

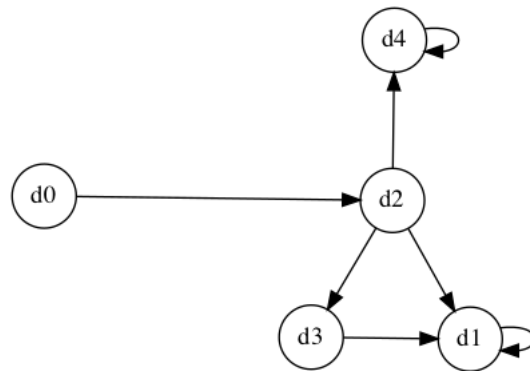
(20/05/2019 - 3 puntos)

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

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:

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.

(1 punto)



Se consideran pesos 1 para las Sustituciones, Inserciones y Borrados.

(0,8 puntos)

[illegible]

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 ($\alpha=0.15$):

```
[ 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: $\pi = [0.030 \ 0.563 \ 0.055 \ 0.046 \ 0.305]$

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)