Helsingin Yliopisto – Tietojenkäsittelytieteen laitos 02.09.12
Tietorakenteiden harjoitustyö loppukesä 2012 / Ohjaaja: Kristiina Paloheimo
Kati Kyllönen Opiskelijanumero: 11539913

Tira harjoitustyö – Sanaindeksointi – Toteutusdokumentti

Sisällysluettelo

1 Tira Harjoitustyö

1.1 Aihe: Sanaindeksointi

Sanaindeksointiohjelma etsii syötteenä saamistaan tekstitiedostoista haettavan merkkijonon esiintymistiheyden, sekä esiintymiskohdat kussakin tiedostossa. Haun suoritettuaan ohjelma tulostaa kaikki tiedostojen ne rivit, joilla haettavat merkkijonot esiintyvät.

Harjoitustyön tarkoituksena on haun toteuttamisen lisäksi tarkkailla ja dokumentoida Trie-hakupuun toteutusratkaisujen välisiä suorituseroja, käyttäen lasten tallettamisen menetelmänä linkitettyä listaa, hakien eroja listan järjestämisajankohdan tai järjestämättömyyden välillä. Tämä idea osoittautui työn edetessä kuitenkin huonoksi, sillä sekä järjestetyn linkitetyn listan, että järjestämättömän linkitetyn listan välillä ei esiintynyt huomattavissaolevia eroja hakuajassa. Millisekunnin tarkkuus ei riittänyt havaitsemaan eroja, ainoastaan mittayksiköllä tuhatosa milllisekunnista toi esiin hienoja eroavaisuuksia.

Tästä johtuen toteutin tämän lisäksi myös Trie rakenteen siten, että solmun lapset talletetaan dynaamiseen tauluun. Tällä hetkellä ohjelma järjestää lapset lisäysjärjestämisen avulla, joka tuottaa saman aikavaativuuden, kuin linkitetyn listan ratkaisukin, eli O(lasten lukumäärä). Lomitusjärjestämismetodi on myös toteutettuna, mutta tämä ei aivan vielä toimi.

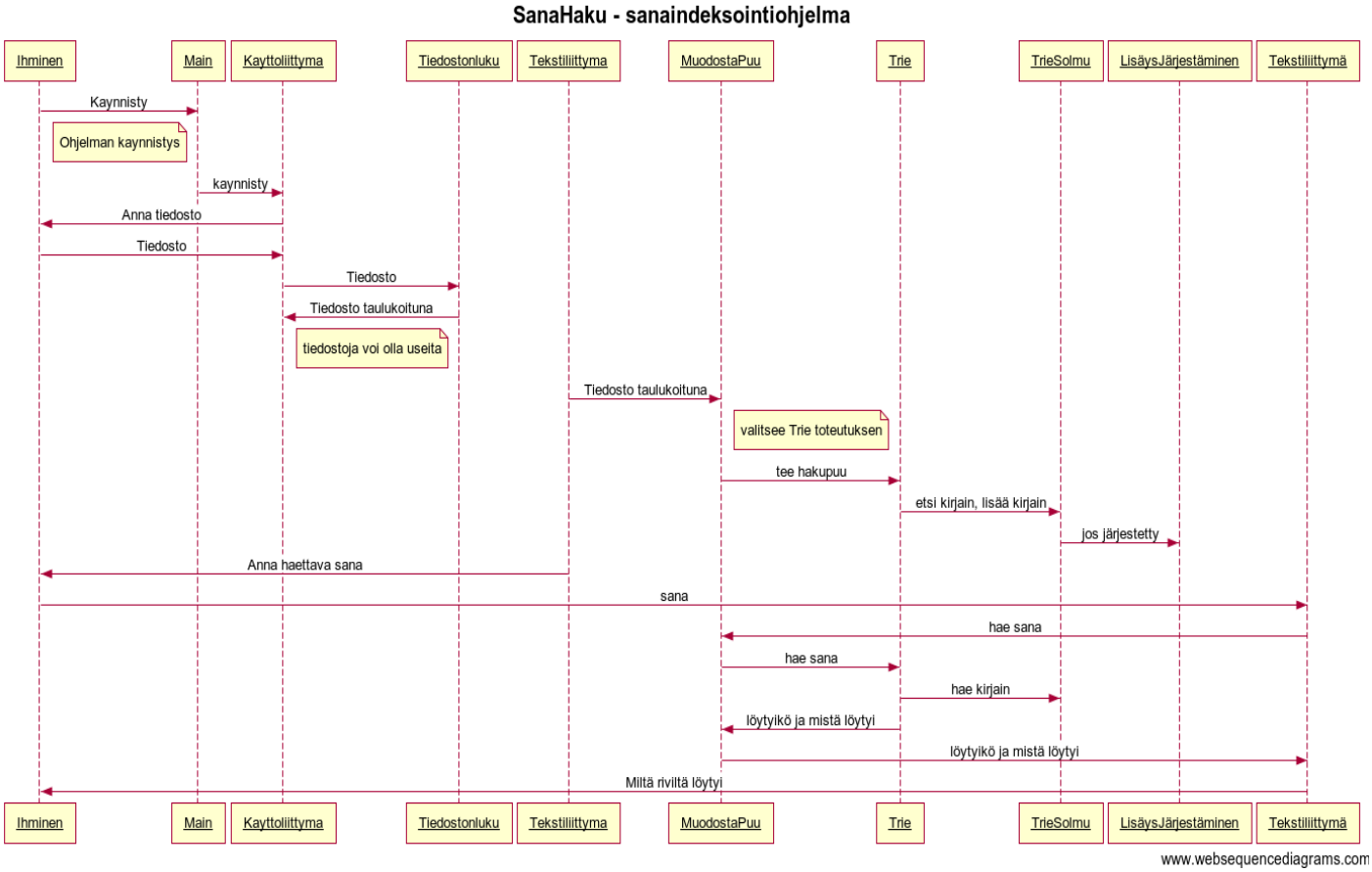
Netbeans kehitysympäristössä on mahdollista kasvattaa käytettävän maksimikeskusmuistin määrää projektin properies-ikkunan kautta, mutta tästä huolimatta maksimirivimääräksi jäi 7000 riviä kehitysympäristön keskusmuistirajoitteiden vuoksi.

2 Toteutusympäristö

2.1 Netbeans 7.2, Java, Windows 7

Ohjelman kehitys tapahtui NetBeans 7.2 kehitysympäristössä, ohjelmointikielenä Java. Käyttöjärjestelmänä oli kehitysprojektin ajan Windows 7.

2.2 Sekvenssikaavio



3 Käyttöohje

3.1 Ohjelman lataus

Ohjelman voi ladata Github-säilöstä, osoitteesta: https://github.com/kxkyll/KkyTira

3.2 Ohjelman suoritus

Ohjelmaa voi suorittaa joko

- NetBeans kehitysympäristössä, jolloin ohjelma täytyy ladata ensin Github-säilöstä tai

- Tietojenkäsittelytieteen laitoksen users koneelta komentoriviltä

  kxkyllon@users:~/tira$ java -jar /home/kxkyllon/tira/SanaHaku.jar

3.3 Ohjelman käyttö

Käynnistyksen jälkeen ohjelma kysyy käyttäjältä ladattavan tekstitiedoston hakupolkua ja nimeä. Huomaa, että tähän tulee antaa koko hakupolku (unix ympäristössä pwd komento kertoo nykyisen hakemistopolun) ja että tiedostoon pitää olla suoritusoikeudet (unix ympäristöissä oikeuden voi muutta chmod +x tiedostonnimi.txt)

Kun ohjelma on ladannut hakemiston tai jos hakemistopolun tai tiedoston nimen antaa väärin ohjelma kysyy uutta hakemistopolkua ja tiedostoa.

Kun käyttäjä ei enää halua ladata tiedostoja lisää tulee painaa enter painiketta siten, että tiedostonnimikenttä on tyhjä.

Seuraavaksi ohjelma kysyy haettavaa sanaa ja ilmoittaa mistä tiedostosta ja miltä riveiltä (tulostaen kyseiset rivit) kyseinen sana löytyy.

Jos sanaa ei löydy, ohjelma ilmoittaa tämän. Kun ohjelman suoritus halutaan lopettaa painetaan enter painiketta siten, että sanakenttä on tyhjä.

3.4 Esimerkkitulostus ohjelman käytöstä

kxkyllon@users:~$ cd tira

kxkyllon@users:~/tira$ ls -l

total 44

-rw-r--r-- 1 kxkyllon tkol 37832 2012-09-02 19:30 SanaHaku.jar

drwx------ 2 kxkyllon tkol 4096 2012-09-02 19:31 tekstit

kxkyllon@users:~/tira$ pwd

/home/kxkyllon/tira

kxkyllon@users:~/tira$ java -jar /home/kxkyllon/tira/SanaHaku.jar

Anna ladattavan tiedoston hakupolku ja nimi (tyhjä merkkijono lopettaa)

Anna ladattavat tiedostot : /home/kxkyllon/tira/tekstit/LKalevala16.txt

Annoit tiedoston: /home/kxkyllon/tira/tekstit/LKalevala16.txt

Tiedoston käsittely vei: 3 millisekuntia

Tiedoston käsittely vei: 3014 mikrosekuntia

Puun muodostaminen vei : 95 millisekuntia

Puun muodostaminen vei : 94993 mikrosekuntia

Anna ladattavat tiedostot : duupa

Annoit tiedoston: duupa

Anna ladattavat tiedostot : duupa.txt

Annoit tiedoston: duupa.txt

Tiedoston avaaminen ei onnistunut

Tiedostonnimi virheellinen

Tiedoston käsittely vei: 0 millisekuntia

Tiedoston käsittely vei: 341 mikrosekuntia

Puun muodostaminen vei : 0 millisekuntia

Puun muodostaminen vei : 47 mikrosekuntia

Anna ladattavat tiedostot : /home/kxkyllon/tira/tekstit/LEng_pg17269.txt

Annoit tiedoston: /home/kxkyllon/tira/tekstit/LEng_pg17269.txt

Tiedoston käsittely vei: 5 millisekuntia

Tiedoston käsittely vei: 5084 mikrosekuntia

Puun muodostaminen vei : 529 millisekuntia

Puun muodostaminen vei : 528922 mikrosekuntia

Anna ladattavat tiedostot : /home/kxkyllon/tira/tekstit/Ltesti1.txt

Annoit tiedoston: /home/kxkyllon/tira/tekstit/Ltesti1.txt

Tiedoston käsittely vei: 0 millisekuntia

Tiedoston käsittely vei: 221 mikrosekuntia

Puun muodostaminen vei : 0 millisekuntia

Puun muodostaminen vei : 229 mikrosekuntia

Anna ladattavat tiedostot :

Anna haettavat sanat: väinämöinen

Haun tulos:

Tiedostosta: /home/kxkyllon/tira/tekstit/LKalevala16.txt löytyivät rivit:

rivi 3: Vaka vanha Väinämöinen, tietäjä iän-ikuinen,

rivi 68: Siitä vanha Väinämöinen tietäjä iän-ikuinen,

rivi 80: Vaka vanha Väinämöinen, tietäjä iän-ikuinen,

rivi 107: Vaka vanha Väinämöinen jo huhuta huikahutti

rivi 120: Vaka vanha Väinämöinen sanan virkkoi, noin nimesi:

rivi 128: Sano totta, Väinämöinen: mi sinun Manalle saattoi?"

rivi 130: Vaka vanha Väinämöinen jo tuossa sanoiksi virkki:

rivi 137: Sano totta, Väinämöinen, sano totta toinen kerta!"

rivi 139: Vaka vanha Väinämöinen sanan virkkoi, noin nimesi:

rivi 148: Tuossa vanha Väinämöinen vielä kerran kielastavi:

rivi 156: "Oi sie vanha Väinämöinen! Jos tahot venettä täältä,

rivi 161: Sanoi vanha Väinämöinen: "Jos vähän valehtelinki,

rivi 177: Sanoi vanha Väinämöinen: "Akka tieltä kääntyköhön,

rivi 183: Itse tuon sanoiksi virkki: "Voi sinua, Väinämöinen!

rivi 188: itse tuon sanoiksi virkki: "Juop' on, vanha Väinämöinen!"

rivi 190: Vaka vanha Väinämöinen katsoi pitkin tuoppiansa:

rivi 196: Sanoi Tuonelan emäntä: "Oi on vanha Väinämöinen!

rivi 200: Sanoi vanha Väinämöinen: "Veistäessäni venoista,

rivi 233: Vaka vanha Väinämöinen sanan virkkoi, noin nimesi:

rivi 247: Siitä vanha Väinämöinen Tuonelasta tultuansa

Hakeminen vei millisekunteja 0

Hakeminen vei mikrosekunteja: 35

Tulostaminen vei millisekunteja: 1

Tulostaminen vei mikrosekunteja: 1161

Hakeminen ja tulostaminen yhteensä: 01 millisekuntia

Hakeminen ja tulostaminen yhteensä: 1196 mikrosekuntia

Anna haettavat sanat: vika

Haettavaa sanaa vika ei löytynyt

Hakeminen vei: 0 millisekuntia

Anna haettavat sanat: haka

Haettavaa sanaa haka ei löytynyt

Hakeminen vei: 0 millisekuntia

Anna haettavat sanat: noista

Haettavaa sanaa noista ei löytynyt

Hakeminen vei: 0 millisekuntia

Anna haettavat sanat: **nimesi**

Haun tulos:

Tiedostosta: /home/kxkyllon/tira/tekstit/LKalevala16.txt löytyivät rivit:

rivi 81: sanan virkkoi, noin nimesi: "Voi poloinen, päiviäni!

rivi 115: Sanan virkkoi, noin nimesi, itse lausui ja pakisi:

rivi 120: Vaka vanha Väinämöinen sanan virkkoi, noin nimesi:

rivi 134: sanan virkkoi, noin nimesi: "Tuosta tunnen kielastajan!

rivi 139: Vaka vanha Väinämöinen sanan virkkoi, noin nimesi:

rivi 143: sanan virkkoi, noin nimesi: "Ymmärrän valehtelijan!

rivi 207: Tuopa Tuonelan emäntä sanan virkkoi, noin nimesi:

rivi 233: Vaka vanha Väinämöinen sanan virkkoi, noin nimesi:

Hakeminen vei millisekunteja 0

Hakeminen vei mikrosekunteja: 15

Tulostaminen vei millisekunteja: 0

Tulostaminen vei mikrosekunteja: 381

Hakeminen ja tulostaminen yhteensä: 00 millisekuntia

Hakeminen ja tulostaminen yhteensä: 396 mikrosekuntia

Anna haettavat sanat: **home**

Haun tulos:

Tiedostosta: /home/kxkyllon/tira/tekstit/LEng_pg17269.txt löytyivät rivit:

rivi 368: Children away from home write to their parents on Mothering Sunday if

rivi 369: unable to get home.

rivi 531: had a garland of flowers put round its horns when driven home at night,

rivi 709: The Harvest Home Suppers are now almost a thing of the past. I went to

rivi 711: when the last load of corn is taken home. This load used to be decorated

rivi 715: Harvest Home! Harvest Home;

rivi 715: Harvest Home! Harvest Home;

rivi 719: We've got our harvest home.

rivi 729: HARVEST HOME.

rivi 739: I've been to Harvest Home all the world over, over, and over,

rivi 945: home brewed ale or mead.

rivi 1018: Crane.:--This game was generally played during the Harvest Home Feast.

rivi 1151: When a cat is taken to a new home its feet should be buttered, and it

rivi 1386: A black cat following anyone into a home brings good luck.

rivi 1482: Bees flying far from their hives and coming home late foretells fine

rivi 1741: Ladybird, ladybird, fly away home,

Hakeminen vei millisekunteja 0

Hakeminen vei mikrosekunteja: 25

Tulostaminen vei millisekunteja: 13

Tulostaminen vei mikrosekunteja: 13741

Hakeminen ja tulostaminen yhteensä: 013 millisekuntia

Hakeminen ja tulostaminen yhteensä: 13766 mikrosekuntia

Anna haettavat sanat: **sa'an**

Haun tulos:

Tiedostosta: /home/kxkyllon/tira/tekstit/LKalevala16.txt löytyivät rivit:

rivi 99: Arvelee, ajattelevi: "Tuolta saan sa'an sanoja,

rivi 244: sa'an saapi taimenia, tuhat emon alvehia,

Hakeminen vei millisekunteja 0

Hakeminen vei mikrosekunteja: 22

Tulostaminen vei millisekunteja: 0

Tulostaminen vei mikrosekunteja: 203

Hakeminen ja tulostaminen yhteensä: 00 millisekuntia

Hakeminen ja tulostaminen yhteensä: 225 mikrosekuntia

Anna haettavat sanat: **tuopilla**

Haun tulos:

Tiedostosta: /home/kxkyllon/tira/tekstit/LKalevala16.txt löytyivät rivit:

rivi 187: toip' on tuopilla olutta, kantoi kaksikorvaisella;

Hakeminen vei millisekunteja 0

Hakeminen vei mikrosekunteja: 21

Tulostaminen vei millisekunteja: 0

Tulostaminen vei mikrosekunteja: 205

Hakeminen ja tulostaminen yhteensä: 00 millisekuntia

Hakeminen ja tulostaminen yhteensä: 226 mikrosekuntia

Anna haettavat sanat:

4 Lähteet

Tietorakenteet kurssin 58131 kevät 2012 kurssimateriaali: Patrik Floréen
http://www.cs.helsinki.fi/u/floreen/tira2012/tira.pdf

Trie rakenteen esittely: Esa Junttila
http://www.cs.helsinki.fi/u/ejunttil/opetus/tiraharjoitus/trie.html

Algoritmit 2, Luento 6: Timo Männikkö
http://users.jyu.fi/~mannikko/algoritmit2/luennot/luento6.pdf

Data Structures and the Java Collections Framework: William J Collins