Bachillerato en Ingeniería en Ciencia de Datos

2023-lll BCD9221 Optativa – Análisis de Datos

Proyecto Final

Profesores:
María Natalia Herrera
Dagoberto Herrera

Alumnos:
Jorge Isaac Barrantes Rojas
Jorge Eduardo Morales Mejía

Diciembre 2023

Contenido

Introducción

Con la clase y dimensiones de información que la Contraloría trabaja, es importante, durante la era de la información, generar análisis de datos por medio de herramientas. Principalmente, la utilización de redes y técnicas estadísticas se ha destacado como medio para entender y mejorar los procesos de contratación pública. La intención de este trabajo es dar un seguimiento sobre las diversas bibliotecas y fragmentos de código para mostrar el potencial de análisis. Además, el análisis exploratorio, identificación de comunidades, jerarquías, correlaciones y redes relacionales que proporcionan valiosa información. Con todo esto podremos ver patrones, mejorar en la transparencia y eficiencia, y así prevenir irregularidades en prácticas de contratación pública.

Contexto

El rol de esta institución es de la supervisión, eficiencia y por supuesto comprobar la legalidad de los recursos públicos. Los recursos del estado siempre deben tener auditorías de la gestión de recursos públicos. Uno de sus muchos objetivos es detectar y prevenir prácticas corruptas dentro de las entidades gubernamentales.

Librerías

Para realizar este reporte decidimos comenzar por analizar algunas de las librerías presentes en el código, omitimos algunas como numpy y pandas puesto que ya estamos algo familiarizados con estas, así que optamos por indagar sobre otras que no hemos usado mucho o que del todo no conocíamos.

NetworkX: esta librería fue construida para análisis de redes ya que ayuda a crear, manipular y analizar redes complejas. Para nosotros los científicos de datos nos ayudan a la construcción de redes a partir de texto o tablas, además de lograr representar las

redes en formatos como matrices o grafos. También nos permite trabajar con aspectos importantes de la red como lo son densidad o clustering, entre otras cosas.

warnings: esta es bastante específica y uno puede casi deducir qué hace con sólo el nombre, pero es la primera vez que escuchamos de esta librería así que por eso la agregamos a nuestro reporte, en resumidas cuentas, ayuda a manejar los warnings, para controlar cómo aparecen en el código.

stats: esta librería se usa tanto para realizar operaciones estadísticas básicas como avanzadas. Ya conocemos un poco de esta librería, pero decidimos agregarla por ser esencial para cualquier científico de datos, puesto que tiene características claves que nos ayudan con la media, mediana, desviación estándar y tests como los t-tests o chi-squared. En resumen, esta librería es clave para el análisis y visualización de datos e incluso Machine Learning.

statsmodels: al igual que la anterior esta se enfoca en estadística, pero más hacia la creación y manejo de modelos, por ejemplo, para probar hipótesis o realizar predicciones. Se usa en trabajos en materia de economía, finanzas y ciencias sociales, entre otros.

Análisis del Código

Preprocesamiento de Datos: Comienza con el preprocesamiento de datos, convirtiendo la columna “contador” del DataFrame a tipo numérico utiliza ‘pd_to_numeric()’.

En este apartado decidimos enfocarnos en algunas funciones y/o métodos utilizados que nos parecieron relevantes, como por ejemplo el “k_clique”, entre otros

Empezamos por ver detenidamente el siguiente trozo de código:

```
df.contador = pd.to_numeric(df.contador)
print("Distribución de relaciones de competencia")
print(df.contador.describe())
print("\nHistograma")
```

```
df.contador.hist()
```

Este es parte de lo que conocemos como el análisis exploratorio de datos y es importante ya que facilita identificar patrones o “banderas rojas” que por métodos tradicionales no son tan sencillos de ver. Podríamos por ejemplo conocer si la red está concentrada, como nos indicaba el profesor Dagoberto en la anterior clase, si existe(n) alguna(s) empresa(s) con “contadores” abultados, no necesariamente es algo malo, pero genera una “alerta” que permite indagar y corroborar. Ya aquí podemos hablar del término “comunidades” y si están aisladas o presentan comportamientos anómalos.

Identificación de Cliques: Utiliza la función ‘k_clique_communities’ de la libreria NetworkX para identificar cliques(subgrafos completamente conectados) con al menos 4 miembros en la red.

Ahora veamos este código:

```
#Identificar cliques con al menos 4 miembros
c = list(nx.community.k_clique_communities(H, 4))
for n in c:
  print(len(n))
  print(list(n))
```

Aquí se utiliza la librería “networkx” que ya comentamos anteriormente y se nos introduce la función “k_clique_communities” devolviendo una lista con los nodos de una clique de tamaño 4. ¿Pero qué significa esto? Según nuestro entendimiento es básicamente un “identificador” de grupos de nodos que se relacionan fuertemente entre sí. De nuevo retomamos el concepto de “comunidades” y con esta función también se pueden realizar predicciones del comportamiento de estos nodos en la red. Desde la perspectiva de redes de licitación pública nos permite no sólo identificar comunidades, sino también analizarlas, esto no necesariamente con un enfoque negativo, por ejemplo una empresa puede ver sus posibles competidores más fuertes, identificar oportunidades de colaboración o mejorar su estrategia y si lo vemos desde el punto de

vista del gobierno como ya lo hemos anotado analizar la estructura de la red, identificar comunidades y predecir comportamientos brinda herramientas para mejorar controles y evitar fraudes.

Creación de DataFrame de Comunidades: Con generación de un DataFrame llamado “communities” para la generación de la información sobre comunidades encontradas en la red de licitaciones.

Luego nos encontramos con este fragmento de código

```
communities=pd.DataFrame(columns=["community","index"])
#communities.columns=["community","node"]
for idx, x in enumerate(c):
   for n in x:
    new_row = {'community':idx, 'index':n}
    communities = communities.append(new_row, ignore_index=True)
communities
nodes=nodes.reset_index()
```

Aquí lo que vemos es la creación de un DataFrame de nombre “communities” para almacenar información de las comunidades en la red de licitación pública encontradas en el paso anterior, permitiendo un análisis más profundo de los nodos que pertenecen a una comunidad específica. En las siguientes líneas de código podemos ver cómo esto nos permite generar “filtros” y agrupar por ejemplo los grupos de empresa por sector (construcción, biomédico, etc).

En resumen, este código permite mostrar la combinación de métodos de análisis estadísticos. Las redes permiten identificar patrones, comunidades y estructuras dentro de las redes de licitaciones pública. Los últimos segmentos de código nos introducen los conceptos de Coherencia y Exclusividad, según nuestro conocimiento e investigación podemos simplificar la definición de coherencia como el grado de fortaleza en que los nodos de una comunidad se contentan entre sí y entendemos por

exclusividad las pocas o bajas conexiones de los nodos de una comunidad con otros nodos fuera de la comunidad. ¿Y esto en qué le sirve a la contraloría? Bueno, desde la perspectiva del Gobierno, todo esto se resume en una herramienta que genera valiosos datos para mejorar la eficiencia y la transparencia en las licitaciones públicas, por ejemplo, detectando patrones de colaboración en las comunidades, detección de prácticas anticompetitivas, entre otras cosas.

Conclusión

El implemento de estas herramientas fortalece la transparencia en procesos de contratación pública. Como ha demostrado en ayudar a optimizar recursos y procesos de contratación para evitar cuellos de botella o detectar ineficiencias. Un grafo es una herramienta sumamente poderosa para la Contraloría General de la República que le permitiría, entre otras cosas, mejorar la toma de decisiones, evitar prácticas ilícitas y/o corrupción, así como mejorar la eficiencia en términos generales de todo el proceso de contratación pública. Definir mejor la jerarquización de los procesos entre entidades, con ellos tener seguimiento de cualquiera de las transferencias.

Bibliografía

- Hagberg, A. A., Schult, D. A., & Swart, P. J. (2003). Exploring network structure, dynamics, and function using NetworkX. In Proceedings of the 7th Python in Science Conference (pp. 11-15).

- Python Software Foundation. (2023). warnings — Handling warnings in Python. Retrieved from https://docs.python.org/3/library/warnings.html

- SciPy Developers. (2023). stats — Statistical functions and distributions. Retrieved from https://docs.scipy.org/doc/scipy/tutorial/stats.html

- Statsmodels Developers. (2023). statsmodels: Statistical modeling and econometrics. Retrieved from https://www.statsmodels.org/v0.10.2/