Beskrivelse af det implementerede pointsystem

Udfordringens del 1: Generelle anbefalinger

Hver film gives 0–10 point og de tre med flest point vises til alle brugere. Det samlede pointtal er en sum af to komponenter, der hver giver 0-5 point:

1. Filmens rating.
2. Et antal point, der er proportionelt med antallet af gange den givne film er blevet købt, og sådan at de(n) mest købte film gives 5 point.

Udfordringens del 2: Individuelle anbefalinger

For hver bruger, der kigger på en film, anbefales tre andre film. Hver film gives 0-25 point; film, som brugeren tidligere har købt samt den film brugeren p.t. kigger på, ekskluderes; og de tre tilbageværende film med flest point vises. Det samlede pointtal er en sum af fem komponenter, der hver giver 0-5 point:

1. Filmens rating.
2. Point baseret på filmens pris. En statistisk test anvendes til at undersøge, om brugeren har en tendens til at vælge film efter deres pris. Hvis deres historik afslører, at de går efter billige film, gives de(n) billigste film 5 point og de(n) dyreste 0 point. Hvis de går efter dyre film,[1] er det omvendt. I begge tilfælde afhænger pointtildelingen til de resterende film lineært af deres pris.
3. Ligesom komponent 2, men for gamle versus nye film.
4. En komponent, der afhænger af filmens genrer. Først beregnes brugerens *genrepræferencer*, der består af er et tal mellem 0 og 1 for hver genre. Det udregnes som det vejede gennemsnit over de film, brugeren har købt, kigget på tidligere, eller kigger på nu, af indikatorfunktion for prædikatet "filmen tilhører denne genre." Vægtene er: $\frac{1}{2}$ for en film, brugeren tidligere har kigget på men ikke købt; 1 for en købt film; og 5 for den film, brugeren aktuelt kigger på.[2] Derefter beregnes pointtallet for en given film som gennemsnittet af brugerens genrepræference for hver genre, som filmen tilhører, multipliceret med en faktor, der sikrer, at mindst én film gives 5 point.

[1] Selvom brugere, der vælger dyre film *fordi* de er dyre, nok er sjældne, antager jeg, at dette er fornuftigt. Grunden er, at det må formodes, at mange brugere vælger film efter kriterier, der er positivt korrelerede med prisen.

[2] Begrundelsen for dette valg er, at filmen, der aktuelt kigges på, må antages at være en *stærk* indikator for brugerens præferencer på det specifikke tidspunkt, og at film, som brugeren har kigget på, også må antages at have *nogen* relevans for afdækning af brugerens præferencer, selv hvis de ikke blev købt, idet genretilhørsforhold nemt kan afkodes fra films titler og posters.

5. Den sidste pointkomponent afhænger af den film f_0, som brugeren aktuelt kigger på. Hver film f (bortset fra f_0) tildeles et pointtal, der er proportionelt med antallet af brugere, der har købt både f_0 og f. Mindst én film får 5 point i denne kategori.

Specialtilfælde

Ovenstående beskriver kun, hvordan point tildeles i typiske tilfælde. Implementeringen tager hensyn til edge cases, hvor det er umuligt at gøre som hidtil beskrevet, men jeg vil ikke gå i detaljer med dem alle. Der er blot ét tilfælde, der bør nævnes, da det opstår i forbindelse med det leverede datasæt. I forbindelse med komponent 5 ovenfor, kan det forekomme, at den film, brugeren aktuelt kigger på, aldrig har været købt. I så fald er det ikke rigtigt, at mindst én film gives 5 point: Alle film gives 0 point. Og dette tilfælde opstår ikke bare en enkelt gang for de leverede data men faktisk for samtlige brugere, der er logget ind. Og det medfører, at komponent 5 slet ikke har haft nogen indflydelse på det faktiske output.

Begrundelse for det valgte pointsystem

...eller rettere: mangel på samme. Ovenfor har jeg givet begrundelser for nogle af mine valg i fodnoter men ikke en mere overordnet ræsonnement for pointsystemet. Man kan med rimelighed spørge, om det er rimeligt at medtage netop de $2 + 5$ komponenter, jeg har medtaget; om det er rimeligt at give dem samme vægt; og om der overhoved bør bruges et system, der blot adderer pointtal fra forskellige komponenter. Jeg har ikke gode svar på disse spørgsmål. Det ville både kræve flere data og mere tid til at foretage statistiske undersøgelser. Og ideelt set ville en anbefalingsfunktion gøre brug af machine learning for at forbedre sig selv over tid.

Output

Sidst i dette dokument kan programmets output ses. Først kommer outputet, når programmet køres i default mode, hvilket består af den tekst, der skal ses af brugerne. Det er også muligt at køre programmet i en log mode ved at tilføje argumentet "-logmode." Outputet fra en sådan kørsel er også inkluderet.