

Projet LINFO1114 – Mathématiques discrètes

« Ranking de réseaux sociaux et pages web : PageRank »

| | |
|-------------------|---|
| Professeur | Marco Saerens marco.saerens@uclouvain.be |
| Bureau | b.102 |
| Adresse | Université catholique de Louvain Place des Doyens 1 1348 Louvain-La-Neuve Belgique |
| Assistants | Pierre Leleux p.leleux@uclouvain.be Flore Vancompernelle Vromman flore.vancompernelle@uclouvain.be |
| Date | 8 Novembre 2023 |

Objectif : Le but de ce projet est d'implémenter et de tester un algorithme permettant de classer (to rank) les noeuds d'un graphe en assignant un score d'importance à chacun de ces noeuds. Cet algorithme est celui de PageRank avec téléportation et vecteur de personnalisation. Vous travaillerez par groupe de *trois* étudiant-e-s impérativement (merci de vous inscrire dans un groupe sur Moodle). Chaque groupe trouvera un vecteur de personnalisation différent sur Moodle, sur lequel il devra se baser pour calculer les scores PageRank. La résolution doit être détaillée dans le rapport en utilisant deux méthodes différentes (1) en résolvant un système d'équations linéaires en Python (voir slides du cours) et (2) en implémentant la "power method" en Python à l'aide de Numpy (bibliothèque Python pour le calcul scientifique). Il faudra suivre rigoureusement ces algorithmes, comme détaillé au cours. Vous trouverez toutes les informations utiles dans les slides du cours mais aussi dans des ouvrages de "link analysis" (par exemple l'ouvrage *Google's PageRank and Beyond* de Amy Langville et Carl Meyer).

Fonctions à implémenter (Python 3) :

Dans le cadre de l'implémentation de la résolution du système d'équations linéaires, la signature de la fonction est :

```
def pageRankLinear(A~: np.matrix, alpha~: float, v~: np.array) -> np.array
```

- **Input :** Une matrice d'adjacence¹ **A** d'un graphe dirigé, pondéré et régulier *G*, un vecteur de personnalisation **v**, ainsi qu'un paramètre de téléportation α compris entre 0 et 1 (0.9 par défaut et pour les résultats à présenter). Toutes ces valeurs sont non-négatives.
- **Output :** Un vecteur **x** contenant les scores d'importance des noeuds ordonnés dans le même ordre que les lignes de la matrice d'adjacence (représentant les noeuds).

Dans le cadre de l'implémentation de la power method, la signature de la fonction est :

```
def pageRankPower(A~: np.matrix, alpha~: float, v~: np.array) -> np.array
```

- **Input :** Une matrice d'adjacence **A** d'un graphe dirigé, pondéré et régulier *G*, un vecteur de personnalisation **v**, ainsi qu'un paramètre de téléportation α compris entre 0 et 1 (0.9 par défaut et pour les résultats à présenter).
- **Output :** Un vecteur **x** contenant les scores d'importance des noeuds ordonnés dans le même ordre que les lignes de la matrice d'adjacence (représentant les noeuds).

Vous devez donc calculer les scores PageRank de deux façons différentes :

- En Python, en résolvant un système d'équations linéaires. Pour cette technique, vous pouvez utiliser les fonctions numpy permettant de résoudre un système d'équations linéaires.
- En Python, en calculant le vecteur propre dominant de gauche de la matrice Google, en utilisant la "power method". Pour cette technique, vous devez implémenter vous-même la power method, et donc la boucle permettant de calculer le vecteur propre dominant de gauche de la matrice Google.

Comme nous considérons une téléportation *non-uniforme*, vous devrez calculer la matrice Google à partir de la matrice de probabilités de transitions du graphe et un vecteur de personnalisation non-uniforme. Notez que, dans l'algorithme basé sur la "power method", il faudra initialiser les scores par le degré entrant (indegree) de chaque noeud (la somme de chaque colonne *j* de la matrice d'adjacence, $\sum_{i=1}^n a_{ij}$).

Ainsi que déjà mentionné, évitez d'utiliser les fonctions permettant de calculer directement des vecteurs propres² car le but du projet est justement de comprendre et d'implémenter la power method. Donc, dans ce cadre, vous ne pouvez utiliser cette librairie que pour effectuer des opérations matricielles simples telles que la transposée, la multiplication matricielle, etc. Par contre, pour la résolution à l'aide d'un système d'équations, nous vous autorisons à utili-

1. Aussi parfois notée **W** (pour Web ou Weight) dans les slides du cours.

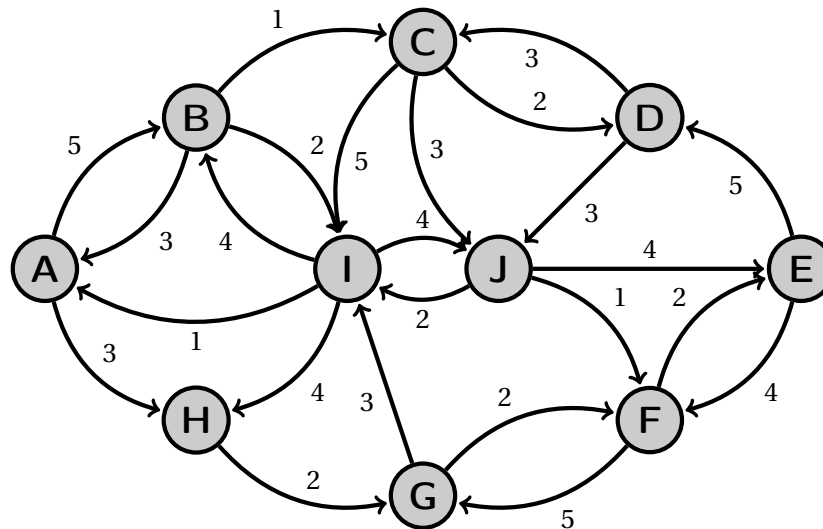
2. Par contre, vous pouvez les utiliser pour vérifier les résultats de votre implémentation.

ser des fonctions spécialisées.

De plus, il vous est demandé d'ajouter à votre code une fonction "main" qui lit un fichier csv contenant la matrice d'adjacence A du graphe présenté ci-dessous (séparez vos valeurs par une virgule et chaque ligne de la matrice par un passage à la ligne), qui exécute le calcul de PageRank (via vos fonctions `pageRankPower` et `pageRankLinear`) et imprime les résultats à l'écran.

Données :

– Un graphe dirigé et pondéré représentant un réseau de 10 noeuds.



– Ainsi qu'un vecteur de personnalisation propre à chaque groupe disponible sur Moodle dans le dossier "Vecteur".

Rapport : Le rapport est un fichier PDF (8 pages maximum tout inclu; écrit en LaTeX) qui se compose de :

- Un rappel théorique expliquant brièvement (avec équations; environ une page) comment calculer le vecteur de scores PageRank (en résolvant un système d'équations linéaires et en implémentant la power method). Discuter également l'impact du paramètre de téléportation α et du vecteur de personnalisation \mathbf{v} sur le score.
- La présentation numérique en LaTeX du système d'équations linéaires permettant de calculer le score PageRank du graphe ci-dessus avec le vecteur de personnalisation qui vous a été assigné.

Code complet : Le code complet comprendra deux fichiers .py, un fichier contenant vos fonctions pour la power method et la résolution du système d'équations linéaires, et un fichier "main" qui lit vos fichiers csv (matrice d'adjacence et vecteur de personnalisation) et calcule les score PageRank en faisant appel à vos fonctions définies dans l'autre fichier. Votre fichier contenant les fonctions doit donc contenir au moins ces deux fonctions :

- **La power method** : Dans le cadre de l'implémentation en Python de la power method, nous vous demandons plusieurs impressions à l'écran lorsque le code est exécuté :
 - de la matrice d'adjacence **A**,
 - de la matrice de probabilités de transition **P**,
 - de la matrice Google **G**,
 - des trois premières itérations de la power method (vecteur de scores), et
 - du résultat final, c'est à dire le score PageRank final après convergence (output de la fonction).
- **La résolution de système d'équations linéaires** : Dans le cadre de l'implémentation en Python de la résolution du système d'équations, nous demandons l'impression à l'écran :
 - du résultat final, c'est à dire le score PageRank final obtenu (output de la fonction).

Attention, respectez bien ces consignes car nous nous baserons sur celles-ci pour calculer la note du projet de manière semi-automatique. N'oubliez pas de bien commenter vos codes!

Langage de programmation : L'implémentation devra impérativement être codée en Python 3 en respectant les consignes et les signatures énoncées ci-dessus. N'imprimez pas de valeurs intermédiaires inutiles.

Comme déjà mentionné, vous devez utiliser une librairie Python externe de calcul et de manipulation matricielle/vectorielle nommée numpy. Cette librairie vous évitera d'implémenter la multiplication matricielle, ou la transposition, de manière à vous concentrer sur l'algorithme proprement dit.

Evaluation et consignes : Le projet est obligatoire et à réaliser par groupes de trois étudiantes et étudiants. L'évaluation portera sur le contenu du rapport (maximum 8 pages) et le code (lisibilité, structure, **commentaires**,...) et comptera pour 3 points sur 20 dans la note finale (le reste des points étant donné par l'examen écrit). Le rapport doit être très professionnel³, du type article scientifique ou technique, et doit contenir les références bibliographiques sur lesquelles vous vous êtes basés.

Les différents fichiers, c'est-à-dire les fichiers de code source (un fichier pour le main qui s'exécute et un fichier contenant les fonctions), les fichiers .csv (la matrice d'adjacence et le vecteur de personnalisation) et le rapport en pdf, tous compressés ensemble (nom du fichier compressé : "groupe" suivi du numéro du groupe (deux chiffres), et suivi par les noms de famille (majuscule à la première lettre) des membres du groupe séparés par des underscores et par ordre alphabétique⁴. Nous ne corrigeons pas les projets qui ne respectent pas cette consigne : vous devrez re-soumettre le projet avec pénalités.), sont à remettre sur Moodle impérativement pour le samedi 23 décembre 2023, avant 23h59. Aucun retard ne sera accepté, dès lors soyez prévoyant et évitez de remettre votre projet dans les dernières minutes. La note sera la même pour tous les membres du groupe.

Bon travail!

3. Par exemple pas de copier/coller d'images de formule mathématique ou algorithme.

4. Par exemple "groupe05_Leleux_Saerens_Vancompernelle.zip".
