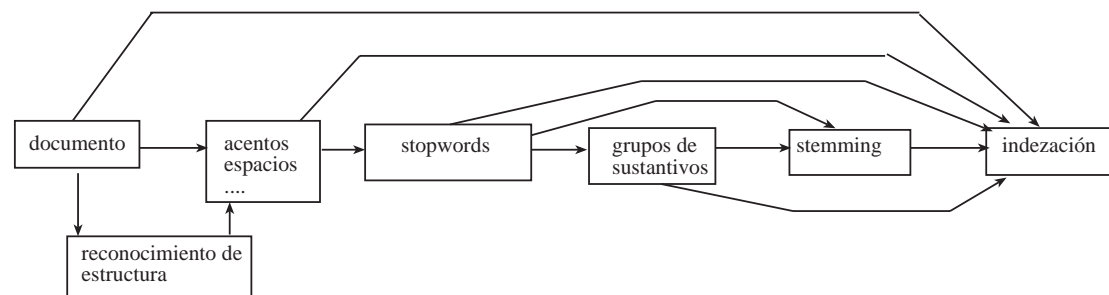


PREPROCESAMIENTO DE TEXTO

El preprocesamiento de texto puede ser visto como un proceso que controla el tamaño del vocabulario (es decir, el número de palabras usadas como claves). Se asume que el uso de un vocabulario controlado lleva a un mejoramiento en el rendimiento de recuperación. Sin embargo, la reducción del vocabulario puede hacer más difícil para el usuario la especificación de una consulta como la interpretación de una respuesta.

El preprocesamiento de un texto puede ser dividido en 5 tipos de operaciones de texto:

- Análisis léxico del texto de manera de tratar dígitos, puntuaciones, guiones, mayúsculas/minúsculas, etc..
- Eliminación de stopwords, con el objeto de filtrar palabras con baja capacidad discriminadora para propósitos de recuperación.
- Stemming, con el objeto de permitir la recuperación de documentos conteniendo términos de una consulta con variaciones sintáticas.
- Selección de palabras claves.
- Construir estructuras de categorización de términos, tales como tesauros.



Análisis Léxico

El análisis léxico convierte una cadena de caracteres en una cadena de palabras. Además de separar palabras por espacios, este análisis debe considerar los siguientes casos:

- Dígitos. En general números no son buenos candidatos de palabras claves. Sin embargo, la normalización de ciertos números en el contexto de ciertas palabras pueden ser relevantes para la recuperación de información.
- Guiones. Este es otra tarea difícil para la discriminación del analizador. Puede que sea o no sea relevante la eliminación de guiones. En general, se adopta una regla y se agregan excepciones.
- Puntuaciones son generalmente removidas.
- Generalmente el texto es transformado a mayúscula o minúscula.

Eliminación de Stopwords

Las palabras que aparecen con frecuencia entre los documentos no son buenas para la recuperación de información. Así palabras que aparecen en más del 80% de documentos no son consideradas y se les llama “stopwords”. Artículos, preposiciones, conjunciones son candidatos naturales a ser stopwords.

Además de mejorar la selección de palabras claves, la eliminación de stopwords reduce el tamaño de los índices (menos de 40% del original). Por otro lado, la eliminación de stopwords puede reducir el *recall*, lo que hace que sea interesante la indexación del texto completo.

Stemming

Stem es lo que queda de una palabra después de eliminar todos los prefijos y sufijos. Stemming también reduce el tamaño del índice ya que el número de distintas palabras también es reducido. Sin embargo, algunos sistemas prefieren no aplicar stemming ya que existen estudios con resultados contradictorios.

Un algoritmo clásico es el *affix removal*, el cual considera que la mayoría de las variantes de una palabra son productor de sufijos (en vez de prefijos). Existen tres bien conocidos algoritmos de sufijos. El algoritmo usa una lista de sufijos para el análisis. Por ejemplo,

sses → ss
ies → i
ss → ss
s → Ø
ed → Ø
ing → Ø
.....

Selección de Palabras Claves

Cuando el texto completo es adoptado, todo se indeza. Para una selección automática, un buen enfoque es el uso de sustantivos. Debido a que es común combinar dos o tres sustantivos en un único concepto, se pueden usar grupos de sustantivos. Un grupo de sustantivos es el cual tiene una distancia sintáctica en el texto que no excede un umbral especificado.

Tesaurus

El propósito de un thesaurus es

- entregar un vocabulario estándar
- ayudar a los usuarios a localizar palabras para la formulación de consultas
- dar una jerarquía de clasificación para modificar la consulta.