Que fait Google ?

Année 2013

ALI MOHAMED Nasrati (Terminale S), ANDRIAMANANJAONA Iris (2nd), ISSABHAY Imraan (2nd), LAMOLY Maeva (Terminale S), MARIAPIN Arjuna (2nd), RIVIERE Séverine (Terminale S), VIRAPIN Anthony (2nd), VIRASSAMY Rachel (Terminale S).

Lycée Jean Hinglo
2 rue des sans soucis - 97420 LE PORT.

Lycée Bellepierre
Avenue Gaston Monnerville - 97475 ST DENIS Cedex

Enseignants :
ANSELMET Jérôme, FUR-DESOUTTER Sophie.

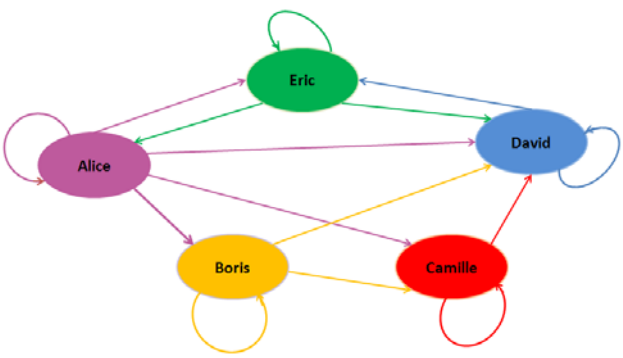
Chercheur : LEGONIDEC Marion, Université de La Réunion.

I- Le sujet

- Lorsqu'on fait une recherche sur Google, comment choisit-il quel lien (non commercial) mettre en premier ?

- Comment Google fait-il pour classer les pages ?

Un problème analogue :
A un groupe de 5 élèves de sport étude, on pose la question suivante : quels sont parmi vous les plus forts en sport ?
Par exemple, le schéma ou graphe ci-dessous représente les votes de chaque élève :

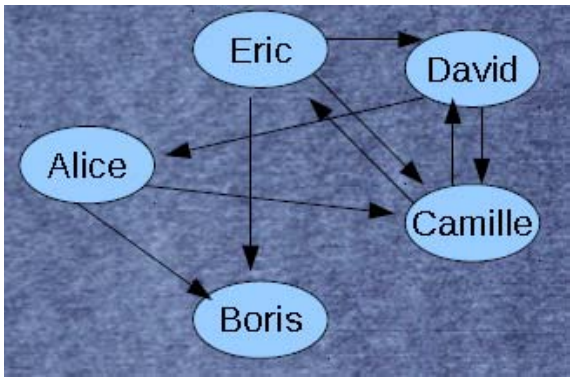


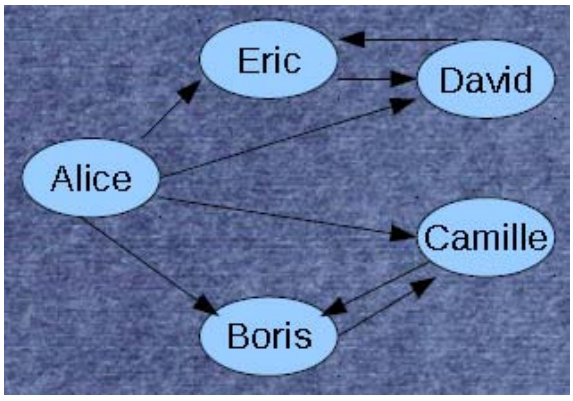
Comment faire un classement qui tienne compte des réponses de chacun ?

On considère que les cinq élèves représentent cinq pages internet et les votes des élèves sont les liens entre les différentes pages.

Ainsi la page 1 (Alice), a un lien vers les pages 1, 2, 3, 4 et 5.
De même pour les autres pages.

Autres cas :





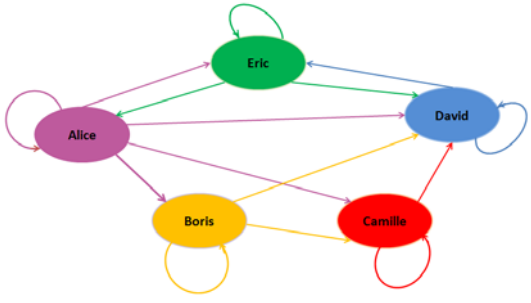
-Peut-on établir un classement ?

-Si des problèmes apparaissent comment les contourner ?

Nous avons essayé de trouver une démarche qui permette d'établir un classement des joueurs (ou des pages).

II- Vers les premières règles

Nous avons considéré le cas suivant :



Les arêtes représentent les votes attribués (ou les liens d'une page vers une autre page web). Ainsi Alice vote pour les cinq joueurs, tandis que David ne vote que pour lui-même et pour Eric.

En comptant le nombre d'arêtes, nous obtenons le tableau suivant :

Alice	Boris	Camille	David	Eric
2	2	3	5	3

Cette méthode ne nous permet pas de conclure, puisque nous obtenons des élèves ex-aequo.

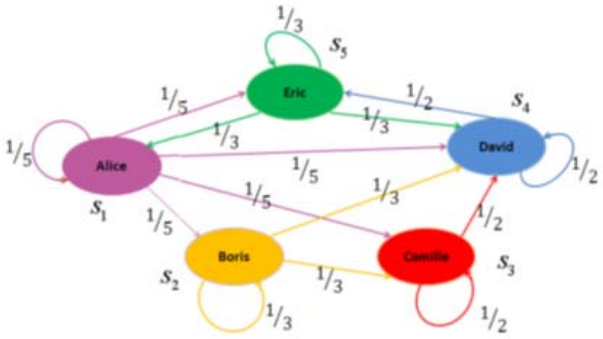
Nous avons choisi d'établir des règles.

Alice vote pour tout le monde, son vote n'est pas tranché, nous estimons donc qu'il aura moins d'importance que le vote de David ou de Camille, qui n'ont voté que pour deux personnes.

Ainsi, nous avons attribué un score, à chacun des joueurs et nous avons considéré que chaque joueur donne une partie de son score à ceux qu'il désigne comme meilleur joueur.

On note :
s_1 le score d'Alice
s_2 le score de Boris,
s_3 le score de Camille,
s_4 le score de David,
s_5 le score d'Eric.

Nous avons alors le graphe suivant :



Alice partage son score entre Boris, Camille, David, Eric et elle-même., c'est-à-dire qu'elle partage son score en 5.

Donc chaque joueur reçoit $\frac{1}{5}$ du score d'Alice.

De la même façon, Alice, David et Eric reçoivent $\frac{1}{3}$ du score d'Eric.
Nous obtenons le système suivant :

$$\begin{cases} s_1 = \frac{1}{5}s_1 + \frac{1}{3}s_5 \\ s_2 = \frac{1}{5}s_1 + \frac{1}{3}s_2 \\ s_3 = \frac{1}{5}s_1 + \frac{1}{3}s_2 + \frac{1}{2}s_3 \\ s_4 = \frac{1}{5}s_1 + \frac{1}{3}s_2 + \frac{1}{2}s_3 + \frac{1}{2}s_4 + \frac{1}{3}s_5 \\ s_5 = \frac{1}{5}s_1 + \frac{1}{2}s_4 + \frac{1}{3}s_5 \end{cases}$$

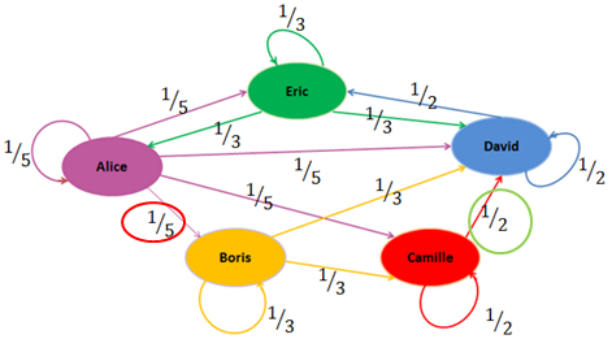
A l'aide d'un logiciel de calcul formel, nous avons pu résoudre ce système.
Les solutions sont données en fonction du paramètre s_1.

$$S = \left\{ s_1 ; s_2 = \frac{3}{10}s_1 ; s_3 = \frac{3}{5}s_1 ; s_4 = \frac{14}{5}s_1 ; s_5 = \frac{12}{5}s_1 \right\}$$

Nous pouvons donc en conclure un classement : David, Eric, Alice, Camille et Boris.

II- Autre façon de résoudre le problème : avec les matrices

Une matrice est avant tout un tableau de nombres.
Nous reprenons la situation précédente.



On traduit ce graphe par un tableau.

	Alice	Boris	Camille	David	Eric
Alice	$\frac{1}{5}$	0	0	0	$\frac{1}{3}$
Boris	$\frac{1}{5}$	$\frac{1}{3}$	0	0	0
Camille	$\frac{1}{5}$	$\frac{1}{3}$	$\frac{1}{2}$	0	0
David	$\frac{1}{5}$	$\frac{1}{3}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{3}$
Eric	$\frac{1}{5}$	0	0	$\frac{1}{2}$	$\frac{1}{3}$

Pour le lire on part de la colonne vers la ligne. Par exemple, le vote d'Alice vers Boris est de $\frac{1}{5}$ (entouré en rouge).
Un autre exemple, le vote de Camille pour David est de $\frac{1}{2}$ (entouré en vert).

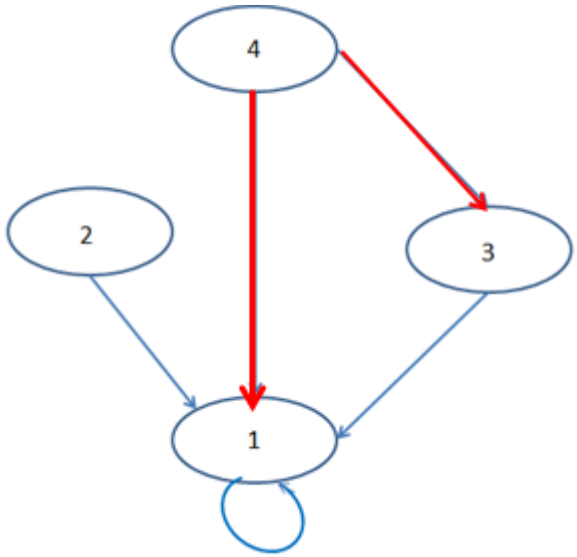
L'écriture mathématique d'une matrice est similaire au tableau avec l'ajout des parenthèses.

$$\begin{pmatrix} \frac{1}{5} & 0 & 0 & 0 & \frac{1}{3} \\ \frac{1}{5} & \frac{1}{3} & 0 & 0 & 0 \\ \frac{1}{5} & \frac{1}{3} & \frac{1}{2} & 0 & 0 \\ \frac{1}{5} & \frac{1}{3} & \frac{1}{2} & \frac{1}{2} & \frac{1}{3} \\ \frac{1}{5} & 0 & 0 & \frac{1}{2} & \frac{1}{3} \end{pmatrix}$$

Tous les autres nombres de la matrice (lus de la colonne vers la ligne) représentent de même la valeur inscrite sur l'arête du graphe.

III- Retour à internet

Considérons quatre pages, dont les liens entre elles sont représentés par les arêtes sur le graphe suivant.



Cette matrice A, représente ce graphe.

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

Probabilité de passer de 4 à 1

Probabilité de passer de 4 à 3

Deux liens partent de la page 4, un vers la page 3 et un vers la page 1, nous avons donc 1 chance sur 2 de passer de la page 4 à la page 1 et de même de la page 4 à la page 3.

Alors imaginons que nous soyons initialement sur la page n°4, la matrice de position sera donc la matrice $P = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}$. Les valeurs de cette matrice correspondent aux probabilités d'être sur une page, ainsi 0 car la probabilité d'être sur la page 1 est nulle, idem pour les pages 2 et 3, en revanche on est sur la page 4 donc la probabilité d'être sur la page 4 est 1.

Lorsque nous multiplions la matrice A et la matrice P nous obtenons la matrice AP. (1)

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

Probabilité d'arriver sur la page 1 en partant de la page 4 en cliquant une seule fois sur un lien.

Les coefficients de cette matrice $\begin{pmatrix} \frac{1}{2} \\ 0 \\ \frac{1}{2} \\ 0 \end{pmatrix}$ sont les probabilités de passer de la page 4 à une autre page si nous cliquons une seule fois sur un lien.
La probabilité que nous arrivions sur la page 1 est de $\frac{1}{2}$ et de même pour la page 3.

Nous ne pouvons pas arriver sur la page 2 en partant de 4 donc la probabilité est nulle, de même pour la page 4.

En réitérant notre calcul, nous avons calculé, $A^2P = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}$. (2)

On obtient une nouvelle matrice position qui correspond bien à la probabilité d'arriver sur une page si nous cliquons sur deux liens.

Nous remarquons donc que nous ne pouvons arriver que sur la page 1, à partir de 4 et en cliquant sur deux liens.
Ce qui correspond bien, à ce que nous observons sur le graphe, il s'agit des chemins : 4-3-1 ou 4-1-1.

Nous avons continué en calculant A^3P. $A^3P = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}$. Nous remarquons qu'en suivant trois liens, nous avons le même résultat que pour le précédent, c'est-à-dire d'atterrir sur la page 1.

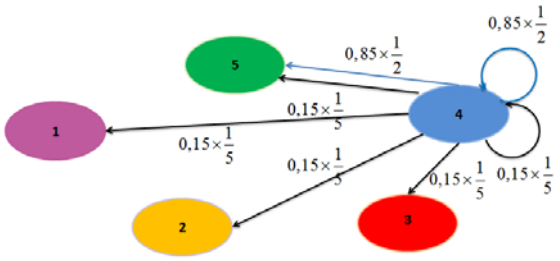
Nous sommes donc « coincés » sur la page 1 à partir de deux clics. Ce qui se visualise sur le graphe puisque la page 1 n'a aucun lien pointant vers une autre page. A partir du moment où on arrive sur la page 1, on y reste.

Ce qui n'est pas réaliste, par rapport à internet.

IV- Vers un modèle plus réaliste

Afin de rendre la situation plus réaliste, nous devons prendre en compte un paramètre aléatoire :
- un internaute passe d'une page à l'autre en suivant un lien dans 85% des cas,
- sinon il passe d'une page à n'importe quelle autre page, (sans obligatoirement suivre un lien), dans les 15% des cas restants.

Nous pouvons établir cette modification en reprenant la situation initiale en considérant 1, 2, 3, 4 ,5 comme cinq pages du web.



La page 4 a un lien vers la page 5 et vers elle-même donc on a 85% de chances de suivre un des deux liens ainsi ces liens sont affectés des coefficients $0,85\times\frac{1}{2}$.

On peut ensuite passer de 5 à n'importe quelle page avec une probabilité de 15% donc on rajoute 5 arêtes sur le graphe. Chaque arête est affectée du coefficient $\frac{1}{5}$ multiplié par la probabilité 0,15.

Cette modélisation illustre mieux le principe de Google puisqu'on ne choisit pas toujours de suivre un lien direct pour passer d'une page à une autre.

Nous obtenons donc une nouvelle matrice, somme de la matrice « des liens » (déterminée précédemment) multiplié par 0,85 et de la matrice illustrant un cheminement aléatoire multiplié par 0,15.

$$A = 0,85\times\begin{pmatrix} \frac{1}{5} & 0 & 0 & 0 & \frac{1}{3} \\ \frac{1}{5} & \frac{1}{3} & 0 & 0 & 0 \\ \frac{1}{5} & \frac{1}{3} & \frac{1}{2} & 0 & 0 \\ \frac{1}{5} & \frac{1}{3} & \frac{1}{2} & \frac{1}{2} & \frac{1}{3} \\ \frac{1}{5} & 0 & 0 & \frac{1}{2} & \frac{1}{3} \end{pmatrix} + 0,15\times\begin{pmatrix} \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} \\ \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} \\ \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} \\ \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} \\ \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} \end{pmatrix}$$

La matrice A est donnée en annexe.

C'est cette nouvelle matrice représentant la situation « réaliste » que l'on va itérer pour déterminer l'état stable.

On remarque que dans la matrice à la puissance 38, ce qui revient à 38 étapes, les résultats de chaque ligne commencent à se stabiliser. (Voir les calculs en annexe). (3)
Nous avons ensuite calculé A^{50} (en annexe).

La matrice reste stable.
A partir de la matrice à la puissance 39, on peut établir le classement : la page 4, la page 5, la page 1, la page 3 et enfin la page 2.

Nous avons ensuite modifié la valeur du paramètre, nous avons choisi 0,20 au lieu de 0,15.
Les places des pages 1 et 3 étaient inversées.
On peut estimer qu'une erreur au niveau du choix des coefficients pourrait perturber le classement.

V- Programmation en Python.

Nous avons ensuite écrit un programme en Python, pour obtenir un classement.
Nous entrons la première ligne de la matrice A « matrice des liens ».

A l'aide d'une boucle, nous construisons la matrice A.

Elle est ensuite modifiée, pour prendre en compte le paramètre aléatoire. La matrice obtenue est la matrice B, qui est élevée à une puissance choisie grande pour la stabiliser.
La matrice obtenue est la matrice qui permet d'obtenir un classement.

```
# -*- coding: utf-8 -*-

# package pour matrices, tableaux
import numpy as np
from fractions import Fraction

# n=le nombre de pages ; e=coefficient ;
# p=puissance de la matrice ;
# A=Matrice initiale, première ligne ;
# D=matrice avec que des "1"

n=int(input("n="))
e=float(input("e="))
p=int(input("p="))
A=input("A=")

d=float(Fraction(1,n))
D=d*np.ones((n,n))

# Reste de la matrice initiale
for i in range (2,n+1):
    L=input("L=")
    A=np.concatenate((A,L))

# Transforme les lignes en colonnes
# pour ensuite, construire la matrice
# avec le coefficient
A=np.transpose(A)

B=(1-e)*A+e*D

# On eleve la matrice obtenue à des
# puissances, pour la stabiliser
for i in range (p):
    B=np.dot(B,B)

# Affichage

print B
```

VI – Conclusion

En introduisant un paramètre aléatoire (15%), nous sommes capables d'obtenir un classement de pages (ou de joueurs) quelle que soit la situation donnée.

Les problèmes d'ex-aequo sont ainsi éliminés.

Nous avons aussi testé notre programme sur les deux autres cas donnés initialement dans le sujet.

Ainsi, à l'aide d'une méthode similaire, Google classe toutes les pages du web, en leur attribuant un score (ou pagerank). Lorsque nous faisons une recherche, il renvoie les pages liées aux mots-clés donnés, classées grâce à leur score.

Annexe

$$A=\begin{pmatrix} \frac{1}{5} & \frac{3}{100} & \frac{3}{100} & \frac{3}{100} & \frac{47}{150} \\ \frac{1}{5} & \frac{47}{150} & \frac{3}{100} & \frac{3}{100} & \frac{3}{100} \\ \frac{1}{5} & \frac{47}{150} & \frac{91}{200} & \frac{3}{100} & \frac{3}{100} \\ \frac{1}{5} & \frac{47}{150} & \frac{91}{200} & \frac{91}{200} & \frac{47}{150} \\ \frac{1}{5} & \frac{3}{100} & \frac{3}{100} & \frac{91}{200} & \frac{47}{150} \end{pmatrix}$$

```
float(A^^38);
```

0.13601959002417	0.13601959002417	0.13601959002417	0.13601959002417	0.13601959002417
0.074125577168523	0.074125577168523	0.074125577168524	0.074125577168525	0.074125577168524
0.12891404724961	0.12891404724961	0.12891404724961	0.12891404724961	0.12891404724961
0.3683657512516	0.3683657512516	0.3683657512516	0.3683657512516	0.3683657512516
0.2925750343061	0.2925750343061	0.2925750343061	0.2925750343061	0.2925750343061

```
float(A^^50);
```

0.13601959002417	0.13601959002417	0.13601959002417	0.13601959002417	0.13601959002417
0.074125577168524	0.074125577168524	0.074125577168524	0.074125577168524	0.074125577168524
0.12891404724961	0.12891404724961	0.12891404724961	0.12891404724961	0.12891404724961
0.3683657512516	0.3683657512516	0.3683657512516	0.3683657512516	0.3683657512516
0.2925750343061	0.2925750343061	0.2925750343061	0.2925750343061	0.2925750343061

Notes de l'édition

(1) Expliquer comment on calcule AP (*voir à ce sujet l'annexe qui suit page 7*)

(2) Mettre que $A^2P = A(AP)$ et remplacer par la valeur obtenue précédemment.

(3) c'est-à-dire qu'à l'étape suivante (39), la matrice aura à peu près les mêmes valeurs.

Première partie : Multiplication par une seule matrice

Pour comprendre les calculs de matrice faits dans l'article, nous allons procéder sur un exemple. Prenons la matrice A = $\begin{pmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \\ -1 & -2 & 1 & 3 \\ 5 & 6 & -4 & -2 \end{pmatrix}$ et la matrice P = $\begin{pmatrix} 1 \\ 0 \\ 1 \\ 2 \end{pmatrix}$.
Pour faire le calcul de AP, nous allons procéder en plusieurs étapes.

D'abord, regardons la première ligne de la matrice A = $\begin{pmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \\ -1 & -2 & 1 & 3 \\ 5 & 6 & -4 & -2 \end{pmatrix}$. Nous allons prendre chaque chiffre de cette ligne et nous allons les multiplier par les chiffres de P puis additionner le tout:

$$1 \times 1 + 2 \times 0 + 3 \times 1 + 4 \times 2 = 12$$

Enfin, nous plaçons le résultat dans la première case de la nouvelle matrice :

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \\ -1 & -2 & 1 & 3 \\ 5 & 6 & -4 & -2 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \\ 1 \\ 2 \end{pmatrix} = \begin{pmatrix} 12 \\ ? \\ ? \\ ? \end{pmatrix}$$

Pour trouver les trois autres chiffres, il ne reste plus qu'à recommencer avec les trois autres lignes :

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \\ -1 & -2 & 1 & 3 \\ 5 & 6 & -4 & -2 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \\ 1 \\ 2 \end{pmatrix} = \begin{pmatrix} 1 \times 1 + 2 \times 0 + 3 \times 1 + 4 \times 2 \\ 5 \times 1 + 6 \times 0 + 7 \times 1 + 8 \times 2 \\ (-1) \times 1 + (-2) \times 0 + 1 \times 1 + 3 \times 2 \\ 5 \times 1 + 6 \times 0 + (-4) \times 1 + (-2) \times 2 \end{pmatrix} = \begin{pmatrix} 12 \\ 28 \\ 6 \\ -3 \end{pmatrix}$$

Ainsi, nous obtenons le résultat :

$$AP = \begin{pmatrix} 12 \\ 28 \\ 6 \\ -3 \end{pmatrix}$$

Il est important de remarquer qu'il faut donc autant de colonnes dans la matrice A que la matrice P n'a de lignes.

Deuxième partie : Multiplication par plusieurs matrices

Dans l'article, les auteurs multiplient parfois par plusieurs matrices. Le plus simple est de faire par étape :

$$A^2P = A \times AP = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \\ -1 & -2 & 1 & 3 \\ 5 & 6 & -4 & -2 \end{pmatrix} \times \begin{pmatrix} 12 \\ 28 \\ 6 \\ -3 \end{pmatrix} = \begin{pmatrix} 1 \times 12 + 2 \times 28 + 3 \times 6 + 4 \times (-3) \\ 5 \times 12 + 6 \times 28 + 7 \times 6 + 8 \times (-3) \\ (-1) \times 12 + (-2) \times 28 + 1 \times 6 + 3 \times (-3) \\ 5 \times 12 + 6 \times 28 + (-4) \times 6 + (-2) \times (-3) \end{pmatrix}$$

Et au final, nous obtenons :

$$A^2P = \begin{pmatrix} 74 \\ 246 \\ -71 \\ 210 \end{pmatrix}$$

Troisième partie : Généralisation

A la fin de l'article, les auteurs calculent des "puissances" de matrice. Commençons par le calcul de $A^2 = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \\ -1 & -2 & 1 & 3 \\ 5 & 6 & -4 & -2 \end{pmatrix} \times \begin{pmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \\ -1 & -2 & 1 & 3 \\ 5 & 6 & -4 & -2 \end{pmatrix}$.

Pour faire cela, nous regardons chaque colonne de la matrice de droite et faisons comme précédemment puis nous accolons les matrices obtenues. Schématiquement, nous obtenons :

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \\ -1 & -2 & 1 & 3 \\ 5 & 6 & -4 & -2 \end{pmatrix} \times \begin{pmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \\ -1 & -2 & 1 & 3 \\ 5 & 6 & -4 & -2 \end{pmatrix} = \begin{pmatrix} 28 & 32 & 4 & 21 \\ 68 & 80 & 32 & 73 \\ 3 & 2 & -28 & -23 \\ 29 & 42 & 61 & 60 \end{pmatrix}$$

Nous pouvons faire les calculs sur un exemple plus petit, multiplions la matrice $\begin{pmatrix} 1 & 2 \\ -3 & 4 \end{pmatrix}$ par la matrice $\begin{pmatrix} -1 & 0 \\ 2 & 3 \end{pmatrix}$:

$$\begin{pmatrix} 1 & 2 \\ -3 & 4 \end{pmatrix} \times \begin{pmatrix} -1 & 0 \\ 2 & 3 \end{pmatrix} = \begin{pmatrix} 1 \times (-1) + 2 \times 2 & 1 \times 0 + 2 \times 3 \\ (-3) \times (-1) + 4 \times 2 & (-3) \times 0 + 4 \times 3 \end{pmatrix} = \begin{pmatrix} 3 & 6 \\ 11 & 12 \end{pmatrix}$$

Encore une fois, il est important de signaler qu'il faut autant de colonnes dans la première matrice que de lignes dans la seconde.