Guia 7: Information Retrieval

1. Considerando a cada línea como un documento y cada palabra como un término construya la matriz de términos x documentos usando TF-IDF. Finalmente indique cuántas dimensiones elegiría para usar LSI justificando adecuadamente su respuesta.

   Salsa Tomate Pizza Muzzarella Tomate
   Fideos Tomate Salsa Fideos Fideos
   Pizza Tomate Tomate
   Fideos Salsa Salsa Salsa

2. Indique en que casos el código unario es mas conveniente que el código gamma o delta para almacenar los punteros de un índice invertido. Justifique adecuadamente

3. Para la siguiente colección y considerando a cada línea como un documento y cada palabra como un término construir un índice invertido usando códigos gama para los punteros y front-coding para el léxico.

   casa casa arbol
   alicia arbol casa
   arbol arbol paisaje alicia
   casa chalet

4. En un índice invertido en el cuál tenemos 2500 documentos y 785 términos estime cuál es una cota superior para el tamaño total del índice.

5. Si se sabe que para una colección de textos p=0.0023, ¿es verdad que la mayoría de los textos deberían hablar sobre temas diferentes?.

6. Que características debe tener una colección de documentos para que Gamma sea mejor que Delta y Unario. No alcanza con decir que debe haber mayoría de distancias en donde Gamma sea mas corto que Delta o Unario, eso ya lo sabemos.

7. Determinar si las siguientes afirmaciones son V / F justificando la respuesta:

   a. Si para un cierto término el código óptimo para representar sus punteros es el código unario entonces podemos decir que el término aparece en muchos documentos.

   b. A medida que aumentamos el valor de “b” los códigos de Golomb crecen de forma cada vez mas suave

   c. En una colección de 1300 documentos el b optimo para un término que aparece en 900 de ellos es 1.

d. Si para una cierta consulta q d1 es mas relevante que d2 entonces luego de aplicar LSI d1 seguirá siendo mas relevante que d2.

e. Si la probabilidad de las distancias 8 a 15 es similar a 1/256 entonces es buena idea representarlas usando código Delta.

f. En los casos en los que hay mas términos que documentos el código unario es óptimo.