BÚSQUEDA CON TOLERANCIA: ÍNDICE DE K-GRAMAS

Se pide responder las siguientes preguntas:

a) Comenta brevemente en qué consiste y cómo se construye un índice de n-gramas.
b) Explica cómo sería la búsqueda de documentos correspondientes a la wildcard query "ca*sa".
c) Si tenemos el siguiente diccionario de bigramas indica que términos devolvería para la consulta ca*sa. Comenta también si todas las palabras devueltas son correctas para la consulta realizada.

$a ➡	acabo	antena	antigua	asar					
$c ➡	camino	comino	camisa	canto	cansa	cena	comida	carcasa	casaca
a$ ➡	cansa	antena	antigua	camisa	carcasa	poca	casaca	comida	cena
an ➡	antigua	cansa	pantano	canto	antena	gusano			
ca ➡	acabo	camisa	cansa	casaca	canto	carcasa	rocas	poca	camino
sa ➡	pasar	carcasa	cansa	pesar	camisa	cosas	casaca	asar	gusano

a. Un índice de n-gramas es un segundo índice que se construye para poder hacer búsquedas con tolerancia. Se calculan los n-gramas de caracteres a partir de los términos que aparecen en los documentos a los que se ha añadido previamente el símbolo "$" al inicio y fin. En el índice de n-gramas, cada n-grama apunta a la lista de términos que contienen ese n-grama.

b. La búsqueda de documentos correspondientes a la wildcard query "ca*sa"se realizaría a partir de la expresión lógica: $c AND ca AND sa AND a$. Utilizando un algoritmo de INTERSECCIÓN de las listas de términos correspondientes a cada bigrama implicado nos devolvería la lista de términos resultante.

c. Términos que devolvería: "camisa", "carcasa", "cansa", "casaca". El término "casaca" es incorrecto para la consulta realizada ya que no acaba en "sa", pero sería devuelto por el índice.