MOOC-näyttökokeen ohjelmointiosio

Näyttökokeen ohjelmointiosiossa luot ohjelman annettujen ohjeiden perusteella. Huomaa, että käytössäsi on **2h 30 min** aikaa tehtävien ratkaisuun. On hyvin mahdollista että et ehdi tekemään tehtäviä kokonaisuudessaan, mikä on täysin hyväksyttävää.

Pyri tekemään mieluummin niin, että toteutat pienemmän määrän hyvin toimivia toiminnallisuuksia, kuin että toteuttaisit ison määrän huonosti toimivia toiminnallisuuksia. Käytännössä vain toimivat osuudet ovat tärkeitä.

Ohjeet

- Avaa Netbeans with TMC
- TMC NetBeans näyttää kirjautumisruudun.
    - Kirjoita oma **Mooc.fi sähköpostisi/käyttäjänimesi** ja salasanasi
    - Mikäli et muista salasanaasi, mene osoitteeseen https://tmc.mooc.fi/ ja nollaa salasanasi sieltä
    - Jos sinulta kysytään organisaatiota, valitse organisaatioksi *MOOC näyttökoe*
    - Valitse kurssi *“MOOC Näyttökoe 2019”*
    - Lataa tarjolla oleva tehtäväpohja “Tehtavapohja”
- Lue seuraavat kohdat huolellisesti ja aloita kokeen tekeminen!
- **HUOM!** Vaikka tehtäväpohja ladataan TMC:llä, tehtävässä ei ole testejä ja testinapin painelemisesta ei ole mitään hyötyä!
- Jos tarvitset teknistä apua, nosta käsi pystyyn

Näppäimistö tietojenkäsittelytieteen laitoksen koneella

Näppäimistökomennot osaston koneilla saattavat toimia eri tavalla kuin olet tottunut:

- Koodin automaattinen täydennys tapahtuu painamalla `ctrl + space`
- Koodin automaattinen sisennys tapahtuu painamalla `alt + shift + f`
- @-merkin saat painamalla `<Alt Gr (oikealla)> + 2`
- `{` ja `}` -merkit saat `<Alt Gr (oikealla)> + 7` ja `<Alt Gr (oikealla)> + 0`

Lopputoimet

- Lähetä tehtävä palvelimelle painamalla testit ajavan "silmän" oikealla puolella olevaa ylöspäin osoittavaa nuolta (submit), palvelimen virheilmoitus voi kertoa osan testeistä epäonnistuneen. Älä kuitenkaan häiriinny tästä!

Tehtävänanto

Näyttökokeessa analysoidaan erään kirjojen arviointiin keskittyvän palvelun tietoja. Käytät kokeessa neljää tiedostoa, jotka löytyvät työpöydällä olevan Data-kansion alla sijaitsevasta kansiosta books. Tiedostot ovat seuraavat:

- Tiedosto **books.csv** sisältää kirjojen tietoja.
- Tiedosto **tags.csv** sisältää käyttäjien kirjoittamia tägejä, joita voidaan liittää kirjoihin.
- Tiedostoa **book_tags.csv** käytetään kirjojen ja tägien yhdistämiseen.
- Tiedosto **ratings.csv** sisältää käyttäjien antamat arvostelut kirjoille.

Tarkemmat tiedostojen ja niihin liittyvien sarakkeiden kuvaukset löytyvät tehtävänannon lopusta.

Näyttökokeen tehtäväpohjassa on mukana CSV-tiedostojen lukemiseen käytettävä OpenCSV-kirjasto (http://opencsv.sourceforge.net/). Ylimääräisiä kirjastoja ei saa lisätä projektiin. CSV-muotoisen tiedoston lukeminen onnistuu OpenCSV-kirjaston avulla seuraavasti:

```
Scanner lukija = new Scanner(System.in);
System.out.println("Mikä tiedosto luetaan?");
String tiedosto = lukija.nextLine();

List<String[]> rivit = new CSVReader(new FileReader(tiedosto)).readAll();
System.out.println("Tiedostossa oli " + rivit.size() + " riviä.");
```

Yllä lista `rivit` sisältää taulukko-olioita, joista kukin vastaa tiedoston yhtä riviä.

Tehtäväpohjassa on valmiina ohjelma, joka kysyy tiedostojen polkua ja tulostaa annetussa polussa olevan tiedoston **books.csv** rivien määrän. Näyttökokeen koneilla polku tiedostoihin on `/home/cs-mooc/Desktop/Data/books/` tai `~/Desktop/Data/books/`

Analyysien suorittaminen tulee aloittaa kun tehtäväpohjassa olevan luokan main-metodi suoritetaan. Toteuttamasi sovelluksen tulee kysyä suorituksen alussa polkua analysoitaviin tiedostoihin. Analyysi tulee tehdä käyttäjän antamassa polussa olevista tiedostoista.

Tehtävien analyysien kuvaukset alkavat seuraavalta sivulta. Toteuta analyysit yksi kerrallaan. Puoliksi toteutettuja toiminnallisuuksia ei arvioida. Mikäli jäät jumiin, voit siirtyä seuraavaan analyysiin. Tehtäviä ei ole ennalta pisteytetty. Voit olettaa, että vaikeammista tehtävistä saa enemmän pisteitä. Mikäli kaksi osallistujaa saa samat analyysit valmiiksi, arvioidaan toteutukset selkeyden ja ylläpidettävyyden perusteella.

Huomaa, että tiedostojen sisällöt voivat muuttua, mutta tiedostojen nimet, sarakkeiden nimet, ja sarakkeiden indeksit pysyvät samoina. Toteuta analyysit aina siten, että ohjelma toimii, vaikka tiedostojen sisältö muuttuisi.

Ensimmäisissä analyyseissä käytetään tiedostoa **books.csv**.

1. Tulosta kaikki kirjat, jotka ovat saaneet yli miljoona kertaa arvosanan 5. Arvosanojen 5 lukumäärä löytyy sarakkeesta `ratings_5`. Kirjoista tulee tulostaa sarakkeen `title` arvo.

   Tulosta ensimmäisen kohdan analyysin tulosten yläpuolelle rivi, joka sisältää merkkijonon "`Analyysi 1:`".

2. Tulosta kaikki kirjat, jotka on arvioitu vähintään satatuhatta kertaa, ja joiden arvioista yli 40% on arvosana 3. Arvosanat saat selville sarakkeista `ratings_1, ratings_2, ...`. Kirjoista tulee tulostaa sarakkeen `title` arvo.

   Tulosta toisen kohdan analyysin tulosten yläpuolelle rivi, joka sisältää merkkijonon "`Analyysi 2:`".

3. Joskus useammalla kirjalla on sama ISBN-numero. Tulosta kirjat, joiden ISBN-numero toistuu listassa vähintään kaksi kertaa (mikäli ISBN-numero puuttuu, jätä rivi huomiotta). Tulosta kirjoista sarakkeen `title` arvo. Käytä ISBN-numerona sarakkeen `isbn13` arvoa.

   Tulosta kolmannen kohdan analyysin tulosten yläpuolelle rivi, joka sisältää merkkijonon "`Analyysi 3:`".

4. Tulosta vanhin kirja. Mikäli riviltä puuttuu julkaisuvuosi (sarake on tyhjä), jätä kyseinen rivi huomiotta. Voit olettaa, että vanhin kirja on yksikäsitteinen (eli vanhimmalla julkaisuvuodella ei ole useampia kirjoja). Tulosta kirjasta sen nimi (sarakkeen `title` arvo) ja julkaisuvuosi. Käytä julkaisuvuotena saraketta `original_publication_year`.

   Tulosta neljännen kohdan analyysin tulosten yläpuolelle rivi, joka sisältää merkkijonon "`Analyysi 4:`".

5. Kullakin kirjalla voi olla yksi tai useampi kirjoittaja. Tulosta kirjoittaja, joka on ollut kirjoittajana useimmassa kirjassa. Käytä kirjoittajan tunnistamiseen sarakkeen `authors` arvoa. Mikäli kirjalla on useampi kirjoittaja, sarakkeen `authors` sisältämät nimet on eroteltu toisistaan pilkulla.

   Voit olettaa, että eniten kirjoja kirjoittanut kirjoittaja on yksikäsitteinen. Tulostuksessa tulee olla sekä kirjoittajan nimi että kirjoittajan yksin ja yhdessä muiden kanssa kirjoittamien kirjojen lukumäärä.

   Tulosta viidennen kohdan analyysin tulosten yläpuolelle rivi, joka sisältää merkkijonon "`Analyysi 5:`".

Seuraavissa analyyseissä käytetään tiedoston **books.csv** lisäksi tiedostoa **book_tags.csv**. Tiedoston **book_tags.csv** sarakkeen `goodreads_book_id` arvo vastaa tiedoston **books.csv** sarakkeen `goodreads_book_id` arvoa. Esimerkiksi, mikäli tiedostossa **book_tags.csv** olevan sarakkeen `goodreads_book_id` arvo on 31337, vastaa se **books.csv** tiedostossa olevaa riviä, jossa sarakkeen `goodreads_book_id` arvo on 31337 -- rivillä on kirjan “Blackwood Farm (The Vampire Chronicles, #9)” tiedot.

6. Laske ja tulosta tiedostosta **book_tags.csv** niiden rivien lukumäärä, joissa sarakkeen `goodreads_book_id` arvo on 31337.

   Tulosta kuudennen kohdan analyysin tulosten yläpuolelle rivi, joka sisältää merkkijonon “`Analyysi 6:`”.

7. Laske ja tulosta tiedostosta **book_tags.csv** niiden rivien sarakkeen count summa, joissa sarakkeen `tag_id` arvo on 30574.

   Tulosta seitsemännen kohdan analyysin tulosten yläpuolelle rivi, joka sisältää merkkijonon “`Analyysi 7:`”.

8. Laske ja tulosta vuonna 1945 julkaistuun kirjaan “1984” liittyvien tägien lukumäärä (eli kyseiseen kirjaan liittyvien rivien lukumäärä tiedostosta **book_tags.csv**). Analyysin tulee toimia, vaikka sarakkeen `goodreads_book_id` arvo muuttuisi. Etsi sarakkeen `goodreads_book_id` arvo kirjan tietojen perusteella tiedostosta **books.csv** ja tarkastele tämän jälkeen tiedostoa **book_tags.csv** löydetyn arvon perusteella. Älä siis kirjoita ohjelmaa, joka tarkastelee vain kirjaa, jonka `goodreads_book_id`:n arvo on 13 -- ohjelman tulee toimia vaikka sarakkeen `goodreads_book_id` arvo olisi jotain muuta.

   Tulosta kahdeksannen kohdan analyysin tulosten yläpuolelle rivi, joka sisältää merkkijonon “`Analyysi 8:`”.

9. Selvitä ja tulosta niiden kirjojen nimet (sarake `title` tiedostosta **books.csv**), joihin liittyy tiedostossa **book_tags.csv** sarakkeen `tag_id` arvo 32837.

   Tulosta yhdeksännen kohdan analyysin tulosten yläpuolelle rivi, joka sisältää merkkijonon “`Analyysi 9:`”.

Seuraavissa analyyseissä käytetään tiedostojen **books.csv** ja **book_tags.csv** lisäksi tiedostoa **tags.csv**. Tiedoston **tags.csv** sarakkeen `tag_id` arvo vastaa tiedoston **book_tags.csv** olevan sarakkeen `tag_id` arvoa.

10. Tulosta tiedostosta **tags.csv** ne sarakkeen `tag_id` arvot, joihin liittyvän sarakkeen `tag_name` arvo alkaa merkkijonolla “history-of”.

    Tulosta kymmenennen kohdan analyysin tulosten yläpuolelle rivi, joka sisältää merkkijonon “`Analyysi 10:`”.

11. Laske ja tulosta niiden kirjojen lukumäärä, joihin on liitetty tägi “classic-poetry”. Selvitä siis ensin tiedostosta **tags.csv** sarakkeen `tag_id` arvo riville, jonka sarakkeen `tag_name` arvo on “classic-poetry”. Laske tämän jälkeen tiedostosta **book_tags.csv** niiden rivien lukumäärä, joissa esiintyy edellä selvittämäsi `tag_id`. Kuten aiemmin, huomaa että tiedostojen sisältö voi muuttua -- toteuta analyysi siis siten, että analyysi toimii vaikka tiedostojen sisältö muuttuisi.

    Tulosta yhdennentoista kohdan analyysin tulosten yläpuolelle rivi, joka sisältää merkkijonon “`Analyysi 11:`”.

12. Tulosta niiden kirjojen nimet (tiedoston **books.csv** sarake title), joihin on liitetty tägi “computer-programming”. Selvitä siis ensin tiedostosta **tags.csv** sarakkeen `tag_id` arvo riville, jonka sarakkeen `tag_name` arvo on “computer-programming”. Selvitä tämän jälkeen tiedostosta **book_tags.csv** rivit, joiden `tag_id` -sarakkeen arvo on aiemmin selvittämäsi arvo. Selvitä näiden rivien sarakkeiden `goodreads_book_id` arvot, ja selvitä niiden perusteella tiedostosta **books.csv** kirjojen otsikot (sarake `title`). Kuten aiemmin, huomaa että tiedostojen sisältö voi muuttua -- toteuta analyysi siis siten, että analyysi toimii vaikka tiedostojen sisältö muuttuisi.

    Tulosta kahdennentoista kohdan analyysin tulosten yläpuolelle rivi, joka sisältää merkkijonon “`Analyysi 12:`”.

13. Selvitä ja tulosta kolme yleisintä kirjaan “The Name of the Wind (The Kingkiller Chronicle, #1)” liittyvää tägiä. Tägien yleisyyden saat selville tiedoston **book_tags.csv** sarakkeesta count.

    Tulosta kolmannentoista kohdan analyysin tulosten yläpuolelle rivi, joka sisältää merkkijonon “`Analyysi 13:`”.

14. Selvitä ja tulosta viisi yleisintä kirjoittajan “Suzanne Collins” kirjoihin liittyvää tägiä. Huomioi tägien yleisyyttä selvittäessä tiedoston **book_tags.csv** sarakkeen count arvo.

    Tulosta neljännentoista kohdan analyysin tulosten yläpuolelle rivi, joka sisältää merkkijonon “`Analyysi 14:`”.

15. Selvitä ja tulosta mihin välillä 1990-2000 julkaistuun kirjaan (tai kirjoihin, mikäli tieto ei ole yksikäsitteinen) liittyy eniten “to-read”-tägejä. Huomioi tägien yleisyyttä selvittäessä tiedoston **book_tags.csv** sarakkeen count arvo.

    Tulosta viidennentoista kohdan analyysin tulosten yläpuolelle rivi, joka sisältää merkkijonon “`Analyysi 15:`”.

Seuraavissa analyyseissä käytetään tiedostoja **books.csv** ja **ratings.csv**.

16. Tulosta käyttäjä (tai käyttäjät), jotka ovat antaneet eniten arvioita. Tulosta käyttäjästä käyttäjän tunnus (sarakkeen `user_id` arvo). Laske kukin **ratings.csv** -tiedoston rivi yhdeksi arvioksi.

    Tulosta kuudennentoista kohdan analyysin tulosten yläpuolelle rivi, joka sisältää merkkijonon “`Analyysi 16:`”.

17. Selvitä ja tulosta käyttäjä tai käyttäjät, jotka ovat arvioineet eniten samoja kirjoja kuin käyttäjä, jonka tunnus (`user_id`-sarakkeen arvo) on `6630`.

    Tulosta seitsemännentoista kohdan analyysin tulosten yläpuolelle rivi, joka sisältää merkkijonon “`Analyysi 17:`”.

18. Selvitä ja tulosta käyttäjä, jolla on eniten samanarvoisia arvioita samoista kirjoista, kuin käyttäjällä, jonka tunnus on `6630`. Mikäli samaa mieltä olleita käyttäjiä on useita, tulosta ne kukin omalle rivilleen.

    Tulosta kahdeksannentoista kohdan analyysin tulosten yläpuolelle rivi, joka sisältää merkkijonon “`Analyysi 18:`”.

19. Listaa edellisen osan perusteella käyttäjälle `6630` kirjoja (sarakkeen `title` arvo), joita käyttäjä `6630` ei ole arvioinut, mutta joita edellisessä osassa tunnistetut käyttäjät ovat pitäneet arvion 5 arvoisena. Listaa korkeintaan 5 kirjaa.

    Tulosta yhdeksännentoista kohdan analyysin tulosten yläpuolelle rivi, joka sisältää merkkijonon “`Analyysi 19:`”.

20. Selvitä kirja, jonka arvioissa on suurin keskihajonta. Tarkastele vain kirjoja, jotka ovat saaneet vähintään miljoona arviota tiedoston **books.csv** sarakkeiden `ratings_1`, `ratings_2`, … mukaan. Käytä arvioiden keskihajonnan laskemiseen tiedostossa **ratings.csv** olevia arvioita. Tulosta eniten mielipiteitä jakaneesta kirjasta (eli kirjasta, jonka arvioiden keskihajonta on suurin) kirjan nimi (sarake `title`) sekä siihen liittyvien arvioiden keskihajonta.

    Tulosta kahdennenkymmenennen kohdan analyysin tulosten yläpuolelle rivi, joka sisältää merkkijonon "`Analyysi 20:`".

Seuraavissa analyyseissä käytetään tiedostoja **books.csv**, **book_tags.csv**, **tags.csv** ja **ratings.csv**.

21. Selvitä viisi yleisintä tägiä (**book_tags.csv** -tiedoston `count` -sarakkeen perusteella tarkasteltuna) kirjalle, jonka arvioissa on pienin keskihajonta. Ota mukaan vain ne kirjat, jotka ovat saaneet vähintään seitsemänsataaviisikymmentätuhatta arviota tiedoston **books.csv** sarakkeiden ratings_1, ratings_2, … mukaan. Käytä arvioiden keskihajonnan laskemiseen tiedostossa **ratings.csv** olevia arvioita. Tulosta lopulta tägeihin liittyvä tiedostosta **tags.csv** löytyvän sarakkeen `tag_name` arvo.

    Tulosta kahdennenkymmenennenensimmäisen kohdan analyysin tulosten yläpuolelle rivi, joka sisältää merkkijonon "`Analyysi 21:`".

22. Selvitä viisi eniten mielipiteitä jakavaa kirjaa, joihin liittyy tägi "fantasy". Saat mielipiteitä jakavat kirjat selville laskemalla niihin liitttyvien arvioiden keskihajonnan (mitä suurempi keskihajonta, sitä suurempi mielipiteiden jakautuminen). Ota mukaan vain ne kirjat, joihin tägi "fantasy" on liitetty yli kymmenentuhatta kertaa (tiedoston **book_tags.csv** sarake `count`), ja jotka on arvioitu vähintään viisituhatta kertaa (tiedoston **books.csv** sarakkeiden ratings_1, ratings_2, … mukaan). Käytä arvioiden keskihajonnan laskemiseen tiedostossa **ratings.csv** olevia arvioita. Tulosta lopulta kirjojen nimet (`title`) ja niiden arvioiden keskihajonnat.

    Tulosta kahdennenkymmenennentoisen kohdan analyysin tulosten yläpuolelle rivi, joka sisältää merkkijonon "`Analyysi 22:`".

Tiedostojen kuvaukset

Jokaisen tiedoston ensimmäisellä rivillä on sarakkeiden nimet. Arvot alkavat toiselta riviltä.

Tiedosto **books.csv** sisältää kirjojen tietoja. Näyttökokeen kannalta oleelliset sarakkeet, indeksit ja sarakkeiden kuvaukset ovat seuraavat:

sarake	**indeksi**	**kuvaus**
book_id	0	Kirjan yksilöivä tunniste. Käytetään yhdessä tiedoston ratings.csv kanssa. Vastaa tiedoston ratings.csv saraketta book_id.
goodreads_book_id	1	Kirjan yksilöivä tunniste. Käytetään yhdessä tiedoston book_tags.csv kanssa. Vastaa tiedoston book_tags.csv saraketta goodreads_book_id.
isbn13	6	Kirjan ISBN-numero.
authors	7	Kirjan kirjoittaja tai kirjoittajat. Mikäli sisältää useamman kirjoittajan, kirjoittajien nimet ovat eroteltu toisistaan pilkuilla.
original_publication_year	8	Kirjan julkaisuvuosi.
title	10	Kirjan otsikko.
ratings_1	16	Kirjan saamien arvosanojen 1 lukumäärä.
ratings_2	17	Kirjan saamien arvosanojen 2 lukumäärä.
ratings_3	18	Kirjan saamien arvosanojen 3 lukumäärä.
ratings_4	19	Kirjan saamien arvosanojen 4 lukumäärä.
ratings_5	20	Kirjan saamien arvosanojen 5 lukumäärä.

Tiedosto **tags.csv** sisältää käyttäjien kirjoittamia tägejä, joita voidaan liittää kirjoihin. Näyttökokeen kannalta oleelliset sarakkeet, indeksit ja sarakkeiden kuvaukset ovat seuraavat:

sarake	**indeksi**	**kuvaus**
tag_id	0	Tägin yksilöivä tunniste. Käytetään yhdessä tiedoston book_tags.csv kanssa. Vastaa tiedoston book_tags.csv saraketta tag_id.
tag_name	1	Tägiä kuvaava konkreettinen merkkijono.

Tiedostoa **book_tags.csv** käytetään kirjojen ja tägien yhdistämiseen. Näyttökokeen kannalta oleelliset sarakkeet, indeksit ja sarakkeiden kuvaukset ovat seuraavat:

sarake	**indeksi**	**kuvaus**
goodreads_book_id	0	Tägin tiettyyn kirjaan yhdistävä tunniste. Käytetään yhdessä tiedoston books.csv kanssa. Vastaa tiedoston books.csv saraketta goodreads_book_id.
tag_id	1	Tägin yksilöivä tunniste. Käytetään yhdessä tiedoston tags.csv kanssa. Vastaa tiedoston tags.csv saraketta tag_id.
count	2	Kertoo kuinka monta kertaa kyseinen tägi on lisätty kyseiselle kirjalle.

Esimerkki tiedoston **book_tags.csv** lukemisesta: mikäli sarakkeen goodreads_book_id arvo on 1, löytyy sitä vastaava kirja tiedostosta **books.csv** riviltä, jonka sarakkeen goodreads_book_id arvo on 1. Kirja on “Harry Potter and the Half-Blood Prince (Harry Potter, #6)”. Vastaavasti, mikäli sarakkeen tag_id arvo on 30574, löytyy sitä vastaava tägin nimi tiedostosta **tags.csv** riviltä, jonka sarakkeen tag_id arvo on 30574. Tägin nimi on “to-read”.

Edellistä tietoa soveltaen tiedostossa **book_tags.csv** oleva rivi, jonka sarakkeen goodreads_book_id arvo on 1, sarakkeen tag_id arvo on 30574, ja sarakkeen count arvo on 167697 kertoo seuraavaa: Kirjalle “Harry Potter and the Half-Blood Prince (Harry Potter, #6)” on lisätty tägi “to-read” yhteensä 167697 kertaa.

Tiedosto **ratings.csv** sisältää käyttäjien antamat arvostelut kirjoille. Näyttökokeen kannalta oleelliset sarakkeet, indeksit ja sarakkeiden kuvaukset ovat seuraavat:

sarake	**indeksi**	**kuvaus**
user_id	0	Tietyn käyttäjän yksilöivä tunniste.
book_id	1	Tietyn kirjan yksilöivä tunniste. Käytetään yhdessä tiedoston books.csv kanssa. Vastaa tiedoston books.csv saraketta book_id.
rating	2	Kertoo minkä arvion kyseinen käyttäjä on antanut kyseiselle kirjalle. Arvio on numeerinen arvo välillä [1, 5]. Mahdolliset muut arviot tulee jättää huomiotta.

Esimerkiksi tiedostossa **ratings.csv** oleva rivi, missä sarakkeen user_id arvo on 32, sarakkeen book_id arvo on 2357, ja sarakkeen rating arvo on 5 tulee tulkita seuraavasti. Käyttäjä 32 on antanut kirjalle “The Botany of Desire: A Plant's-Eye View of the World” arvosanan 5.