[redacted]

Prova 1

1. A ideia principal desse tipo de filtragem é de se aproveitar da "wisdom of the crowds". Usar apenas a rating média nos dá alguns problemas:

a. Podemos ter um item 1 com 5 avaliações máximas; tendo uma média de 5. Ao mesmo tempo, podemos ter um item 2 com 20 mil avaliações cuja média é 4,89. O segundo item é mais popular e, mesmo assim, retém boa avaliação. - Ele deveria ser o recomendado.

b. Os usuários não avaliam da mesma forma. Alguns são mais gentis (avaliações mais altas), enquanto outros são mais críticos (avaliações mais baixas). Se levarmos em conta a média do item, sem levar em conta esse viés, não vamos ter uma boa representação das preferências.

2. a. $m = 10$ e $n = 1.00.000$

Item-based: a combinação de poucos usuários e muitos itens gera vetores muito esparsos. Posteriormente, temos a vantagem de computar a lista de vizinhos para cada item de forma offline.

b. $m = 1.000.000$ e $n = 10$.

User-based: como há pouca quantidade de itens os vetores de usuários não são esparsos

c. $m = 1.000.000$. e $n = 1.000.000$

Item-based: podemos aproveitar o fato de que os vetores de item são menos esparsos e mais estáveis. Essa estabilidade permite a pré-computação das similaridades p/ todos os pares de itens. Posteriormente, não é necessário manter todos os vizinhos de modo a aumentar eficiência.

[redacted]

[redacted]

3.

Vantagens:

1. Dados futuros não vazam. Ex. não saberemos que um usuário u assistiu a Star Wars Ep. V antes de ver o Ep. IV (porte que ele assistiu a trilogia na ordem cronológica)
2. Mix realístico de cold e não-cold start cases.
3. É possível verificar consumos sazonais com o uso de janelas deslizantes.

Desvantagens

1. O tempo em que o corte será feito pode afetar negativamente os resultados
2. Podemos fazer uma partição que sofra de efeitos sazonais, fato que nos forçaria a utilizar o método de janelas deslizantes para a correção.

4. Meu usuário: 116

↳ queremos: item-based / similaridade cosseno / K = 3

- Cálculo da similaridades (116 avaliou 1, 3, 5)

$s(0,1) = 0,19$	$s(2,1) = 0,43$	$s(4,1) = 0,17$
$s(0,3) = 0,38$	$s(2,3) = 0,48$	$s(4,3) = 0,16$
$s(0,5) = 0,55$	$s(2,5) = 0,19$	$s(4,5) = 0,36$

- Aplicando esses valores em $N_{ui} = N_{oi} = [1, 3, 5]$

$$\hat{r}_{00} = \frac{0,19 \cdot 4 + 0,38 \cdot 3 + 0,55 \cdot 3}{0,19 + 0,38 + 0,55} = 3,17$$

$$\hat{r}_{02} = \frac{0,43 \cdot 4 + 0,48 \cdot 3 + 0,19 \cdot 3}{0,43 + 0,48 + 0,19} = 3,38$$

$$\hat{r}_{04} = \frac{0,17 \cdot 4 + 0,16 \cdot 3 + 0,36 \cdot 3}{0,17 + 0,16 + 0,36} = 3,24$$

[redacted]

Portanto,

$\hat{r}_{00} = 3,17$ | $\hat{r}_{02} = 3,38$ | $\hat{r}_{04} = 3,24$

5. $RMSE_0$ → $n = 3$ tal que o ranking é $[2, 4, 0]$

$$\sqrt{\frac{1}{3}\left[\underset{1}{(r_{00} - \underset{3,17}{\hat{r}_{00}})^2} + \underset{2}{(r_{02} - \underset{3,38}{\hat{r}_{02}})^2} + \underset{4}{(r_{04} - \underset{3,24}{\hat{r}_{04}})^2}\right]}$$

$4,7089 \qquad 1,9044 \qquad 0,5776$

$RMSE_0 = 1,55$

$$DCG_0 = \left(\frac{2^{r_{02}} - 1}{\log_2(2)} + \frac{2^{r_{04}} - 1}{\log_2 3} + \frac{2^{r_{00}} - 1}{\log_2 4}\right)$$

$DCG_0 = 12,96$