Álgebra Lineal Computacional

1er Cuatrimestre 2025

Trabajo Práctico N° 1: Museos en red.

Introducción

La Noche de los Museos es un evento anual que tiene lugar en la Ciudad Autónoma de Buenos Aires (entre otros lugares). El evento consiste de una actividad nocturna en la cual los habitantes pueden recorrer distintos museos, lugares históricos, centros educativos y otros. La propuesta consiste en moverse de un punto al otro, recorriendo las ubicaciones entre las 7PM y las 3AM de un día en particular. En este TP, nos proponemos estudiar los potenciales viajes de los visitantes de un museo al otro, teniendo en cuenta su ubicación geográfica.

Para estudiar el movimiento de un museo al otro, la propuesta es construir una red que vincule a los distintos establecimientos de la ciudad. Informalmente, una red se define a partir de un conjunto de elementos (llamados vértices o nodos) y un conjunto de relaciones que vinculan los elementos de a pares (llamadas aristas o ejes). Mientras que el ejemplo moderno más común son las redes sociales, donde los nodos representan usuarios y las conexiones representa vínculos entre ellos (seguir, retwittear, ser amigo, etc), este concepto puede extender a muchas otras cosas como relaciones humanas, páginas de internet, interacciones entre organizaciones, cadenas alimenticias, paquetes de software; o por qué no, museos. En la Figura 1 pueden ver una posible versión de la red que conecta a los museos de CABA. Las conexiones de la red vinculan a un museo con los tres museos más cercanos a este. Pueden observar que hay museos relativamente aislados en el borde de la ciudad, mientras que la mayoría se concentra en la zona céntrica. Esto define a su vez donde ocurren la mayoría de las conexiones, ya que en la zona céntrica, lo más común es que la relación sea reciproca (no así para los establecimientos que se encuentran en la periferia).

Representación Algebraica

Una vez que se ha construido la red, podemos representarla matemáticamente como una matriz, denominada *matriz de adyacencia*. La matriz de adyacencia $A \in \mathbb{R}^{N \times N}$ tiene tantas filas y columnas como nodos hay en la red (es decir, museos). El elemento A_{ij} representa la relación entre los museos i y j. En nuestro ejemplo de la Figura 1, tenemos que $A_{ij} = 1$ si el museo j se encuentra entre los 3 más cercanos al museo i, y $A_{ij} = 0$ en otro caso. Noten que la matriz A en general no es simétrica ($A_{ij} \neq A_{ji}$), por lo que decimos que representa una red *dirigida*. En general, A_{ij} puede tomar valores distintos de 0 o 1. En ese caso, se dice que la red está *pesada*.

A la hora de pensar en movimientos sobre la red, podemos transformar a la matriz A en una matriz de transiciones, donde ahora el elemento (i, j) representa la probabilidad de moverse desde el museo i al museo j. Para esto, definimos la matriz de grado

$$K_{ij} = \begin{cases} \sum_{h=1}^{N} A_{ih}, & i = j \\ 0, & i \neq j \end{cases} \tag{1}$$

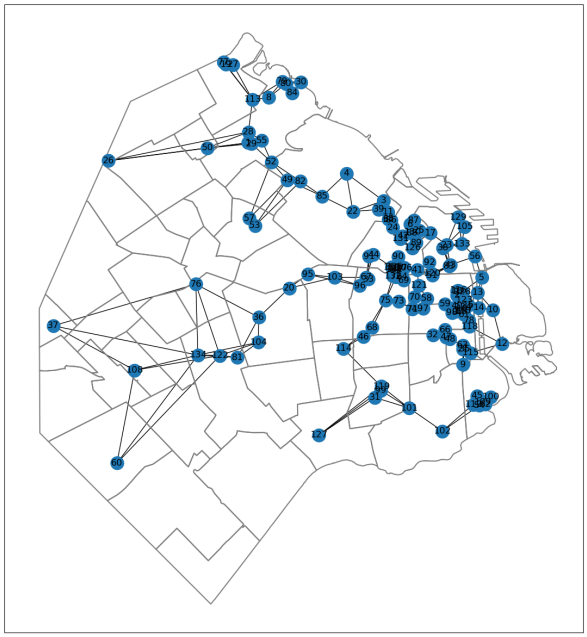


Figura 1: Red de museos en la Ciudad Autónoma de Buenos Aires. Un museo se conecta con sus tres museos más cercanos.

y entonces la matriz de transiciones C queda definida como

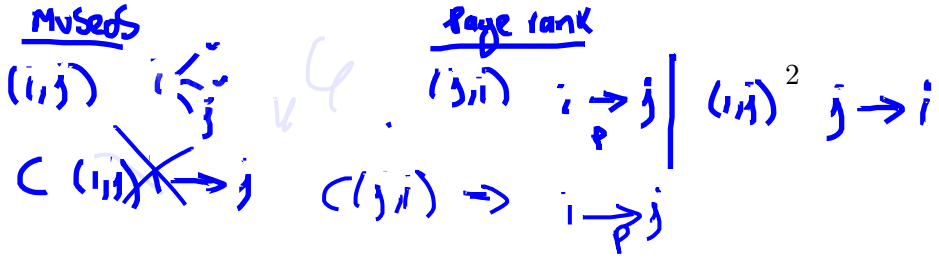
$$C = A^T K^{-1} \tag{2}$$

donde A^T es la matriz A, traspuesta. La matriz C es *estocástica*[1] ya que sus columnas suman 1. Esto permite usarla para estudiar como evolucionan distribuciones en el tiempo. Por ejemplo, consideremos el vector $v_0 \in \mathbb{R}^N$ que representa la distribución de museos que son la primer visita de los visitantes, y tiene en su elemento i la cantidad de visitantes que tienen al museo i como su primer opción. Entonces $v_1 = Cv_0$ representa el número esperado de visitantes en cada museo en el siguiente paso, si C_{ji} representa la probabilidad de moverse al museo j luego de haber visitado el i. En general, luego de k pasos, esperamos ver la distribución $v_k = C^k v_0$.

Page Rank y relevancia de cada museo

En 1999, L. Page, uno de los creadores de Google, buscaba un algoritmo para poder identificar las páginas más relevantes para un usuario en la web. Mientras que en su momento los buscadores se basaban en encontrar páginas que incluyesen los términos de búsqueda, Page tuvo la idea de rankear las páginas en base a la estructura de la red de hipervínculos que las conecta. En particular, la intuición era que si una página era incluida usualmente entre los hipervínculos de páginas relevantes, esta sería a su vez relevante con mayor probabilidad. Para estimar esto genero la siguiente heurística. Supongamos que tenemos un vector $\mathbf{p}$ de tamaño N que tiene en la posición i la relevancia de la página i, de forma tal que $\mathbf{p}_i > 0 \ \forall i, ||\mathbf{p}||_1 = 1$. Además, sea C la matriz que indica en su posición (j, i) la probabilidad de seguir un hipervínculo que llegue a la página j desde la página i (de forma análoga a como

[1] Pueden leer más al respecto en el apunte de la materia



definimos la matriz C para los museos), entonces se busca que el ranking $\mathbf{p}$ cumpla con la ecuación:

$$\mathbf{p} = (1-\alpha)\, C\, \mathbf{p} + \frac{\alpha}{N}\mathbf{1} \tag{3}$$

donde $\mathbf{1}$ representa a un vector de unos de $\mathbb{R}^N$, y $\alpha \in [0,1]$ se denomina *factor de amortiguamiento*. Desglocemos lo que dice la ecuación 3. Empecemos considerando $\alpha = 0$. Tendríamos que $\mathbf{p} = C\mathbf{p}$. Es decir, el valor de $\mathbf{p}_j = \sum_{i=1}^{N} C_{ji}\mathbf{p}_i$, y por lo tanto $\mathbf{p}_j$ es igual a la suma de los $\mathbf{p}_i$ que le apuntan, pesados por la probabilidad de moverse de i hacia j. Los valores de $\alpha > 0$ se introducen para representar un corte en esta caminata aleatoria. En el contexto de navegación de páginas web, α representa la probabilidad de cortar la caminata aleatoria entre páginas web, y comenzar desde un nodo elegido al azar. Es por esto que el término va acompañado por $\frac{1}{N}\mathbf{1}$, tal que $||\frac{1}{N}\mathbf{1}||_1 = 1$, y el peso de cada página es igual a $1/N$. Se puede mostrar que el vector $\mathbf{p}$ representa la distribución de probabilidad de llegar a la página en cuestión luego de surfear un largo tiempo por la web.

Volviendo a la red de museos, podemos hacer una analogía entre el *surfer* de internet y los visitantes de los museos: el Page Rank representa la probabilidad de llegar a un dado museo después de recorrerlos por un rato. Si pensamos en la combinación de todas las cadenas de visitas (una por cada visitante), podemos hacer un paralelo entre α (la probabilidad de reiniciar la caminata desde una nueva página web) y la longitud típica de las caminatas: α representaría la probabilidad de que una serie de visitas se termine en un dado instante, y por lo tanto $1/\alpha$ es del órden del número de museos distintos visitados durante una noche de los museos[2].

Una matriz de transiciones ligeramente más general

En la construcción de la matriz de adyacencia anterior consideramos que los viajes sólo se dan entre museos que estén entre los m más cercanos. Esta es una reducción útil para simplificar el número de opciones entre las cuales tiene que elegir. Sin embargo, un caminante podría moverse entre cualesquiera dos museos. Lo que tiene sentido es pensar que las transiciones más comunes serán aquellas que involucren museos cercanos. Para modelar esto, podemos pensar que hay una función $f(d_{ji})$, que toma la distancia entre los museos i y j y nos dice cuán tentador es el museo j para el i. De esta forma, podemos calcular la probabilidad de moverse desde el museo i al j como

$$C_{ji} = P(i \to j) = \frac{f(d_{ij})}{\sum_{k=1,k\neq i}^{N} f(d_{ik})} \tag{4}$$

Existen muchas funciones posibles a ser empleadas para este propósito, que puede ser justificadas en mayor o menor medida. Para nuestro TP vamos a emplear la más sencilla, que es tomar a $f(d_{ji}) = d_{ji}^{-1}$. Es decir, que la preferencia por j decae linealmente con la distancia. Noten que esto da lugar a una matriz C donde todas las transiciones tienen algún nivel de plausibilidad (es decir, no es rala). Desde la perspectiva del grafo, esto equivale a pensar que el caminante se mueve en una red pesada, donde los elementos de la matriz de adyacencia son iguales a $1/d_{ij}$. Noten que bajo nuestro modelo del caminante, el caminante no puede moverse del museo i al i nuevamente, y por lo tanto $C_{ii} = 0$.

[2] Quizá tengan que esperar hasta cursar probabilidad para conocer la binomial negativa

Precalentamiento

Habiendo leído **atentamente** la introducción, abran el notebook `TP1_template.ipynb`. En él encontraran la descarga de la red de museos usando `geopandas`, el cálculo de la matriz de distancias y la visualización de los museos y la red. Prueben construir el grafo para distintas cantidades de conexiones entre museos (parámetro m). ¿Cómo cambia la apariencia del mismo? ¿Cuántas conexiones deben considerarse para que la periferia (zona norte, oeste y sur) se conecten entre sí directamente, además de a través del centro? Abran también el archivo `template_funciones.py`, donde encontrarán una plantilla de las funciones a completar que se describen en el enunciado, descripto en la sección siguiente.

Enunciado

En este TP buscaremos trabajar con la idea de las visitas a los museos usando el modelo del caminante aleatorio. Como parte **obligatoria** se pide implementar la factorización LU para resolver el sistema de ecuaciones 3, y obtener el ranking de los museos. **NO** esta permitido utilizar una función de inversión de alguna librería de Python. Se podrá solo utilizar la función `solve_triangular` de la librería `SciPy` de `Python` para resolver sistemas triangulares.

A continuación, se detalla una serie de puntos a atender:

1. Partiendo de la ecuación 3, muestre que el vector de rankings $\mathbf{p}$ es solución de la ecuación $M\mathbf{p} = \mathbf{b}$, con $M = \frac{N}{\alpha}(I - (1-\alpha)C)$ y $\mathbf{b} = \mathbf{1}$.

2. ¿Qué condiciones se deben cumplir para que exista una única solución a la ecuación del punto anterior? ¿Se cumplen estas condiciones para la matriz M tal como fue construida para los museos, cuando $0 < \alpha < 1$? Demuestre que se cumplen o dé un contraejemplo.

3. Usando la factorización LU implementada, encuentre el vector $\mathbf{p} = M^{-1}\mathbf{b}$ en los siguientes casos:

   a. Construyendo la red conectando a cada museo con sus $m = 3$ vecinos más cercanos, calculen el Page Rank usando $\alpha = 1/5$. Visualizen la red asignando un tamaño a cada nodo proporcional al Page Rank que le toca

   b. Construyendo la red conectando a cada museo con sus m vecinos más cercanos, para $m = 1, 3, 5, 10$ y usando $\alpha = 1/5$.

   c. Para $m = 5$, considerando los valores de $\alpha = 6/7, 4/5, 2/3, 1/2, 1/3, 1/5, 1/7$.

   Usando los valores de $\mathbf{p}$ obtenidos para cada caso,

   a. Identifiquen los 3 museos más centrales (para cada m y cada α) y grafiquen sus puntajes (valores de Page Rank) en función del parametro a variar (es decir, en función de m o de α). ¿Son estables las posiciones en el ranking? Describa los distintos patrones que observa, identificando qué ubicaciones son relevantes en cada caso. ¿Hay museos que sólo son relevantes en redes con pocas conexiones?¿O museos que se vuelven más relevantes mientras más conexiones aparecen?

   b. Construyan visualizaciones del mapa, usando el Page Rank para representar el tamaño de cada museo. ¿Qué regiones se vuelven más predominantes al aumentar α? ¿Y al aumentar m?

4. Supongan que cada persona realiza r visitas antes de abandonar la red de museos. Si el número total de visitas que recibió cada museo está dado por el vector $\mathbf{w}$, tal que $\mathbf{w}_i$ tiene el número total de visitantes que se recibieron en el museo i, muestre que el vector $\mathbf{v}$, que tiene en su componente $\mathbf{v}_i$ el número de personas que tuvo al museo i como punto de entrada a la red, puede estimarse como:

$$\mathbf{v} = B^{-1}\mathbf{w} \tag{5}$$

con $B = \sum_{k=0}^{r-1} C^k$. *Tip:* Recuerden que si $\mathbf{v}$ da el número de visitantes que entraron a la red en cada museo, entonces luego de k pasos podemos esperar la distribución $C^k\mathbf{v}$ sobre el total de museos.

5. Usando la Eq. 5, y suponiendo que las personas dan $r = 3$ pasos en la red de museos, calcular la cantidad total de visitantes que entraron en la red, $||\mathbf{v}||_1$, a partir del vector $\mathbf{w}$ provisto en el archivo `visitas.txt`. Usar para esto la matriz de transiciones definida por la Eq. 4. Para esto:

   - Construya una función `calcula_matriz_C_continua` que reciba la matriz de distancias entre museos D y retorne la matriz C definida en la Eq. 4.
   - Construya una función `calcula_B(C,r)` que reciba la matriz C y el número de pasos r como argumento, y retorne la matriz B de la Eq. 5.
   - Utilice la función `calculaLU` para resolver la Eq. 5.

6. Supongan que se enteran de que el número total de visitantes tiene un error del 5 %, y necesitan estimar cómo se propaga ese error a la estimación del número total de visitantes. Llamemos $\tilde{w}$ y $\tilde{v}$ son los valores reales para el total de visitas y el total de primeras visitas respectivamente. Si expresamos este problema usando el número de condición tenemos que

$$\frac{||v - \tilde{v}||_1}{||v||_1} \leq \text{cond}_1(B)\frac{||w - \tilde{w}||_1}{||w||_1} \tag{6}$$

Calcule el número de condición de B y estime la cota para el error de estimación de v.

Entrega y lineamientos

La entrega se realizará a través del campus virtual de la materia con las siguientes fechas y formato:

- Fecha de entrega: hasta el viernes **25 de abril** a las 23:59 hs.
- Formato: Jupyter Notebook del template-alumnos. Archivo template-funciones.py completo.

Prestar especial atención a las siguientes indicaciones:

- El TP1 se realizará en grupos de tres personas. Deberán inscribir su grupo en el foro 'Foro de Grupos de TP' destinado para tal fin, dentro de la sección Laboratorio/TP1 del campus de la materia.

  Importante: es indispensable realizar la inscripción previa del grupo para poder hacer el envío a través del campus. Los grupos o personas no inscriptas en grupos no estarán habilitadas en el formulario de carga del TP. No se aceptarán envíos por email.

- Leer el enunciado completo antes de comenzar a generar código y sacarse todas las dudas de cada ítem antes de implementar. Para obtener un código más legible y organizado, pensar de antemano qué funciones deberán implementarse y cuáles podrían reutilizarse.

- El código debe estar correctamente comentado. Cada función definida debe contener un encabezado donde se explique los parámetros que recibe y qué se espera que retorne. Además las secciones de código dentro de la función deben estar debidamente comentados. Los nombre de las variables deben ser explicativos.

- Las conclusiones y razonamientos que respondan los ejercicios, o cualquier experimentación agregada, debe estar debidamente explicada en bloques de texto de las notebooks (markdown cells), separado de los bloques de código. Aprovechen a utilizar código LaTeX si necesitan incluir fórmulas.

- Gráficos: deben contener título, etiquetas en cada eje y leyendas indicando qué es lo que se muestra.