1. Considera el documento siguiente:

   > Peer-to-peer (P2P) computing is the sharing of computer resources and services by direct collaboration between client systems. These resources and services often include the exchange of information (Napster, Freenet, etc.), processing cycles (distributed.net, SETI@home, etc.), and disk storage for files (OceanStore, Farsite, etc.). Peer-to-peer computing takes advantage of existing desktop computing power and networking connectivity, allowing off-the-shelf clients to leverage their collective power beyond the sum of their parts.
   > Current research on P2P has evolved from very different research areas. Among others, P2P has attracted the attention of researchers working on classical distributed computing, mobile agents, parallel computing, or communications. Very interestingly, the P2P paradigm is different from those studied in all these areas. For instance, while in some sense peer-to-peer computing is very similar to classical distributed computing (as opposed to the client-server paradigm), some new characteristics emerge. These include the clear and present danger of malicious peers, high churn rate (peers joining and leaving the system), among others.

   Considera que los términos de índice son los que aparecen en la siguiente tabla, donde además se da la frecuencia de cada término en una colección de documentos con un millón de documentos.

Term	Document ferq
computing	300901
network	200019
computer	109789
system	110990
clients	80921
agents	42003
p2p	20909

   (a) Escribe su representación en el modelo booleano.
   (b) Escribe su representación en el modelo vectorial.

2. En una colección de alrededor de medio millón de documentos se han observado las siguientes frecuencias de aparición de términos:

Term	Document freq
eyes	213312
kaleidoscope	87009
marmalade	107913
skies	271658
tangerine	46653
trees	316812

Recomienda un orden de procesamiento para las interrogaciones siguientes y justifica la respuesta.

(a) tangerine AND marmelade

(b) (eyes OR kaleidoscope) AND skies

(c) (tangerine OR trees) AND (marmelade OR skies) AND (kaleidoscope OR eyes)

(d) (NOT trees) AND tangerine AND (NOT marmelade)

(e) trees AND ((NOT kaleidoscope) OR (NOT tangerine))

3. Considera la siguiente colección con 4 documentos

Doc 1: Information Retrieval Systems

Doc 2: Information Storage

Doc 3: Digital Speech Synthesis Systems

Doc 4: Speech Filtering, Speech Retrieval

Considera el modelo vectorial (utilizando pesos tf-idf).

(a) Calcula todas las componentes no nulas del vector que representa Doc 1.

(b) Calcula el orden de todos los docuementos en la coleción respecto de la interrogación "Speech Systems"?

(c) Calcula la similaridad entre (a) docs 1 y 2 (b) docs 3 y 4.

4. Considera la misma colección de 4 documentos del apartado anterior, y considera ahora el modelo probabilístico.

- Escribe la representación vectorial de los cuatro documentos.
- Calcula el el orden de todos los docuementos en la coleción respecto de la interrogación "Speech Systems"? después de un paso de cálculo de la probabilidad.
- Calcula el el orden de todos los docuementos en la coleción respecto de la interrogación "Speech Systems"? después de un máximo de 4 iteraciones en el cálculo de la probabilidad.

5. Considera el siguiente fragmento de un índice invertido, en el que en lugar de almacenar la frecuencia del término de índice en el documento se guarda información posicional de las ocurrencias, en este caso la posición se refiere al carácter del texto a partir del cual se puede encontrar el término de índice. El primer número en cada línea (antes de :) identifica el documento.

the

3:34,38,55;
5:12,16,25,44;
7:67,87,90,101;
10:33,39,45,62;

undiscovered

3:12,15,19;
5:3,5,17,41,45,96;
6:21,25,55,62;
7:4,68,70,85,110;
10:15,34,40,65,81;

country

3:22,26;
5:18,46,52,65;
7:5,69,91,105;
8:32,42,65,93;
10:32,44,75,83;

(a) ¿Cuantas veces aparece la frase "the undiscovered country" en cada documento?

(b) ¿Qué documentos satisfacen la interrogación "undiscovered AND country"?

(c) ¿Qué documentos satisfacen la interrogación "undiscovered AND country WITHIN 3"? ¿Cuantas veces se satisface?

6. Considera los términos del vocabulario asociados a la colección de documentos del ejercicio 3 (todos los términos).

(a) Obtiene los codigos de Huffman de las palabras del vocabulario.

(b) Describe la versión comprimida de los documentos de la colección.

(c) ¿Cómo se formula internamente la interrogación "Speech Systems?" cuando los documentoos están comprimidos?

(d) ¿Qué diferencias hay entre los documentos recuperados?