CAIM, examen final

11 de gener de 2019. Temps: 2 hores 30 minuts

Exercici 1 (2 puntos). Tenemos una colección de D documentos, y tras preprocesar los documentos nos quedamos con 6 términos. La siguiente tabla contiene el número de documentos que contienen dichos términos.

ave	mamífero	reptil	perro	gato	pájaro
0.1 D	0.05 D	0.03 D	0.02 D	0.01 D	0.01 D

1. Calcula la similitud entre los documentos $D_1 =$ " pájaro pájaro ave gato " y $D_2 =$ " ave perro gato ave pájaro ". Utiliza pesos tf-idf y la similitud del coseno. Puedes utilizar calculadora.

2. Ahora añadimos $5D$ documentos a la colección (que tendrá un total de $6D$ documentos). Nos dicen que los documentos vienen de la misma fuente y tienen aproximadamente las mismas longitudes y distribuciones de palabras que los documentos iniciales. Si recalculamos la similitud entre los documentos D_1 y D_2 en la nueva colección, ¿obtendremos el mismo resultado que antes? Justifica tu respuesta brevemente.

Exercici 2 (2 puntos). Considera el siguiente grafo no dirigido con la siguiente matriz de adyacencia:

$$\begin{pmatrix} 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 1 & 1 & 0 \end{pmatrix}$$

Calcula el pagerank de cada nodo utilizando un "damping factor" $\lambda = 4/5$. Además, calcula el coeficiente de clustering local y la "betweenness" de cada nodo.

Exercici 3 (2 puntos) Somos los propietarios de una red social similar a "Facebook" y queremos calcular, para cada par de usuarios, el número de amigos comunes que tienen. La información sobre la red viene dada en ficheros que contienen las listas de amigos de cada usuario, por ejemplo:

```
marta:  balqui, ricard, ramon, larri, javier
ricard:  ramon, marta, toni
...
```

Lo cual quiere decir que Marta está conectada con Balqui, Ricard, Ramon, Larri y Javier, y que Ricard está conectado con Ramon, Marta y Toni. Las relaciones en esta red son simétricas, por lo cual si una persona x aparece en la lista de amigos de y, entonces y aparece también en la lista de x.

Explica como calcular el número de amigos comunes para aquellas parejas de usuarios que tienen amigos comunes en el formalismo de mapreduce. Indica cuántas fases de mapreduce necesitas y da pseudocódigo las funciones de map y reduce de cada fase.

Exercici 4 (2 puntos). Describe brevemente un par de posibles aplicaciones de locality sensitive hashing en el contexto de information retrieval.

Exercici 5 (2 puntos).

1. Dada la posting list:

   $[10, 1, 15, 3, 22, 2, 23, 4, 34, 1, 44, 1, 50, 2, 58, 8, 90, 1, 101, 1, 112, 2]$

   (que podemos interpretar como que el término asociada a la posting list aparece 1 vez en el documento 10, 3 veces en el documento 15, etc.) calcula su codificación si la comprimimos con "self-delimiting unary" para las frecuencias y gap-compression + Elias' Gamma para los identificadores de documento.

2. Decodifica el siguiente string:

   000010101100010001010001000100011011001000110

   que ha sido codificado utilizando el esquema del apartado anterior.

   (Nota: a modo de ejemplo, la codificación en unario self-delimiting de 3 es 110, y el código Elias Gamma de 4 es 00100).

Nota 2: Este ejercicio será utilizado para calcular la nota de la "Competència Transversal Aprenentatge Autònom".