# TIETORAKENTEIDEN HARJOITUSTYÖ, KESÄ 2012

Ohjaaja: Kristiina Paloheimo
Ryhmä: 1, ajanjakso 14.05.2012-01.06.2012
Työn tekijä: Kari Korpinen
Opiskelijanumero: 012218686
Työn aihe: Sanaindeksi

Määrittelydokumentti

## Aihe

Tehtävänä on toteuttaa sanaindeksointiohjelma, joka lukee useita tekstitiedostoja ja muodostaa niiden sanoista hakupuun. Ohjelmalla voidaan tehdä lukemisen jälkeen useita sanahakuja, joihin ohjelma vastaa luettelemalla tiedostojen nimet, rivinumerot ja rivit, joilla etsitty sana esiintyy. Ohjelman tulee tukea myös alkuosahakua, jossa etsitään kaikkien annetulla merkkijonolla alkavien sanojen esiintymät. Lisäksi ohjelman on tuettava useamman sanan hakemista kerrallaan useammasta tiedostosta.

## Ohjelman yleiskuvaus

Ohjelma käsittelee merkkimuotoista dataa. Ohjelman käsittelemien tiedoston nimet syötetään sisään yhden tiedoston kautta, jossa lukee ohjelmaan syötettävien tiedostojen nimet..Tiedostot syötetään ohjelmaan raakadatana, josta ne muokataan merkkimuotoisiksi. Ohjelman pitäisi käsitellä ja tallentaa hakurakenteeseen vain riveittäin luettavaa kirjoitusmerkkimuotoista dataa, ei näkymättömiä ohjausmerkkejä kuten rivin päättymismerkki, rivinvaihtomerkki, tabulaattorimerkki, tiedoston loppumerkki jne. Tyhjä merkki jakaa riveillä olevan materiaalin sanoiksi. Sanat tallennetaan tietorakenteeseen. sanan ensimmäisen ja sitä seuraavien merkkien mukaiseen järjestykseen.

## Algoritmit ja tietorakenteet

Alkutoteutuksessa käytetään javan valmista TreeSet luokkaa. Luokka on toteutettu puna-mustapuun tapaan. Toteutus poikkeaa punamustapuusta siten, että lapsisolmujen ja vanhempien välillä ei ole yhteyttä. TreeSet:issä kaikki avaimet ovat koko ajan järjestyksessä. Puna-mustapuussa lisäys ja haku tapahtuu O(log n) ajassa. Tilantarve on O(n)

Tietorakenne toteutetaan javalla NetBeans IDE 7.1 ympäristössä

## Linkkejä

http://stackoverflow.com/questions/1298144/what-are-the-pros-and-cons-of-a-treeset
http://docs.oracle.com/javase/1.4.2/docs/api/java/util/TreeSet.html
http://docs.oracle.com/javase/1.4.2/docs/api/java/io/FileInputStream.html
http://docs.oracle.com/javase/1.4.2/docs/api/java/io/DataInputStream.html
http://docs.oracle.com/javase/1.4.2/docs/api/java/io/BufferedReader.html
http://docs.oracle.com/javase/1.5.0/docs/api/java/lang/String.html
http://en.wikipedia.org/wiki/Red%E2%80%93black_tree