07

Mãos à obra!

Vamos começar importando o `pandas`. Para isso, vamos criar uma célula no colaboratory e digitar o código:

```
import pandas as pd
```

Com o `pandas` importado, vamos buscar os dados dos filmes pela URI e renomear as colunas do `Dataframe`:

```
uri_filmes = 'https://raw.githubusercontent.com/oyurimatheus/clusterirng/master/movies/movies.csv'
filmes = pd.read_csv(uri_filmes)

filmes.columns = ['filme_id', 'titulo', 'generos']
```

Podemos ver os cinco primeiros filmes invocando o método `head` do `Dataframe` de filmes:

```
filmes.head()
```

Com isso, podemos extrair os dummies da coluna de gêneros. Logo, vamos falar para o nosso Dataframe de `filmes` pegar a coluna gêneros como string (`str`) e pegar os dummies (`get_dummies`):

```
generos = filmes.generos.str.get_dummies()
```

Isso retorna para gente um Dataframe com os dummies dos gêneros.

Podemos pegar este Dataframe e pedir para o `pandas` concatená-lo (`concat`) com o de `filmes` com as colunas (`axis=1`).

```
dados_dos_filmes = pd.concat([filmes, generos], axis=1)
```

Conseguimos ver os cinco primeiros dados do novo Dataframe utilizando o mesmo método `head()`:

```
dados_dos_filmes.head()
```

Por fim, temos que reescalar os dummies para saber quais dos gêneros mais influenciam os filmes. Portanto vamos importar o escalador da biblioteca `sklearn` e criar um objeto a partir da classe `StandardScaler`:

```
from sklearn.preprocessing import StandardScaler

scaler = StandardScaler()
```

Vamos falar para o `scaler` aprender com os dummies e transformá-los (`fit_transform`) para que, dessa forma, tenhamos mais informações sobre como os gêneros influenciam o filme:

```
generos_escalados = scaler.fit_transform(generos)
```

Podemos ver os novos gêneros reescalados colocando a variável como última instrução no bloco do documento do *colaboratory*:

```
generos_escalados
```