MovieLens Gagnagrunnur

Haukur Hlíðberg, Jón Kr. Helgason, Kristófer Reykjalín,
Róbert B. Ólafsson, Vladimir Omelianov,

T-316-GAVI Gagnavinnsla

Kennari: Eyjólfur Ingi Ásgeirsson

4. Desember 2015

1 Inngangur

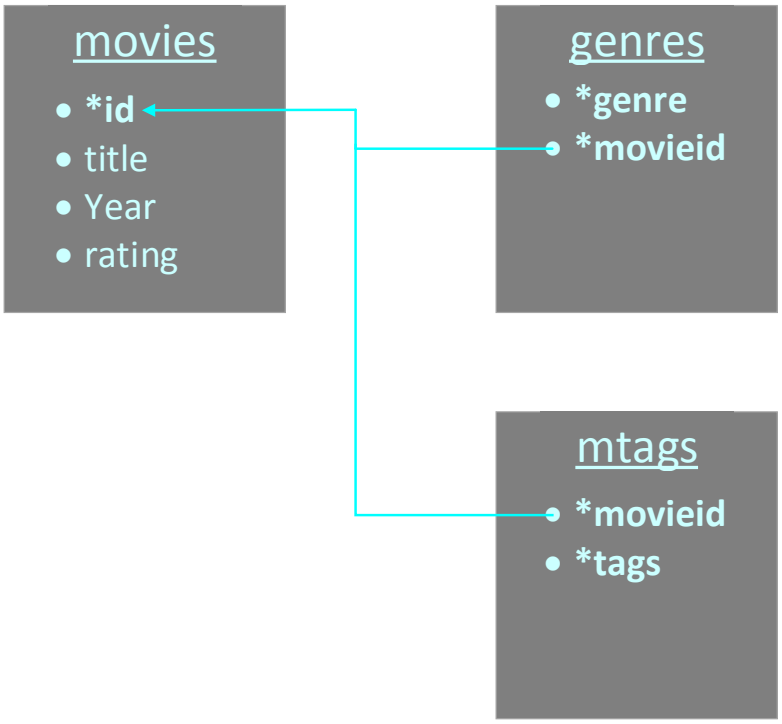
Markmið verkefnisins er að nota gögn frá Grouplens [1] og ná þar í gagnasett sem inniheldur 10 milljón einkunnagjafir og 100 þúsund 'tags' á 10 þúsund kvikmyndum frá 72 þúsund notendum. Nemendur áttu svo að útbúa SQL gagnagrunn byggðan á þessum gögnum og skrifa python forrit sem talar við gagnagrunninn.

Forritið á að virka á þann hátt að þegar notandi slær inn heiti á einni eða fleiri kvikmyndum eiga að birtast tillögur að svipuðum myndum.

2 Framkvæmd

Byrjað var að nýta minni útgáfu af Grouplens gagnasettinu sem inniheldur "aðeins" 100 þúsund einkunnargjafir á 1700 kvikmyndum. Þetta gagnasett var notað til þess að hægt væri að þróa forrit án þess að keyrslutíminn yrði of langur.

- Gögn lesinn inn og forunnin með python
  - Ár tekið frá myndum og sett í sér dálk
  - Gögn úr ratings notuð til að reikna meðaleinkunn hverrar myndar. Meðaleinkunn smellt aftan á upplýsingar um myndir
  - Genres við hverja mynd aðskilin og sett upp í töflu þar sem key er (genre, mynd), sbr. Mynd 1
  - Tög (e. tag) við hverja mynd aðskilin og sett upp í töflu þar sem key er (mynd, tag), sbr. Mynd 1
- Unnin gögn sett upp og hlaðið inn í SQL gagnagrunn á því formi sem sést á Mynd 1
- GUI búið til í Qt4 Designer [2]
- Queries búnar til fyrir gagnagrunn
  - Leit sem finnur genre myndar út frá nafni og ári
  - Leit sem finnur tög myndar út frá nafni og ári
  - Leit að svipuðum myndum út frá genres og tags. Meðal-rating "input" kvikmynda notað til að sía niðurstöður
- Gluggi sem gerir notanda kleift að slá inn upplýsingar um gagnagrunn búinn til



Mynd 1: Database schema

3 Aðferð

3.1 Hönnun

Það var ákveðið að hafa GUI (Graphical user interface) og markmiðið var að hafa það sem einfaldast og notendavænast. Því hefur aðalgluggi forritisins (sbr. Mynd 2) tvo dálka, eina leitarstiku og fimm hnappa.

Notast var við Qt4 Designer [2] til að hanna GUI, það var valið með það í huga að það er auðvelt í notkun og við yrðum mun fljótari að hanna GUI á þann hátt.

3.2 Virkni leita

Á tveimur mismunandi stigum í forritinu er haft samband við gagnagrunninn. Annars vegar þegar leitað er að myndum sem notandinn slær inn, og hins vegar þegar leitað er að tillögum að myndum sem notandanum gæti líkað við.
Finna myndir sem notandi slær inn, sbr. Query 1 í Appendix A:

- Finna allar myndir þar sem hluti titilsins eða ársins inniheldur leitarstrenginn

Finna tillögur að myndum sem notanda gæti líkað við:

- Finna genres, sbr. Query 2 í Appendix A

  - Finna genres kvikmynda/r sem notanda líkar við

- Finna tög, sbr. Query 3 í Appendix A
  - Finna tög sem eiga við myndir sem notanda líkar við
- Finna tillögur að myndum, sbr. Query 4 í Appendix A
  - Finna hvert og eitt tilvik af mynd sem tilheyrir einu af þeim genres eða hefur sama tag og einhver af myndunum sem notandanum líkar við, auk þess sem rating þarf að vera hærra en meðal-rating þeirra mynda sem notandanum líkar við, mínus hálfur.
  - Raða myndum eftir því hversu oft þær koma fyrir
  - Þær 10 myndir sem koma oftast fyrir er síðan raðað eftir rating

ATH: Töflur í gagnagrunni verða að heita þeim nöfnum sem sjást á Mynd 1 annars virka leitarquery-in ekki.

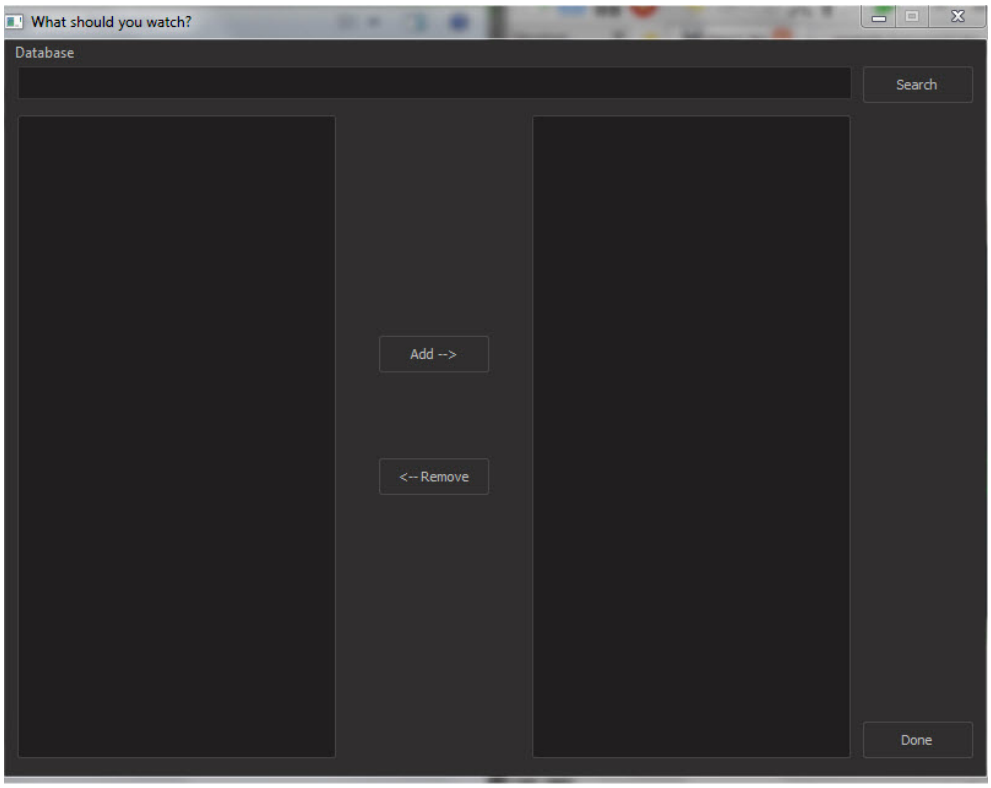
3.3 Virkni forrits

Þegar gui_test.py skráin er keyrð opnast aðalgluggi forritisins (Mynd 2). Til að tengjast gagnagrunni er hægt að ýta á Ctrl + E (slaufa + E á mac) eða smellt á "Database" og valið "Edit info...". Þá opnast sprettigluggi þar sem upplýsingar um gagnagrunn eru settar inn, sjá Mynd 3. Eftir að ýtt er á "Ok" í sprettiglugganum birtist annar sprettigluggi sem lætur notandann vita hvort tekist hafi að tengjast gagnagrunninum.

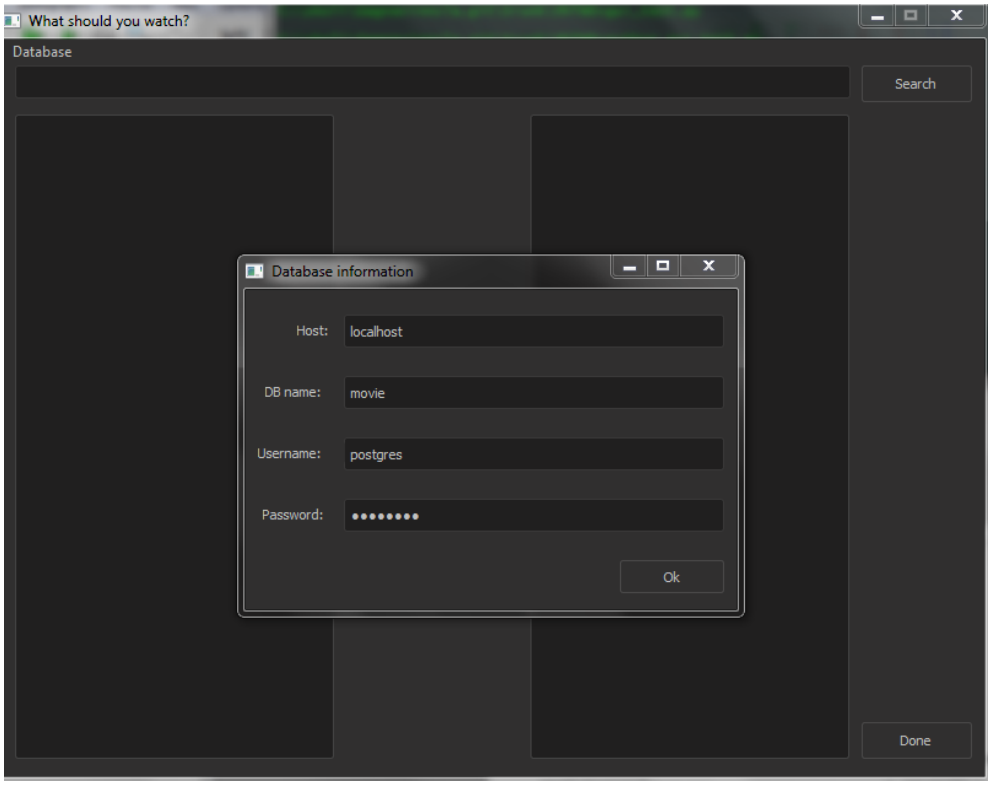
Þegar notandinn hefur tengst gagnagrunni getur hann leitað að mynd með því að skrifa nafn, eða hluta nafns, í leitarstikuna og ýtir svo á search (eða Return). Einnig er hægt að leita að útgáfuári, en þá er slegið ár inní leitarstikuna og fást þá allar myndir gefnar út á því ári.

Niðurstöður þeirrar leitarinnar birtast í vinstri dálknum. Ef notandanum líst á einhverja af myndunum sem koma upp við leitina getur hann bætt henni við á listann yfir myndir sem notandanum líkar við með því að ýta á á "add" hnappinn og þá færist valda myndin úr vinstri dálkinn yfir í hægri dálkinn, sjá Mynd 4. Ef notandinn vill taka myndir af listanum getur hann ýtt á "Remove" hnappinn og þá færist myndin úr hægri dálknum yfir í vinstri dálkinn aftur.

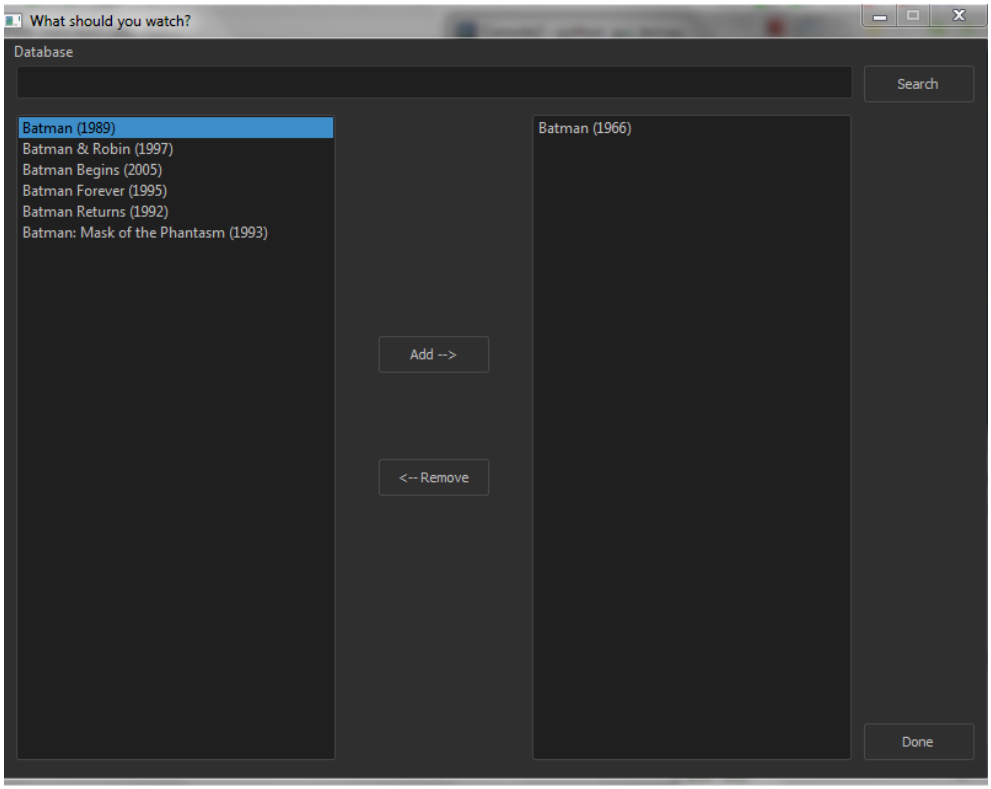
Þegar ýtt er á "done" hnappinn sprettur upp annar sprettigluggi sem kemur með tillögur að myndum sem notandanum gæti líkað við, sjá Mynd 5. Eftir það er hægt að prófa aðrar myndir með því að ýta á "yes" hnappinn, annars ýtir notandinn á "no" hnappinn og þá er forritinu lokað.



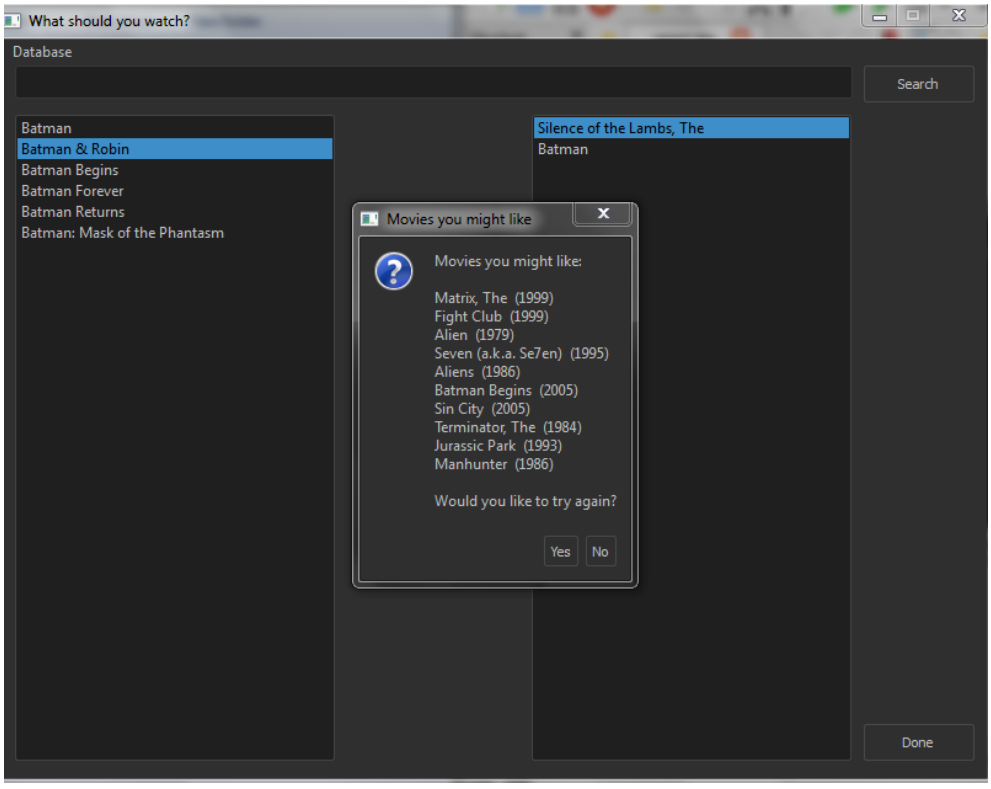
Mynd 2: Aðalgluggi forritsins



Mynd 3: Sprettiglugginn með upplýsingum útfylltum



Mynd 4: Myndaleit skilar niðurstöðum í vinstri dálk



Mynd 5: Svipaðar myndir og þær sem settar voru í hægri dálkinn birtast í sprettiglugga

4 Niðurstöður

Að mati hópsins virkar forritið merkilega vel. Forritið skilar stöðugt svipuðum myndum og þeim sem notanda líkar við. Hins vegar á forritið erfiðara með að finna líkar myndir ef myndir sem notanda líkar við eru mjög ólíkar. Þetta verður að teljast eðlilegt þ.s. forritið þarf þá að skoða mun fleiri gögn. Þrátt fyrir það eru tillögurnar samt sem áður góðar (hafa gott rating), en eru ekki endilega mjög líkar þeim myndum sem notanda líkar við.

Annað sem má nefna um niðurstöðurnar er að tögin eiga það til að gera niðurstöður frekar bjagaðar þegar um er að ræða vinsælar myndaseríur, þá sér í lagi teiknimyndir frá Pixar eða Disney. Þær valda því að meirihluti tillaga eru í þeim dúr.

Ef notandi setur inn 5 eða færri myndir sem honum líkar við tekur almennt minna en 3 sekúndur að finna tillögur að myndum. Eftir því sem myndirnar aukast tekur meiri tíma að finna tillögur en almennt tekur það ekki meira en 10 sekúndur.

Prófað var að setja inn allar myndir frá árinu 2000 og tók 1 mínútu og 22 sekúndur að finna tillögur að myndum. Líklega væri hægt að besta SQL query-in og láta python vinna minni vinnu við tilbúning þeirra, en ákveðið var að hafa kóðann eins og hann er vegna tíma og læsileika á það hvað er að gerast.
Nokkur dæmi um notkun forritisins:

- Myndir sem notanda líkar við:
  - *Batman*, *Batman Begins*, *Batman Forever*
- Tillögur að myndum sem notanda gæti líkað við:
  - *Star Wars: Episode V - The Empire Strikes Back* (1980), *Lord of the Rings: The Fellowship of the Ring, The* (2001), *Pulp Fiction* (1994), *Snatch* (2000), *Incredibles, The* (2004), *X2: X-Men United* (2003), *Spider-Man* (2002), *Spider-Man 2* (2004), *X-Men* (2000), *Batman Returns* (1992)
- Myndir sem notanda líkar við:
  - *Iron Man*, *Princess Mononoke*, *A Bug's Life*, *The Prestige*
- Tillögur að myndum sem notanda gæti líkað við:
  - *Spirited Away (Sen to Chihiro no kamikakushi)* (2001), *Lord of the Rings: The Fellowship of the Ring, The* (2001), *Lord of the Rings: The Return of the King, The* (2003), *Blade Runner* (1982), *Nausicaä of the Valley of the Winds (Kaze no tani no Naushika)* (1984), *Howl's Moving Castle (Hauru no ugoku shiro)* (2004), *Incredibles, The* (2004), *Toy Story* (1995), *Kiki's Delivery Service* (Majo no takkyûbin) (1989), *Finding Nemo* (2003)

- Myndir sem notanda líkar við:
  - *Lucky Number Slevin*, *The Shawshank Redemption*, *Back to the Future*, *WALL-E*, *The Sixth Sense*, *Psycho*, *Ocean's Eleven*, *Teenage Mutant Ninja Turtles*
- Tillögur að myndum sem notanda gæti líkað við:
  - *Star Wars: Episode IV - A New Hope (a.k.a. Star Wars)* (1977), *Silence of the Lambs, The* (1991), *Princess Bride, The* (1987), *Blade Runner* (1982), *Eternal Sunshine of the Spotless Mind* (2004), *Seven (a.k.a. Se7en)* (1995), *Clockwork Orange, A* (1971), *Toy Story* (1995), *Sin City* (2005), *Jurassic Park* (1993)

A SQL Queries

Query 1: Finna myndir sem notanda líkar út frá leitarstreng

```
select title, year, rating
from movies
where lower(title) like '%leitarstrengur%' or year like '%leitarstrengur
    ↪ %'
order by title;
```

Query 2: Finna genres út frá titlum og ári titla

```
select distinct lower(g.genre)
from genres g, movies m
where m.id = g.movieid
and ( lower(m.title) = 'mynd1' and m.year = 'ar1'
or lower(m.title) = 'mynd2' and m.year = 'ar2'
or ...
)
order by lower(g.genre)
```

Query 3: Finna tags út frá titlum og ári titla

```
select distinct lower(mt.tag)
from genres g, movies m, mtags mt
where m.id = g.movieid and m.id = mt.movieid
and ( lower(m.title) = 'mynd1' and m.year = 'ar1'
or lower(m.title) = 'mynd2' and m.year = 'ar2'
or ...
)
order by lower(mt.tag)
```

Query 4: Finna líkar myndir

```
select m.title, m.year, m.rating
from movies m
where (m.title, m.year) in (
select m.title, m.year
from genres g, movies m, mtags mt
where m.id = g.movieid and m.id = mt.movieid
and m.rating >= (avg_input_rating - 0.5)
and ( lower(g.genre) = 'genre1' or lower(g.genre)='genre2'... )
and ( lower(mt.tag) = 'tag1' or lower(mt.tag) = 'tag2' ... )
and ( lower(m.title) != 'mynd1' and m.year != 'ar1'
and lower(m.title) != 'mynd2' and m.year != 'ar2'
...
)
group by m.title, m.year
```

```
order by count(m.title) desc
limit 10
)
order by m.rating desc;
```

References

[1] GroupLens, *Movielens*, 2015. [Online]. Available: `http://grouplens.org/datasets/movielens/`.

[2] The Qt Company Ltd., *Qt designer manual*, 2015. [Online]. Available: `http://doc.qt.io/qt-4.8/designer-manual.html`.