CURSO: CMP 5002 - DATAMINING
COLEGIO: POLITÉCNICO
Semestre: 1er Semestre 2023/2024

Proyecto 2: Ejercicio usando la técnica *PageRank*

Modalidad: Trabajo en equipo. Por tanto, cada integrante debe preparar su parte (asignada por el líder) y dominar la técnica evaluada completamente.

Problema:

1. Un representante del departamento de marketing de una trasnacional desea incluir en su punto de análisis de la reunión, el nivel de importancia de 15 páginas Web dentro de una sección WWW seleccionada aleatoriamente. Para dar respuesta a esta necesidad, usted como ingeniero propone utilizar la técnica *PageRank* sobre la sección WWW seleccionada.

Nota: Como punto de partida, se constará con un grafo dirigido de 15 nodos (ejemplo reducido en la fig. 1), el cual debe incluir 5 trampas (2 spider traps y 3 dead ends o viceversa). Informáticamente, para simplificar el punto de partida, el grafo se tradujo a una matriz de adyacencia donde cada celda representa la cantidad de *in-links* al nodo activo (ver ejemplo de matriz).

Nodos	**N1**	**N2**	**…**	**N15**
N1	0	0	…	0
N2	1	1	…	0
…	…	…	…	…
N15	0	0	…	0

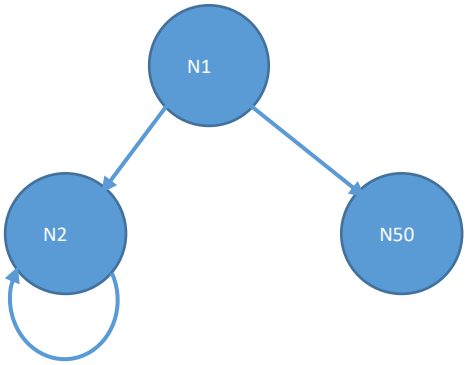

Fig. 1 Ejemplo reducido del grafo de 15 nodos, conteniendo 1 spider trap y 1 dead end.

Requisitos de evaluación:

- Mostrar la matriz y el grafo (dibujado y legible) seleccionado por usted.
- Es obligatorio el uso de la filosofía *PageRank* sobre un grafo dirigido.
- El algoritmo *random walker* debe incursionar de forma aleatoria (selección de caminos basado en la probabilidad sobre los *links* de salida del nodo actual). (**2 puntos)**
- Es obligatorio presentar la trazabilidad del vector **r** calculada con el *power iteration* sobre el tiempo ***t, t+1, t+2,…, t+n***; hasta que ***n*** sea el estado donde el algoritmo *random walker* haya explorado todos los nodos del grafo. Sugerentemente, presentar las probabilidades p(t) de salida del nodo actual en el tiempo ***t*** para algún nodo en el tiempo ***t+1*** y mostrar en ese instante el vector **r** calculado (pueden presentar una matriz donde se vaya actualizando e incrementando dinámicamente el vector **r** por cada momento ***t***). (**4 puntos)**
- Es obligatorio que en la trazabilidad del vector **r** se pongan de manifiesto las trampas y el *teleport* del *power iteration* para continuar con su desempeño. Se sugiere, ¿que a cada iteración se le pregunte al usuario si desea o no teletransportarse? (**3 puntos**)
- Cargar al D2L los códigos implementados (fichero compactado que incluye el ejecutable ej: el .JAR de java) y un PDF con la matriz inicial de adyacencia M, así como el grafo dibujado (ej: ver Fig.1) dentro del plazo de entrega.
- En la defensa del proyecto los estudiantes deben mantener un lenguaje técnico y con dominio del conocimiento sobre la técnica evaluada. (**1 punto**)