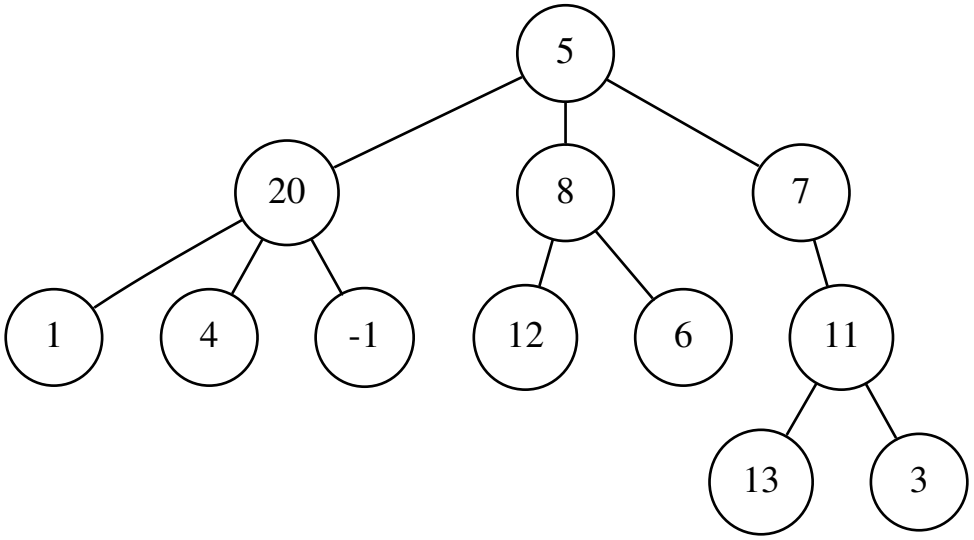
TiLa II
Analyzing words in a text file

Anni Järvenpää
014338836

20. joulukuuta 2015



Kuva 1: Esimerkki puusta, johon on tallennettu kokonaislukuja.

1 Johdanto

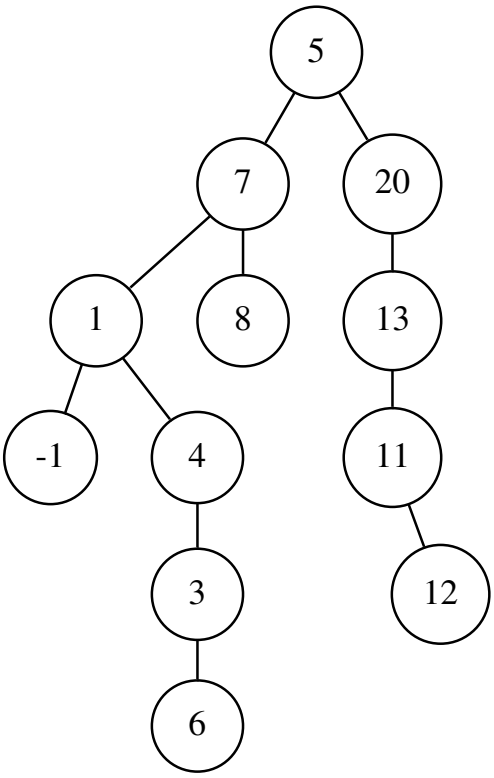
Tavoitteena on laskea tiedostossa olevien sanojen toistokertojen määrä. Tällaisia niinkutsuttuja frekvenssilistoja käytetään paljon kielitieteessä ja erityisesti korpuslingvistiikassa. Niistä on hyötyä muun muassa tekstiä koskevien hypoteesien muodostamisessa sekä tehtyjen oletusten tarkastamisessa. Lisäksi eri teksteistä saatuja frekvenssilistoja voidaan vertailla toisiinsa. [2]

2 Menetelmät

Sanojen laskemisessa hyödynnetään binäärihakupuuta. Puu on solmuista koostuva tietorakenne, jossa jokaiseen solmuun voidaan tallentaa tietoa ja jokaisella solmulla on 1 tai 0 vanhempaa sekä n lasta missä $n \in \mathbb{N}_0$. Puu voidaan esittää suunnattuina yhtenäisinä verkkoina, joissa jokaisesta solmusta on kaari jokaiseen lapseensa. [1]

Puussa on aina tasan yksi solmu, jolla ei ole vanhempaa ja tätä solmua kutsutaan juureksi. Solmuja, joihin ei tule kaarta mistään toisesta solmusta kutsutaan lehdiksi. Solmun korkeus on kaarien määrä pisimmällä polulla solmusta lehteen. Usein puhutaan myös puun korkeudesta, jolla tarkoitetaan puun juurisolmun korkeutta. Esimerkiksi kuvassa 1 on esitettynä puu, jonka solmuihin on tallennettu kokonaislukuja. Puun juurisolmun arvo on 5 ja lehtisolmuissa on arvot 1, 4, -1, 12, 6, 13 ja 3 sekä korkeus 3 (kaaret $5 \rightarrow 7$, $7 \rightarrow 11$ ja $11 \rightarrow 13$ tai $11 \rightarrow 3$). [1]

Binäärihakupuussa jokaisella solmulla on 0, 1 tai 2 lasta ja kunkin solmun vasemmasta lapsesta lähtevässä alipuussa on vain arvoltaan solmun arvoa pienempiä arvoja ja oikeassa



Kuva 2:

lapsessa lähtevässä alipuussa vain solmun arvoa suurempia arvoja. Näin etsittäessä tiettyä solmua, voidaan kunkin solmun kohdalla sulkea pois toinen solmun lapsista, jolloin joudutaan tutkimaan korkeintaan puun korkeuden verran solmuja.

Eräs toteutus

3 Toteutus

4 Tulokset

5 Johtopäätökset

Viitteet

[1] Thomas H. Cormen. *Introduction to algorithms*. MIT Press, Cambridge (MA), 2001.

[2] Hanna Tuomisto. Xterm-korpuskyselykielen kehittäminen ja korpuskyselykielten vertailu. `https://tampub.uta.fi/bitstream/handle/10024/83713/gradu06022.pdf?sequence=1`. Luettu 20.12.2015.