Kurssi: Tietorakenteet ja algoritmit 2020
Tehtävä: Harjoitustyö
Opiskelija: Tomi Heinonen (1650860)
tomi.heinonen@nondest.fi
theinonen@student.oulu.fi

Työselostus

1. Ongelman ratkaisu

Tehtävänä oli etsiä tekstitiedostosta sen sata yleisintä sanaa, C-kieltä käyttäen.

Tehtävä toteutettiin lukemalla käyttäjän syöttämä tekstitiedosto ohjelman main-funktiossa ja tallentamalla kukin tekstissä esiintyvä sana avain-arvo parina hajautustauluun. Taulukko alustettiin nollilla init_hash_table()-funktiossa ja arvoa kasvatettiin yhdellä sanojen esiintymistymiskertojen mukaan. Sanojen sijoittaminen hajautustauluun toteutettiin laskemalla kullekin yksittäiselle sanalle sen hash-arvo, hash_func():ssa ja tallentamalla sana hajautustauluun sen hash-arvoa vastaavalle paikalle. Sijoittaminen suoritettiin hash_table_insert()-funktiossa, avointa hajautusta ja lineaarista kokeilua käyttäen.

Lopuksi taulukon alkiot järjestettiin quick_sort()-funktion avulla sanojen esiintymiskertojen mukaan ja tulostettiin sata yleisintä sanaa print_table()-funktion avulla suuruusjärjestykseen.

Ohjelman tietorakenteet ja funktiot parametreineen:

```
struct sanasto{
   char *key;
   int value;
} sanasto;

struct sanasto* hash_table[SIZE];

unsigned int hash_func(char *sana)
void init_hash_table()
void print_table()
bool hash_table_insert(struct sanasto *p)
void quick_sort(int alku, int loppu)
int partition(int alku, int loppu)
int main()
```

2. Ratkaisun analyysi

Ongelman ratkaisu on sikäli onnistunut, että ohjelma toimii kohtalaisen hyvin ja tehokkaasti varsinkin pienillä tekstitiedostoilla. Hajautustaulun operaatiot vievät aikaa O(m), missä m on kokeilujonon pituus. Pikalajittelun aikavaativuus on keskimäärin O(n log n), jossa n on lajiteltavien alkioiden määrä.

Isoilla tekstitiedostoilla ohjelma toimii hyvin ensimmäisellä suorituskerralla, mutta suorituskyky hiipuu välittömästi tämän jälkeen, johtuen ilmeisesti epäonnistuneesta muistin vapautuksesta, joka taas johtuu monimutkaisesta tietorakenteen toteutuksesta. Ongelma olisi ratkaistavissa pätevämmän ohjelmoijan toimesta, mutta itseltäni tämä jäi nyt suorittamatta. Tästä johtuen myös tehtäväksi annettu ohjelman suoritusaikojen mittaus jäi tekemättä.

Muita ohjelmassa esiintyviä puutteita voi havaita ohjelman tulosteessa, johon on jäänyt yksittäisiä kirjaimia ja sanojen loppuja (strtok()-metodin epätäydellisestä toteutuksesta johtuen), sekä jonkin verran duplikaatteja, mikäli sana on esiintynyt sekä isolla että pienellä alkukirjaimella tekstissä.