

Harjoitustyö

Harjoitustyö

- Tehtävä: ohjelmoi Java-kielellä tekstikokoelmia käsittelevä olioperustainen ohjelma.
 - Dokumentit säilötään Javan *LinkedList*-luokasta peritylle omalle listalle.
- Työ tehdään pääosin itse.
 - Ideoita voi vaihtaa – koodia ei.
 - Henkilökohtaista ohjausta mikroluokissa.
 - Ohjaajilta voi kysyä sähköpostilla.
- Palautus WETO-järjestelmään noin kahden kuukauden kuluttua.
 - Testaus automaattisesti julkisilla ja salaisilla testeillä.
- Lopullisen version palautus keskiviikkona 29.4.

Harjoitustyö

- Torstaina 9.4. on välipalautuspiste, johon mennessä täytyy palauttaa WETOn osa harjoitustyön koodista (ainakin luokkahierarkia ja oma lista).
 - Välipisteessä tarkistuksen pääroolissa on WETOn automaattinen testaus. Opettaja lukee koodia vain poikkeustapauksissa.
- Tarkka tehtävänanto julki tällä viikolla tai ensi viikon alussa.
- Varaa työlle aikaa, aloita ajoissa, tee työtä säännöllisesti ja käy ohjauksissa jo niiden alkuvaiheessa, jos tarvitset apua.
 - Ensimmäisillä ohjauskerroilla opettajalla on enemmän aikaa yhdelle oppilaalle ja päinvastoin.
 - Toissakertaisen kurssitoteutuksen opiskelijat ($n = 151$) arvioivat harjoitustyön tekemiseen menneen aikaa näin: keskiarvo 44 tuntia, keskihajonta 35 h ja mediaani 35 h.

Harjoitustyö

- Harjoitustyöstä on kaksi versiota: suppea ja laaja.
- Laajassa työssä on enemmän toiminnallisuutta ja vain sillä voi saada arvosanan 5.
 - Suppealla työllä voi saada korkeintaan arvosanan 4.
- Arvosanassa huomioidaan myös
 - työn WETO-testitulokset sekä väli- että lopullisessa palautuksessa samoin kuin
 - palautusaikataulussa pysyminen sekä väli- että lopullisessa palautuksessa.
- Arvostelu kuvataan tarkemmin tarkassa tehtävänannossa, jossa selitetään myös mitä kaikkea laajaan työhön vaaditaan.

Tekstikokoelmat

- Tekstikokoelma (myös korpus) koostuu tekstimuotoisista dokumenteista.
- Dokumenteissa on usein tunniste ja muuta metatietoa, esimerkiksi päivämäärä.
- Tekstikokoelmat ovat yleensä hyvin suuria.
- Harjoitustyössä käytetään hyvin pieniä kokoelmia, joissa on korkeintaan joitakin kymmeniä dokumentteja.
- Harjoitustyö ohjelma osaa käsitellä kahdentyyppisiä kokoelmia.
 - Lajiteltuja vitsejä ja päivämäärän sisältäviä uutisia.
- Molempien tyyppisissä dokumenteissa on yksilöivä kokonaislukutunniste (> 0).

Tekstikokoelmat

- Dokumentit on annettu tekstitiedostoissa, joissa yhden dokumentin tiedot ovat aina yhdellä rivillä.
- Dokumenttien järjestys on satunnainen.
- Vitsejä ja uutisia sisältävien tiedostojen nimet alkavat vastaavasti *jokes*- ja *news*-sanoilla.
 - *jokes_oldies.txt*
 - *news_reuters.txt*
- Dokumentit ovat englanniksi.
- Kokoelmien voi olettaa aina olevan kunnossa.
 - Esimerkiksi tunnisteita ei puuttuu ja dokumenteissa on aina tekstiä.
 - Rivien aluissa tai loppuissa ei ole välilyöntejä.
- Dokumentin tiedot erotetaan toisistaan merkkijonolla "///".
 - Tunniste, erotin, vitsin laji, erotin, vitsin teksti.
 - Tunniste, erotin, uutisen päivämäärä, erotin, uutisen teksti.

Tekstikokeelmat

Alla on annettu *jokes_oldies.txt*-tiedoston sisältöä. Kuvan ulkopuolelle jatkuvat rivit on esitetty kolmella pisteellä. Rivit ovat luonnollisesti tiedostossaan täysimittaisia. Esimerkiksi ensimmäisellä rivillä on tunnisteella 4 yksilöity sekalaiseen luokkaan (miscellaneous) kuuluva vitsi, jonka kuuluu: *This may be the wine talking, but I really, really, really, really love wine.*

```
4///miscellaneous///This may be the wine talking, but I really, really, really, really love wine.
11///computing///These two strings walk into a bar and sit down. The bartender says, ...
9///computing///If you put a million monkeys at a million keyboards, one of them will ...
3///miscellaneous///Nothing sucks like a Hoover.
5///computing///Q. What is the biggest lie in the entire universe? A. "I have read and agree ...
10///computing///To understand what recursion is, you must first understand recursion.
8///computing///A SQL query goes into a bar, walks up to two tables and asks, "Can I join you?"
2///miscellaneous///Did you hear about the semi-colon that broke the law? He was given two ...
1///miscellaneous///Q. What's the difference between ignorance and apathy? A. I don't know and I don't care.
6///computing///Q. How does a computer get drunk? A. It takes screenshots.
7///computing///Q: How many programmers does it take to change a light bulb? A: None. It's a hardware problem.
12///computing///When your hammer is C++, everything begins to look like a thumb.
```

Tekstikokoelmat

Alla *news_reuters.txt*-tiedoston sisältöä. Tämän tiedoston dokumentit ovat pitempiä ja kaikki rivit menevät siksi kuvan ulkopuolelle. Kokoelman ensimmäinen uutisen tunniste on 20255 ja uutinen on julkaistu kaukana menneisyydessä 1.3.1987. Tämä uutinen käsittelee silloisia kansainvälisiä suhteita.

```
20255///1.3.1987///U.S. Secretary of State George Shultz is likely to visit Moscow soon for talks ...
17033///23.4.1987///Special Japanese envoy Shintaro Abe said in a brief interview with Reuters that ...
22996///9.3.1987///Union Bank of Finland is issuing a 10 billion yen eurobond due March 30, 1992 paying ...
13876///19.10.1987///Six black miners have been killed and two injured in a rock fall three km underground ...
3225///9.3.1987///The Ghana Cocoa Board said it purchased 1,323 tonnes of cocoa in the 21st week, ended ...
13876///7.4.1987///Pyramid Technology Corp said it has reached an agreement to sell Oracle Corp's Oracle ...
1023///3.3.1987///Sandoz AG said it planned a joint venture to produce herbicides in the Soviet Union. ...
```


Luokkahierarkia

- Kokoelmia mallinnetaan konkreettisilla *Vitsi*- ja *Uutinen*-luokilla, joilla on abstrakti *Dokumentti*-yliluokka.
- Yliluokalla on kaksi dokumenteille yhteistä tietoa.
 - Dokumentin tunniste: **int**-tyyppinen *tunniste*-attribuutti (> 0).
 - Dokumentin teksti: *String*-tyyppinen *teksti*-attribuutti (ei saa olla **null** tai tyhjä merkkijono "").
- *Vitsi*-luokan omalle tiedolle tehdään *String*-tyyppinen *laji*-attribuutti.
 - Ei saa olla **null** tai tyhjä merkkijono.
- *Uutinen*-luokan tiedolle tehdään puolestaan *LocalDate*-tyyppinen *päivämäärä*-attribuutti.
 - *LocalDate* on Javan API:n *java.time*-pakkauksessa oleva luokka.
 - Ei saa olla **null**.

Luokkahierarkia

- Kätke attribuutit ja tee aksessorit harjoitusten tapaan.
- Korvaa luokissa *toString*-metodi mielellään ketjuttaen.
 - Merkkijonoesitykset ovat täsmälleen samanlaiset kuin tiedostossa olevat rivit.
 - *Dokumentti*: palauttaa tunnisteiden ja tekstin toisistaan erotettuna. Paluuarvo esimerkiksi: "3///Nothing sucks like a Hoover."
 - *Vitsi*: lisää ylliluokassa korvatun metodin palauttaman merkkijonon keskelle laji. Paluuarvo esimerkiksi "3///miscellaneous///Nothing sucks like a Hoover."
 - *Uutinen*: lisää ylliluokassa korvatun metodin palauttaman merkkijonon keskelle päivämäärä. Paluuarvo esimerkiksi "20255///1.3.1987///U.S. Secretary of State George Shultz is likely to visit Moscow soon for talks following new Soviet arms control proposals, U.S. Officials said. The officials told Reuters no decision had been reached on when the trip might take place, but it was likely to be within a month. Kremlin leader Mikhail Gorbachev proposed on Saturday that talks on medium range missiles be separated from other arms issues. Schultz is currently visiting Peking."

Luokkahierarkia

- Korvaa *Dokumentti*-luokassa *equals*-metodi siten, että vertaillet vain ja ainoastaan dokumenttien tunnisteita.
- Toteuta *Dokumentti*-luokassa *Comparable*-rajapinta ja vertaile myös *compareTo*-metodissa tunnisteita.
 - Palauta korvatusta operaatiosta -1, 0 tai 1.
 - Kiinnitä tyyppiparametrin arvoksi *Dokumentti*-tyyppi.
- Tee luokille parametrilliset rakentajat.
 - *Dokumentti*: **int**- ja *String*-tyyppiset parametrit.
 - *Vitsi*: **int**-, *String*- ja *String*-tyyppiset parametrit. Keskimmäinen parametri on luokan omalle tiedolle.
 - *Uutinen*: **int**-, *LocalDate*- ja *String*-tyyppiset parametrit.
- Ketjuta rakentajat: kutsu aliluokkien rakentajista ylliluokan rakentajaa.

Luokkahierarkia

- Rakentajat samoin kuin asettavat aksessorit heittävät *IllegalArgumentException*-poikkeuksen, jos parametreissa havaitaan virhe.
 - Poikkeus heitetään, jos tunnisteparametri on arvoltaan yhtä pienempi, viitetyyppinen parametri on **null**-arvoinen tai *String*-tyyppinen parametri on tyhjä merkkijono.
- Opettaja antaa valmiina *harjoitustyo.apulaiset*-pakkauksessa kolme rajapintaa, joista yksi on toteutettava hierarkiassa.
 - *Dokumentti*-luokka toteuttaa *Tietoinen<T>*-rajapinnan, jossa koko hierarkialle yhteisiä metodeja.
 - Pakkaus julkaistaan Hervannan kampuksen hallinnoimassa GitLabissa.
- Pakkaa luokkahierarkia *harjoitustyo.dokumentit* pakkaukseen.
- Luokkahierarkiasta kerrotaan lisää tarkassa tehtäväannossa.

Oma lista

- Peri Javan *LinkedList<E>*-luokasta geneerinen *OmaLista<E>*-luokka.
- Listaan on kirjoitettava vähintään yksi opettajan antamassa *Ooperoiva<E>*-rajapinnassa määritelty listaoperaatio.
 - Alkion lisäys siten, että alkioden välille muodostuu kasvava järjestys.
 - Myös *Ooperoiva<E>* kuuluu myöhemmin julkaistavaan *harjoitustyo.apulaiset*-pakkaukseen.
- Operaatioiden on oltava “puhtaita”: niissä ei saa esiintyä esimerkiksi *Dokumentti*-tyyppiä.
- Sijoita oma lista *harjoitustyo.omalista*-nimiseen pakkaukseen.

Muista luokista

- Tee omat luokka kokoelmalle (*Kokoelma*) ja käyttöliittymälle ja tee erillinen *main*-metodin sisältävä ajoluokka *Oope2HT*.
 - Luokilla tulee olla selkeät vastuut: kokoelma vastaa ohjelman logiikasta, kun taas käyttöliittymä vuorovaikuttaa.
- Kokoelma toteuttaa komennot omaa listaa käsittelemällä.
 - *Kokoelma*-luokalla on *OmaLista<Dokumentti>*-tyyppinen ja *dokumentit*-niminen attribuutti. Listan viitteet liittävät dokumentit kokoelmaansa.
- Kaikki käyttäjän antamat tiedot luetaan käyttöliittymässä.
 - Pääsilmukan luonteva paikka on käyttöliittymän luokka.
 - Käyttöliittymässä pyritään tekemään valtaosa tulostuksista. *Kokoelma*-luokan metodit voivat palauttaa tulosteensa esimerkiksi rivinvaihtoja sisältävinä merkkijonoina.

Muista luokista

- Ajoluokka luokka lähinnä luo käyttöliittymäolion ja käynnistää sen pääsilmutkan.
- Muita luokkia voi tehdä tarpeen mukaan.
- Muista aina pohtia tarkkaan mikä suhde luokkiesi välillä on luontevin. Erityisesti periytymissuhde on *IsA*-sellainen, kun taas koostesuhde on luonteva, kun voidaan sanoa *HasA* is *PartOf*.
 - Älä esimerkiksi peri *Kokoelma*-luokkaa *OmaLista*<*E*>-luokasta, vaan valitse näiden luokkien välille koostesuhde.

Ohjelman komennot

- **print**: tulostaa näytölle tietyn dokumentin merkkijonoesityksen tai kaikki dokumentit listan järjestyksessä. Tulostettava dokumentti valitaan sen tunnisteiden avulla.
- **add**: lisää kokoelmaan uuden dokumentin. Lisäys epäonnistuu, jos kokoelmassa on jo annetulla tunnisteella yksilöity dokumentti tai vitsien kokoelmaan yritetään lisätä uutista tai päinvastoin.
- **remove**: poistaa dokumentin kokoelmasta. Dokumentti valitaan sen tunnisteiden avulla.
- **find**: tulostaa näytölle niiden dokumenttien tunnisteet, jotka sisältävät kaikki komennolle annetut hakusanat. Hakua alkaa kokoelman ensimmäisestä dokumentista.

Ohjelman komennot

- **polish**: koko kokoelman esikäsittelyä. 1) poistaa kaikista dokumenteista mahdollisesti annetut välimerkit, 2) muuttaa kirjaimet pieniksi ja 3) poistaa dokumenteista niin sanottuja sulkusanoja, jotka ovat niin yleisiä ettei niistä ole apua tiedonhaussa.
- **reset**: lataa dokumenttitiedoston uudelleen ja poistaa kaikki aiemmin tehdyt muutokset.
- **echo**: käynnistää tai lopettaa komentojen kaiutuksen.
- **quit**: lopettaa ohjelman.

Ohjelman komennot

- **freqs**: tulostaa joko tietyn dokumentin tai kokoelman sanojen frekvenssit. Vain laajaan työhön kuuluva **valinnainen** komento.
- **sort**: lajittelee kokoelman kolmella eri tavalla: minkä tahansa kokoelman tunnisteiden mukaan, vitsikokoelmassa vitsien lajin mukaan ja uutiskokoelmassa päivämäärän mukaan. Kaksi viimeistä lajittelua epäonnistuvat, jos ne kohdistetaan väärään tyyppiseen kokoelmaan. Vain laajaan työhön kuuluva **valinnainen** komento.
- **pprint**: rivittää tulosteen siten, että tunnisteella yksilöity dokumentti tai kaikki dokumentit tulostetaan annetun levyiseen sarakkeeseen. Pohdi tätä ajan kanssa; saattaa yllättää. Vain laajaan työhön kuuluva **valinnainen** komento.

Ohjelman komennot

- Laajassa työssä on kaikki edellä mainitut komennot. Suppeassa työssä ei ole **freqs-**, **sort-** ja **pprint**-komentoja.
- Muuta:
 - Kaikista komennoista kerrotaan lisää tarkassa tehtäväannossa.
 - Pienet ja isot kirjaimet katsotaan eri merkeiksi.
 - Virheen tarkistus liittyy pitkälti parametreihin, jotka voivat esimerkiksi olla arvoltaan virheellisiä tai niitä voi olla väärä määrä.
 - Kaikkiin virheisiin reagoidaan tulostamalla merkkijono "Error!"

Ohjelman käynnistäminen

- Ohjelma lataa käynnistyessään dokumenttitiedosto ja sulkusanat sisältävän tiedoston.
 - Ohjelma päättelee tiedoston nimestä onko kokoelmassa vitsejä vai uutisia.
- Tiedostojen nimet annetaan ohjelmalle komentorivi-parametreina siten, että kokoelman sisältävä tiedosto on ensimmäinen parametri.
 - Esimerkiksi: `java Oope2HT jokes_oldies.txt stop_words.txt`
 - Ohjelma tulostaa virheilmoituksen ja pysähtyy, jos parametreja on väärä määrä tai tiedostoista jompaakumpaa tai molempia ei voida avata.
 - Tiedoston avaus voi epäonnistua, kun se on jossain toisessa hakemistossa tai jos tiedoston nimi on kirjoitettu väärin.

Esimerkkiajo

```
Welcome to L.O.T.  
Please, enter a command:  
print 6  
6///computing///Q. How does a computer get drunk? A. It takes screenshots.  
Please, enter a command:  
print 13  
Error!  
Please, enter a command:  
add 13///computing///This code was generated by a tool.  
Please, enter a command:  
print 13  
13///computing///This code was generated by a tool.  
Please, enter a command:  
find The  
9  
11  
Please, enter a command:  
find understand recursion  
10
```

Esimerkkiajo

```
Please, enter a command:
find bar
11
Please, enter a command:
remove 11
Please, enter a command:
find bar
Please, enter a command:
print 8
8///computing///A SQL query goes into a bar, walks up to two tables and asks, "Can I join you?"
Please, enter a command:
polish ,.:?"'
Please, enter a command:
find bar
8
Please, enter a command:
print 8
8///computing///sql query goes bar walks tables asks join
Please, enter a command:
reset
```

Esimerkkiajo

```
Please, enter a command:
print 8
8///computing///A SQL query goes into a bar, walks up to two tables and asks, "Can I join you?"
Please, enter a command:
print 10
10///computing///To understand what recursion is, you must first understand recursion.
Please, enter a command:
freqs 10
To 1
first 1
is, 1
must 1
recursion 1
recursion. 1
understand 2
what 1
you 1
A total of 10 words.
Please, enter a command:
polish ,.:?"'
```

Esimerkkiajo

```
Please, enter a command:
print 10
10///computing///understand recursion understand recursion
Please, enter a command:
freqs 10
recursion 2
understand 2
A total of 4 words.
Please, enter a command:
reset
Please, enter a command:
sort type
Please, enter a command:
pprint 50
5///computing///Q. What is the biggest lie in the
entire universe? A. "I have read and agree to the
Terms & Conditions."
6///computing///Q. How does a computer get drunk?
A. It takes screenshots.
```


Esimerkkiajo

7///computing///Q: How many programmers does it take to change a light bulb? A: None. It's a hardware problem.

8///computing///A SQL query goes into a bar, walks up to two tables and asks, "Can I join you?"

9///computing///If you put a million monkeys at a million keyboards, one of them will eventually write a Java program. The rest of them will write Perl programs.

10///computing///To understand what recursion is, you must first understand recursion.

11///computing///These two strings walk into a bar and sit down. The bartender says, "So what'll it be?" The first string says, "I think I'll have a beer quag fulk boorg jdk^CjfdLk jk3s d#f67howe%^U r89nvy owmc63^Dz x.xvcu." "Please excuse my friend," the second string says, "He isn't null-terminated."

12///computing///When your hammer is C++, everything begins to look like a thumb.

Esimerkkiajo

```
1///miscellaneous///Q. What's the difference  
between ignorance and apathy? A. I don't know and  
I don't care.  
2///miscellaneous///Did you hear about the  
semi-colon that broke the law? He was given two  
consecutive sentences.  
3///miscellaneous///Nothing sucks like a Hoover.  
4///miscellaneous///This may be the wine talking,  
but I really, really, really, really love wine.  
Please, enter a command:  
quit  
Program terminated.
```