ACTO2 – SAR

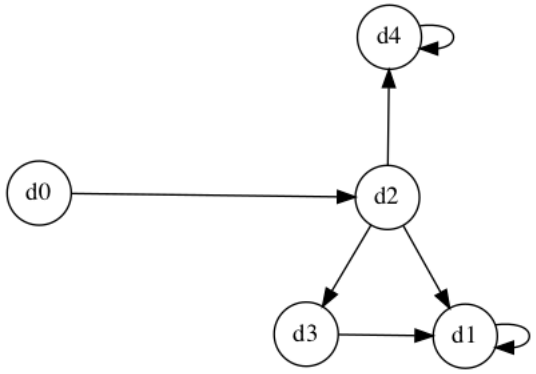
(20/05/2019 - 3 puntos)

(IMPORTANTE: Se deben justificar las respuestas, cálculos redondeados a 3 decimales)

1) Se pide obtener la postings list a partir de la siguiente secuencia de bits codificada utilizando codificación variable en bytes: **(0,3 puntos)**
   00000001 00010100 10001101 00000110 10010100 10001010

2) Partiendo de una colección de documentos como la estudiada como ejemplo en la asignatura, se pide pronunciarse sobre la veracidad o falsedad de las afirmaciones siguientes, justificando las respuestas: **(0,9 puntos)**
   a. La eliminación de stop words en el preproceso de construcción del índice invertido afecta mucho al tamaño del diccionario.
   b. La eliminación de stop words en el preproceso de construcción del índice invertido afecta mucho al tamaño de las postings lists de un índice no posicional.
   c. La eliminación de stop words en el preproceso de construcción del índice invertido afecta más al tamaño de las postings lists de un índice posicional que a las de un índice no posicional.

3) Dadas las siguientes páginas web y los enlaces entre ellas representadas como un grafo, se pide calcular el pagerank de cada página. Se debe calcular: i) la matriz de enlaces, ii) la matriz de probabilidades de transición, iii) la matriz de probabilidades de transición con teletransporte (utiliza un α=0,15 para el teletransporte), iv) todas las iteraciones para calcular el pagerank. Realiza como máximo cinco iteraciones. **(1 punto)**



4) Se quiere buscar el patrón “VASO” en el texto “ESBASOKIBASICO”. Para ello se hace una búsqueda aproximada para obtener aquellos segmentos cuya distancia al patrón es menor o igual que 1. Se pide construir la matriz que corresponde al algoritmo de búsqueda aproximada e indicar las soluciones, es decir, los segmentos del texto que son resultados de la búsqueda.
   Se consideran pesos 1 para las Sustituciones, Inserciones y Borrados. **(0,8 puntos)**

Solución:

1)
Los valores codificados son [(1*128*128)+(20*128)+13=18.957, (6*128)+20=788, 10]. Recordando que los valores codificados corresponden a los gaps, la postings list es [18.957, 19745, 19755]

2)
a. FALSO. La eliminación de stop words en el preproceso de construcción de un índice invertido afecta muy poco al tamaño del diccionario ya que se elimina un conjunto muy reducido de términos del diccionario (generalmente entre 30 y 150 términos) en comparación al tamaño del mismo (unos 400.000 términos en el caso de la colección RVC1).
b. VERDADERO. La eliminación de stop words en el preproceso de construcción de un índice invertido afecta mucho al tamaño de las postings list de un índice no posicional ya que son los términos más frecuentes, y por tanto, a los que les corresponden las postings lists más largas. De hecho, en general, son términos que aparecen en todos los documentos de la colección por lo que sus postings lists contienen todos los documentos de la colección y su eliminación supone eliminar las postings list más largas del índice. (un 30% de reducción en el caso de la colección RVC1 y una lista de stop words de 150 términos)
c. VERDADERO. La eliminación de stop words en el preproceso de construcción del índice invertido afecta más al tamaño de las postings list de un índice posicional que a las de un índice no posicional ya que son términos que además de ocurrir en todos los documentos de la colección, son muy frecuentes en cada uno de ellos. Por tanto, la eliminación de las stop words y de sus postings lists reduce mucho más un índice posicional que uno no posicional. (un 47% de reducción en el caso de la colección RVC1 y una lista de stop words de 150 términos).

3)

Matriz de enlaces:

[0 0 1 0 0]
[0 1 0 0 0]
[0 1 0 1 1]
[0 1 0 0 0]
[0 0 0 0 1]

Matriz de probabilidades de transición inicial:

[0.000 0.000 1.000 0.000 0.000]
[0.000 1.000 0.000 0.000 0.000]
[0.000 0.333 0.000 0.333 0.333]
[0.000 1.000 0.000 0.000 0.000]
[0.000 0.000 0.000 0.000 1.000]

Matriz de probabilidades de transición con teletransporte (α=0.15):

[0.030 0.030 0.880 0.030 0.030]
[0.030 0.880 0.030 0.030 0.030]
[0.030 0.313 0.030 0.313 0.313]
[0.030 0.880 0.030 0.030 0.030]
[0.030 0.030 0.030 0.030 0.880]

Calculo del Pagerank:

[1 0 0 0 0]
[0.030 0.030 0.880 0.030 0.030]
[0.030 0.330 0.055 0.279 0.305]
[0.030 0.563 0.055 0.046 0.305]

Pagerank: π = [0.030 0.563 0.055 0.046 0.305]

4)

O	4	4	3	3	3	2	1	2	3	4	3	2	2	3	3
S	3	3	2	3	2	1	2	3	3	3	2	1	2	3	3
A	2	2	2	2	1	2	2	2	2	2	1	2	2	2	2
V	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		E	**S**	**B**	**A**	**S**	**O**	**K**	**I**	**B**	**A**	**S**	**I**	**C**	**O**

Subcadenas:

BASO (3-6)
ASO (4-6)