Testausdokumentti

Muutin ohjelmaa siten, että tiedostonkirjoitus on erotettu pakkaus- ja purkualgoritmeilla, niin että testiajat mittaavat pelkkää LZW-algoritmia. Tein testit seitsemällä Project Gutenbergista otetulla tekstitiedostolla ja kolmella testaustavalla: ensimmäiset testaukset tein Javan valmiillta ArrayList- ja HashMap-tietorakenteilla ja kaksi jälkimmäistä testiä omilla tietorakenteilla. Jälkimmäiseen testiin vaihdoin pakkauksen hajautusalgoritmia. Ensin hajautusalgoritmi tallensi sanakirjaan merkkijonot/tavut kaavalla a+b+c+d jne. jaettuna alkuluvulla. Paransin algoritmia yksinkertaisesti niin, että sanakirjaan tallennettuihin merkkeihin lisättiin kumulatiivinen kerroin kaavalla a*1+b*2+c*3+d*4+...+x*n, joka jaetaan myös alkuluvulla. Tällä jälkimmäisellä tavalla pakkausnopeus puolittui ja läheni Javan HashMap-rakennetta. Varsinkin suuremmissa tiedostoissa tehokkuus parani huomattavasti.

Tein ohjelmaan Java-luokan, jolla voidaan testata näitä seitsemää tekstitiedostoa kymmenellä toistokerralla. Testaus edellyttää, että samannimiset dokumentit ovat samassa kansiossa kuin ohjelma. Laitoin kyseiset dokumentit GitHubiin.

<u>Aleksis Kivi, Seitsemän veljestä (suomeksi)</u>

Tiedostokoko: 671611 tavua
Pakkauskoko: 263140 tavua
Pakkaustehokkuus (pakkauksen koko alkuperäiseen verrattuna): 39,2 %

Javan valmiilla tietorakenteilla:
Keskimääräinen pakkausaika: 1427,4 millisekuntia
Keskimääräinen purkuaika: 424,6 millisekuntia

Omilla tietorakenteilla:
Keskimääräinen pakkausaika: 4455,3 millisekuntia
Keskimääräinen purkuaika: 818,8 millisekuntia

Parannetulla hajautusalgoritmilla:
Keskimääräinen pakkausaika: 2174,7 millisekuntia

<u>Shakespeare, Hamlet (englanniksi)</u>

Tiedostokoko: 180710 tavua
Pakkauskoko: 78344 tavua
Pakkaustehokkuus: 43,4 %

Javan valmiilla tietorakenteilla:
Keskimääräinen pakkausaika: 361,9 millisekuntia
Keskimääräinen purkuaika: 120,1 millisekuntia

Omilla tietorakenteilla:
Keskimääräinen pakkausaika: 686,4 millisekuntia
Keskimääräinen purkuaika: 199,8 millisekuntia

Parannetulla hajautusalgoritmilla:
Keskimääräinen pakkausaika: 435,4 millisekuntia

<u>Shakespeare, Hamlet (suomeksi)</u>

Tiedostokoko: 194041 tavua
Pakkauskoko: 80932 tavua
Pakkaustehokkuus: 41,7 %
Javan valmiilla tietorakenteilla:
Keskimääräinen pakkausaika: 369,6 millisekuntia
Keskimääräinen purkuaika: 127,9 millisekuntia

Omilla tietorakenteilla:
Keskimääräinen pakkausaika: 706,7 millisekuntia
Keskimääräinen purkuaika: 249,5 millisekuntia

Parannetulla hajautusalgoritmilla:
Keskimääräinen pakkausaika: 506,9 millisekuntia

Shakespeare, Hamlet (ranskaksi)

Tiedostokoko: 326477 tavua
Pakkauskoko: 136001 tavua
Pakkaustehokkuus: 41,7 %

Javan valmiilla tietorakenteilla:
Keskimääräinen pakkausaika: 683,4 millisekuntia
Keskimääräinen purkuaika: 204,4 millisekuntia

Omilla tietorakenteilla:
Keskimääräinen pakkausaika: 1421,1 millisekuntia
Keskimääräinen purkuaika: 376,1 millisekuntia

Parannetulla hajautusalgoritmilla:
Keskimääräinen pakkausaika: 859,2 millisekuntia

Täysin satunnainen tiedosto

Tiedostokoko: 200000 tavua
Pakkauskoko: 256060 tavua
Pakkaustehokkuus: 41,7 %

Javan valmiilla tietorakenteilla:
Keskimääräinen pakkausaika: 917,1 millisekuntia
Keskimääräinen purkuaika: 377,6 millisekuntia

Omilla tietorakenteilla:
Keskimääräinen pakkausaika: 3920,2 millisekuntia
Keskimääräinen purkuaika: 444,5 millisekuntia

Parannetulla hajautusalgoritmilla:
Keskimääräinen pakkausaika: 2408,6 millisekuntia

James Joyce, Ulysses (englanniksi)

Tiedostokoko: 1553455

Pakkauskoko: 657176

Pakkaustehokkuus: 42,3 %

Javan valmiilla tietorakenteilla:
Keskimääräinen pakkausaika: 3337 millisekuntia
Keskimääräinen purkuaika: 1054,6 millisekuntia

Omilla tietorakenteilla:
Keskimääräinen pakkausaika: 18751 millisekuntia
Keskimääräinen purkuaika: 1954,6 millisekuntia

Parannetulla hajautusalgoritmilla:
Keskimääräinen pakkausaika: 7892 millisekuntia

Hanshangmengren, 風月夢 (kiinaksi)

Tiedostokoko: 463960 tavua
Pakkauskoko: 218466 tavua
Pakkaustehokkuus: 47,1 %

Javan valmiilla tietorakenteilla:
Keskimääräinen pakkausaika: 833 millisekuntia
Keskimääräinen purkuaika: 331 millisekuntia

Omilla tietorakenteilla:
Keskimääräinen pakkausaika: 5859,5 millisekuntia
Keskimääräinen purkuaika: 321,4 millisekuntia

Parannetulla hajautusalgoritmilla:
Keskimääräinen pakkausaika: 1923,5 millisekuntia