

**Faculté des Sciences Économiques Et de Gestion de Nabeul**

Rapport de projet

**« Réseaux sociaux »**

Thème :

**Political blogosphere**

Élaborée par :

Harzallah Soumaya

1. **Collecte des données:**

* Les données sur l’orientation politique proviennent des annuaires de blogs, comme indiqué. Quelques blogs ont été étiquetés manuellement, en fonction des liens et des messages entrants et sortants à l’époque de l’élection présidentielle de 2004. Les étiquettes dérivées d’un répertoire sont sujettes à des erreurs .Les étiquettes manuelles le sont encore plus.

1. Identifier une source de données en ligne :

Blogosphère politique févr. 2005.

Données compilées par Lada Adamic et Natalie Glance.

Les données sont sous forme d’un fichier GML.

1. Identifier les entités (nœuds) et les relations entre elles (liens):

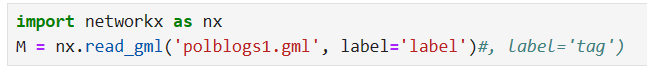
* Un réseau dirigé d’hyperliens entre des blogs sur la politique américaine.
* Les nœuds : les blogs.
* Les liens entre les blogs ont été automatiquement extraits d’un crawl de la première page

de la page du blog.

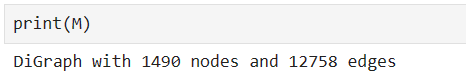
1. Identifier les informations additionnelles valables (comme les attributs des nœuds, les poids des liens, etc) :

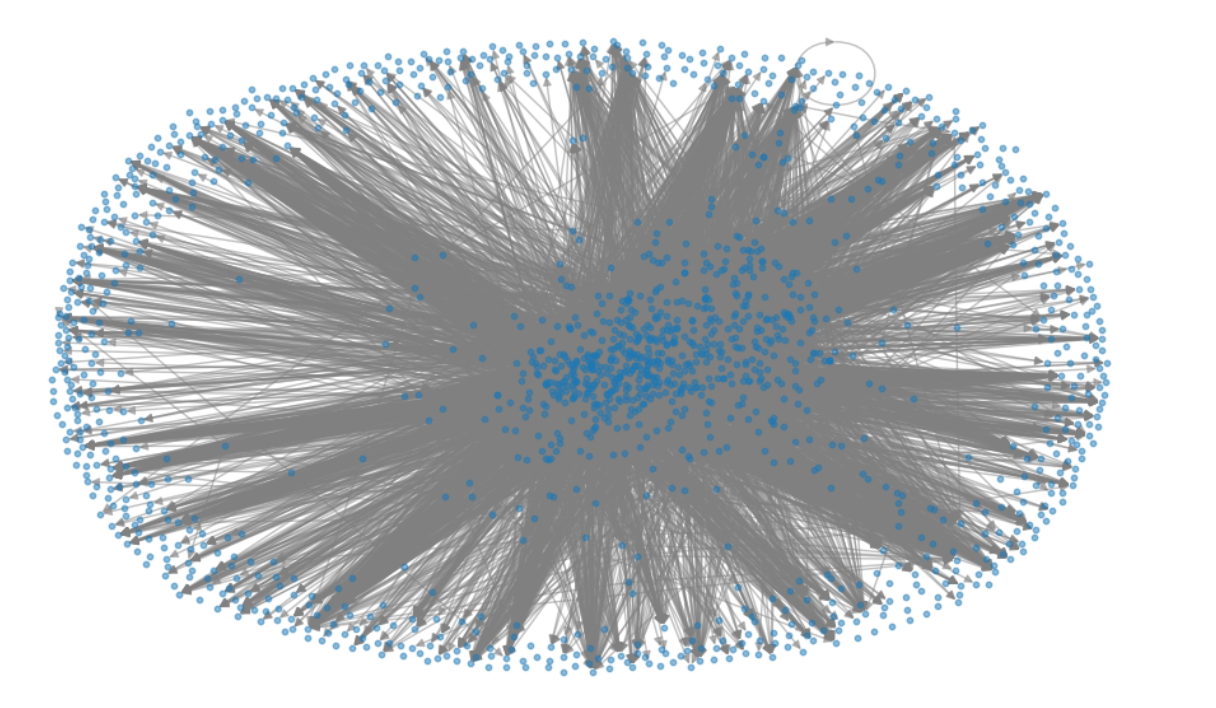
Les attributs de « valeur » du nœud indiquent l’orientation politique en fonction de : 0 (gauche ou libéral) ou 1 (droite ou conservatrice)

1. Obtenir les données à partir de la source de données sélectionnée :



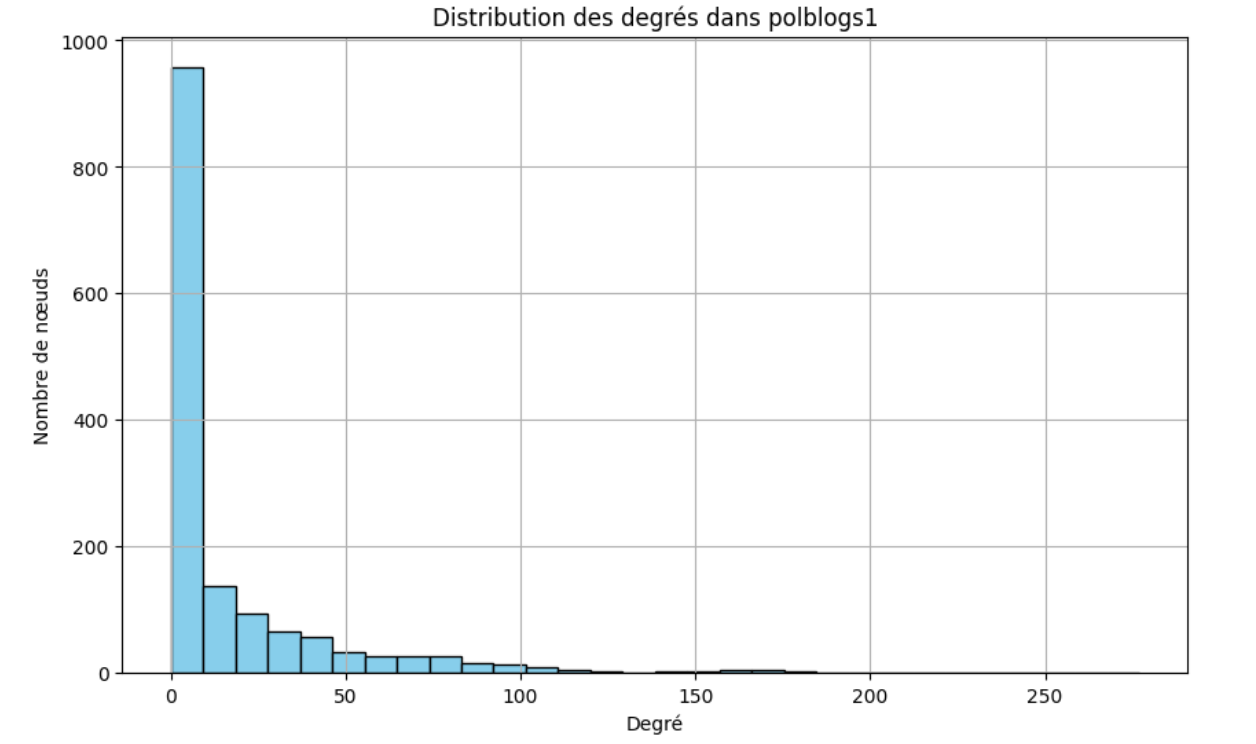
1. Construire un réseau à partir des données:



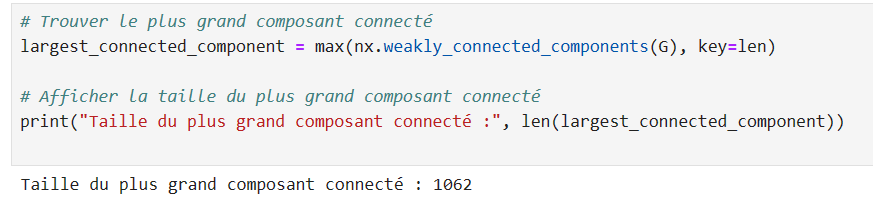


1. **Analyse du réseau:**

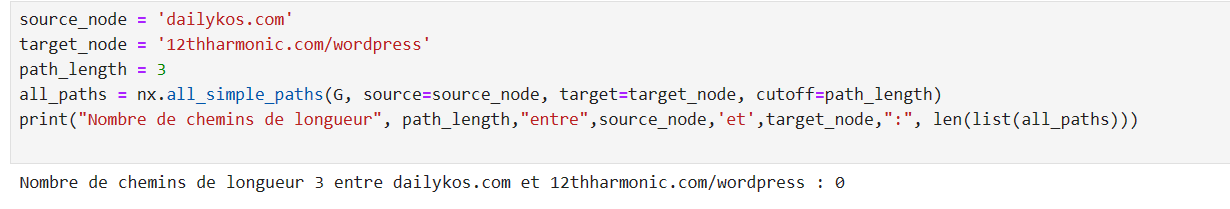
 L’analyse de la distribution des degrés:

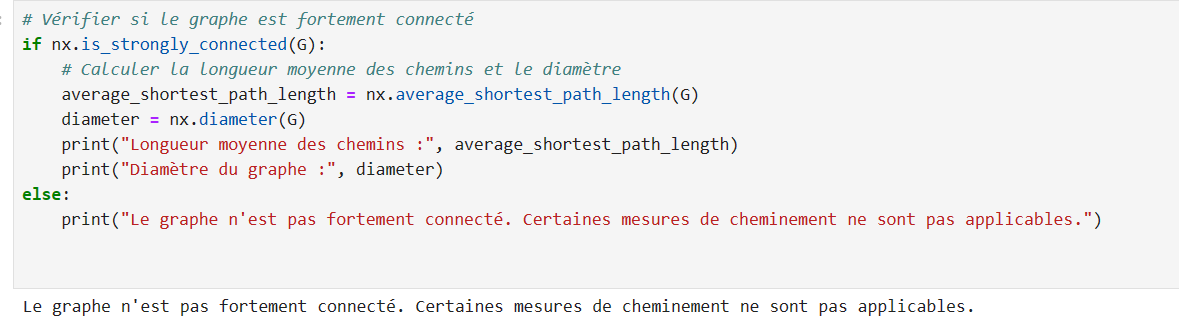


 L’analyse des composants connectés :

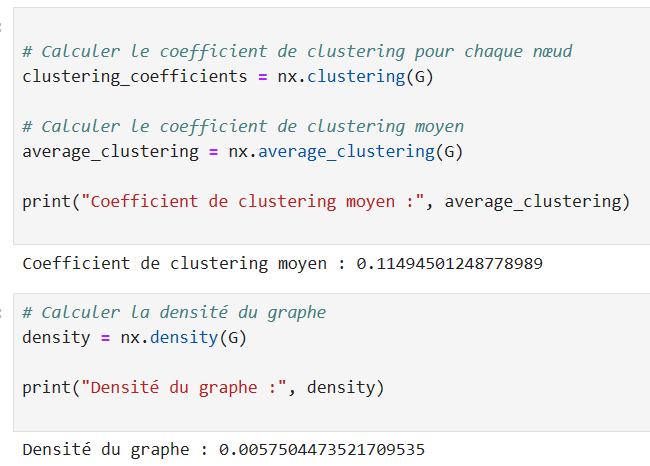


 L’analyse des chemins :

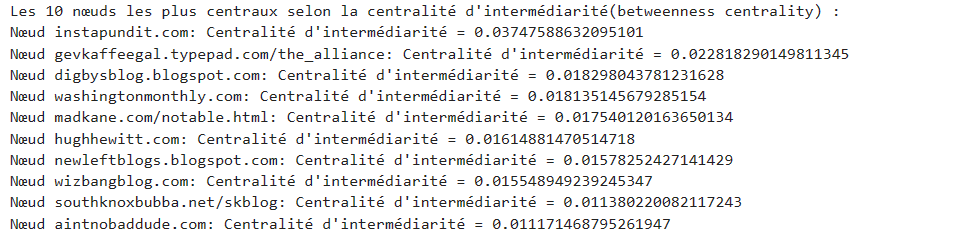




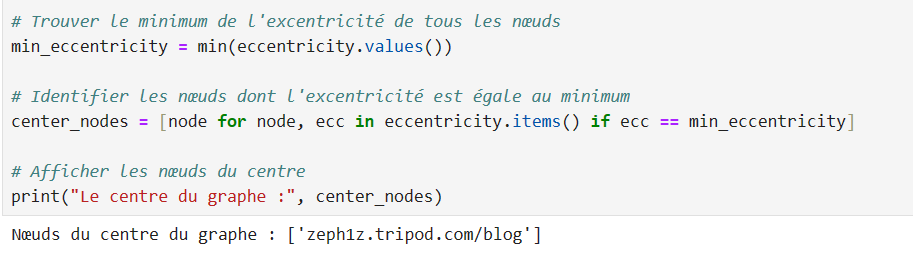
 Le coefficient de clustering et l’analyse de la densité:



 L’analyse de la centralité:

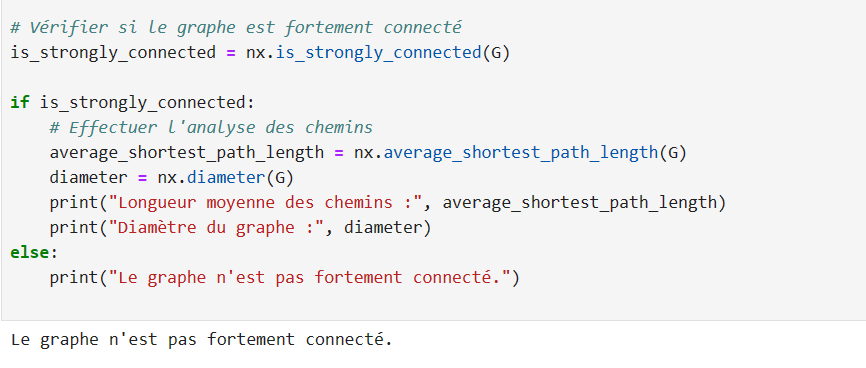


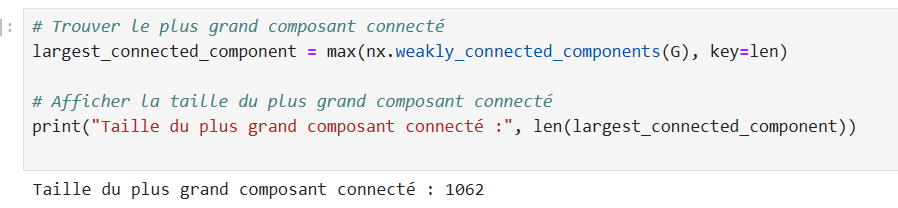
* Autres mesures:







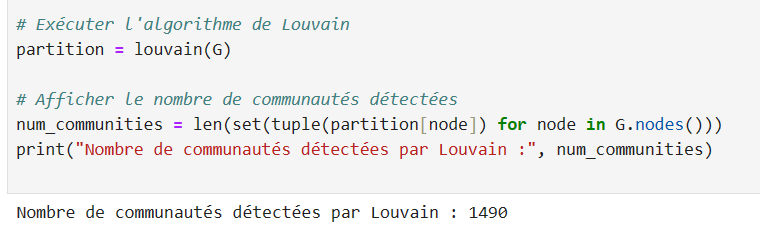




1. **Identification des communautés:**
2. **Algorithme K-clique:**



1. **Algorithme de Louvain:**

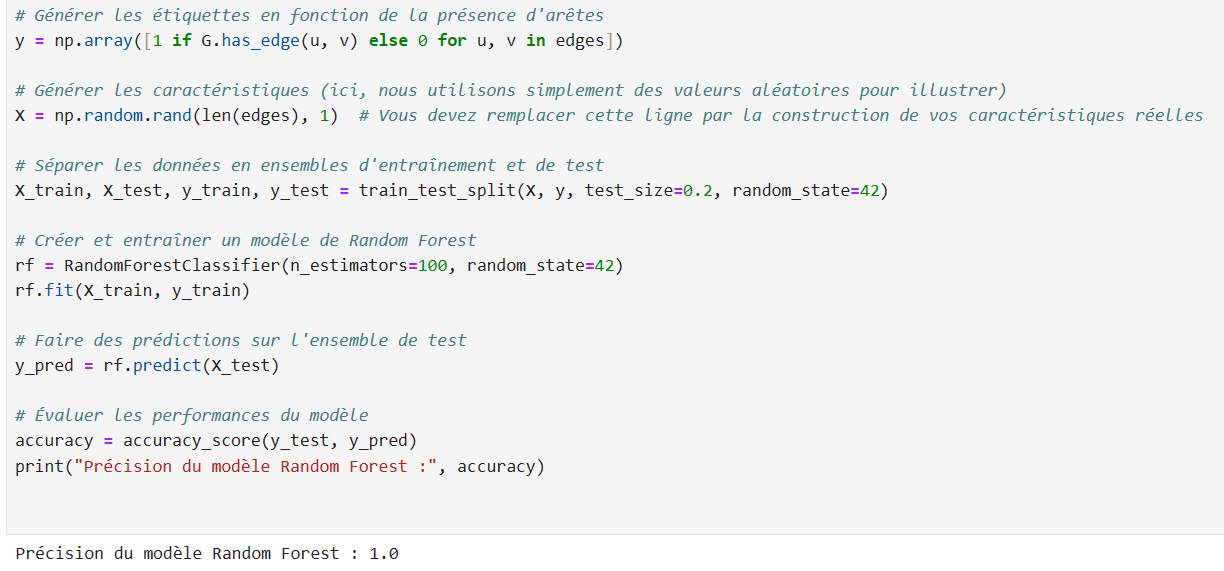


1. **Algorithme basé sur la modularité:**



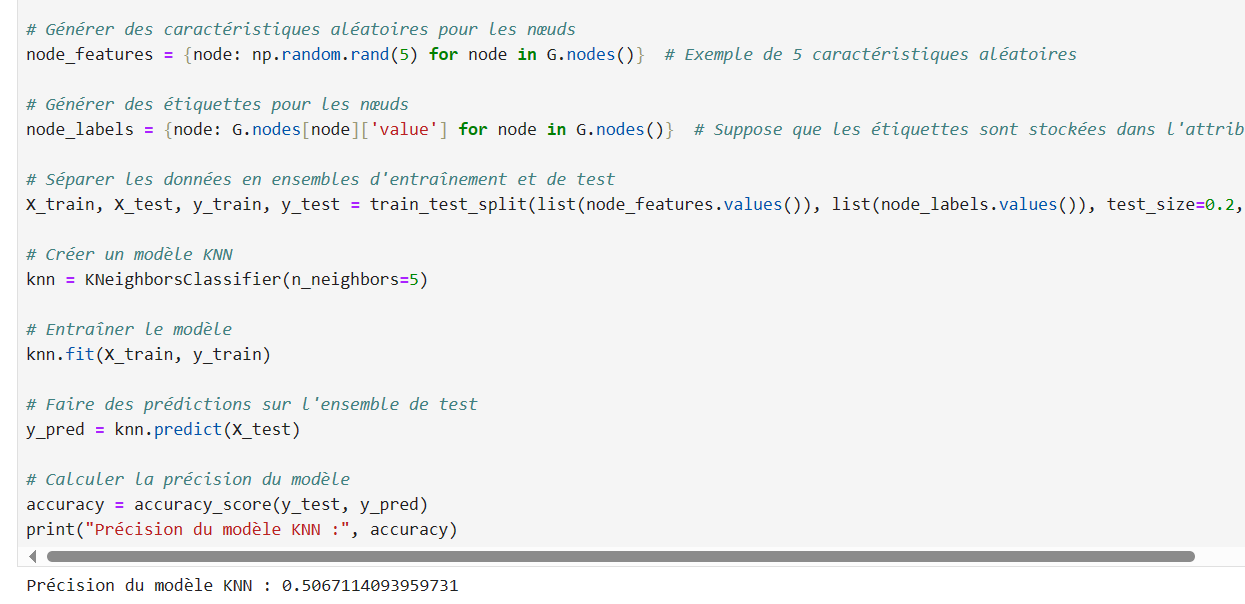
1. **Prédiction des liens:**
2. Forêts aléatoires :

* Les forêts aléatoires sont des ensembles d'arbres de décision où chaque arbre est formé sur un échantillon de données et avec un sous-ensemble aléatoire des caractéristiques. Les prédictions sont agrégées à partir de tous les arbres.



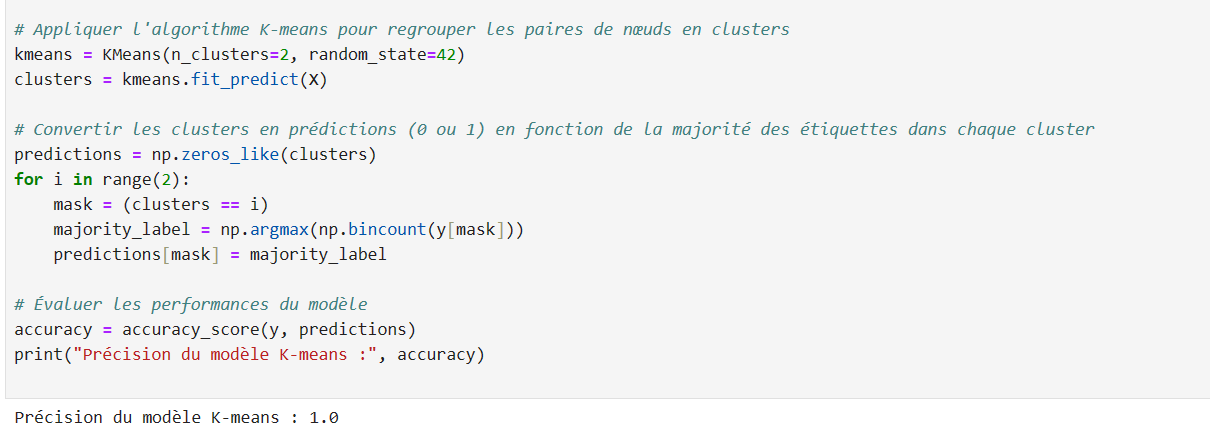
1. k-nearest neighbors (KNN) :

* L'algorithme KNN (K-Nearest Neighbors) est un algorithme d'apprentissage automatique utilisé pour la classification et la régression. Il fonctionne en trouvant les k voisins les plus proches d'un point de données non étiqueté dans l'espace des caractéristiques, puis en utilisant la classe majoritaire (dans le cas de la classification) ou la moyenne (dans le cas de la régression) des étiquettes des voisins pour prédire l'étiquette du point de données non étiqueté.



1. K-means :

* L'algorithme K-means est une méthode de clustering largement utilisée en apprentissage automatique non supervisé. Son objectif est de diviser un ensemble de données en k groupes en fonction de leurs caractéristiques. Chaque cluster est représenté par son centroïde, qui est la moyenne des données dans ce cluster.



Interprétation: Le modèle KNN n’est pas approprié dans ce problème.

1. Prédiction des liens manquants avec la similarité de Jaccard:



Interprétation: Il y’a des liens qui ne sont pas prédits mais réellement peuvent exister.