Lista 2 - Redução de dimensão e clusterização

João Paixão

1. Dado uma tabela usuários por filmes (matriz):

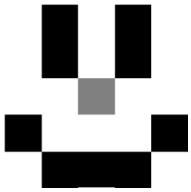
Ratings Matrix	Titanic	Rocky	The Hobbit	Fight Club	Jurassic Park
User A	0.82	0.18	0.5	0.1	0.26
User B	0.74	0.26	0.5	0.2	0.32
User C	0.34	0.69	0.5	0.7	0.62
User D	0.58	0.42	0.5	0.4	0.44
User E	0.1	0.9	0.5	1	0.8

   (a) Calcule o SVD com 2 componentes para a **transposta** dessa matriz (filmes por usuários).

   (b) Desenhe os filmes em dimensão dois (pode ser na mão ou em Python ou Julia) baseado no resultado da questão anterior.

   (c) Qual filme você recomendaria para quem gostou de Titanic usando os itens anteriores?

2. Considere a seguinte imagem:



   Insira esta imagem em uma matriz 5×5. Suponha que as sombras que você vê são apenas 0, 0.5 e 1 como fizemos na lista anterior. Faça o SVD com 1, 2, 3, 4 e 5 componentes e calcule o erro para cada caso. Discuta o que está acontecendo.

3. Encontre (ou desenhe) uma imagem A na internet, com mais ou menos 200×200 pixels (pode ser um pouco maior), que

   (a) exija mais que 3 componentes e menos que 6 no SVD para recuperar pelo menos 99% de sua norma total.

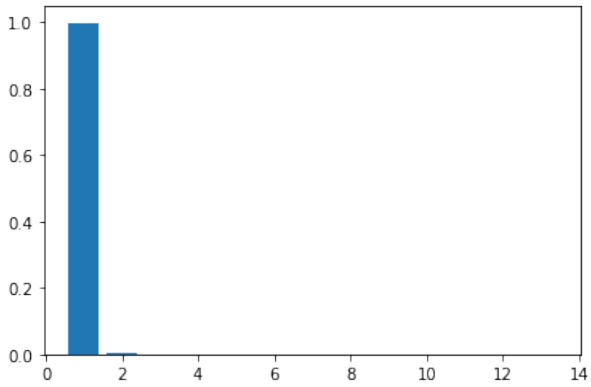
   (b) exija mais que 40 componentes para recuperar pelo menos 99% de sua norma total.

Dicas: muitos ícones tem o tamanho 200x200. Experimente buscar pelos termos "200x200 png" no Google Imagens. Veja os links abaixo:

(a) `https://www.google.com/search?q=200x200%20png&tbm=isch&hl=pt-BR&tbs=ic:gray`

(b) `https://imagepng.org/`

4. Abaixo vemos uma redução de dimensionalidade realizada para o dataset de vinhos, onde o eixo x representa cada componente e o eixo y a norma individual daquele componente. Responda:



(a) A redução de dimensionalidade foi realizada com o Colab `https://colab.research.google.com/drive/1C5mQnsMfIaMIEsKqw0Q78ML4IPmcTslv`. De acordo com o código e o resultado apresentado, podemos considerar essa redução incorreta. Justifique. Qual foi o provável erro cometido?

(b) Utilize o dataset de vinhos e faça a redução de dimensionalidade correta. Quantas componentes são necessárias para que atinjamos 80% da norma total do dataset?

5. Implemente o algoritmo K-means da maneira que apresentamos na aula (parecido com o Alternating Least Squares).

6. Escolha um dado qualquer (se quiser pode ser um dado sintético) e faça uma redução de dimensionalidade para duas dimensões e depois use o K-means para clusterizar.