תכנות פרוצדורלי

מטלת בונוס

כתוב תוכנית בשם huffman לדחיסה של קבצי טקסט וגם פתיחת קבצים דחוסים.

<u>הפעלת התוכנית</u>

[ahmad@mars~]$ huffman *option file*

כאשר option :

c- דחיסה של קובץ הטקסט file ויצירה של קובץ חדש דחוס בשם file.huf

d- פתיחה של קובץ file . התוכנית תקבל קובץ file.huf ותשחזר קובץ מקור תחת השם file.

<u>דחיסה – אלגוריתם הופמן</u>

נניח שנתון לנו קובץ טקסט בשם f1 והוא מכיל את הטקסט הבא:

go go gophers

התוכנית תחשב עבור תוי הקובץ את השכיחויות שלהם:

g:3

o:3

space:2

p:1

h:1

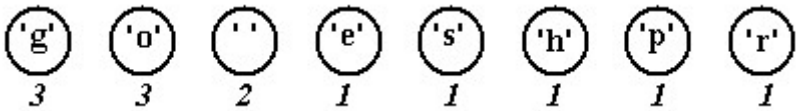
e:1

r:1

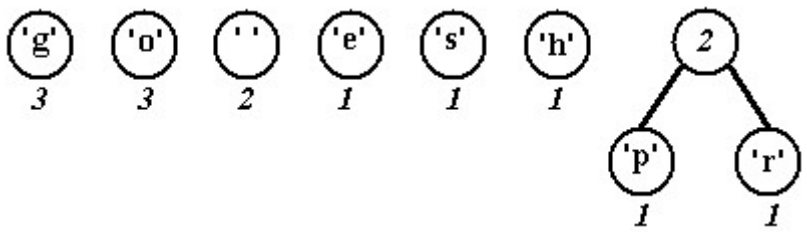
s:1

התוכנית תיצור קוד לכל תו על ידי בניית עץ באופו הבא:

ניצור תחילה את העלים של העץ. הם יהיו התווים השונים עם השכיחויות שלהם.

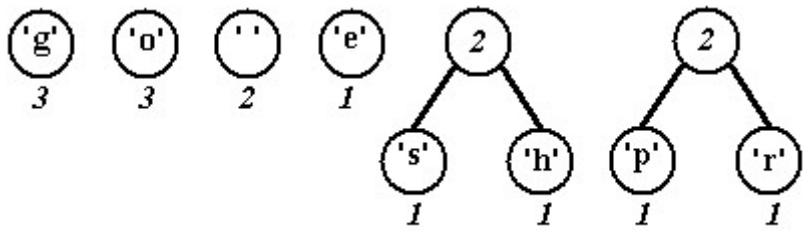


נבחר את שני התווים בעלי השכיחויות הקטנות ביותר. כשיש יותר משני מועמדים אין חשיבות במי בוחרים. נבחר את p,r בגלל שיש להם שכיחות של 1 כל אחד.
לשני העלים האלה נבנה צומת חדש שהם יהיו הבנים שלו. נקבל:

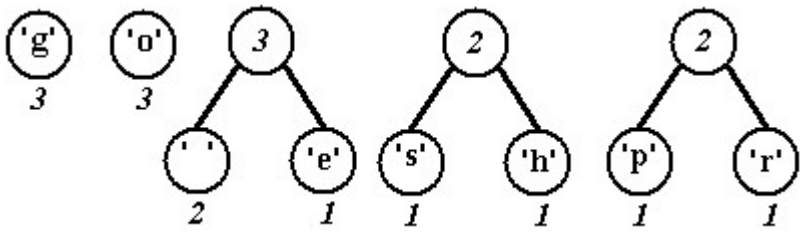


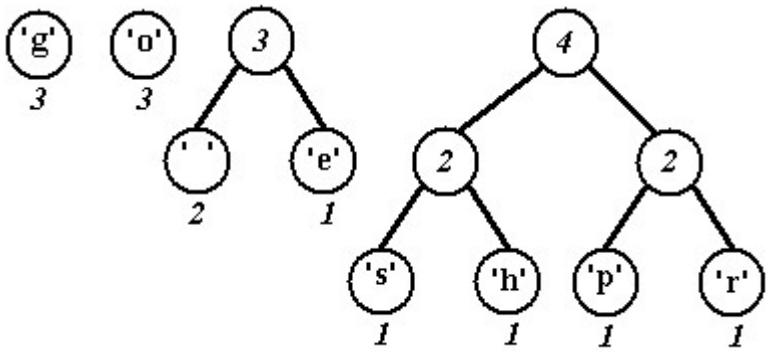
שימו לב שהצומת החדש מכיל את סכום השכיחויות של שני בניו.

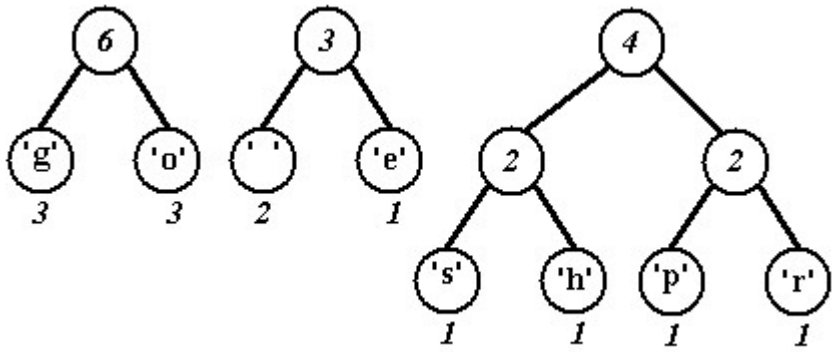
נחזור על התהליך של בחירת שני צמתים בעלי שכיחות מינימלית ובניית צומת אב משותף עד שנגיע לשורש העץ.
שימו לב שכל צומת נבחר פעם אחת.
שימו לב שצומת חדש שנוצר יוכל להבחר.
לכן השלב הבא יהיה לבחור את s,h כי לשניהם יש 1. נקבל :

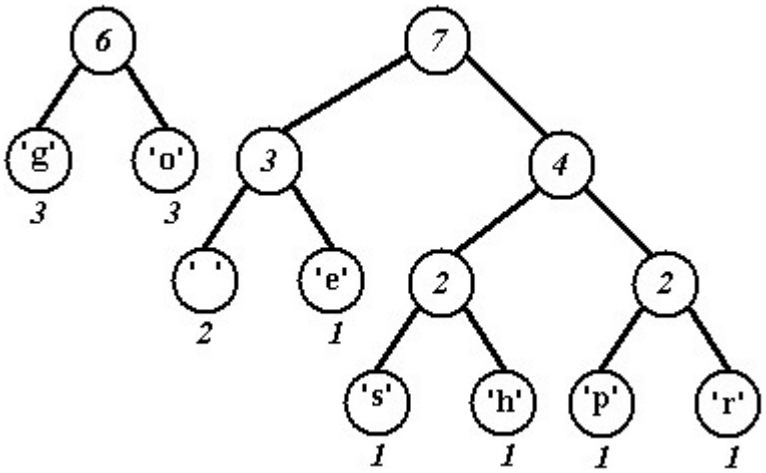


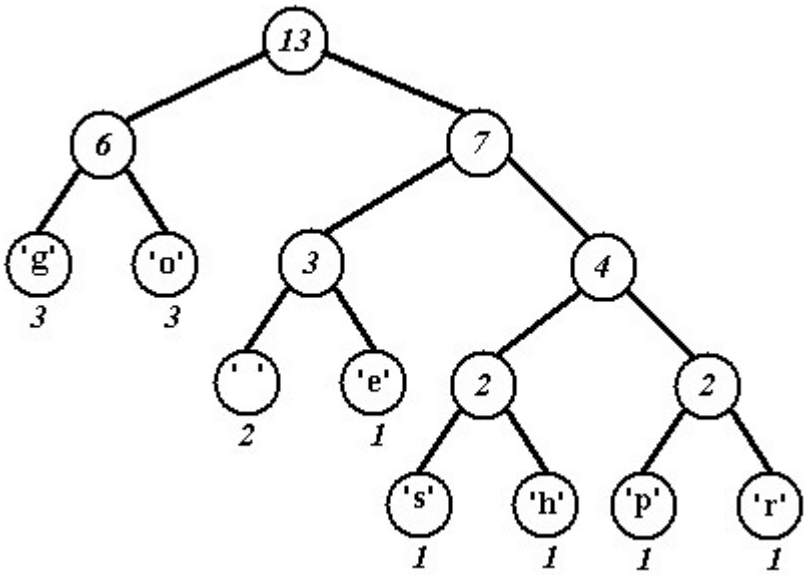
לאחר מכן נקבל :











לאחר השלמת הבניה של העץ ניתן לקבל לכל תו את הקוד שלו. למשל, קבלת הקוד עבור התו s :

נתחיל מהשורש, כל פעם שנלך שמאלה נשרשר אפס וכל פעם שנלך ימינה נשרשר אחד.

המסלול של s : ימינה, ימינה, שמאלה,שמאלה ולכן הקוד יהיה : 1100

להלן טבלה של הקודים השונים:

char	binary
'g'	00
'o'	01
'p'	1110
'h'	1101
'e'	101
'r'	1111
's'	1100
' '	100

לכן הקובץ f1 יהיה מורכב מהרצף הבא:

00 01 100 00 01 100 00 01 1110 1101 101 1111 1100

סך הכל 37 סיביות שכדי שנשמור אותן בקובץ יספיקו 5 בתים (5 * 8 ונקבל 40 סיביות). נשתמש בפעולות על סיביות כדי לשמור בבית אחד 8 סיביות כל פעם ואז את הבית הזה נכתוב לתוך קובץ. שימו לב שהבית האחרון יש בו 5 סיביות כי שלוש הסיביות הנותרות לא רלוונטיות.

<u>מידע נוסף לשמירה</u>

לא מספיק לשמור בתים דחוסים בקובץ. צריך לזכור שלצורך פתיחת הקובץ הדחוס יש צורך שיהיה לנו הקודים של התווים כדי שנוכל לשחזר. ועל כן, חובה לשמור בתוך הקובץ הדחוס גם מידע נוסף שיאפשר לנו לשחזר את הקודים של התווים או את העץ של הקודים. אפשרות אחת היא לשמור כל תו ואת השכיחות שלו. אפשרות שניה היא לשמור את תוכן העץ (בסריקת inorder) .

מידע זה יישמר בתחילת הקובץ הדחוס.

כלומר הקובץ הדחוס ייראה כך:

Header info.

Compressed text.

חובה גם לדעת להפריד בין המידע ששייך ל- header לבין הטקסט הדחוס.

<u>פתיחת קובץ דחוס</u>

התוכנית תקבל כפרמטרים את האופציה d– וקובץ דחוס והיא תייצר את קובץ הטקסט המקורי.

בשלב ראשון התוכנית תקרא את ה- header של הקובץ הדחוס וממנו תשחזר את העץ .

בשלב שני התוכנית תקרא את הבתים מתוך הקובץ ותעבור על הסביות בתוך הבתים. כמובן מתחילים מהשורש, כל פעם שהתוכנית תקרא 0 היא תלך שמאלה בעץ ואם תקרא אחד היא תלך ימינה עד שנגיע לעלה בעץ ומשם נקח את התו. ושוב נחזור לשורש ונמשיך עם הסיבית הבאה עד ששוב נגיע לעלה.

הערה:

יש לכם חופש פעולה בכל דבר שלא תואר בתרגיל.