

COMP90038

Algorithms and Complexity

Assignment Project Exam Help
<https://powcoder.com>

Lecture 21: Huffman Encoding for Data Compression
(with thanks to Harald Søndergaard & Michael Kirley)

Add WeChat powcoder

Andres Munoz-Acosta

munoz.m@unimelb.edu.au

Peter Hall Building G.83

- The SES is open now. Please take a time to review this subject. All feedback is greatly appreciated.

Assignment Project Exam Help

- The final exam has been scheduled for 1:15PM on Thursday 8th of November at Wilson Hall
 - It is a closed book exam. No calculators are permitted.
 - Reading time will be 15 minutes. Exam duration will be 3 hours.
 - Answers must be provided in the exam paper in the space allocated.
 - The reverse side of the pages can be used for rough work.
 - All questions should be attempted. Some are easier than others.
 - Any unreadable parts will be considered wrong. Be neat in your answers.

- A sample exam paper will be provided this week.
 - We have instructed the tutors NOT to provide hints on the sample exam
- Assignment 2 is due next Sunday at 11:59PM.
 - We will provide sample answers on during the lecture and through the LMS.
- Next week we will use both lectures for a quick review of the content.
 - Only examinable topics will be discussed in the review.

Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder

Recap

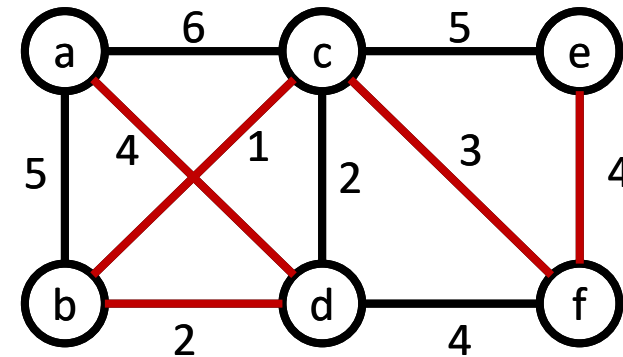
- We discussed **greedy algorithms**:
 - A problem solving strategy that takes the **locally best** choice among all feasible ones. Such choice is **irrevocable**.
 - Usually, **locally best** choices do not yield **global best** results.
 - In some exceptions a greedy algorithm is **correct and fast**.
 - Also, a greedy algorithm can provide good approximations.
- We applied this idea to two graph problems :
 - Prim's algorithm for finding **minimum spanning trees**
 - Dijkstra's algorithm for **single-source shortest path**

What is a Minimum Spanning Tree?

- A **minimum spanning tree** of a weighted graph $\langle V, E \rangle$ is a tree $\langle V, E' \rangle$ where E' is a subset of E , such that the connections have the lowest cost

Assignment Project Exam Help

- We use Prim's algorithm to find the minimum spanning tree.
 - It constructs a sequence of subtrees T , by **adding to the latest tree the closest node not currently on it.**



Prim's Algorithm

- We examined the complete algorithm, that uses priority queues:

```
function PRIM( $\langle V, E \rangle$ )  
  for each  $v \in V$  do  
     $cost[v] \leftarrow \infty$   
     $prev[v] \leftarrow nil$   
  pick initial node  $v_0$   
   $cost[v_0] \leftarrow 0$   
   $Q \leftarrow \text{INITPRIORITYQUEUE}(V)$  ▷ priorities are cost values  
  while  $Q$  is non-empty do  
     $u \leftarrow \text{EJECTMIN}(Q)$   
    for each  $(u, w) \in E$  do  
      if  $weight(u, w) < cost[w]$  then  
         $cost[w] \leftarrow weight(u, w)$   
         $prev[w] \leftarrow u$   
         $\text{UPDATE}(Q, w, cost[w])$  ▷ rearranges priority queue
```

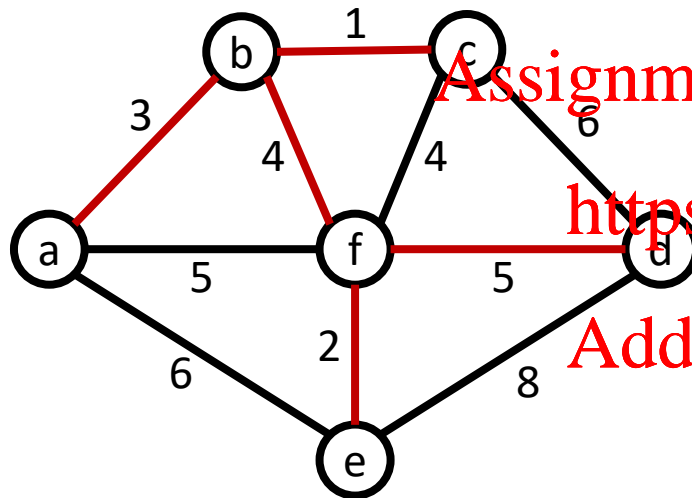
Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder

Another example

- Let's work with the following graph:



- What would happen if we start on b?
 - The sequence will be different, but the edges may be the same
- How many different trees can we have?
 - If there are ties, the tie breaking has influence

Tree T		a	b	c	d	e	f
	cost	∞	∞	∞	∞	∞	∞
	prev	nil	nil	nil	nil	nil	nil
	cost	0	∞	∞	∞	∞	∞
	prev	nil	nil	nil	nil	nil	nil
a	cost		3	∞	∞	6	5
a	prev		a	nil	nil	a	a
a,b	cost			1	∞	6	4
a,b	prev			b	nil	a	b
a,b,c	cost				6	6	4
a,b,c	prev				c	a	b
a,b,c,f	cost				5	2	
a,b,c,f	prev				f	f	
a,b,c,f,e	cost				5		
a,b,c,f,e	prev				f		
a,b,c,f,e,d	cost						
a,b,c,f,e,d	prev						

Dijkstra's Algorithm

- **Dijkstra's algorithm** finds all shortest paths from a fixed start node. Its complexity is the same as that of Prim's algorithm.

Assignment Project Exam Help

```
function DIJKSTRA( $\langle V, E \rangle, v_0$ )
```

```
  for each  $v \in V$  do
```

```
     $dist[v] \leftarrow \infty$ 
```

```
     $prev[v] \leftarrow nil$ 
```

```
   $dist[v_0] \leftarrow 0$ 
```

```
   $Q \leftarrow \text{INITPRIORITYQUEUE}(V)$ 
```

▷ priorities are distances

```
  while  $Q$  is non-empty do
```

```
     $u \leftarrow \text{EJECTMIN}(Q)$ 
```

```
    for each  $(u, w) \in E$  do
```

```
      if  $dist[u] + weight(u, w) < dist[w]$  then
```

```
         $dist[w] \leftarrow dist[u] + weight(u, w)$ 
```

```
         $prev[w] \leftarrow u$ 
```

```
         $\text{UPDATE}(Q, w, dist[w])$ 
```

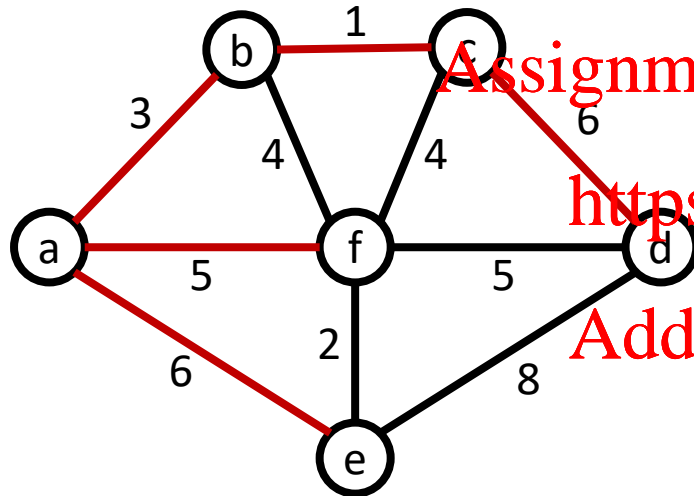
▷ rearranges priority queue

<https://powcoder.com>

Add WeChat powcoder

Another example

- Let's work with this graph again:



- What would happen if we start on b?
 - It is possible to end up with a different tree
- How many different trees can we have?
 - Ties can also influence the final tree.

Tree T		a	b	c	d	e	f
	cost	∞	∞	∞	∞	∞	∞
	prev	nil	nil	nil	nil	nil	nil
	cost	0	∞	∞	∞	∞	∞
	prev	nil	nil	nil	nil	nil	nil
a	cost		3	∞	∞	6	5
a	prev		a	nil	nil	a	a
a,b	cost			4	∞	6	5
a,b	prev			b	nil	a	a
a,b,c	cost				10	6	5
a,b,c	prev				c	a	a
a,b,c,f	cost				10	6	
a,b,c,f	prev				c	a	
a,b,c,f,e	cost				10		
a,b,c,f,e	prev				c		
a,b,c,f,e,d	cost						
a,b,c,f,e,d	prev						

Data compression

- From an information-theoretic point of view, most computer files contain much redundancy.

Assignment Project Exam Help

- Compression is used to store files in less space.
 - For text files, savings up to 50 are common.
 - For binary files, savings up to 90 are common.

<https://powcoder.com>

Add WeChat powcoder

- Savings in space mean savings in time for file transmission.

Run-Length Encoding

- For a text with long runs of **repeated characters**, we could compress by counting the runs. For example:

Assignment Project Exam Help

AAAABBBAAABBBBCCCCCCCCDABCBAAABBBBCCCD

<https://powcoder.com>

- can then be encoded as: Add WeChat powcoder

4A3BAA5B8CDABCB3A4B3CD

- This is not useful for normal text. However, for **binary files** it can be very effective.

Run-Length Encoding

Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder

Assignment Project Exam H

<https://powcoder.com>

Add WeChat powcoder

Variable-Length Encoding

- Fixed-length encoding uses a static number of symbols (bits) to represent a character.
 - For example, the ASCII code uses 8 bits per character.
- Variable-Length encoding assigns shorter codes to common characters.
 - In English, the most common character is **E**, hence we could assign **0** to it.
 - However, no other character code can start with **0**.
- That is, no character's code should be a prefix of some other character's code (unless we somehow put separators between characters, which would take up space).

Variable-Length Encoding

- Suppose our alphabet is {A,B,C,D,E,F,G}
- We analyzed a text and found the following number of occurrences
- The last column shows some sensible codes that we may use for each symbol
 - Symbols with higher occurrence have shorter codes

SYMBOL	OCCURRENCE	CODE
A	28	11
B	4	0000
C	14	011
D	5	0001
E	27	10
F	12	010
G	10	001

Assignment Project Exam Help

<https://powcoder.com>

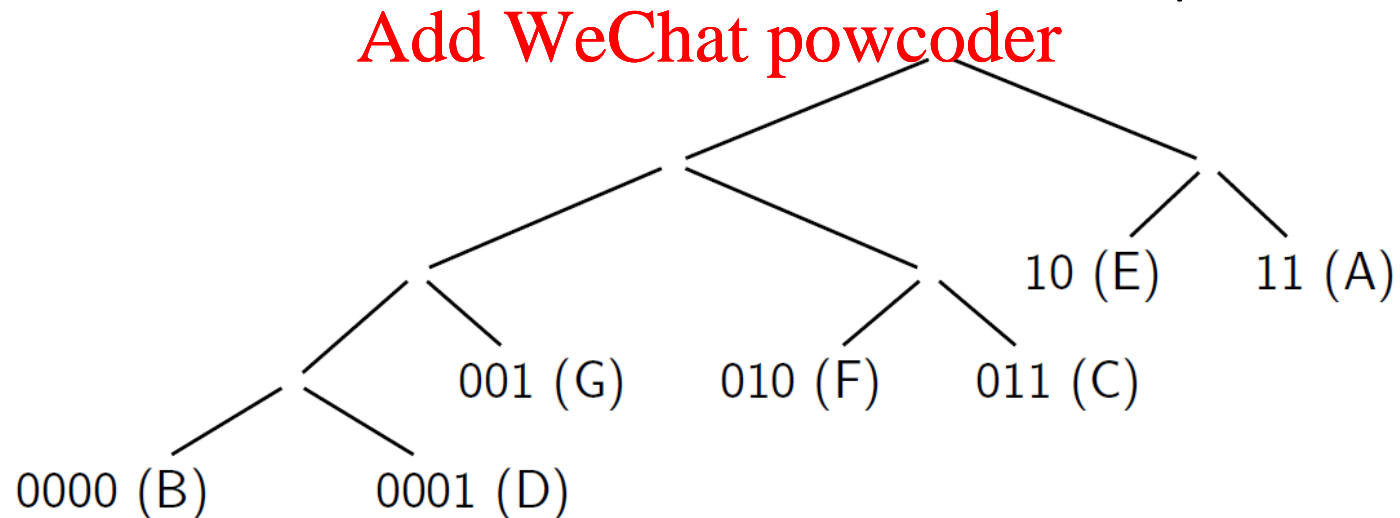
Add WeChat powcoder

Tries for Variable-Length Encoding

- A **trie** is a binary tree used on search applications
- To search for a key we look at individual bits of a key and descend to the **left** whenever a bit is **zero** and to the right whenever it is **one**

<https://powcoder.com>

- Using a trie to determine codes means that no code will be the prefix of another



Encoding messages

- To encode a message, we just need to concatenate the codes. For example:

Assignment Project Exam Help
<https://powcoder.com>
Add WeChat powcoder

F A C E
010 11 011 10

B A G G E D
0000 11 001 001 10 0001

SYMBOL	CODE
A	11
B	0000
C	011
D	0001
E	10
F	010
G	001

- If we were to assign three bits per character, FACE would use 12 bits instead of 10. For BAGGED there is no space savings

Decoding messages

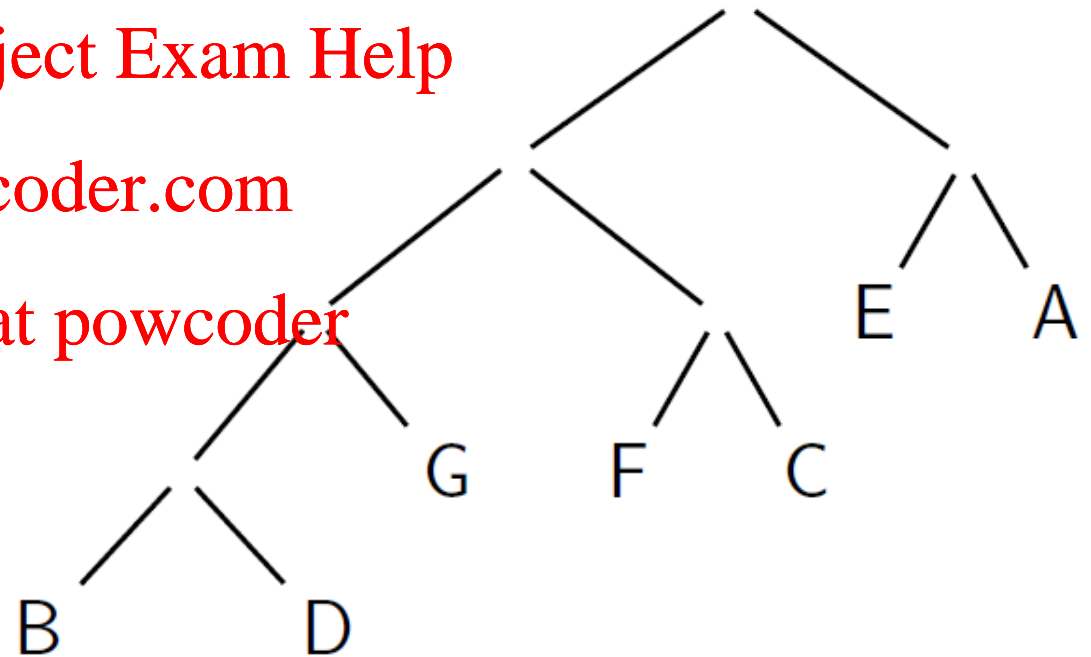
- Try to decode **0001100111010** and **000011000100110** using the trie

Assignment Project Exam Help

- Starting from the root, print each symbol found as a leaf <https://powcoder.com>
- Repeat until the string is completed

Add WeChat powcoder

- Remember the rules: Left branch is 0, right branch is 1



Huffman Encoding: Choosing the Codes

- Sometimes (for example for common English text) we may know the frequencies of letters fairly well.

Assignment Project Exam Help

- If we don't know about frequencies then we can still count all characters in the given text as a first step

<https://powcoder.com>

Add WeChat powcoder

- But how do we assign codes to the characters once we know their frequencies?
 - By repeatedly selecting the two smallest weights and fusing them.
- This is **Huffman's algorithm** – another example of a **greedy method**.
 - The resulting tree is a **Huffman tree**.

Huffman Trees (example)

Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder

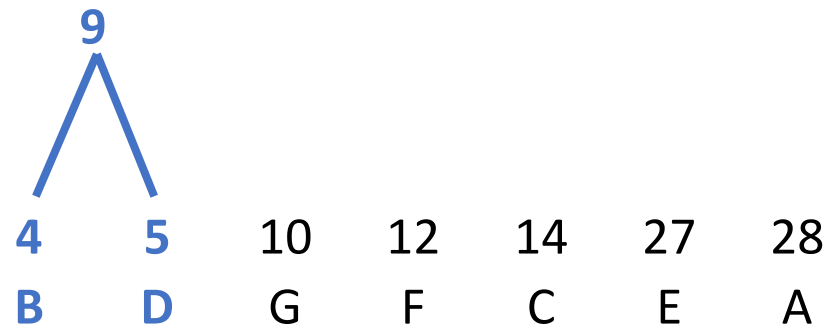
4	5	10	12	14	27	28
B	D	G	F	C	E	A

Huffman Trees (example)

Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder

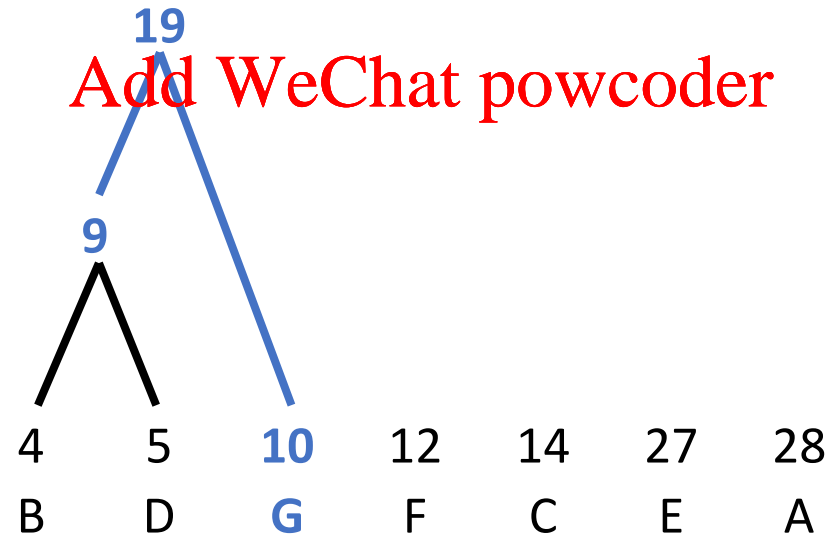


Huffman Trees (example)

Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder

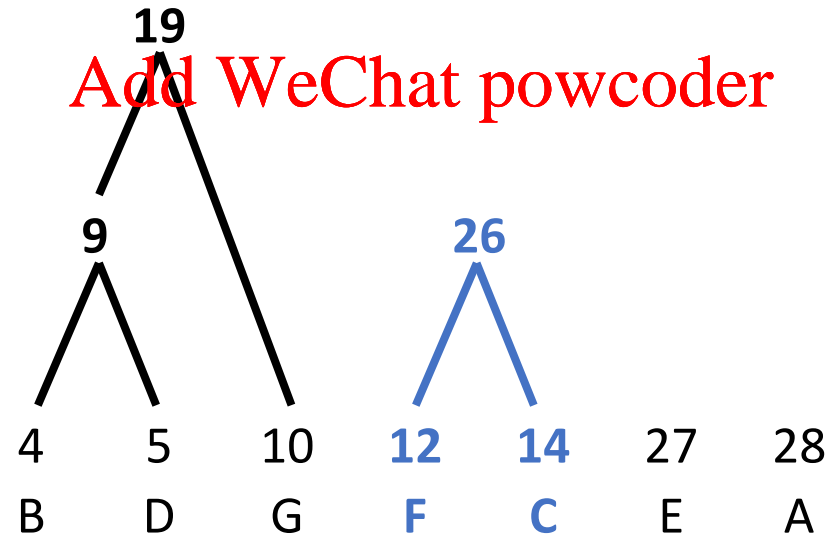


Huffman Trees (example)

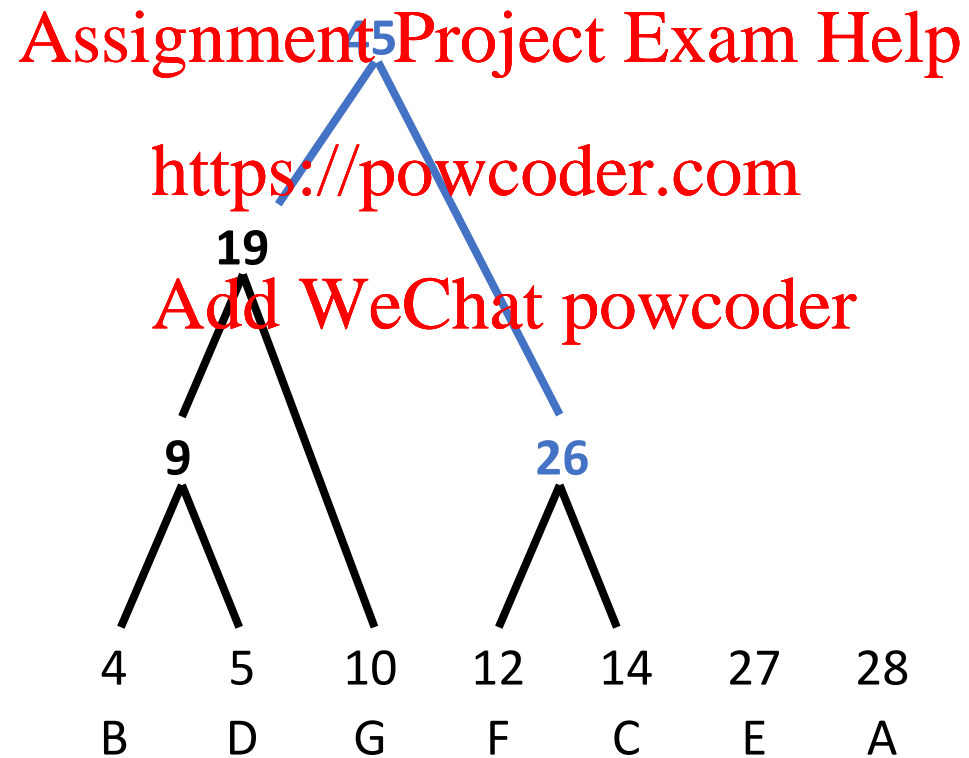
Assignment Project Exam Help

<https://powcoder.com>

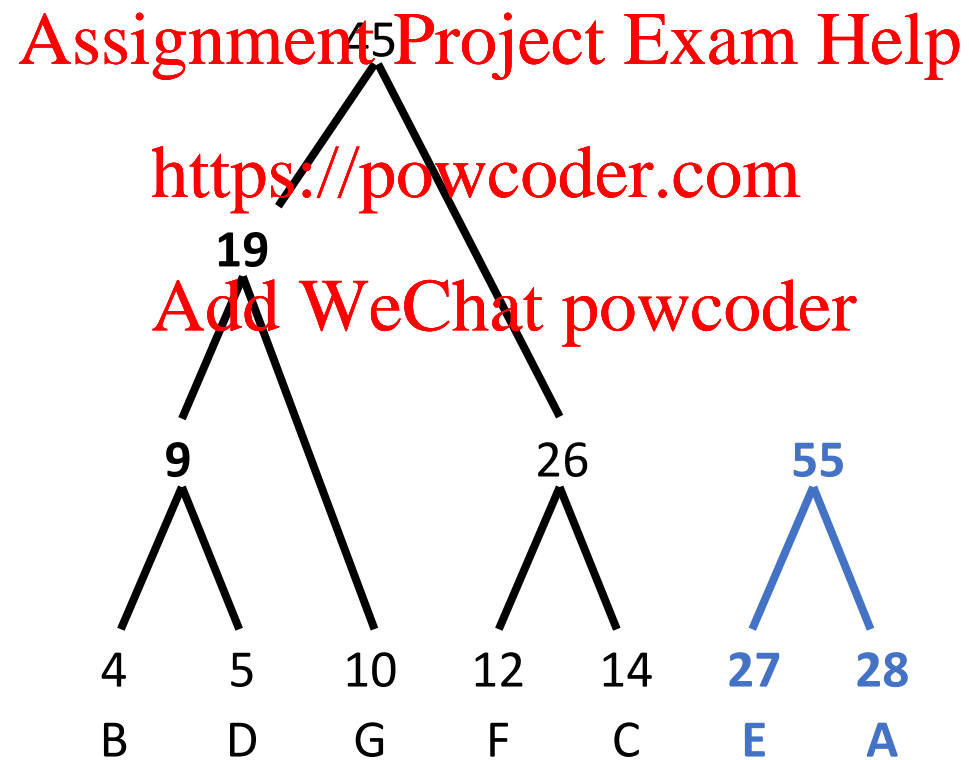
Add WeChat powcoder



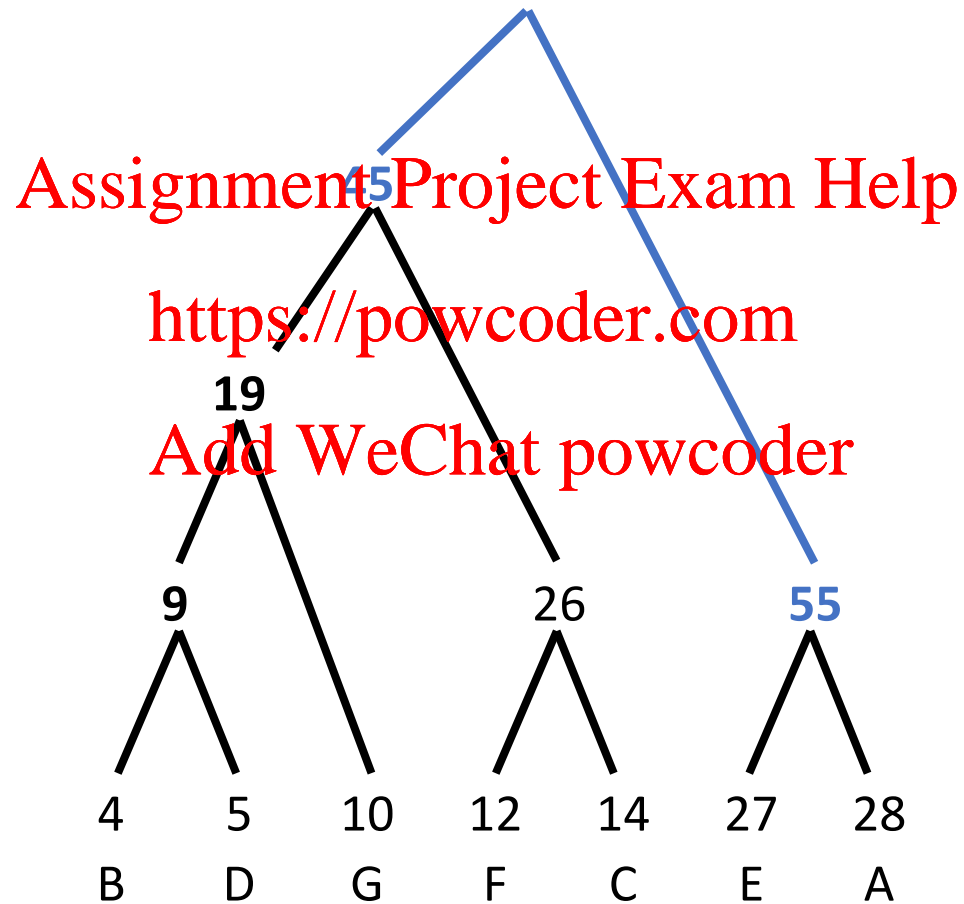
Huffman Trees (example)



Huffman Trees (example)



Huffman Trees (example)



An exercise

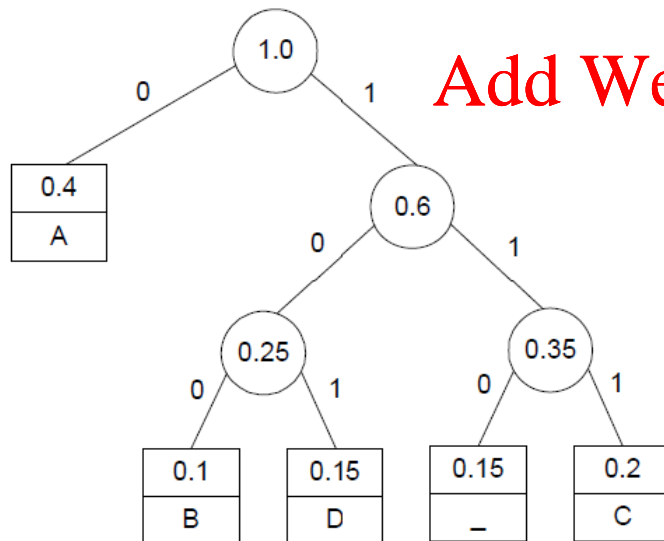
- Construct the Huffman code for data in the table, placing in the tree from left to right [A,B,C,D,_]
- Then, encode **ABACABAD** and decode **100010111001010**
- **0100011101000101** / BAD_ADA

Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder

SYMBOL	FREQUENCY	CODE
A	0.40	0
B	0.10	100
C	0.20	111
D	0.15	101
_	0.15	110



Compressed Transmission

- If the compressed file is being sent from one party to another, the parties must agree about the codes used.
 - For example, the trie can be sent along with the message.
- For long files this extra cost is negligible.
- Modern variants of Huffman encoding, like **Lempel-Ziv compression**, assign codes not to individual symbols but to sequences of symbols.

Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder

Next lecture

Assignment Project Exam Help

- We briefly discuss complexity theory, NP-completeness and approximation algorithms

<https://powcoder.com>

Add WeChat powcoder

- On the final week we will devote time to review all the content