Trabajo Práctico 1: Noche de los museos y PageRank

Álgebra Lineal Computacional

Lic. en Ciencias de Datos
Facultad de Ciencias Exactas y Naturales
Universidad de Buenos Aires

1er Cuatrimestre 2025

Contexto - Motores de búsqueda

¿Qué debe hacer un motor de búsqueda? ¿Qué tareas debe poder resolver?

Contexto - Motores de búsqueda

¿Qué debe hacer un motor de búsqueda? ¿Qué tareas debe poder resolver?

- Explorar la red e identificar todas las páginas con acceso público.
- Almacenar la información obtenida, para realizar búsquedas eficientemente.
- Determinar un **orden** de las páginas según su **importancia**, para presentar la información con un **orden de relevancia**.

En el TP nos centraremos en este último ítem.

Cómo determinar un orden de importancia

- ¿Qué significa que una página sea importante?

Cómo determinar un orden de importancia

- ¿Qué significa que una página sea importante?

  "Automated search engines that rely on keyword matching usually return too many low quality matches."

Cómo determinar un orden de importancia

- ¿Qué significa que una página sea importante?

  "Automated search engines that rely on keyword matching usually return too many low quality matches."

- ¿Como podemos utilizar la estructura de la web para ayudarnos?

Cómo determinar un orden de importancia

- ¿Qué significa que una página sea importante?

  "Automated search engines that rely on keyword matching usually return too many low quality matches."

- ¿Como podemos utilizar la estructura de la web para ayudarnos?

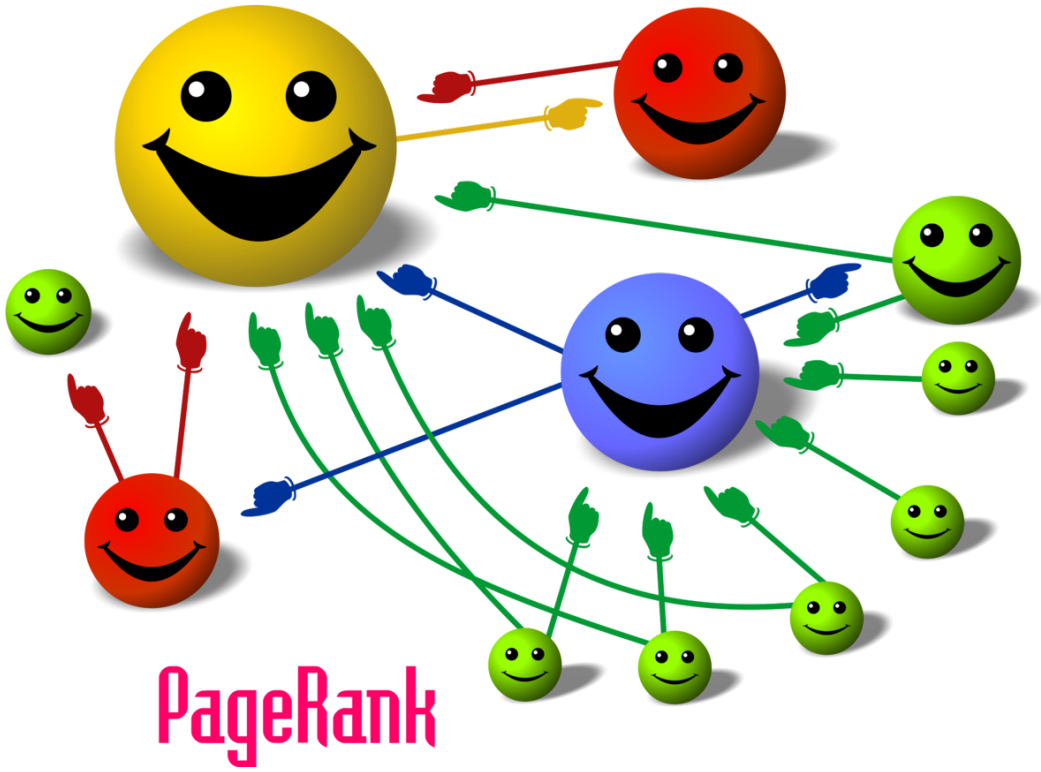
  "The citation (link) graph of the Web is an important resource that has largely gone unused in existing Web search engines."

Usando el grafo de la web

- Vamos a usar los links de la web para ayudarnos
- ¿Importará la cantidad de links? ¿Cómo? ¿Todos los links *valen* los mismo?

Usando el grafo de la web

- Vamos a usar los links de la web para ayudarnos
- ¿Importará la cantidad de links? ¿Cómo? ¿Todos los links *valen* los mismo?
- Contaremos la **cantidad** y **calidad** de los links que apuntan a una determinada página.



Pagerank - El modelo

- Sea el conjunto de páginas web: $\{1, \ldots, N\}$.
- Definimos la **matriz de conectividad** $\boldsymbol{A} \in \{0,1\}^{N \times N}$ como:

$$A_{ij} = \begin{cases} 1 & \text{si la página } i \text{ tiene un link hacia la página } j \\ 0 & \text{si no} \end{cases}$$

- Además, ignoramos los autolinks, $A_{ii} = 0 \ \forall i \in A$
- Informalmente: "En la fila i están las webs apuntadas por i, y en la columna j están las webs que apuntan a j."

Pagerank - El modelo

- Sea el conjunto de páginas web: $\{1, \ldots, N\}$.
- Definimos la **matriz de conectividad** $\boldsymbol{A} \in \{0,1\}^{N \times N}$ como:

$$A_{ij} = \begin{cases} 1 & \text{si la página } i \text{ tiene un link hacia la página } j \\ 0 & \text{si no} \end{cases}$$

- Además, ignoramos los autolinks, $A_{ii} = 0 \ \forall i \in A$
- Informalmente: "En la fila i están las webs apuntadas por i, y en la columna j están las webs que apuntan a j."
- Para una página i, definimos su **grado** como:

$$k_i = \sum_{j=1}^{N} A_{ij}$$

Es decir, la cantidad de links que *salen* de i.

¿Cómo definimos puntajes?

- Llamamos $\mathbf{p}_i$ al **puntaje** asignado a la página i.

¿Cómo definimos puntajes?

- Llamamos $\mathbf{p}_i$ al **puntaje** asignado a la página i.
- Dadas $i, j \in \{1 \dots N\}$, el **aporte** del link $j \longrightarrow i$ a la página i como:

$$\frac{\mathbf{p}_j}{k_j} A_{ji}$$

.

- Es decir, j le aporta a i su puntaje ponderado por cuántos links salientes tiene. Si no hay link o $k_j = 0$ le aporta cero.

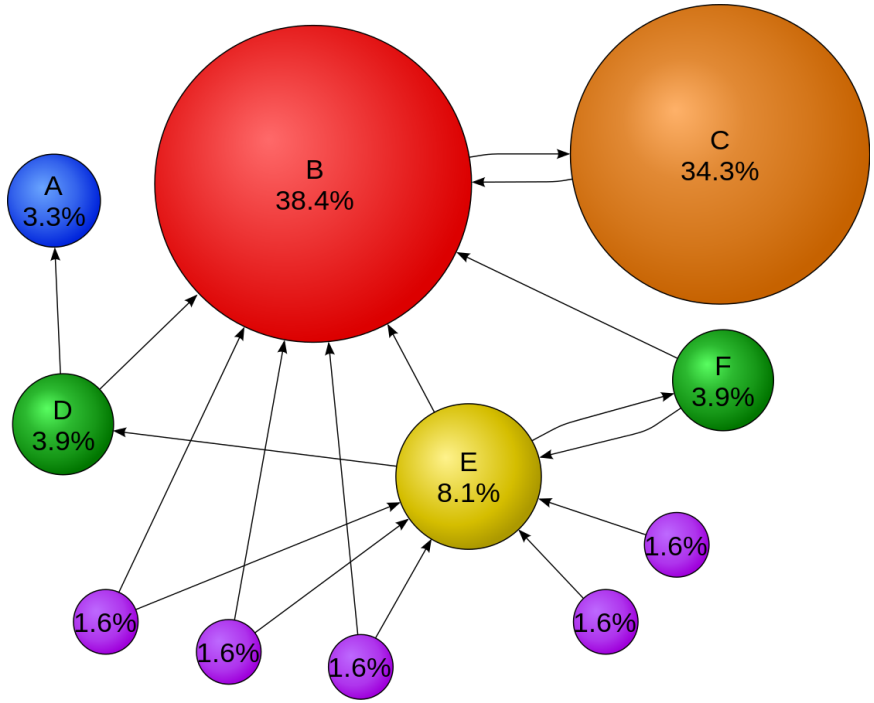
¿Cómo definimos puntajes?

- Llamamos $\mathbf{p}_i$ al **puntaje** asignado a la página i.
- Dadas $i, j \in \{1 \dots N\}$, el **aporte** del link $j \longrightarrow i$ a la página i como:

$$\frac{\mathbf{p}_j}{k_j} A_{ji}$$

.

- Es decir, j le aporta a i su puntaje ponderado por cuántos links salientes tiene. Si no hay link o $k_j = 0$ le aporta cero.



¿Cómo calculamos puntajes?

El puntaje de la página i debe ser igual al puntaje aportado por todas las páginas que le puntan:

$$\mathbf{p}_i = \sum_{j=1}^{N} \frac{\mathbf{p}_j}{k_j} A_{ji}$$

El puntaje de una página depende del puntaje de otras.

¿Cómo calculamos puntajes?

El puntaje de la página i debe ser igual al puntaje aportado por todas las páginas que le puntan:

$$\mathbf{p}_i = \sum_{j=1}^{N} \frac{\mathbf{p}_j}{k_j} A_{ji}$$

El puntaje de una página depende del puntaje de otras.
Luego, se define la matriz de transferencia $C = A^t K^{-1}$, donde K^{-1} es una matriz diagonal de la forma:

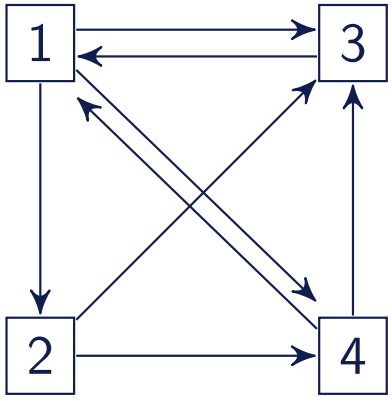
$$K^{-1}_{ii} = \begin{cases} 1/k_i & \text{si} \quad k_i \neq 0 \\ 0 & \text{si} \quad k_i = 0 \end{cases},$$

Lo cual nos permite calcular el ranking de todas las páginas como:

$$\boxed{C\ \mathbf{p} = \mathbf{p}} \quad \text{donde } \mathbf{p} = (p_1, \ldots, p_i, \ldots, p_N)^t$$

Finalmente, tomaremos como solución al vector **p** normalizado para que sume 1.

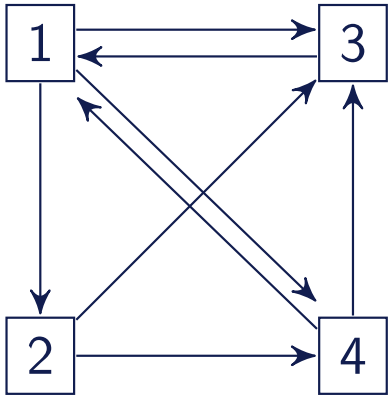
Veamos un ejemplo



$$C = \begin{bmatrix} 0 & 0 & 1 & \frac{1}{2} \\ \frac{1}{3} & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{3} & \frac{1}{2} & 0 & 0 \end{bmatrix} \qquad \mathbf{p} = \begin{bmatrix} \mathbf{p}_1 \\ \mathbf{p}_2 \\ \mathbf{p}_3 \\ \mathbf{p}_4 \end{bmatrix}$$

$$k_1 = 3,\ k_2 = 2,\ k_3 = 1,\ k_4 = 2$$

Veamos un ejemplo



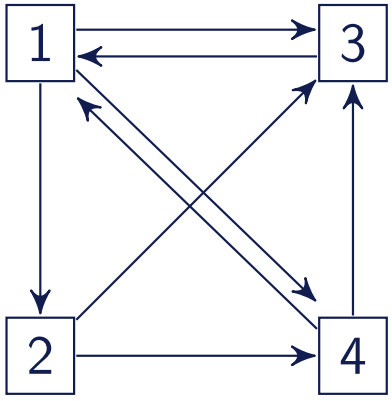
$$C = \begin{bmatrix} 0 & 0 & 1 & \frac{1}{2} \\ \frac{1}{3} & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{3} & \frac{1}{2} & 0 & 0 \end{bmatrix} \qquad \mathbf{p} = \begin{bmatrix} \mathbf{p}_1 \\ \mathbf{p}_2 \\ \mathbf{p}_3 \\ \mathbf{p}_4 \end{bmatrix}$$

$$k_1 = 3,\ k_2 = 2,\ k_3 = 1,\ k_4 = 2$$

Si **p** es el vector con los puntajes de las páginas:

$$(C\mathbf{p})_i = fila_i(C)\mathbf{p}$$

Veamos un ejemplo



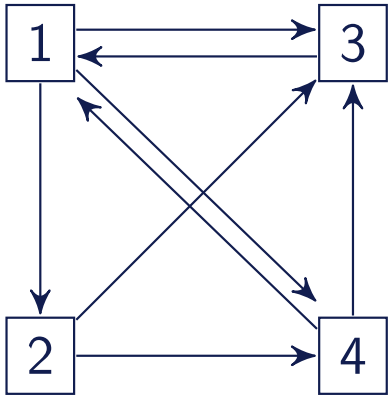
$$C = \begin{bmatrix} 0 & 0 & 1 & \frac{1}{2} \\ \frac{1}{3} & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{3} & \frac{1}{2} & 0 & 0 \end{bmatrix} \qquad \mathbf{p} = \begin{bmatrix} \mathbf{p}_1 \\ \mathbf{p}_2 \\ \mathbf{p}_3 \\ \mathbf{p}_4 \end{bmatrix}$$

$$k_1 = 3,\ k_2 = 2,\ k_3 = 1,\ k_4 = 2$$

Si **p** es el vector con los puntajes de las páginas:

$$(C\mathbf{p})_i = fila_i(C)\mathbf{p} = \sum_{j=1}^{N} \frac{\mathbf{p}_j}{k_j} A_{ji}$$

Veamos un ejemplo



$$C = \begin{bmatrix} 0 & 0 & 1 & \frac{1}{2} \\ \frac{1}{3} & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{3} & \frac{1}{2} & 0 & 0 \end{bmatrix} \qquad \mathbf{p} = \begin{bmatrix} \mathbf{p}_1 \\ \mathbf{p}_2 \\ \mathbf{p}_3 \\ \mathbf{p}_4 \end{bmatrix}$$

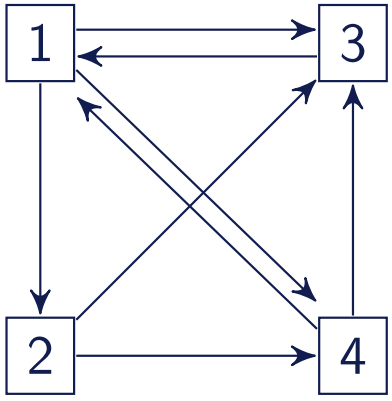
$$k_1 = 3,\ k_2 = 2,\ k_3 = 1,\ k_4 = 2$$

Si **p** es el vector con los puntajes de las páginas:

$$(C\mathbf{p})_i = fila_i(C)\mathbf{p} = \sum_{j=1}^{N} \frac{\mathbf{p}_j}{k_j} A_{ji}$$

$$(C\mathbf{p})_i = \mathbf{p}_i \quad \forall i$$

Veamos un ejemplo



$$C = \begin{bmatrix} 0 & 0 & 1 & \frac{1}{2} \\ \frac{1}{3} & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{3} & \frac{1}{2} & 0 & 0 \end{bmatrix} \qquad \mathbf{p} = \begin{bmatrix} \mathbf{p}_1 \\ \mathbf{p}_2 \\ \mathbf{p}_3 \\ \mathbf{p}_4 \end{bmatrix}$$

$$k_1 = 3,\ k_2 = 2,\ k_3 = 1,\ k_4 = 2$$

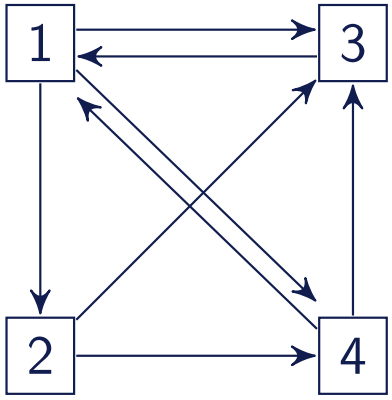
Si **p** es el vector con los puntajes de las páginas:

$$(C\mathbf{p})_i = fila_i(C)\mathbf{p} = \sum_{j=1}^{N} \frac{\mathbf{p}_j}{k_j} A_{ji}$$

$$(C\mathbf{p})_i = \mathbf{p}_i \quad \forall i$$

Luego, la solución al sistema es $\mathbf{p} = [\frac{12}{31}, \frac{4}{31}, \frac{9}{31}, \frac{6}{31}]^t$

Veamos un ejemplo



$$C = \begin{bmatrix} 0 & 0 & 1 & \frac{1}{2} \\ \frac{1}{3} & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{3} & \frac{1}{2} & 0 & 0 \end{bmatrix} \qquad \mathbf{p} = \begin{bmatrix} \mathbf{p}_1 \\ \mathbf{p}_2 \\ \mathbf{p}_3 \\ \mathbf{p}_4 \end{bmatrix}$$

$$k_1 = 3,\ k_2 = 2,\ k_3 = 1,\ k_4 = 2$$

Si $\mathbf{p}$ es el vector con los puntajes de las páginas:

$$(C\mathbf{p})_i = fila_i(C)\mathbf{p} = \sum_{j=1}^{N} \frac{\mathbf{p}_j}{k_j} A_{ji}$$

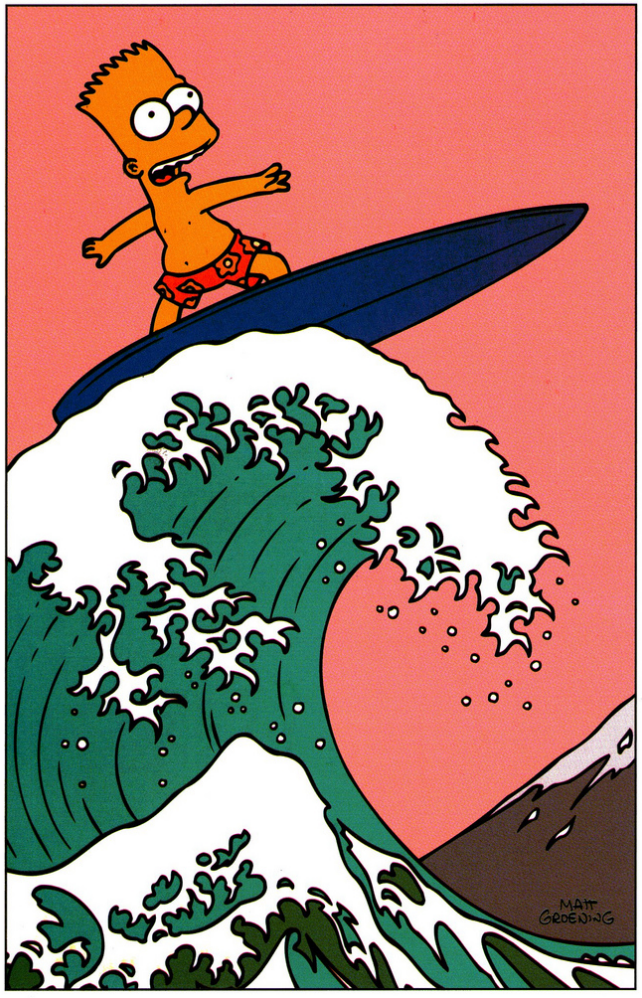
$$(C\mathbf{p})_i = \mathbf{p}_i \quad \forall i$$

Luego, la solución al sistema es $\mathbf{p} = [\frac{12}{31}, \frac{4}{31}, \frac{9}{31}, \frac{6}{31}]^t$

Para pensar...

- La página 3 es apuntada por todas las demás pero sin embargo no tiene el mayor puntaje. ¿Por qué?
- ¿Qué pasa si una página j no tiene links salientes ?

Intuición: Navegante aleatorio



- El modelo anterior tiene un problema y es que no logra capturar el comportamiento errático del usuario mientras surfea la red redes.
- Un enfoque alternativo es considerar el modelo del *navegante aleatorio*.
- Pagerank se basa en la idea de un **navegante aleatorio**. Este inicia en una página cualquiera y *navega* por la web mediante los links, moviéndose de página en página con cierta probabilidad.

Navegante aleatorio

Recorrido aleatorio de páginas

- En cada página j que visita el navegante elige:
  1. con probabilidad $\alpha \in (0, 1)$, si va a pasar a otra página cualquiera,
  2. con probabilidad $(1 - \alpha)$ si va a seguir uno de sus links.
- Caso 1: Si decidió seguir un link de la página j elige uno al azar con probabilidad $1/k_j$,
- Caso 2: si decidió pasar a otra página cualquiera entonces elige una al azar con probabilidad $1/N$.
- Cuando la página j no tiene links salientes ($k_j = 0$), elige al azar una página cualquiera del conjunto.

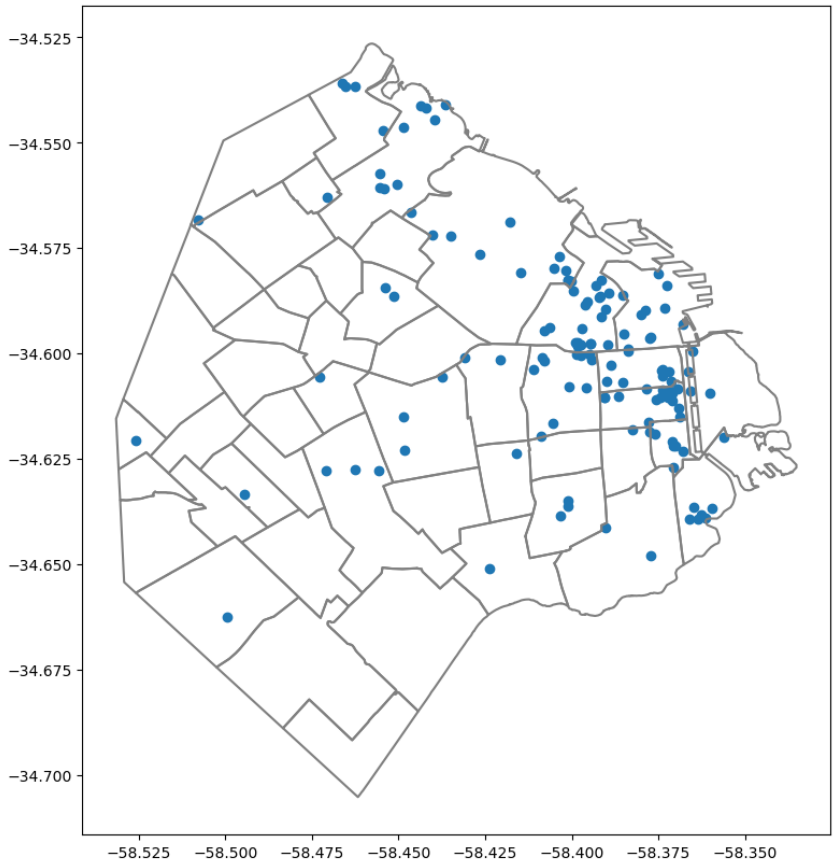
Navegante aleatorio

- Con todo lo dicho, podemos re-escribir la ecuación para los puntajes de las páginas combinando los dos escenarios.
- El Page Rank es la solución al sistema:

$$\boxed{\mathbf{p} = (1-\alpha)C\mathbf{p} + \tfrac{\alpha}{N}\mathbf{1}}$$

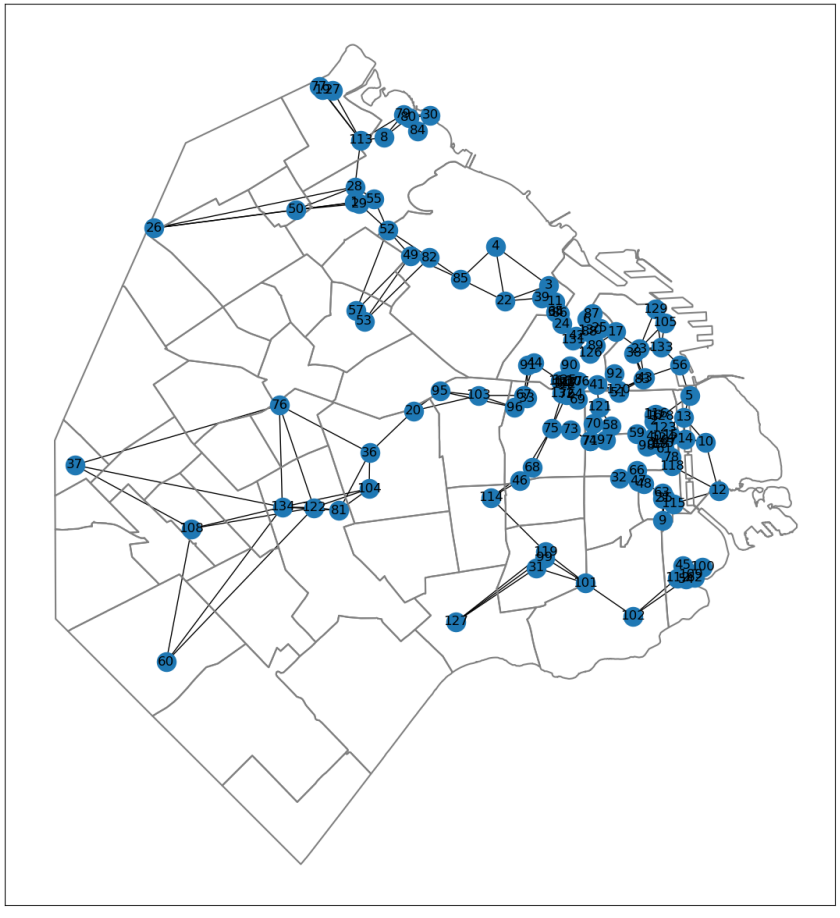
  que cumple $\mathbf{p}_i \geq 0$ y $\sum_i \mathbf{p}_i = 1$.
- En el TP, vamos a hallar $\mathbf{p}$ como la solución a un sistema lineal equivalente.

Los museos en CABA



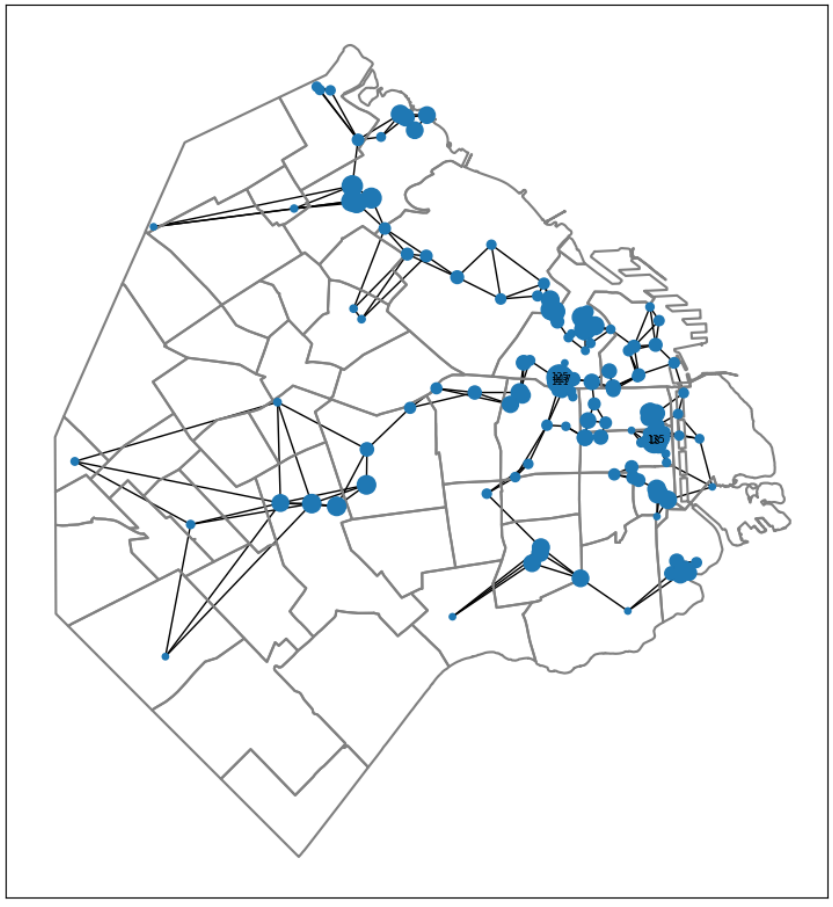
- Para este TP, vamos a reemplazar las páginas web con museos de CABA.
- Si bien no es obvio cómo, es claro que ciertos museos apuntarán a otros con mayor frecuencia.
- Como *proxy* a esa idea, vamos a tomar el vínculo más simple: proximidad geográfica.

Los museos en CABA



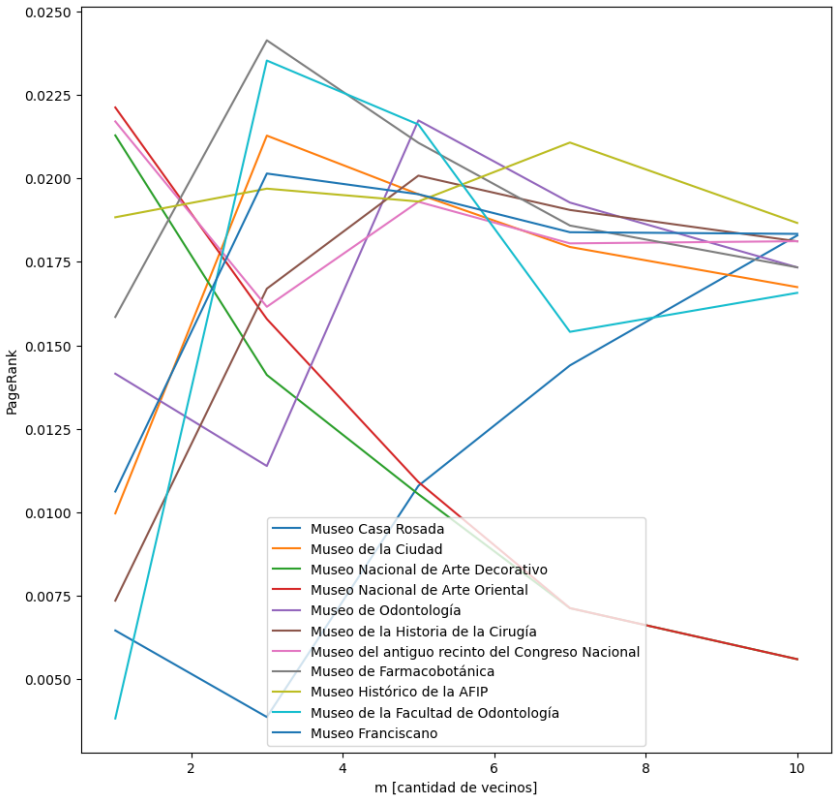
- Tomamos para un dado museo sus **3 museos más cercanos**
- La matriz A tiene $A_{ij} = 1$ si el museo está entre los 3 más cercanos (y 0 si no).
- Noten que la relación no tiene por qué ser reciproca. Nadie apunta a los museos más perifericos.

Los museos en CABA



- Al calcular **p** vemos que los museos más centrales geográficamente también tienen mayor ranking.
- El cálculo anterior asume $m = 3$ vecinos conectados, y $\alpha = 1/5$.
- ¿Qué esperan que cambie al modificar m? ¿Y α?

Los museos en CABA



- Al cambiar m hay museos que pierden relevancia y otros que ganan.
- Noten como los museos de arte se vuelven menos relevantes mientras más museos consideramos
- Al contrario, el Franciscano y el de Odontología ganan relevancia.

Ahora ustedes

- En el TP van a tener que programar ustedes funciones para calcular la inversa de la matriz y obtener el Page Rank.
- La inversa la vamos a obtener mediante la factorización LU, que toca la próxima clase.
- También van a tener que calcular número de condición, *así que no salteen este laboratorio.*

Bibliografía

- Sergey Brin and Lawrence Page, The anatomy of a large-scale hypertextual Web search engine, *Computer Networks and ISDN Systems*, April 1998.
- Kurt Bryan and Tanya Leise, The linear algebra behind google. *SIAM Review*, 2006.
- Sepandar D. Kamvar, Taher H. Haveliwala, Christopher D. Manning, and Gene H. Golub, Extrapolation methods for accelerating pagerank computations. In *Proceedings of the 12th international conference on World Wide Web*, New York, NY, USA, 2003. ACM.