Graphes : chaînes de Markov
Problème de recherche : le classement des pages web

Un petit web contient trois pages A, B, C. La page A contient des liens vers B et C, la page B vers A et C, la page C vers A seulement. Un internaute part d'une page au hasard puis, à chaque étape, clique au hasard sur l'un des liens de la page où il se trouve. On note π_n la loi de la page où il se trouve après n clics (vecteur ligne, ordre A, B, C), avec $\pi_0 = \left(\frac{1}{3} \quad \frac{1}{3} \quad \frac{1}{3}\right)$.

La mission

1. Représenter la situation par un graphe probabiliste et écrire la matrice de transition M.
2. Avec Python, calculer π_n pour n allant jusqu'à 50. Conjecturer la limite et la vérifier par le calcul. Classer les trois pages par ordre d'importance.
3. Un internaute réel ne clique pas indéfiniment : il lui arrive de taper une adresse au hasard. On suppose qu'à chaque étape, avec la probabilité 0,85, il clique au hasard sur un lien de la page, et avec la probabilité 0,15, il va sur l'une des trois pages choisie au hasard. Démontrer que la nouvelle matrice de transition est $G = 0{,}85M + 0{,}05J$, où J est la matrice carrée d'ordre 3 dont tous les coefficients valent 1.
4. Avec Python, déterminer une valeur approchée de la loi stable pour G. Le classement change-t-il ? Vérifier que la loi $\left(\frac{74}{171} \quad \frac{40}{171} \quad \frac{1}{3}\right)$ est stable.

Coups de pouce

Bloqué ? Des indices progressifs sont disponibles en ligne. N'en ouvrir qu'un à la fois, et seulement après avoir cherché !

Éléments de correction

1. La matrice

$$M = \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} \\ 1 & 0 & 0 \end{pmatrix}$$

(graphe : $A \to B$ et $A \to C$ de poids $\frac{1}{2}$, $B \to A$ et $B \to C$ de poids $\frac{1}{2}$, $C \to A$ de poids 1).

2. Sans amortissement

```
M = [[0, 0.5, 0.5], [0.5, 0, 0.5], [1, 0, 0]]
pi = [1/3, 1/3, 1/3]
for n in range(50):
    pi = [sum(pi[i] * M[i][j] for i in range(3)) for j in range(3)]
print(pi)
```

On obtient environ $\begin{pmatrix} 0{,}4444 & 0{,}2222 & 0{,}3333 \end{pmatrix}$; on conjecture la limite $\pi = \begin{pmatrix} \frac{4}{9} & \frac{2}{9} & \frac{3}{9} \end{pmatrix}$.
Vérification : $\pi M = \begin{pmatrix} \frac{1}{9}+\frac{3}{9} & \frac{2}{9} & \frac{2}{9}+\frac{1}{9} \end{pmatrix} = \pi$.
Classement : A (la plus visitée, car B et C pointent vers elle), puis C, puis B.

3. L'amortissement

D'après la formule des probabilités totales, la probabilité d'aller de la page i à la page j est $0{,}85 \times m_{ij} + 0{,}15 \times \frac{1}{3} = 0{,}85\, m_{ij} + 0{,}05$: c'est le coefficient (i, j) de $0{,}85M + 0{,}05J$. Par exemple

$$G = \begin{pmatrix} 0{,}05 & 0{,}475 & 0{,}475 \\ 0{,}475 & 0{,}05 & 0{,}475 \\ 0{,}9 & 0{,}05 & 0{,}05 \end{pmatrix}.$$

Tous les coefficients sont strictement positifs : on peut aller de n'importe quelle page à n'importe quelle autre en un clic.

4. Avec amortissement

On définit `G` par la ligne ci-dessous, puis on reprend la même boucle avec `G` :

```
G = [[0.85*M[i][j] + 0.05 for j in range(3)] for i in range(3)]
```

On obtient environ $\begin{pmatrix} 0{,}4327 & 0{,}2339 & 0{,}3333 \end{pmatrix}$. Le classement ne change pas (A, puis C, puis B), mais les écarts sont plus faibles.
Vérification avec $\pi = \begin{pmatrix} \frac{74}{171} & \frac{40}{171} & \frac{57}{171} \end{pmatrix}$: comme la somme des coefficients de π vaut 1, $\pi J = \begin{pmatrix} 1 & 1 & 1 \end{pmatrix}$ et $\pi G = 0{,}85\,\pi M + 0{,}05 \begin{pmatrix} 1 & 1 & 1 \end{pmatrix}$. On a

$$\pi M = \frac{1}{171}\begin{pmatrix} 20+57 & 37 & 37+20 \end{pmatrix} = \frac{1}{171}\begin{pmatrix} 77 & 37 & 57 \end{pmatrix},$$

puis

$$\pi G = \frac{1}{171}\begin{pmatrix} 0{,}85\times 77 + 8{,}55 & 0{,}85\times 37 + 8{,}55 & 0{,}85\times 57 + 8{,}55 \end{pmatrix},$$

c'est-à-dire $\pi G = \frac{1}{171}\begin{pmatrix}74 & 40 & 57\end{pmatrix} = \pi$ ($0{,}05 \times 171 = 8{,}55$; $65{,}45 + 8{,}55 = 74$; $31{,}45 + 8{,}55 = 40$; $48{,}45 + 8{,}55 = 57$).

Enrichissement historique : ce principe est celui de l'algorithme PageRank, présenté en 1998 par Sergey Brin et Lawrence Page, fondateurs de Google, avec la probabilité 0,85. Le fait que la loi stable existe, soit unique et soit la limite de (π_n) est admis ici.

Ce qu'on a utilisé du chapitre

— Chaîne de Markov à 3 états, matrice de transition.

— Loi après n étapes ; loi stable $\pi M = \pi$.