Les matrices creuses

Sparse Matrix

Simon OUELLET octobre 2020

Matrices à très grandes dimensionnalités

Comment les manipuler

Comment identifier une matrice creuse.

Comment gérer une matrice creuse.

Comment explorer les relations d’une matrice à grandes dimensions grâce à TSNE.

Sparse Matrix

Qu’est-ce qu’une matrice creuse?

- Matrice creuse :
  - Valeurs manquantes ;
  - Valeurs à 0 ;
  - Valeurs à False ;
- Matrice dense

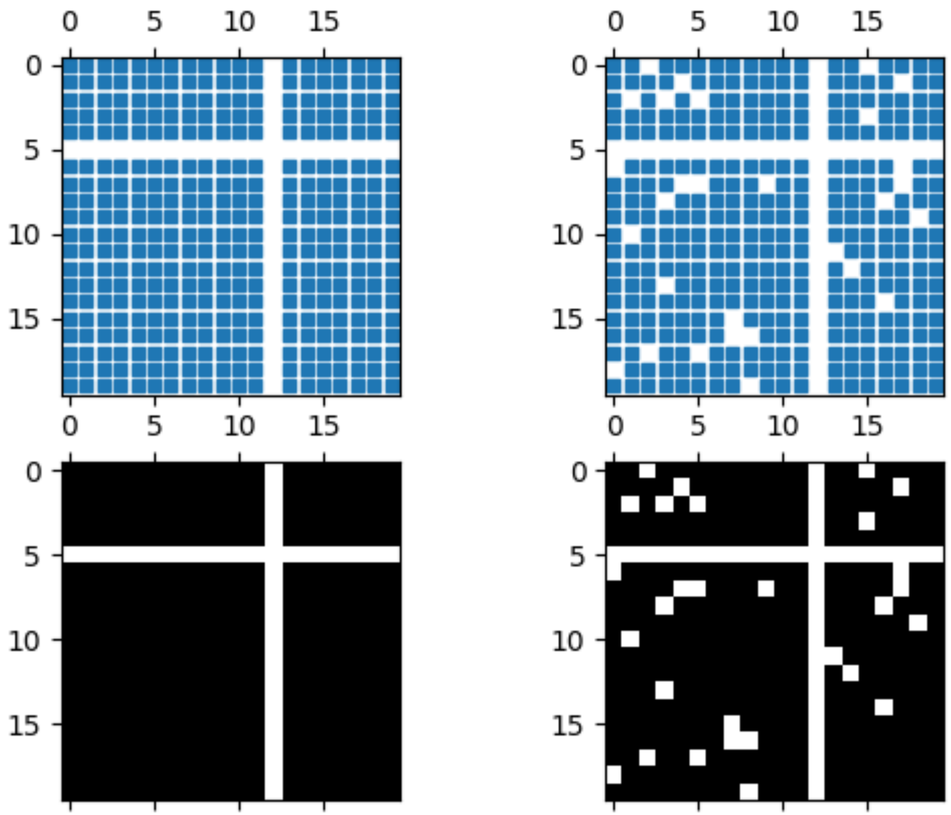
	Item_1	Item_2	Item_3	Item_4	...	Item_m
User_1	2.5	?	4.0	?	...	?
User_2	?	4.5	?	?	...	2.0
User_3	2.5	?	3.0	?	...	?
⋮	⋮	⋮	⋮	⋮	⋮	⋮
User_n	4.5	?	0.0	?	...	2.0

Visualiser

Matrice Creuse

- Python : matplotlib SPY

```
import matplotlib.pyplot as plt


fig, axs = plt.subplots(2, 2)
ax1 = axs[0, 0]
ax2 = axs[0, 1]
ax3 = axs[1, 0]
ax4 = axs[1, 1]

x = np.random.randn(20, 20)
x[5, :] = 0.
x[:, 12] = 0.

ax1.spy(x, markersize=5)
ax2.spy(x, precision=0.1, markersize=5)

ax3.spy(x)
ax4.spy(x, precision=0.1)

plt.show()
```



Manipuler

Pandas get_dummies

- Transformer une variable catégorielle.

```
pandas.get_dummies(
df_n,
columns=['glid_num'],
prefix="book",
sparse=True)
```

```
[47]: df_n.head()
```

[47]:

people_id	glid_num
348D0AA4DC5E40C4B439DBF4C86C877D	1023
EAD4649FC143416CBA4F0302DF6251FD	1023
348D0AA4DC5E40C4B439DBF4C86C877D	880
348D0AA4DC5E40C4B439DBF4C86C877D	720
348D0AA4DC5E40C4B439DBF4C86C877D	644

```
[46]: pd.get_dummies(df_n, columns=['glid_num'], prefix="book", sparse=True)
```

[46]:

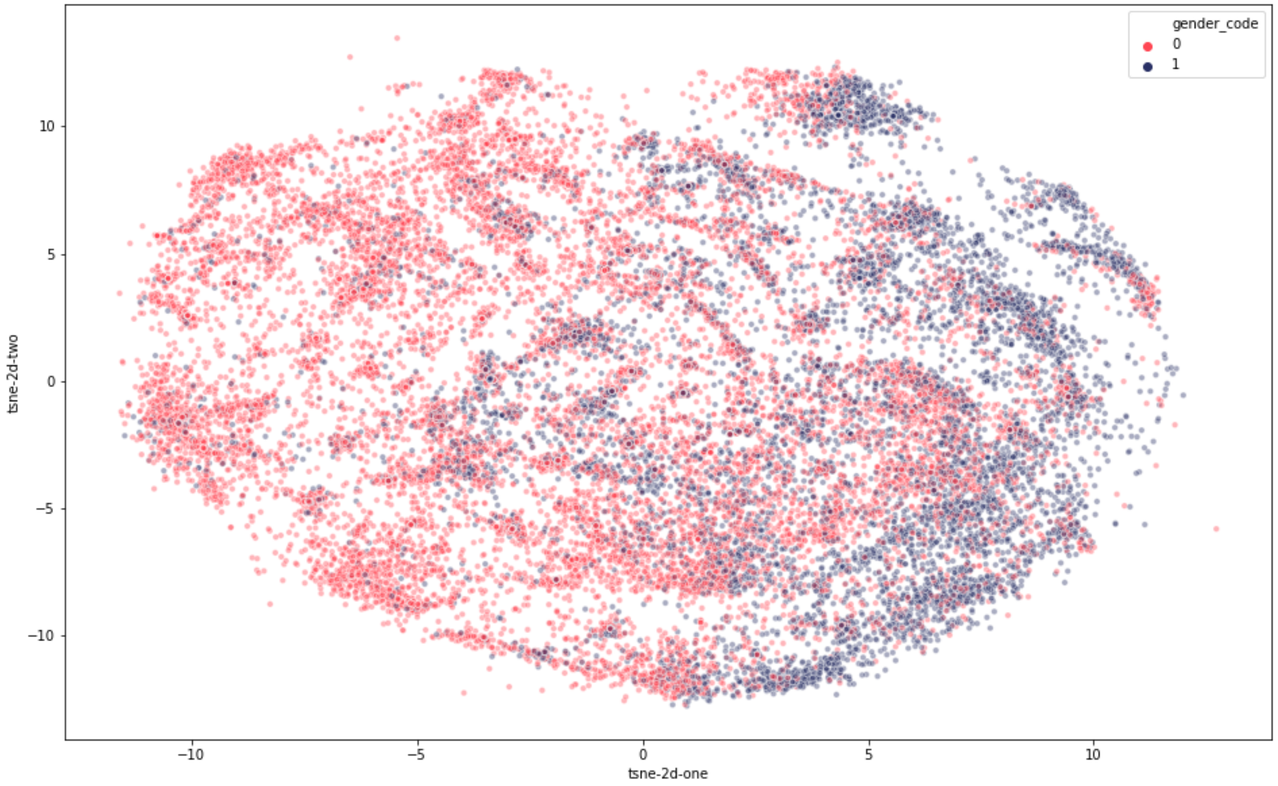
people_id	book_-1	book_0	book_1	book_2	book_3	book_4	book_5	b
348D0AA4DC5E40C4B439DBF4C86C877D	0	0	0	0	0	0	0	
EAD4649FC143416CBA4F0302DF6251FD	0	0	0	0	0	0	0	
348D0AA4DC5E40C4B439DBF4C86C877D	0	0	0	0	0	0	0	
348D0AA4DC5E40C4B439DBF4C86C877D	0	0	0	0	0	0	0	
348D0AA4DC5E40C4B439DBF4C86C877D	0	0	0	0	0	0	0	
...	...	...	...	...	...	...	...	
348D0AA4DC5E40C4B439DBF4C86C877D	0	0	0	0	0	0	0	
EAD4649FC143416CBA4F0302DF6251FD	0	0	0	0	0	0	0	
EAD4649FC143416CBA4F0302DF6251FD	0	0	0	0	0	0	0	
EAD4649FC143416CBA4F0302DF6251FD	0	0	0	0	0	0	0	
EAD4649FC143416CBA4F0302DF6251FD	0	0	0	0	0	0	0	

1241 rows × 1209 columns

t-SNE

t-Distributed Stochastic Neighbor Embedding

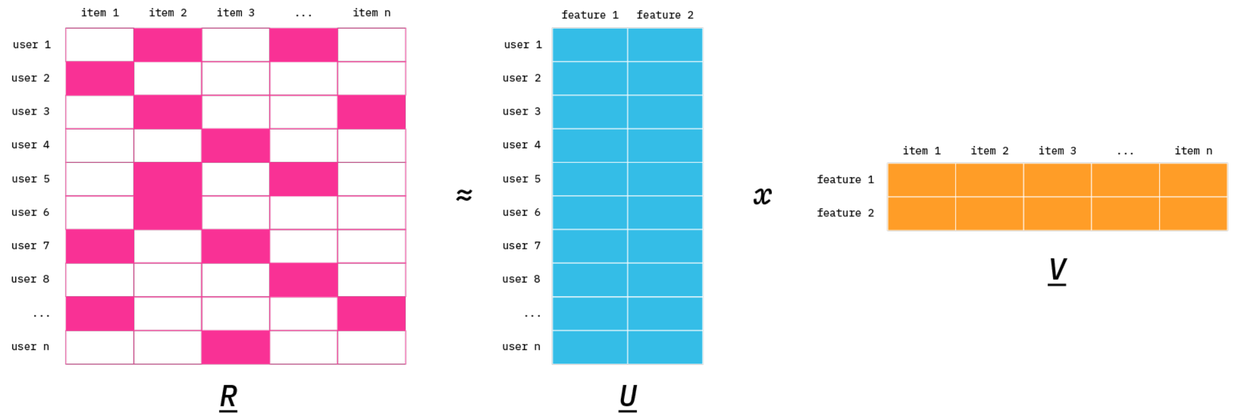
- Réduction de dimensions
- Visualisation
- Identifier s'il existe des relations entre deux variables dont au moins une est catégorielle.



Exemple Gleeph

Est-ce que nos patterns sont genrés?

- Nous avons une matrice latente de 70 dimensions (U) issue d'un calcul BPR (bayesian personal ranking).
- Une dimension représente un pattern de livres clivant.
- Les utilisateurs sont projetés dans cette matrice en fonction du contenu de leur bibliothèque.



T-SNE

Exemple