Verkefni 3c (uppfært)

Hópur 25
Steinn E. Sigurðarson
Björgvin Vilbergsson
Gísli Freyr Brynjarsson

Rannsóknarvinna (gagnasafn, reiknirit)

Gagnasafnið

Notum MovieLens 100k gagnasafnið. Gögnunum var safnað gegnum MovieLens síðuna á 7 mánaða tímabili á árunum 1997-1998.

- Það er gagnasafn með 100.000 einkunum (frá 1 til 5) á 1682 kvikmyndum frá 943 notendum.
- Hver notandi hefur gefið a.m.k. 20 kvikmyndum einkunn.
- Inniheldur einnig upplýsingar um notendur (aldur, kyn, starf, zip(usa)) ← Reiknum ekki með að nota þessar upplýsingar, en hver veit.

Pearson fylgnistuðullinn (Reiknirit)

Aðferðin sem við notum hér leitast við að reikna hve líkir aðrir notendur (i) eru notanda a (sá sem er að leita eftir meðmælum) og skilar okkur fylgnistuðlum milli þeirra og a. Þessir fylgnistuðular er svo notaðir til þess að vigta þær einkunnir sem notendur hafa gefið og fá þannig notendur sem eru líkastir a mesta vægið. Vegnar einkunnir fyrir kvikmyndir sem notandi a hefur ekki séð eru lagðar saman og skalaðar til með summu fylgnistuðlanna (k).

Úr gagnasafninu fáum við fylkið $v_{i,j}$ sem stendur fyrir einkunn á kvikmynd j frá notanda i. I_i er fjöldi þeirra kvikmynda sem notandi i hefur gefið einkunn, má þá finna meðaltal v_i með:

$$\bar{v}_i = \frac{1}{|I_i|} \sum_{j \in I_i} v_{i,j}$$

(eins farið að með v_a)

Fylgni milli notenda a og i er fundinn með:

$$w(a,i) = \frac{\sum_j (v_{a,j} - \bar{v}_a)(v_{i,j} - \bar{v}_i)}{\sqrt{\sum_j (v_{a,j} - \bar{v}_a)^2 \sum_j (v_{i,j} - \bar{v}_i)^2}}$$

þar sem summan yfir j er yfir þær myndir sem bæði notandi a og i hafa gefið einkunn. Og að sjálfsögðu $v_{a,j}$ einkunn á kvikmynd j sem a hefur gefið.
Til að finna svo vigt einkunna notum við:

$$p_{a,j} = \bar{v}_a + k \sum_{i=1}^{n} w(a,i)(v_{i,j} - \bar{v}_i)$$

þar sem n er fjöldi notenda í gagnagrunninum sem hafa e-h vigt og k skölun með summu fylgnistuðlanna, skilgreind sem:

$$k = \frac{1}{\sum_{i=1}^{n} |w(a,i)|}$$

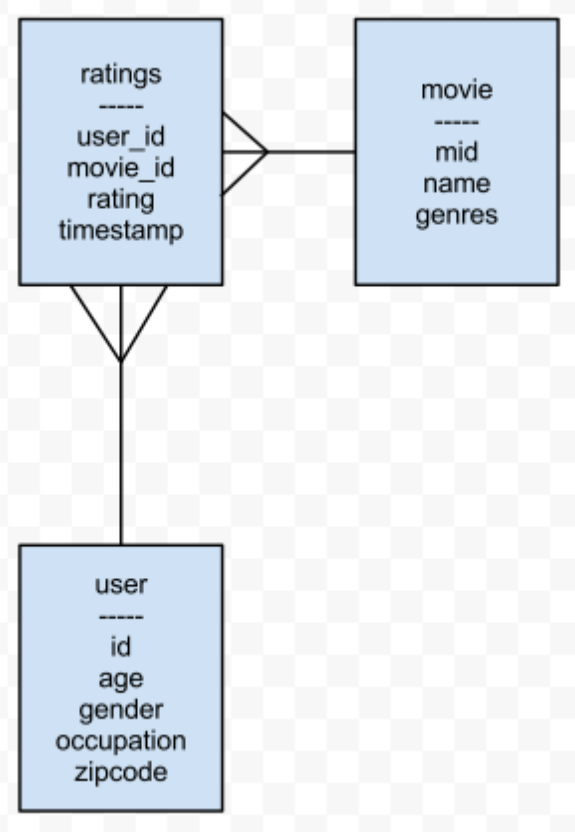
Þar með hefur verið leiðrétt fyrir kvikmyndum sem margir hafa gefið einkunn.

Þetta reiknirit varð fyrir valinu þar sem það ætti að gefa frekar góðar niðurstöður m.v. flækjustig útfærslu.

Útfærsla

Fyrsta útfærsla kerfisins mun aðeins útfæra reikniritið og lesa inn gagnasafnið af staðalinntaki, og vinna með reikniritið í minni forrits. Hægt verður að biðja um meðmæli m.v. ID notenda, eða kvikmynda í kerfinu. Kjarni forritsins verður innlestur og uppstilling gagnasafnsins í minni, en gagnasafnið samanstendur af notendum og kvikmyndum. Þegar gagnasafninu hefur verið uppstillt mun það samanstanda af User og Movie hlutum, ásamt hjálparföllum til að ná í eina tiltekna kvikmynd, eða einn tiltekinn notanda útfrá ID, og keyra föllin Rate (frá notanda á kvikmynd) og getSimilar, sem skilar í báðum tilfellum safni af kvikmyndum útfrá annaðhvort myndi eða notanda (háð samhengi).

Gagnagrind



Lokauppfærsla hönnunarlýsingar

Það sem breyst hefur er að nú eru gögnin geymd i SQL gagnagrunni. Notast var við sqlite3 og útbúið var lítið forrit sem les gögn inn í gagnagrunninn úr data skránum. Datasettið okkar er enn ml-100. Data hlutinn okkar var refactoraður til að lesa úr SQL þegar forritið initilaizar sig og þegar notandi notar command line forritið til að bæta við notanda, bíómynd eða einkunn, þá skrifar data hlutinn í SQL grunninn. Auk þessara breytinga var cmd-line forritið betrumbætt.

Klasarit:

Við vorum ekki vissir um hvernig við ættum að tengja þetta rétt, þ.e. hvernig örvarnar eiga að vísa, en við settum inn gagnamótin eins best og við gátum með mikilvægustu tilviksbreytum og stefum.

Einnig tókum við fram breytutegundir eins og þær koma okkur fyrir sjónir í python.