Määrittelydokumentti - Sanaindeksi

Ohjelman on tarkoitus lukea yksi tai useita tekstitiedostoja ja rakentaa näistä trie-puu. Hakutoiminnossa puulle annetaan hakusana tai -sanat, joiden perusteella ohjelma hakee tekstistä rivejä, jotka sisältävät ko. sanat. Useamman hakusanan käyttö etsii rivejä, jotka sisältävät kaikki hakusanat.

Perusrakenne on trie-puu, jossa jokainen solmu tietää missä tiedostoissa ja millä riveillä niissä esiintyy. Trie-puun rakennusvaiheessa kukin teksti tiedosto siirretään String-taulukkoon, jossa jokaisessa taulukon indeksissä on järjestyksessä yksi teksin rivi Stringinä. Tämän tilavaativuus on O(rivien lkm). Tätä varten rakennetaan dynaaminen taulukko, joka täyttyessään kaksinkertaistaa kokonsa, jolloin tilavativuus on ehkä enemmänkin luokkaa O(2* rivien lkm).

Trie-puu on puurakenne, jossa kussakin solmussa oleva arvo on vain tieto miten se poikkeaa vanhemmastaan, ts. solmun arvo on sen oman arvon ja sen esivanhempien arvojen kooste. Solmulla ei voi olla kahta lasta, joissa olisi sama arvo.

Eritietolähteistä on poimittu seuraavaa tietoa mahdollisista aikavaativuuksista

Trie-puun rakennus vie aikaa luokkaa O(n), missä n on kirjainten lukumäärä.

Jos Solmun lapset ovat järjestetty, niin ne voidaan hakea ajassa O(log n), missä n on lapsien lukumäärä, muuten se on O(n).

Sanahaku trie-puusta riippuu sanan pituudesta z sekä solmun lastenhaun tehokkuudesta.

Toteutunut tila- ja aikavaativuusanalyysi projektin toteutusdokumentissa.

Lähteitä

Trie: http://www.cs.helsinki.fi/u/ejunttil/opetus/tiraharjoitus/trie.html Trien tulostus: http://www.cs.helsinki.fi/u/ejunttil/opetus/tiraharjoitus/treeprint.txt Trie: http://en.wikipedia.org/wiki/Trie