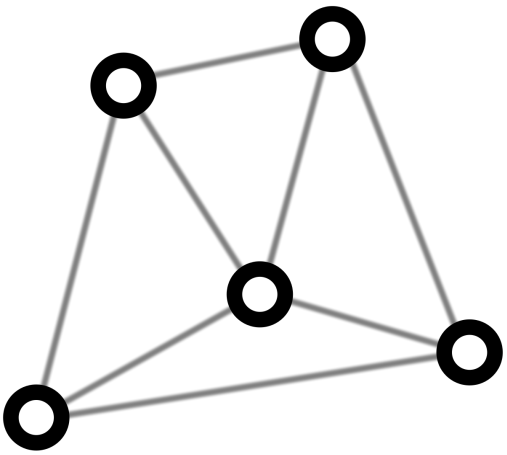
Projet Big Data

Page Rank



DORE Martin
MAILLE Augustin

Tuteur : Maxence Vandromme

Sommaire

Sommaire **1**
Introduction **1**
PageRank Simple **2**
PageRank personnalisé **4**
Conclusion **6**

Introduction

WikiSpeedia est un jeu dont le but est de partir d'une page aléatoire de l'encyclopédie en ligne Wikipedia et d'atteindre uniquement en cliquant sur des liens vers d'autres pages, une page donnée. L'objectif de ce projet est d'utiliser les données des parties réussies de *WikiSpeedia* pour classer les pages du site en utilisant différentes méthodes.

PageRank Simple

Dans un premier temps, nous implémentons la méthode de PageRank sur le jeu de données *path_finished.csv*. Voici différents résultats pour différents Dumping Factor β. Pour chaque essai, on affiche les dix premières pages avec leurs scores associés.

- β = 0.85

L'algorithme converge au bout de **9** itérations et donne le résultat suivant :

Nom de la page	**Score**
United_States	0.031251865432075185
Europe	0.01490611372686973
United_Kingdom	0.013235928189212098
England	0.011582936125988597
Africa	0.00971789999274262
Earth	0.00846155599558927
World_War_II	0.008104412816698062
Germany	0.006154445397740782
North_America	0.005805040282128763
France	0.00547608421318531

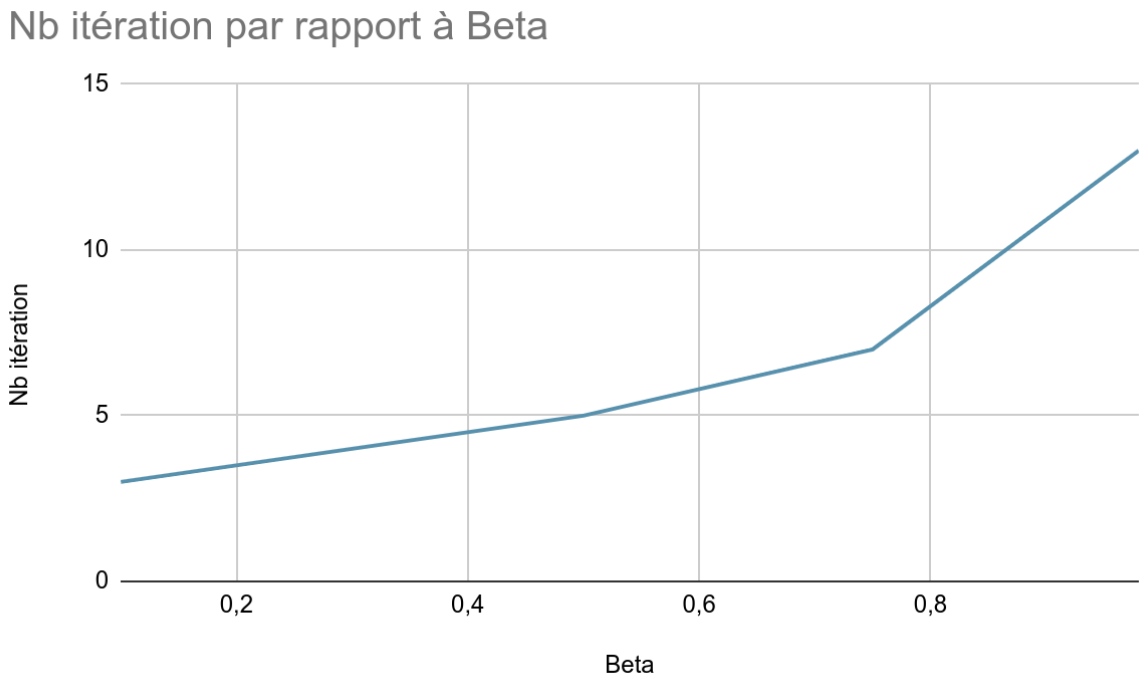
- $\beta = 0.5$

L'algorithme converge au bout de **5** itérations et donne le résultat suivant :

Nom de la page	**Score**
United_States	0.02246299133803719
Europe	0.010259714320667856
England	0.00926041897211565
United_Kingdom	0.009215791745954322
'Africa	0.00584046099157259
World_War_II	0.00549170135012363
Earth	0.004641879964669297
France	0.004334980206666213
Germany	0.004116310440345009
North_America	0.004042953502689022

Entre ces deux itérations de PageRank, on peut voir que les dix pages les plus importantes restent les mêmes, seul l'ordre change.
On remarque que les pages les plus importantes sont toujours sur un thème assez générique comme *Earth* ou *Europe*. Ce résultat est dû au fait que pour atteindre leur objectif les joueurs doivent passer par des pages dites *hub*. Ces pages ne contiennent pas d'information particulière mais permettent d'accéder à de nombreuses autres pages. Par exemple, on peut facilement visualiser que la page *France* permet à la fois d'accéder à de nombreux faits historiques, des lieux touristiques, des personnages célèbres, de la gastronomie, etc… dès l'instant que ces derniers ont un lien avec le pays. La page *World War II* quant à elle figure dans le classement car il s'agit d'un des évènements les plus importants de l'Histoire récente. On peut donc imaginer que l'encyclopédie conserve de nombreux éléments à ce sujet et que de nombreuses pages traitent en réalité d'un thème proche de la seconde guerre mondiale.

On décide de regarder l'évolution du nombre d'itération par rapport au dumping factor choisi.



PageRank personnalisé

On décide désormais d'implémenter l'algorithme en spécifiant des pages par lesquelles les joueurs doivent passer.
Une bonne manière de visualiser cet algorithme est de spécifier des pages qui abordent un thème identifiable.

- PageRank Perso en indiquant les pages *Zinc* et *Salt:*

L'algorithme converge au bout de 12 itérations et donne :

Nom de la page	**Score**
Zinc	0.10194119556069817
Salt	0.10164642132153409
Periodic_table	0.03582408488247967
Edible_salt	0.02173948059726679
United_States	0.015074437929477548
Chemical_element	0.012917612800323393

Brain	0.01190173694133985
Sodium	0.008908625464408307
Chemistry	0.008858209478201894
Cell_%28biology%29	0.008741215440799487

On remarque dans un premier temps que les deux pages personnalisées sont évidemment les plus visitées. De plus la plupart des pages traitent d'un sujet en rapport avec la chimie (car le sel et le zinc sont deux éléments chimiques).
La seule page qui sort du lot est la page *United States* qui sert de nouveaux de *hub* aux joueurs pour atteindre leur réel objectif.

- idem mais en spécifiant la page *Virus* :

L'algorithme converge également au bout de 12 itérations.

Nom de la page	**Score**
Virus	0.2103188906594143
AIDS	0.04196060024369515
Common_cold	0.02844989941265026
United_States	0.019083575291487336
Rabies	0.01863167158455767
Bacteria	0.01849841036348958
Plant	0.011452453744103993
Human	0.011097451221505288
Cell_%28biology%29	0.010709662789563553
Animal	0.01033815314175664

On peut globalement tirer les mêmes conclusions que l'étude précédente ci ce n'est que le thème abordé est celui de la biologie.

Bonus

Nous avons fait une version prenant en compte les chemins non finis.
Voici les résultats trouvés pour page rank avec un beta 0.85.

Nom de la page	**Score**
United_States	0.0381601877980165
'Europe'	0.018756409290101177
'United_Kingdom'	0.01714662697847813
England	0.013577733846819578
'Earth'	0.01110156072505985
'Africa'	0.010287710135224823
'North_America'	0.008329857728646247
'World_War_II'	0.007738788443065872
'Human'	0.007107339646667465
'Animal'	0.006952859980405481

Les résultats diffèrent un peu en bas de tableau de la version avec seulement les chemins complets. Ce qui est normal car on a maintenant plus d'informations.
On peut vérifier les mêmes résultats avec le page rank personnalisé.

Conclusion

Pour conclure, l'utilisation de la méthode de PageRank sur les données de WikiSpeedia permet d'identifier les pages les plus importantes de l'encyclopédie. La méthode personnalisée quant à elle permet de mettre l'accent sur une page ou un thème en particulier.
Certaines pages notamment *United States* figurent dans quasiment tous les classements, peu importe le thème choisi car elles servent de carrefour au joueur pour faire le lien entre deux thèmes.