Agrégation marocaine de Mathématiques, session 2024
Épreuve de modélisation et Calcul Scientifique

Le jury n'exige pas une compréhension exhaustive du texte. Vous êtes libres d'organiser votre discussion comme vous l'entendez. Des suggestions de développement, largement indépendantes les unes des autres, vous sont proposées en fin de texte. Ce ne sont que des suggestions et vous n'êtes pas tenu(e) de les suivre. Il vous est conseillé de mettre en lumière vos connaissances à partir du fil conducteur constitué par le texte. Le jury appréciera que la discussion soit accompagnée d'exemples traités sur ordinateur

Marche aléatoire sur un graphe et Algorithme de Page rank

I. Marche aléatoire sur un graphe

1.1 Définitions et Notations

Definition 1 *Un graphe orienté G est un ensemble $G = (S, A)$ où S est un ensemble fini dont les éléments s'appellent les sommets du graphe G et où $A \subset S^2$. Les éléments de A s'appellent les arêtes orientées du graphe G. Si $a = (s, s') \in A$, a est l'arête orientée reliant le sommet s au sommet s'. Si $(s, s') \in S^2$, on dit que s' est un sommet voisin de s s'il existe une arête orientée reliant s à s', c'est-à-dire si $(s, s') \in A$. On note que s peut être un sommet voisin de lui-même (si $(s, s) \in A$) et que s' peut être un voisin de s alors que s n'est pas un voisin de s' (si $(s, s') \in A$ alors que $(s', s) \notin A$)*

Definition 2 *Soit $P = (p_1, \ldots, p_n)$ un vecteur de $\mathbb{R}^n$. On dit que P est une distribution de probabilité si pour tout $1 \leq i \leq n$, $p_i \geq 0$ et $p_1 + \ldots + p_n = 1$.*

Definition 3 *On dit qu'une matrice $M \in \mathcal{M}_n(\mathbb{R})$ dont tous les coefficients sont positifs ou nuls est stochastique si $Me = e$ où $e = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}$ le vecteur colonne n'ayant que des 1 (**on garde cette définition de e dans toute la suite**).*

Definition 4 *Une distribution de probabilité X est dit invariant par une matrice M si $XM = X$.*

1.2 Préliminaire

On considère un graphe orienté fini dont les sommets sont numérotés de 1 à n. Un point se déplace aléatoirement d'un sommet à un autre de ce graphe en suivant les arêtes orientées du graphe. Le nombre d'étapes de cette marche aléatoire peut tendre vers l'infini. À chaque étape, le point se déplace du sommet où il se trouve vers l'un de ses sommets voisins de façon équiprobable. Ceci entraine notamment que la

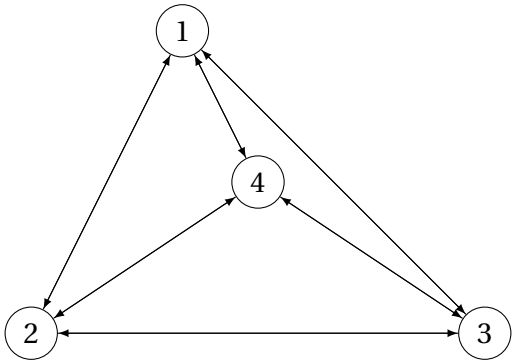


FIGURE 1 – graphe 1 : Marche aléatoire sur un tétraèdre

probabilité de passer du sommet i au sommet j ne dépend pas du rang de l'étape. Pour $1 \leq i, j \leq n$, on note $t_{i,j}$, la probabilité que le point passe du sommet i au sommet j ; en particulier, s'il n'y a pas d'arête reliant i à j, $t_{i,j} = 0$. La matrice dont le coefficient de la ligne i et de la colonne j est égal à $t_{i,j}$ est notée T. Cette matrice s'appelle la matrice de transition du graphe. Pour $k \in \mathbb{N}$, on note $P^{(k)}$ le vecteur ligne $(p_1^{(k)}, p_2^{(k)}, \ldots, p_{n-1}^{(k)}, p_n^{(k)})$, où, pour $1 \leq i \leq n, p_i^{(k)}$ est la probabilité que le point soit sur le sommet i à l'étape de rang k.

Theorem 5 *Le vecteur $P^{(k)}$ est une distribution de probabilité pour tout $k \in \mathbb{N}$.*
$P^{(k+1)} = P^{(k)}T$.
La matrice de transition d'un graphe est une matrice stochastique.

1.3 Exemple : Marche aléatoire sur un tétraèdre

Dans cet exemple, on considère le graphe orienté $G = (S, A)$ où $S = \{1, 2, 3, 4\}$ et
$A = \{(1,2), (2,1), (1,3), (3,1), (1,4), (4,1), (2,3), (3,2), (2,4), (4,2), (3,4), (4,3)\}$.

La figure 1 représente ce graphe. Les sommets sont représentés par des cercles et l'arête orientée reliant le sommet s au sommet s', par une flèche de s vers s'. Par la suite, nous utiliserons la représentation simplifiée dans laquelle si le graphe comporte les deux arêtes orientées (s, s') et (s', s) elles sont représentées par un seul trait avec une flèche à chaque extrémité.

Theorem 6 *la suite de matrices $(T^k)_{k\geq 1}$ est convergente. La suite de vecteur $(P^{(k)})$ est convergente vers un vecteur P stochastique et vérifiant $PT = P$ et ceci quelque soit le vecteur initial stochastique $P^{(0)}$.*

1.4 Matrice stochastique et distributions de probabilités

Soit $M \in \mathcal{M}_n(\mathbb{R})$ une matrice stochastique. Soit λ est une valeur propre de M et V un vecteur propre associé de coordonnées $(u_i)_{1\leq i\leq n}$. Soit $h \in \{1, \ldots, n\}$ tel que $|u_h| = max_{1\leq i\leq n}|u_i|$. Soit $\delta = \min_{1\leq i\leq n} m_{i,i}$. On note $m_{i,j}^{(k)}$ le coefficient de la matrice M^k situé à la ligne i et la colonne j. On pose $\epsilon = \min_{1\leq i,j\leq n} m_{i,j}$. Pou tout $j \in \|1, n\|$, on pose $\alpha_j^{(k)} = \min_{1\leq i\leq n} m_{i,j}^{(k)}$ et $\beta_j^{(k)} = \max_{1\leq i\leq n} m_{i,j}^{(k)}$ avec ses notations on a les deux propositions suivantes :

Proposition 7 *1.*

$$\boxed{|\lambda - m_{h,h}| \leq 1 - m_{h,h}} \text{ et } |\lambda| \leq 1$$

2. $|\lambda - \delta| \leq 1 - \delta$ *Si de plus* $m_{i,i} > 0, \forall i$ *, alors 1 est la seule valeur propre de M de module 1 .*
3. *Ker*$(M - I_n) = Vect(U)$. *Il existe au plus une distribution de probabilité invariante par* M.
4. *Il existe* $(i_0, j_0) \in \|1, n\|^2$ *tel que :* $\alpha_j^{(k+1)} - \alpha_j^{(k)} \geq m_{i_0,j_0}^{(k)}(\beta_j^{(k)} - \alpha_j^{(k)})$
5. *Il existe* $(i_1, j_1) \in \|1, n\|^2$ *tel que :* $\beta_j^{(k+1)} - \beta_j^{(k)} \geq m_{i_1,j_1}^{(k)}(\beta_j^{(k)} - \alpha_j^{(k)})$
6. $\beta_j^{(k+1)} - \alpha_j^{(k+1)} \leq (1 - 2\varepsilon)(\beta_j^{(k)} - \alpha_j^{(k)})$

Proposition 8 *1. la suite matricielle* (M^k) *tend vers la matrice* $B = \begin{pmatrix} b_1 & b_1 & \cdots & b_1 \\ b_2 & b_2 & \cdots & b_2 \\ & & \vdots & \\ b_n & b_n & \cdots & b_n \end{pmatrix}$.

2. B *est stochastique vérifiant* $\forall 1 \leq j \leq n, b_j > 0$.
3. *Soit* $P^{(0)}$ *une distribution de probabilité quelconque.*
   La suite $(P^{(k)} = P^{(0)}M^k)_k$ *converge donc vers UN VECTEUR* P^∞ *unique distribution de probabilité invariante par* M.

II. Méthode de la puissance. Algorithme de Page rank

2.1 Méthode de la puissance

Cette méthode approxime la valeur propre dominante d'une matrice carré (sous certaines conditions). Par exemple dans le cas où A soit une matrice carrée d'ordre n diagonalisable et dont les p valeurs propres vérifient :

$$|\lambda_1| > |\lambda_2| > \ldots > |\lambda_p|.$$

La valeur propre dominante de A c'est λ_1 et un vecteur popre associé sont calculé par la méthode itérative suivante :
Étant donné un vecteur initial arbitraire $x^{(0)} \in \mathbb{R}^n$. On pose $y^{(0)} = \frac{x^{(0)}}{||x^{(0)}||}$ on calcule pour $k = 1, 2 \ldots$

$$\begin{cases} x^{(k)} = Ay^{(k-1)} \\ y^{(k)} = \frac{x^k}{||x^k||} \\ \lambda^{(k)} = (y^{(k)})^T Ay^{(k)} \end{cases}$$

2.2 Algorithme

1. Input : $A \in \mathbf{R}^{n \times n}$ un $u^{(0)} \in \mathbf{R}^n$,
2. Output : La plus grande valeur propre λ_1 et un vecteur propre associé
3. Pour $k = 1, 2, \cdots$ (jusqu'à convergence)
$w^{(k)} = Au^{(k-1)}$, $u^{(k)} = \frac{w^{(k)}}{\|w^{(k)}\|}$
$\lambda^{(k)} = u^{(k)T}\left(Au^{(k)}\right)$

2.3 Analyse de convergence

Theorem 9

$$\lim_{k\to+\infty} u^{(k)} = \frac{u_1}{\|u_1\|} \text{ et } \lambda^{(k)} = u^{(k)T}\left(Au^{(k)}\right) \to \lambda_1$$

avec u_1 un vecteur propre associé à la valeur propre λ_1.

Remark 10 *on peut montrer que pour la méthode de la puissance appliquée à une matrice symétrique, on a l'estimateur d'erreur à l'itération k suivant :*

$$|\lambda^k - \lambda_1| \leq C(A)|\frac{\lambda_2}{\lambda_1}|^{2k}$$

où $C(A)$ est une constante qui dépend de A.

2.4 La matrice de Google :

A un moment donné, on peut considérer que le web est une collection de $N \in \mathbb{N}$, avec N très grand (de l'ordre 10^{10} en octobre 2005). La plupart de ces pages incluent des liens hypertextes vers d'autres pages. On dit qu'elles " pointent " vers ces autres pages. L'idée de base utilisée par les moteurs de recherche pour classer les pages par ordre de pertinence décroissant consiste à considérer que plus une page est la cible de liens venant d'autres pages, c'est-à-dire plus il y a de pages qui pointent vers elle, plus elle a de chances d'être fiable et intéressante pour l'utilisateur final, et réciproquement . Il s'agit donc de quantifier cette idée , c'est-à-dire d'attribuer un rang numérique ou score de pertinence à chaque page.

On se donne donc un ordre arbitraire sur l'ensemble des pages que l'on numérote ainsi $i = 1$ à $i = N$. La structure de connectivité du web peut alors être représentée par une matrice de taille $N \times N$ telle que $c_{ij} = 1$ si la page j pointe sur la page i, $c_{ij} = 0$ sinon . Les liens d'une pages sur elle-même ne sont pas significatifs, on pose donc $c_{ii} = 0$. On observe que la ligne contient tout les liens significatifs qui pointent sur la page i , alors que la colonne j contient les liens significatifs sur la page j.

On souhaite attribuer à chaque page i un score $r_i \in \mathbb{R}_+^*$ de façon à pouvoir classer l'ensemnle des pages par score décroissant et présenter à l'utilisateur une liste ainsi classée des pages correspondants à sa requête.

2.5 Algorithme PageRank :

L'algorithme PageRank part du principe qu'un lien de la page j pointant sur la page i contribue positivement au score de cette dernière, avec une pondération par le score r_j de la page dont est isuu le lien - une page ayant un score élevé a ainsi plus de poids qu'une n'ayant qu'un score médiocre - et par le nombre total de liens présents sur la dite page $N_j = \sum_{k=1}^{N} c_{kj}$. On introduit donc la matrice Q définie par

$$q_{ij} = c_{ij}/N_j \quad \textit{si } N_j \neq 0 q_{ij} = 0 sinon$$

. La somme des coefficients des colonnes non nulles de Q vaut toujours 1. L'application des principes ci-dessus conduit à une équation pour le vecteur $r \in \mathbb{R}^N$ des scores des pages de la forme

$$r_i = \sum_{j=1}^{N} q_{ij} r_j \quad \text{c'est-à-dire } r = Qr$$

Le problème de classement des pages du Web se trouve ainsi ramené à la recherche d'un vecteur propre d'une énorme matrice, associé à la valeur propre 1 !

Il peut arriver que la matrice Q n'admette pas la valeur propre 1 ce qui invalide quelque peu la philosophie originale de l'algorithme. Pour remédier à ce défaut, on considère alors $e =^t (11\dots1) \in R^N$ et $d \in R^N$ tel que $d_j = 1$ si $N_j = 0$, $d_j] = 0$ sinon. La matrice

$$P = Q + \frac{1}{N} e^t d$$

est maintenant la transposée d'une matrice stochastique : ses coefficients sont tous positifs et la somme des coefficients de chaque colonne vaut 1 (remarquons qu'il s'agit d'une " petite " correction de Q car N est très grand). Du fait que ${}^t eP =^t e$, on voit que P admet bien la valeur propre 1.

Comme cette valeur propre est en général multiple, on efectue une dernière modification en choisissant un nombre $0 < \alpha < 1$ et on posant

$$A = \alpha P + (1-\alpha)\frac{1}{N} e^t e.$$

Theorem 11 *la matrice $A = \alpha P + (1-\alpha)\frac{1}{N} e^t e$ est stochastique dont tous les coefficients sont strictement positive.*

Le but est donc de calculer un vecteur propre $r \in \mathbb{R}^N$ tel que r^T est une distribution de probabilité vérifiant : $r = Ar$ ce qui revient à trouver un vecteur dont les N composantes fournissent le classement recherché des page du Web. Le problème est équivalent à trouver un vecteur P^∞ stochastique et vérifiant $P^\infty A = P^\infty$

2.6 Aspects algorithmiques :

2.6.1 Calcul de A^k :

On pourra donner une approximation du vecteur r en approchant P^∞ par $Q^k = QA^k$ pour k assez grand. On pourra utiliser l'algorithme d'exponentiation rapide :

$$\begin{cases} A^0 &= I_N \\ A^k &= (A^2)^{\frac{k}{2}} \ \textit{si } k \ \textit{est pair} \\ A^k &= A(A^2)^{\frac{k-1}{2}} \ \textit{si } k \ \textit{est imppair} \end{cases}$$

2.6.2 Méthode de la puissance :

1. **Input :** $u^{(0)} = e$,
2. **Output :** La plus grande valeur propre λ_1 et un vecteur propre associé
3. Pour $k = 1, 2, \cdots$ (jusqu'à convergence)
$w^{(k)} = Au^{(k-1)}$, $u^{(k)} = \frac{w^{(k)}}{\|w^{(k)}\|}$
$\lambda^{(k)} = u^{(k)T}\left(Au^{(k)}\right)$

La difficulté de la méthode de la puissance dans le cas de la matrice A vient de sa taille gigantesque. La matrice de départ Q est une matrice très creuse pour laquelle des produits matrice-vecteur sont envisageables en pratique. Par contre, la matrice A est pleine, elle contient de l'ordre de 10^{20} éléments et il est hors de question de l'assembler explicitement et encore moins de l'utiliser pour calculer des produits matrice-vecteur !

2.6.3 Algorithme PageRank amélioré :

L'algorithme PageRank exploite le faite que la matrice de départ Q est une matrice très creuse pour laquelle des produits matrice-vecteur sont envisageables en pratique et le calcul suivant :

$$r = Ar = \alpha Pr + (1-\alpha)\frac{1}{N}ee^T r = \alpha Qr + \alpha\frac{1}{N}ed^T r + (1-\alpha)\frac{1}{N}ee^T r \quad (1)$$

La normalisation dans l'algorithme de la méthode de la puissance est inutile :
puisque $\|r^1\|_1 = e^T r^1 = 1$, alors $\|q^2\|_1 = e^T q^2 = 1$
En utilisant la méthode de la puissance pour calculer le PageRank on peut montrer également r peut être calculé par

$$r = \alpha Qr + \frac{1}{N}e - \|\alpha Qr\|_1 e$$

Remarque : Nous n'avons pas besoin de construire d On aboutit de la sorte à un algorithme simple :

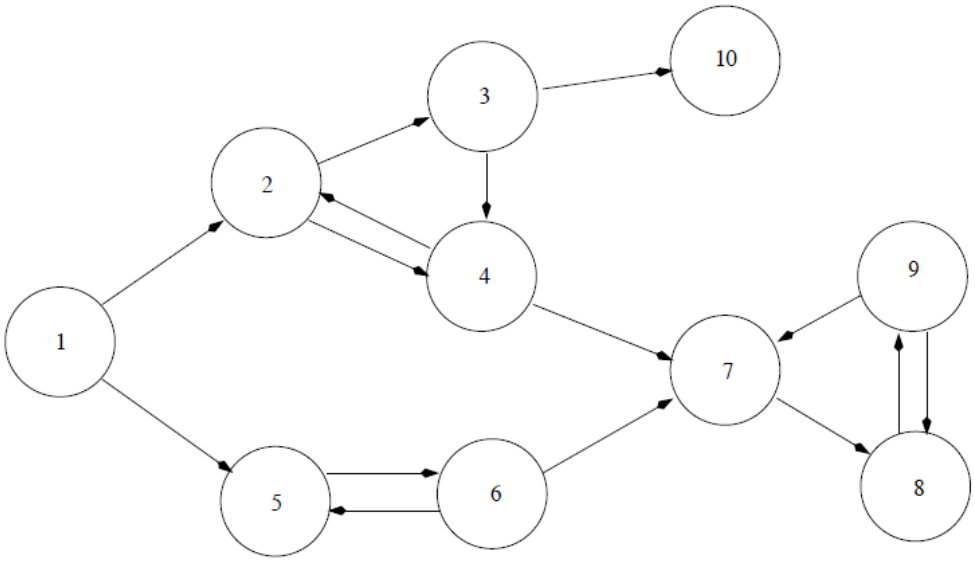
Algorithm 12 - *Choisir* r, *avec* $\| r \|_1 = 1$.

- *Tant que* $s > tol$ *Faire*
  - $\hat{r} = \alpha P^T r$
  - $\beta = 1 - \| \hat{r} \|_1$
  - $r = \hat{r} + \frac{\beta}{N}e$
  - $s = \| r - \hat{r} \|_1$
  - $r = \hat{r}$

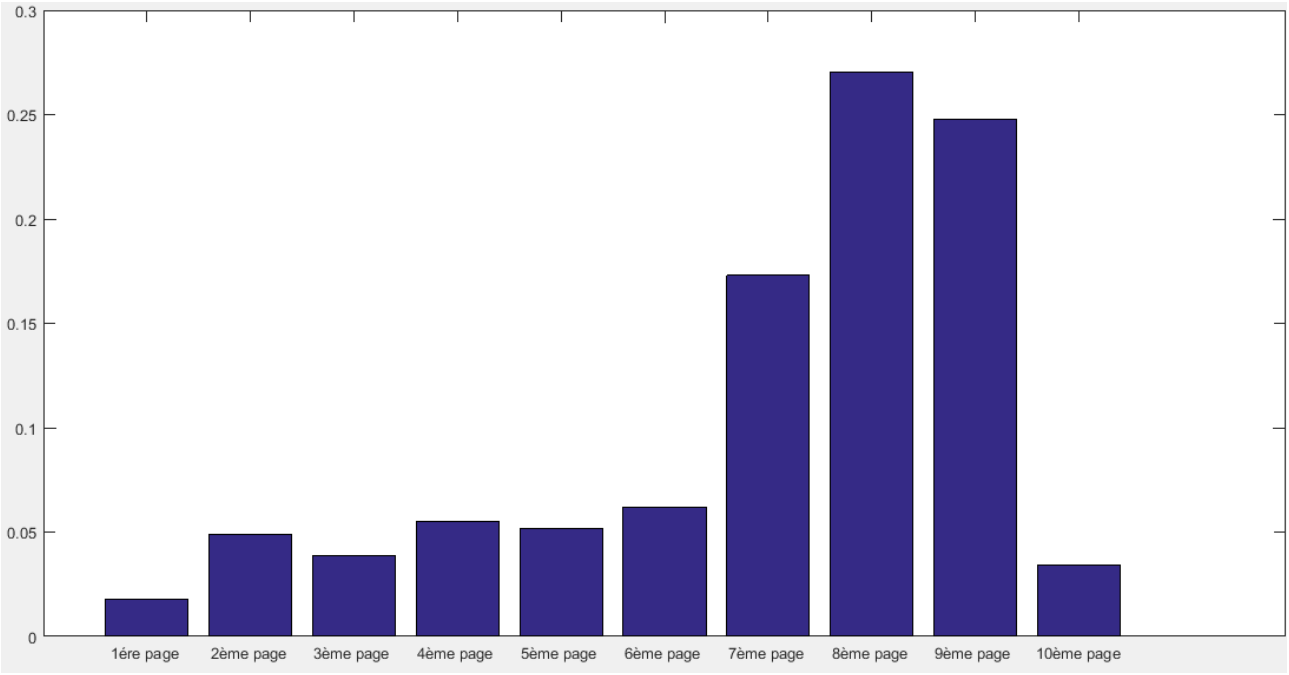
- *Retourner* r

2.7 Simulation numérique à petite échelle

On considère le graphe du web suivant :



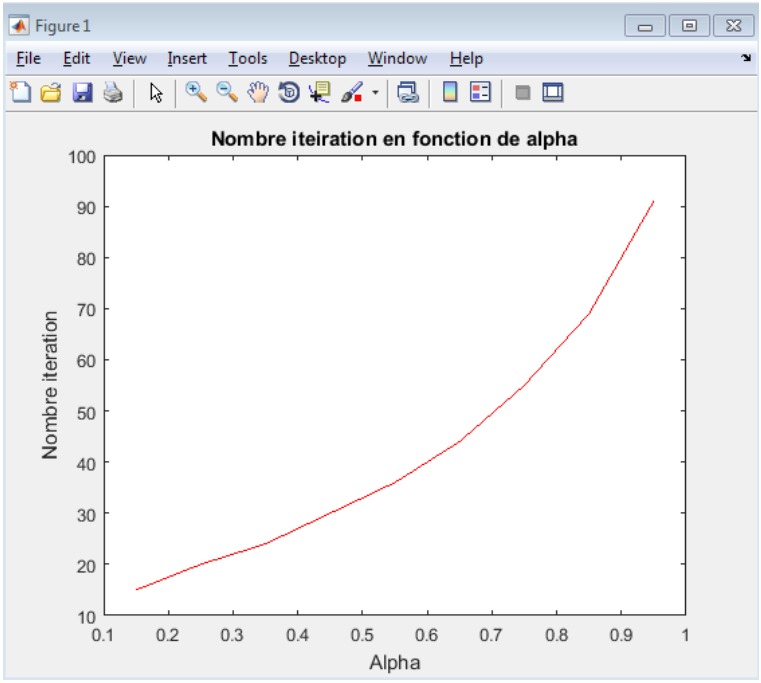
Après programmation de l'algorithme de la puissance pour ce graphe de web dernier qui contient 10 pages on trouve l'ordre d'importance de ces pages comme suite :



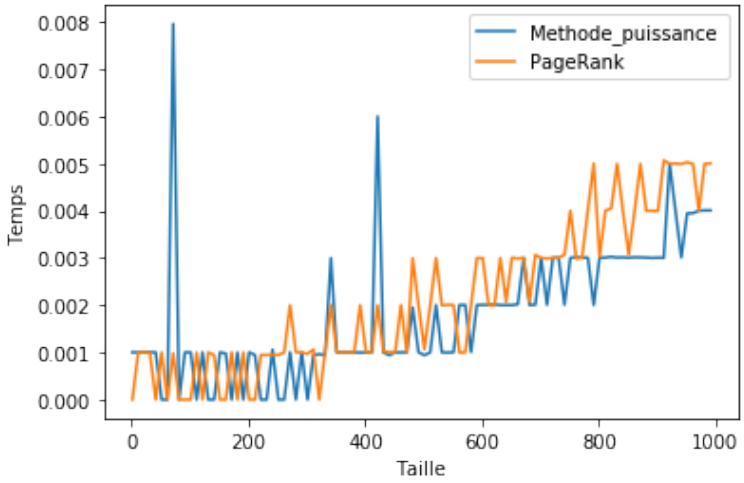
Ce qui nous ramène au classement suivant :

```
Command Window
La 1ére page dans le classement est: la page  8
La 2éme page dans le classement est: la page  9
La 3éme page dans le classement est: la page  7
La 4éme page dans le classement est: la page  6
La 5éme page dans le classement est: la page  4
La 6éme page dans le classement est: la page  5
La 7éme page dans le classement est: la page  2
La 8éme page dans le classement est: la page  3
La 9éme page dans le classement est: la page  10
La 10éme page dans le classement est: la page  1
fx >>
```

Dans cet exemple on prend le même graphe et on calcule le nombre d'itération nécessaire pour obtenir le vecteur PageRank à une précision près, en changeant les valeurs de α.



La figure ci-dessus montre que si les valeurs de α proche de 0 le nombre d'itération devient plus petit ; par exemple il faut seulement 15 itérations pour obtenir le vecteur PageRank avec une précision de 10^{-14} lorsque $\alpha = 0.15$. Dans les mêmes conditions il faut 91 itérations pour $\alpha = 0.95$. par conséquent si α proche de 0 la méthode de la puissance devient très rapide en exécution.



La figure ci-dessus compare le temps de calcul en fonction de la taille de la matrice des algorithms de la puissance et du PageRank.

III. Suggestion pour le développement

Ce paragraphe ne contient qu'un petit nombre de suggestions. Vous pouvez choisir d'étudier certains points seulement, de fa ?çon plus ou moins approfondie, et pas nécessairement dans l'ordre. Vous pouvez aussi vous poser d'autres questions que celles indiquées ci dessous. Il est indispensable que vos investigations comportent une simulation numérique, et si possible, retrouver les représentations graphiques tracés dans le texte.

3.1 Aspect mathématique

Montrer le Théorème 9

Montrer le Théorème 11

Démontrer la Proposition. 7

Démontrer le Théorème. 8

Démontrer le Théorème.6

3.2 Aspect modélisation, calcul numérique et algorithmique

Implémenter la méthode de la puissance et le PagRank et retrouver les simulations données dans le text dans le cas du web paragraphe 2.7

Implémenter la méthode de calcul du vecteur stable par l'exponentiation rapide