Sanaindeksi
Toteutusdokumentti

Patrik Ahvenainen*
Helsingin yliopisto
Fysiikan laitos

22. elokuuta 2012

1 Yleisrakenne

Ohjelman tietorakenteina käyetään Trie- ja punamustaa puuta. Kummatkin puut toteuttavat Tree-rajapinnan. Lisäksi Trie toteuttaa PartialTree-rajapinnan, joka mahdollistaa myös sanojen alkujen etsimisen. Rajapinnat on toteutettu Pythonissa ABCMeta-luokan [viite] avulla. Puut ovat vastuussa indeksoinnista ja sanan (tai sen alun) löytämisestä indeksistä. Searcher-luokka taas on vastuussa hakujen suorittamisesta käyttäen Puu-rajapintaa.

Sanojen lukemiseen tiedostosta käytetään WordReader-luokkaa, joka vastaa myös sanojen muokkaamisesta hyväksyttävään muotoon (lähinnä Trien indeksointia varten). Tiedosto-operaatiota varten on DataHandling-luokka ja sanojen esiintymien tallentamista varten on toteutettu yksinkertainen LinkedList-luokka, joka on kahteen suuntaan linkitetty lista, jonka loppuun voi lisätä arvoja (jono).

Tiedostojen indeksointia voi testata pysanaindeksi-skriptillä, jonka avulla voi testata kumpaakin puuta sekä tehdä hakuja Materials-kansiosta löytyviin tiedostoihin.

2 Aika- ja tilavaativuudet

3 Suorituskyky

4 Parannusehdotuksia

Nyt toteutus toimii vain UTF-8 –tyyppisiin tiedostoihin. Muut fonttikoodaukset voitaisiin luettaessa muuntaa tämän tyyppisiksi, jolloin luettava tiedosto voisi olla minkä tyyppinen tahansa.

*patrik.ahvenainen@alumni.helsinki.fi / 013326292

Aikavertailun kannalta olisi kätevää, jos Trien indeksoinnin voisi toteuttaa myös punamustaan puuhun verrattavalla tavalla, eli ilman sanojen alkujen indeksointia.

Searcher-luokka voisi vaatia, että hakija toteuttaa Finder-rajapinnan, jossa on määritelty etsimismetodin kutsu. Sinäänsä Searcherin ei nimittäin tarvitsisi vaatia hakijan olevan juuri hakupuu, kunhan hakumetodi löytyy.

Lähteet