Analítica de datos
2do semestre 2023
Proyecto
Prof. Ana Aguilera
Ayudante Camila Araya
12/9/2023

Resultados de aprendizaje: CE3.N2.RA1. CE3.N2.RA2. CE3.N2.RA3.

ANÁLISIS DE REDES SOCIALES
CENTRALIDAD Y DISPERSIÓN DE INFLUENCIA

Para esta actividad se utilizará un dataset de su elección de la Stanford Large Network Dataset Collection (SPAN)[1] que contenga grafos dados como listas de aristas en archivos de texto plano.

Sin perder de vista el ámbito y dominio de aplicación del dataset, se pide:
- Hacer un análisis descriptivo de la red, usando medidas de centralidad y centralización.
- Hacer un análisis correlacional, comparando:
  * Distintas medidas de centralidad clásicas entre sí.
  * La dispersión de influencia entre los nodos más relevantes obtenidos para cada medida.
- Visualizar e interpretar resultados obtenidos, en el contexto del dominio de aplicación del dataset.

Más concretamente, se pide:

1) Depurar el dataset escogido para su correcta manipulación en los análisis (10%)
2) Calcular medidas de grado (in-degree y out-degree, si la red es dirigida), cercanía (closeness), intermediación (betweenness) y PageRank sobre todos los nodos de la red (20%)
3) Implementar el modelo de dispersión de influencia de Linear Threshold model (20%)
4) Calcular la dispersión de influencia sobre los top-10 nodos obtenidos por cada medida de centralidad. Esto implica 4 (o 5, si la red es dirigida) ejecuciones. Puede modificar el dataset incluyendo una función peso y/o de etiquetado, en caso de ser necesario y que tenga sentido con el dominio del problema (20%)
5) Aplicar el coeficiente de correlación de Spearman para comparar los rankings obtenidos para medida de centralidad. Visualizar los datos en una matriz de correlaciones, indicando únicamente los valores con p-value <0.05 (20%)
6) Explicar en un párrafo cómo se interpretan los resultados obtenidos en 4) y 5) en el contexto del dominio de aplicación del dataset escogido (10%)

Formato de entrega:
Por el aula virtual se entrega Cuaderno de Júpiter que contenga celdas de :
1) Texto que guíen la lectura explicando cada paso
2) Código
3) Salidas de ejecución

Fecha de entrega:
21/11/2023

Rúbrica			
Ítem	***Logrado***	***Medianamente Logrado***	***No Logrado***
Depuración de dataset (10%)	(10%) Limpieza y transformación correcta del dataset, para aplicación de medidas de centralidad y dispersión de influencia	(5%) Limpieza y transformación incompleta o con errores	(0%) No realiza limpieza ni transformación
Cálculo de medidas de centralidad (20%)	(20%) Calcula correctamente todas las medidas de centralidad solicitadas	(10%) Sólo calcula correctamente la mitad de las medidas de centralidad solicitadas	(0%) No calcula las medidas solicitadas, o bien lo hace sin errores para menos de la mitad de ellas
Implementar modelo de dispersión (20%)	(20%) Implementa correctamente el modelo de dispersión de influencia	(5%) Implementa parcialmente el modelo de dispersión de influencia, o con errores (que igualmente permiten su ejecución)	0%) No implementa el modelo de dispersión de influencia, o bien tiene errores que impiden su ejecución
Cálculos de dispersión de influencia (20%)	(20%) Calcula correctamente la dispersión de influencia sobre los top-10 nodos obtenidos para cada medida de centralidad	(10%) Calcula correctamente la dispersión de influencia sobre los top-10 nodos obtenidos para al menos la mitad de las medidas de centralidad, o bien para todas pero con algunos errores	0%) No realiza los cálculos solicitados, o bien solo los hace para menos de las mitad de las medidas solicitadas
Correlación de medidas de centralidad (20%)	(20%) Obtiene correctamente la matriz de correlaciones para todas las medidas, indicando además p-value obtenidos a	(10%) Obtiene parcialmente la matriz de correlaciones, o bien no indica los p-value obtenidos	0%) No realiza la matriz de correlaciones
Interpretación de resultados (10%)	(10%) Interpreta correctamente los resultados en el contexto del dominio del dataset, considerando la matriz de correlaciones y los resultados de la dispersión de influencia	(5%) Interpreta parcialmente los resultados, olvidando el contexto, o bien sin considerar algunos de los resultados solicitados anteriormente	0%) No interpreta los resultados obtenidos

Referencias

1. Leskovec, J.; Krevl, A. (2014). SNAP Datasets: Stanford Large Network Dataset Collection. https://snap.stanford.edu/data/
2. NetworX documentation - Network analysis in Python https://networkx.org/documentation/stable/index.html
3. Pandas https://pandas.pydata.org/
4. Presentación de análisis de redes sociales