Indicaciones Proyecto 3

Rodrigo Zepeda

Otoño 2020

Objetivo

El objetivo de este proyecto es determinar, a través de muestreo, la estructura de una red social.

> Análisis estadístico en redes es algo súper interesante. Si quieres una introducción simple checa este código; ya para algo más serio te recomiendo este libro o este otro. A la fecha no hay una *buena* forma de hacer modelos estadísticos en redes y existen muchas preguntas abiertas. La que a mí me parece más interesante es como sigue: dada información de *otras* redes $(R_1, R_2, \ldots, R_n)$ tal que para una función se conocen los valores $f(R_1), f(R_2), \ldots, f(R_n)$, determinar una nueva red R_* que cumpla que $f(R_*) = k$ para algún k deseado. Por ejemplo se puede tener una red de consumidores y $f(R)$ es el gasto promedio de dicha red. Si se desea construir una red de consumidores *nueva* con un gasto promedio k ¿cuál debe ser la estructura de esta nueva red?

Fecha de entrega

Noviembre 5. Si requieres un cambio escríbele al profesor.

Indicaciones

1. El proyecto debe ser entregado en pdf o html.
2. El proyecto involucra realizar muestreo en una red social (digital) como Facebook, Twitter, LinkedIn, TikTok, etc. **Se espera que investigues cómo hacerlo**
3. El proyecto debe usar algún esquema de muestreo aleatorio simple sea muestreo aleatorio simple sin reemplazo, muestro aleatorio simple con reemplazo, muestreo aleatorio simple Bernoulli, muestreo aleatorio simple Poisson, o cualquier otro esquema de muestreo aleatorio diseñado específicamente para redes.
4. El proyecto puede escribirse en R, Python, Julia, Matlab, Octave, C, C++, Java, Javascript, LISP ó Prolog. En caso de no hacerse en `R` especificar la versión (ej Python 2 vs Python 3) y el compilador (ej cmake vs gcc); en ese caso, deben asegurarse que corra en Mac ó Linux (no tengo Windows).
5. Editar el código usando las gúias de estilo de Google o bien la de Matlab para el caso de Matlab/Octave, la de Prolog ó la de Julia en sus respectivos casos.
6. Se sugiere usar una API como puede ser rtweet, tuber para Youtube, vosonSML para Reddit, Youtube y Twitter, la API de tik tok, la de Linkedin Linkedin, Facebook. **Ojo: Algunas API como las de Twitter no muestrean todo sino sólo un porcentaje. Tu mecanismo de análisis / muestreo debe considerar esto.**

   No se permite usar datos premuestreados a menos que puedas explicar claramente todo el proceso de muestreo usado y sus limitaciones
7. El proyecto debe responder a tres de las siguientes preguntas *a través del muestreo* (tú eliges cuáles responder):

a. ¿Cuántos nodos (usuarios/videos/posts) hay en total en la red?

b. ¿Cuántos ejes (conexiones entre usuarios / conexiones entre videos / conexiones entre posts) hay en total en la red?

c. Un usuario en promedio cuántas conexiones tiene.

d. En caso de que la red se base en *posts* (como pueden ser imágenes / videos / tweets / comentarios) ¿cuál es el tamaño de un *post* ?

e. En caso de que la red se base en *posts* (como pueden ser imágenes / videos / tweets / comentarios) ¿cuál es la cantidad de *posts* totales que hay?

f. En caso de que la red se base en *posts* (como pueden ser imágenes / videos / tweets / comentarios) ¿cuántos *posts* promedio hace un usuario al día?

g. En caso de que sea una red asimétrica (donde que A siga a B no implica que B siga a A), estimar la diferencia esperada entre seguidos y seguidores para cualquier usuario (*i.e.* `Followers - Following`).

h. En caso de que la red contenga información de los nodos (usuarios), estimar la proporción de usuarios que cumplen una característica: *ej* proporción de profesionistas en la red, proporción de personas con cabello castaño, etc.

i. Dado un nodo elegido por ti con más de 100K seguidores (digamos la página de Facebook, del ITAM), estimar alguna característica determinante de sus seguidores (por ejemplo, la edad promedio) mediante muestreo.

j. Determinar la distancia promedio entre dos nodos (cualesquiera). Es decir, dados dos nodos elegidos *al azar* en la red, ¿por cuántos otros nodos se tiene que pasar para lograr una comunicación entre ellos? Dicho de otra manera, si A es un nodo aleatorio y manda un mensaje a uno de sus nodos relacionados (*amigos*) y luego su amigo manda un mensaje a uno de *sus* amigos (nodos relacionados del amigo de A), ¿por cuántos *amigos de los amigos* (en promedio) se tiene que pasar para que el mensaje de A llegue a B?

k. ¿Qué es lo que hace que cierto contenido se vuelva viral? Determinar puntos en común entre distintas formas de contenido viral (puede ser *nodos de origen* o puede ser algo en función del *contenido*). Para ello, definir qué es *viral* y muestrear sobre contenidos *virales*

l. En total ¿cuántos usuarios se interesan por un tema? Por ejemplo, cuántos usuarios de la red se interesan por la Liga Mx (OJO: Hay muchos usuarios interesados que no necesariamente son seguidores de la página: por ejemplo mucha gente le interesa saber qué pasa con el presidente pero no necesariamente siguen al presidente en redes).

m. Si se te ocurre otra cantidad de interés que se pueda estimar mediante muestreo puedes hacerlo: consulta con tu profe.

8. Se debe especificar cuáles estimadores usaste para cada caso; si son insesgados, si tienen varianza (y estimarla). Se deben incluir intervalos de confianza al menos de 80% (puede ser más) para responder las preguntas con un error que tú consideres aceptable.

   Para nuestro ejercicio, la muestra debe ser al menos de tamaño $n = 100$ (sin límite superior) y ser una muestra aleatoria, no un censo.

9. El proyecto debe incluir una justificación de los resultados.

10. Puedes usar referencias o código de cualquier parte de Internet. Sólo ¡CÍTALO! y asegúrate de explicar lo que hace hasta donde lo entiendes. Ejemplo: `este código entiendo que accede a Twitter y descarga Tweets elegidos por número de Tweet de manera uniforme a partir del 2020`.

11. **PARA TU SEGURIDAD** Si usas una API muchas veces necesitas poner *tu usuario y contraseña de la red social* . Esas cosas que son información personal **NO LA COMPARTAS CON TUS**

COMPAÑEROS NI CON EL PROFE. Aclara dónde es que yo, profe, debo rellenar mi usuario y contraseña en el código (en caso de que fuera necesario) y para correrlo yo lo haré así. ¡NO COMPARTAS TU INFORMACIÓN CON NADIE!

REPITO: NO ME COMPARTAS TU USUARIO Y CONTRASEÑA DE NADA.

12. *Sugerencia* En algunos casos se te pide hacer una cuenta. Lo que se sugiere es que entre el equipo hagan una cuenta de correo entre todos (digamos `aplicada_1_2020@gmail.com` y usen *esa* para entrar a la red social, no usen la propia).