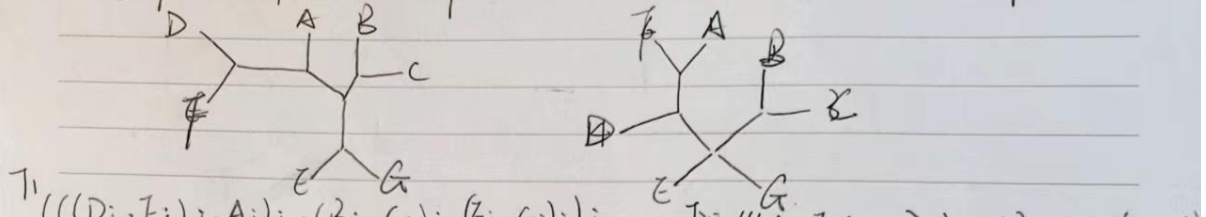
Interpretation of my implementation of Day's algorithm 1985:
Author: Zhang Leyi, Master's in Aarhus University BiRC
2022-04-20

A:B Project 4 : Day's Algorithm.

Compute RF Distance for two unrooted trees. In newick format.



T1: (((D:,F:):,A:):,(B:,C:):,(E:,G:):); T2: (((A:,F:):,D:):,(B:,C:):,(E:,G:):);

① Store these two unrooted trees. O(n)

Using adjacency list. + stack.

def unrootReader(newick)

Read each symbol and letter i from right to left in newick

if i == ')' # should push in.

using na to store the strings I read.

if na == "", only push ')' into s1.

else push ')' into s1 and push na into s2.

elif i == ',' # A, B

same as ')'

elif i == '('

push in

pop until ')' # pop both s1 and s2. and all strings poped are in the same layer (A, B) => A B

else

na = i + na.

In doing elif i == '(', we can put (A:, B:) C: into adjancy

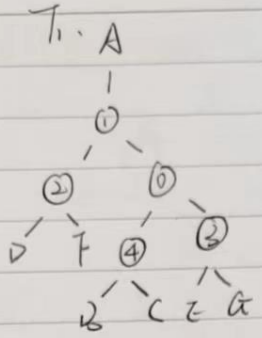
A -> C. C -> A, B -> C, C -> B

In order to DFS later.

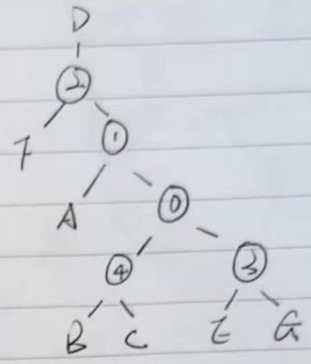
② Since two trees should have same names in leaves, select one leaf as rut and DFS to transfer them into ruted tree.

O(n)

After DFS T_1:



T_2:



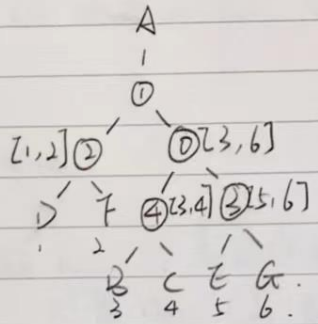
Just DFS.

③ Relabel new leaves by using numbers. O(n)

D	7	B	C	E	G
1	2	3	4	5	6

still DFS.

④ Calculate the scales O(n)



Using dics to make map on ~~leaves~~ nodes and number

Here it used minscale, maxscale, size.

DFS to search for every node "u" and find father.

minscale[father] = min(minscale[u], minscale[f

maxscale ——.

See line 401

⑤ Radix sort these scales. O(n)

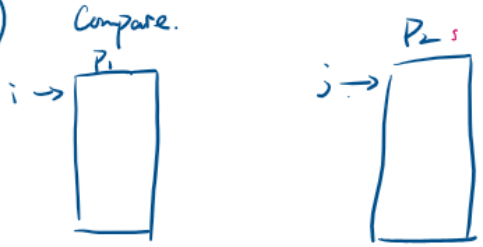
Right now we have scales of T_1 and T_2

⇒ [a,b]: a is the first key, ~~be~~ b is the second key.

[1,2]
[3,4]
[3,6]
[5,6]

⑥ Find shares $RT = |T_1| + |T_2| - 2 \cdot shares$ O(n).

⑥



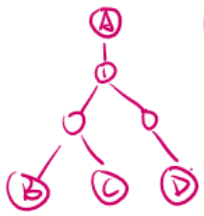

```
if( P1(i) > P2(j) )
    j += 1  # P1(i) could possibly exist in P2 later.
elif. P1(i) == P2(j)
    share += 1.
    i += 1 ; j += 1.
else # P1(i) < P2(j) # could not be existed
    i += 1 .
until i == len(P1) or j == len(P2)
return share.
```

⑦ $|T_1| + |T_2| - 2 \cdot share$

$|T_1|$ and $|T_2|$ are internal nodes:

$T_1 = len(nodes) - len(leaves) - 2$

Why minus 2:



A and 1 are not seen as internal nodes here.

RF-distance = $T_1 + T_2 - 2 \times shares$