Heidi Jauhiainen
011512192
Ohjaaja: Kristiina Paloheimo
Tietorakenteiden harjoitustyö, loppukesä 2012
Helsingin yliopisto, Tietojenkäsittelytieteen laitos
9.8.2012

Määrittelydokumentti

Sanaindeksi Trie-puuta käyttäen

Sanaindeksi on ohjelma, joka saa syötteenä yhden tai useamman tekstitiedoston ja josta pystyy hakemaan annettuja sanoja tai sananalkuja sisältävät rivit. Trie-puu on tehokas tietorakenne tällaisten String-muotoisten avainten hakemiseen. Triessä sanan jokainen kirjain talletetaan eri solmuun. Sanan ensimmäinen kirjain on tyhjän juurisolmun lapsi ja toinen kirjain tämän ensimmäisen solmun lapsi ja niin edelleen. Jokaisella solmulla voi olla lapsia niin paljon kuin käytetyssä aakkostossa on kirjaimia (plus mahdollisesti joitain muita merkkejä). Samanalkuiset sanat jakavat alkupään solmut.

Tässä harjoitustyössä toteutetaan sanaindeksi, joka tulee tukemaan useamman sanan tai sen osan hakemista kaikista tai yhdestä tiedostosta kerrallaan. Indeksi toteutetaan trie-puulla, johon eri tiedostojen sanat tallennetaan. Tieto yksittäisen solmun lapsisolmuista tallennetaan järjestettyyn taulukkoon, josta tietyn kirjaimen sisältämä solmu haetaan binäärihaulla. Jokaiseen solmuun tallennetaan tieto riveistä, joilla sen kirjain ja kirjaimen prefiksi sijaitsee. Rivinumerot tallennetaan dynaamiseen taulukkoon. Koska sanojen tallennus tapahtuu tiedostosta järjestyksessä, solmussa olevat rivinumerot ovat myös järjestyksessä. Ennen rivinumeron lisäämistä taulukkoon, tarkistetaan binäärihaulla onko kyseinen rivi jo taulussa. Jokaisen sanan viimeisen kirjaimen kohdalle tallennetaan rivinumero myös toiseen dynaamiseen taulukkoon. Jälkimmäistä taulukkoa käytetään kokonaisten sanojen hakemiseen ja ensimmäistä sananosien (tekstipohjaisessa käyttöliittymässä erotetaan sanan perään lisättävällä *-merkillä). Ohjelma säilyttää kerran sinne syötettyjen tekstitiedostojen datan, mutta sille voi myös syöttää uusia tiedostoja. Tiedostojen rivit tallennetaan dynaamiseen taulukkoon. Solmuihin tallennettavat rivinumerot ovat tämän taulun indeksejä, joten riven printtaaminen on nopeaa sen jälkeen kun sanahaulla on saatu lista riveistä, joilla haettu sana tai sen osa sijaitsee. Tieto siitä millä tekstit sisältävän taulun rivillä kunkin tiedoston rivit ovat säilytetään erillisessä matriisissa tiedoston nimen kanssa.

Sanan tai sen osan haun aikavaativuus tästä sanaindeksistä on O(m log n), jossa m on sanan kirjainten määrä ja n solmun lasten määrä. Tämä aikavaativuus koostuu seuraavista tapahtumista: Main-metodi kutsuu Hakija-olennon printtaa-metodia, joka puolestaan kutsuu joko Puu-olennon haeOsa- tai haeSana-metodia. Kumpikin näistä metodeista käy läpi haetun sanan jokaisen kirjaimen (O(m)). Kumpikin metodi myös kutsuu jokaisen kirjaimen kohdalla PuuSolmu-olennon getLapsi-metodia, joka hakee binäärihaulla lapsen jolla on haettu kirjain. Binäärihaun aikavaativuus on O(log n), jossa n on solmun lasten määrä. Koska lasten määrä on rajallinen, korkeintaan käytössä olevien kirjainten määrä (muitakin merkkejä voi olla, toisaalta harvoin sanassa kirjainta voi seurata mikä tahansa kirjain), binäärihaun aikavaativuus on useimmilla kielillä hyvin pieni. Koska sanojen sisältämien kirjaintenkin määrä on rajallinen, harvemmin yli 50 kirjainta [wikipedia, Longest words], ei tuo O(m log n) yleensä kasva kovin suureksi pahimmassakaan tapauksessa.

Lähteet:

Wikipedia, Longest words. http://en.wikipedia.org/wiki/Longest_words.