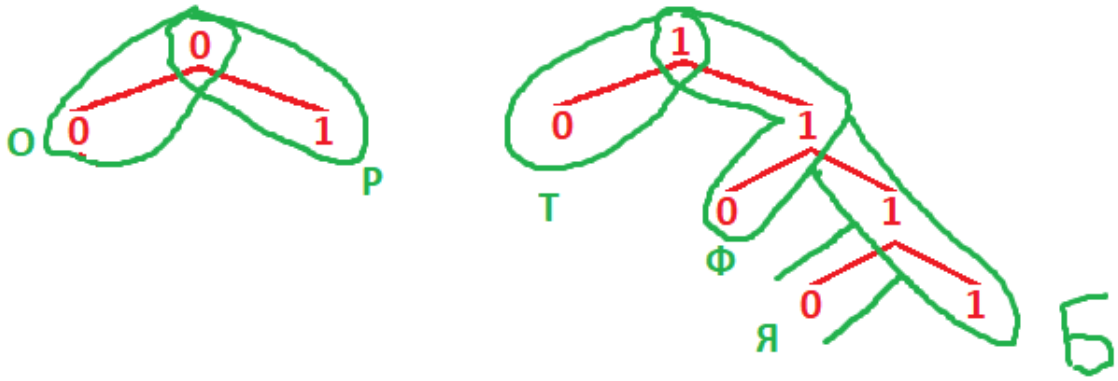
т.к. теперь все последовательности, начинающиеся с 01 и 110, больше не удовлетворяют условию Фано, нет смысла добавлять потомков для Р и Ф.

5. Т.к. в условии спрашивается лишь минимальная длина, то нет смысла обращать внимание на само значение кодового слова.
6. Нельзя брать значения 0, 1 и 11, т.к. тогда заданные по условию коды 01 и 110 перестают удовлетворять условию Фано. Визуально в дереве "обрубаются" все ветви и потомки на последнем значении-узле, когда мы выбираем его в качестве кодового слова для буквы.
7. Мы могли бы ошибочно, ориентируясь лишь на заданное слово, соотнести коды 00 для О, 10 для Т и 111 для Б. Но тогда для Я было бы невозможно подобрать кодовое слово, удовлетворяющее условию Фано (визуально на дереве не осталось свободной ветви). Решение было бы неверным:



8. Не забываем про таблицу из п.2 и задачу закодировать слово минимальным количеством символов. Смотря на дерево, перед нами встает вопрос, какой длины кода какой букве соотнести? Если для кода О очевидна наименьшая возможная длина 2, то для того, чтобы оставить

хотя бы одну ветвь в дереве для Я, может возникнуть вопрос: лучше Т взять длины 3, или Б длины 4?
Если обратиться к частоте появления в слове, то становится понятно, что удлинение Т на 1 символ (по сравнению с вариантом из п.7) удлиняет длину закодированного слова на 2 символа, а удлинение Б - на 1 (соответственно частоте появления). Поэтому конечное дерево Фано будет выглядеть так:



Ветвь для Я остается как бы не закрытой, т.к. этой букве может соответствовать абсолютно любое кодовое слово, являющееся потомком 1110... . Следовательно, если бы помимо Я были ещё буквы в наборе, кодовая длина которых не влияла бы на решение, то им всем могла бы соответствовать одна эта ветвь с бесконечным числом потомков. Например, для Я 11101, для Ю 111000, для Э 1110001 и тд.

9. Заполняем таблицу и находим ответ:

Буква	Кодовое слово	Встречается в слове	Длина кодового слова
О	00	4	2
Т	10	2	2
Р	01	1	2
Ф	110	1	3
Б	1111	1	4
Я	1110...	0	>4

4∗2 + 2∗2 + 2∗1 + 3∗1 + 4∗1 = **21**

Ответ: 21