Procesamiento de Texto

UCR – ECCI

CI-2414 Recuperación de Información

Prof. Kryscia Daviana Ramírez Benavides

Aspectos Generales

- El *procesamiento 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.

Aspectos Generales (cont.)

- No todas las palabras son igualmente significativas para representar la semántica de un documento.
- En lengua escrita, algunas palabras llevan más significado que otras.
- Generalmente, las palabras sustantivas (o los grupos de palabras sustantivas) son las que en la mayoría representan el contenido del documento.
- Y estos términos se pueden utilizar como términos del índice, se debe tener algún proceso/pasos para obtener términos del índice.

Aspectos Generales (cont.)

- El procesamiento previo de los documentos en la colección se puede ver simplemente como un proceso de controlar el tamaño del vocabulario (es decir, el número de las palabras distintas usadas como los términos del índice).
- **Importante:**
  - A pesar de una mejora potencial en el funcionamiento de la recuperación, las transformaciones del texto hechas pueden hacer más difícil la interpretación de las necesidades del usuario.
  - Realmente, algunos motores de búsqueda en el Web están haciendo operaciones sobre el texto y están indexando todas las palabras del texto, usando ahora búsqueda con texto completo.

Aspectos Generales (cont.)

- Las operaciones del texto incluyen:
  - Normalización del texto.
  - Construcción de un tesauro.
  - Compresión del texto.
  - Cifrado.
- La normalización del texto y la construcción de un tesauro son estrategias dirigidas para mejorar la precisión de los documentos recuperados.
- Para reducir tiempo de respuesta ante la necesidad de información del usuario, uno puede considerar la compresión del texto.

Aspectos Generales (cont.)

- Un buen algoritmo de compresión puede reducir el texto a 30-35% de su tamaño original.
- El texto comprimido requiere menos espacio de almacenaje y toma menos tiempo para ser transmitido sobre un enlace de comunicaciones.
- La principal desventaja de la compresión es el tiempo de comprimir y descomprimir el texto.
- Otra operación al texto es el cifrado, debido a la popularización rápida de los servicios en el Web, las preguntas dominantes (y viejas) con respecto a seguridad y el aislamiento que han emergido otra vez.

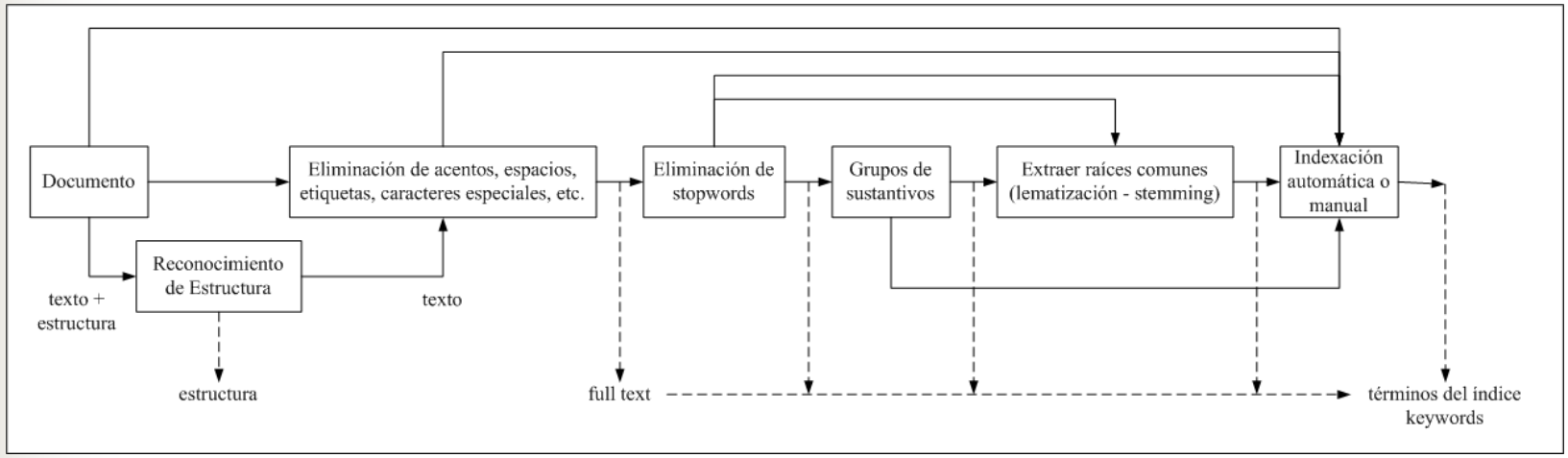
Procesamiento de Documentos

- El procesamiento sobre el texto puede ser dividido en cinco tipos de operaciones de texto (o transformaciones):
  - *Análisis léxico del texto*, con el objetivo de tratar dígitos, guiones, los signos de puntuación, mayúsculas/minúsculas, etc.
  - *Eliminación de stopwords*, con el objetivo de filtrar palabras con baja capacidad discriminadora para propósitos de recuperación.
  - *Selección de términos índice*, con el objetivo de seleccionar los términos del índice, para determinarse que palabras (o los grupos de palabras) serán utilizadas como elementos de la indexación de direcciones. Generalmente, la decisión si una palabra particular será utilizada como término del índice se relaciona con la naturaleza sintáctica o a la palabra. De hecho, las palabras sustantivas llevan más semántica que adjetivos, adverbios, y verbos.

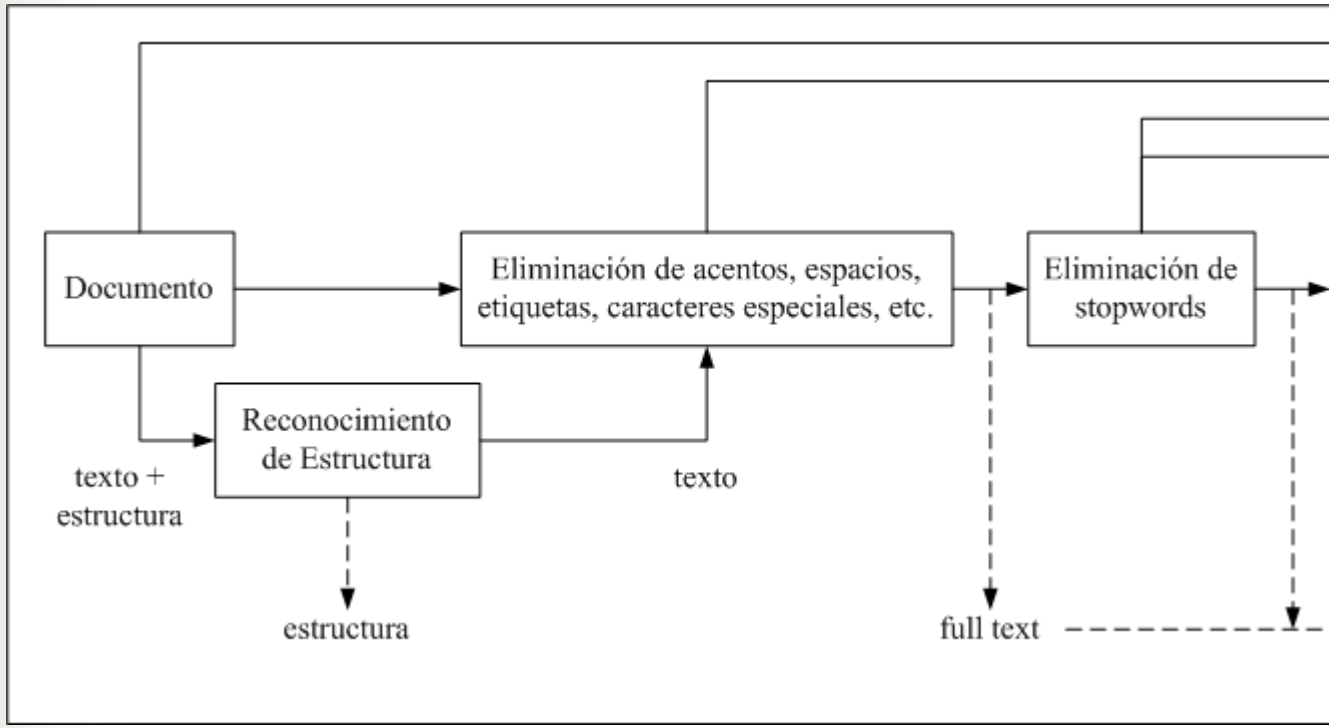
Procesamiento de Documentos (cont.)

- El procesamiento sobre el texto puede ser dividido en cinco tipos de operaciones de texto (o transformaciones):
  - *Stemming*, con el objetivo de quitar afijos (es decir, prefijos y sufijos) y de permitir la recuperación de los documentos que contienen variaciones sintácticas de los términos de la necesidad de información.
  - *Construir estructuras de categorización (clasificación) de términos*, tales como tesauros, o extracción de la estructura representada directamente en el texto, para permitir la extensión de la necesidad de información original con los términos relacionados (un procedimiento generalmente útil).

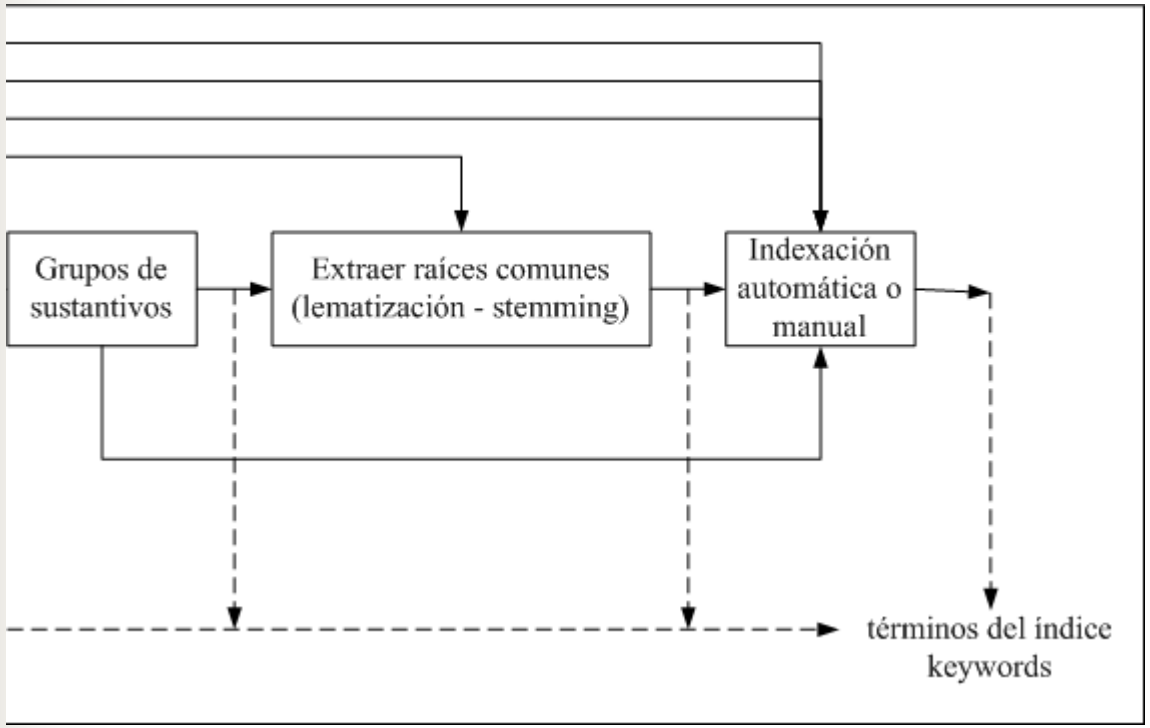
Aspectos Generales (cont.)



Aspectos Generales (cont.)



Aspectos Generales (cont.)



Análisis Léxico

- El análisis léxico es el proceso de convertir un conjunto de caracteres (el texto de los documentos) en un conjunto de palabras (las palabras candidatas que se adoptarán como los términos del índice).
- Convierte una cadena de caracteres en una cadena de palabras.
- Así, uno de objetivos importantes de la fase del análisis léxico es la identificación de las palabras en el texto y el reconocimiento de:
  - **Espacios**, como separadores de palabras.
  - **Dígitos**, estos no son generalmente buenos términos índice porque sin un contexto, son intrínsecamente vagos.

Análisis Léxico (cont.)

- **Guiones**, al encontrarse palabras escritas con guión no puede ser útil debido a la inconsistencia del uso. Por ejemplo, esto permite tratar “estado-del-arte" y de "estado del arte" idénticamente. Sin embargo, hay las palabras que incluyen guiones como parte integral: cerda-borde, B-49, etc.
- **Signos de puntuación**, normalmente se quitan por completo en el análisis léxico. Aunque algunos signos de puntuación son una parte integral de la palabra (' 510B.C '), quitarlos no tienen un impacto en el buen funcionamiento de la recuperación porque el riesgo de la interpretación en este caso es mínimo.
- **Letras mayúsculas/minúsculas**, no es generalmente importante para la identificación de los términos índice. Casi siempre, un analizador léxico convierte generalmente todo el texto a minúscula o a mayúscula.

Análisis Léxico (cont.)

- Generalmente realiza lo siguiente:
  - *Dígitos*. La normalización de ciertos números en el contexto de ciertas palabras pueden ser relevantes para la recuperación de información, se establece un rango.
  - *Guiones*. Puede que sea o no sea relevante la eliminación de guiones. En general, se adopta una regla y se agregan excepciones.
  - *Tildes y caracteres especiales*. Puede que sea o no sea relevante la eliminación de tildes y caracteres especiales. En general, se adopta una regla y se agregan excepciones.
  - *Los signos de puntuación* son generalmente removidos.
  - *Generalmente el texto es transformado a mayúscula o minúscula.*

Eliminación de *Stopwords*

- En un primer paso todas las palabras son buenos candidatos.
- 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*:
  - Los artículos, los pronombres, las preposiciones, y las conjunciones son candidatos naturales.
  - Algunos verbos, adverbios, y adjetivos se podían tratar como *stopwords*.
  - Los términos específicos de un dominio se podían tratar como *stopwords*.

Eliminación de *Stopwords* (cont.)

- Se suele tener una lista de palabras que no son buenos términos de indexación llamada *STOPLIST*, *Lista de Palabras Vacías* o *Diccionario Negativo*.
- La salida del analizador léxico es comprobada con la STOPLIST y se eliminan los términos que aparecen en ella.
- Incorporar la eliminación de las palabras vacías en el analizador léxico:
  - Es más eficiente.
  - No suele ser necesario en la mayoría de los casos.

Eliminación de *Stopwords* (cont.)

- Ventajas:
  - La indexación es más rápida
  - Las palabras vacías aparecen mucho y su lista de referencias es muy grande:
    - Si las quitamos el archivo invertido será más pequeño.
    - El archivo invertido se reduce en un 30% ó 40%.
  - Aumenta la eficiencia, ya que mejora la selección de palabras claves.
- Desventajas:
  - Por otro lado, la eliminación de *stopwords* puede reducir el recall, lo que hace que sea interesante la indexación del texto completo. Por ejemplo, si el usuario está buscando documentos que contienen la frase "ser o no ser", la eliminación de *stopwords* hace casi imposible reconocer correctamente los documentos que contienen la frase especificada.

Selección de Palabras Claves

- Cuando el texto completo es adoptado, todo se indexa.
- Es alternativo adoptar una visión más abstracta en la cual no todas las palabras se utilicen como términos índice.
- Para una selección automática, un buen enfoque es el uso de sustantivos. Aunque, también se usan los adjetivos, verbos y muchas veces los adverbios.
- 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.

Selección de Palabras Claves (cont.)

- Dos maneras posibles de lograr esto:
  - La palabra clave es seleccionada por un especialista.
  - Se seleccionan las palabras claves automáticamente usando un programa:
    - Una oración del lenguaje natural se compone generalmente de sustantivos, de pronombres, de artículos, de verbos, de adjetivos, de adverbios, y de conectores. Mientras que las palabras en cada clase gramatical se utilizan con un propósito particular, la mayoría de la semántica la tienen los sustantivo. Así, una estrategia intuitiva para seleccionar automáticamente términos índice es utilizar los sustantivos en el texto.
    - Puesto que es común combinar dos o tres sustantivos en un solo componente, tiene sentido agrupar los sustantivos que aparecen cerca en el texto en un solo componente de indexación (o concepto). Así, utilizamos grupos de sustantivos, el cual es un sistema de sustantivos cuya distancia sintáctica en el texto (medido en términos del número de palabras entre 2 sustantivos) no excede a un umbral predefinido (por ejemplo, 3).

Lematización o *Stemming*

- Con frecuencia, el usuario especifica una palabra en una consulta pero solamente una variante de esta palabra está presente en un documento relevante.
- Un *stem* (lema) es la porción que queda de una palabra después de retirar los afijos (es decir, los prefijos y los sufijos).
- Un ejemplo típico es la palabra conecta que es el *stem* para las variantes: conectadas, conectando, conexión y conexiones.
- Consiste en convertir todas las palabras parecidas a una forma común (literalmente "obtención del tronco"), no es hallar la raíz léxica.

Lematización o *Stemming* (cont.)

- Se pretende agrupar términos en un solo término de indexación, se pretende agrupar:
  - Plurales y género.
  - Formas del gerundio.
  - Sufijos de tiempo para los verbos.
  - Prefijos, como de negación (in), ya no pertenece (ex), interno (intra), externo (inter), etc.
- La substitución de las palabras por sus *stems* respectivos.
- Obtención mediante patrones.
- Sin embargo, algunos sistemas prefieren no aplicar *stemming*, ya que existen estudios con resultados contradictorios.

Lematización o *Stemming* (cont.)

- Técnicas:
  - Búsqueda en una tabla que tiene todas las derivaciones de un término común.
    - Sencillo y simple.
    - Problemas:
      - Hay que construir la tabla.
      - Es difícil para palabras específicas a un dominio.
      - Dependiente al lenguaje.
      - Requiere espacio de almacenamiento considerable.
      - No es práctico.

Lematización o *Stemming* (cont.)

- Técnicas:
  - Obtención de la variedad de sucesores:
    - Está basado en determinar los límites del morfema.
    - Propiedad estructural de la mayoría de los lenguajes.
    - Las terminaciones de las palabras siguen determinadas pautas.
    - No es necesario construir una tabla pues se construye a partir de una colección
    - Consiste en agrupar palabras con la misma "raíz" .
      - Ej.: **disco**, **disco**s, **disco**teca, **disco**grafía.
    - Es más complejo que los algoritmos de eliminación de afijos.

Lematización o *Stemming* (cont.)

- Técnicas:
  - N-gramas:
    - No pretende obtener una forma común, sino determinar clases o grupos de términos.
    - Es heurístico
    - Se buscan los que comparten un nº mayor de n-gramas, basado en identificar bigramas y trigramas.
    - Es más un proceso de *clustering* de términos que un proceso de *stemming*.

Lematización o *Stemming* (cont.)

- Técnicas:
  - Algoritmos de eliminación de afijos:
    - Es intuitivo, simple y puede ser implementado eficientemente.
    - No son reglas heurísticas, son reglas que aplicadas a las palabras nos dan su forma común, se basan en reglas gramaticales aplicadas al revés.
    - La parte más importante es eliminar sufijos porque existen más variantes de una palabra, que son generadas por la introducción de sufijos
    - El más conocido y usado es el algoritmo de Porter:
      - 30 – 40 reglas.
      - Sólo elimina sufijos.
    - Ventajas:
      - Con un número pequeño de reglas obtengo una buena eficiencia.
      - Ante una nueva palabra puedo sacar su raíz.
    - Desventajas:
      - Hay que construir la tabla de reglas.
      - Dependen del idioma.

Lematización o *Stemming* (cont.)

- Ventajas:
  - Reduce el tamaño del índice, ya que el número de palabras también es reducido.
  - Mejora la importancia de la recuperación porque las variantes de la misma palabra se reducen a un concepto común.

Tesauros

- En los sistemas RI los tesauros se usan para coordinar los procesos básicos de indización y recuperación de documentos.
- Los tesauros RI contienen típicamente:
  - Una lista de términos (siendo cada término una palabra o una frase).
  - Las relaciones entre ellos.
- Proporcionan un vocabulario común, preciso y controlado que ayuda a la coordinación entre la indización y la recuperación.
- Con este objetivo, los tesauros habrán de diseñarse para áreas específicas y serán dependientes del dominio.

Tesauros (cont.)

- El propósito de un tesauros 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.

Tesauros (cont.)

- Un tesauro diseñado cuidadosamente puede ser de gran valor:
  - En la indización se debe poner especial cuidado en la elección de los términos tesaurales más apropiados para representar a los documentos
  - En la búsqueda, el usuario podrá utilizar el tesauro para diseñar la estrategia de búsqueda más adecuada
    - Si la búsqueda no recupera suficientes documentos se puede usar el tesauro para expandir la petición siguiendo los variados enlaces entre términos.
    - Si la búsqueda recupera demasiados documentos el tesauro puede sugerir términos de búsqueda más específicos.

Tesauros (cont.)

- Por tanto, el tesauro es muy valioso en la reformulación de la estrategia de búsqueda:
  - Es muy común proporcionar tesauros en línea que simplifican el proceso de reformulación de la petición.
  - Además, estas modificaciones de la petición pueden ser llevadas a cabo por el sistema en vez de hacerlo el usuario.
  - Aunque ello conlleve bastante complejidad algorítmica, dado que el sistema debe saber:
    - Cómo reformular la petición?
    - Cuándo hacerlo?

Tesauros (cont.)

- Existe una buena literatura sobre los principios, metodología y problemas implicados en la construcción de tesauros:
  - Sin embargo, solo una pequeña parte está dedicada a la construcción automática de tesauros:
    - Esto refleja el actual estado de la cuestión, marcado por una abundancia de tesauros generados de forma manual.
    - De hecho, existe mucho escepticismo en cuanto a la posibilidad de automatizar completamente su proceso de construcción.
  - Esto se debe a que los tesauros manuales son estructuras altamente complejas que exhiben una amplia variedad de relaciones:
    - Jerárquicas.
    - No-jerárquicas.
    - De equivalencia.
    - Asociativas.

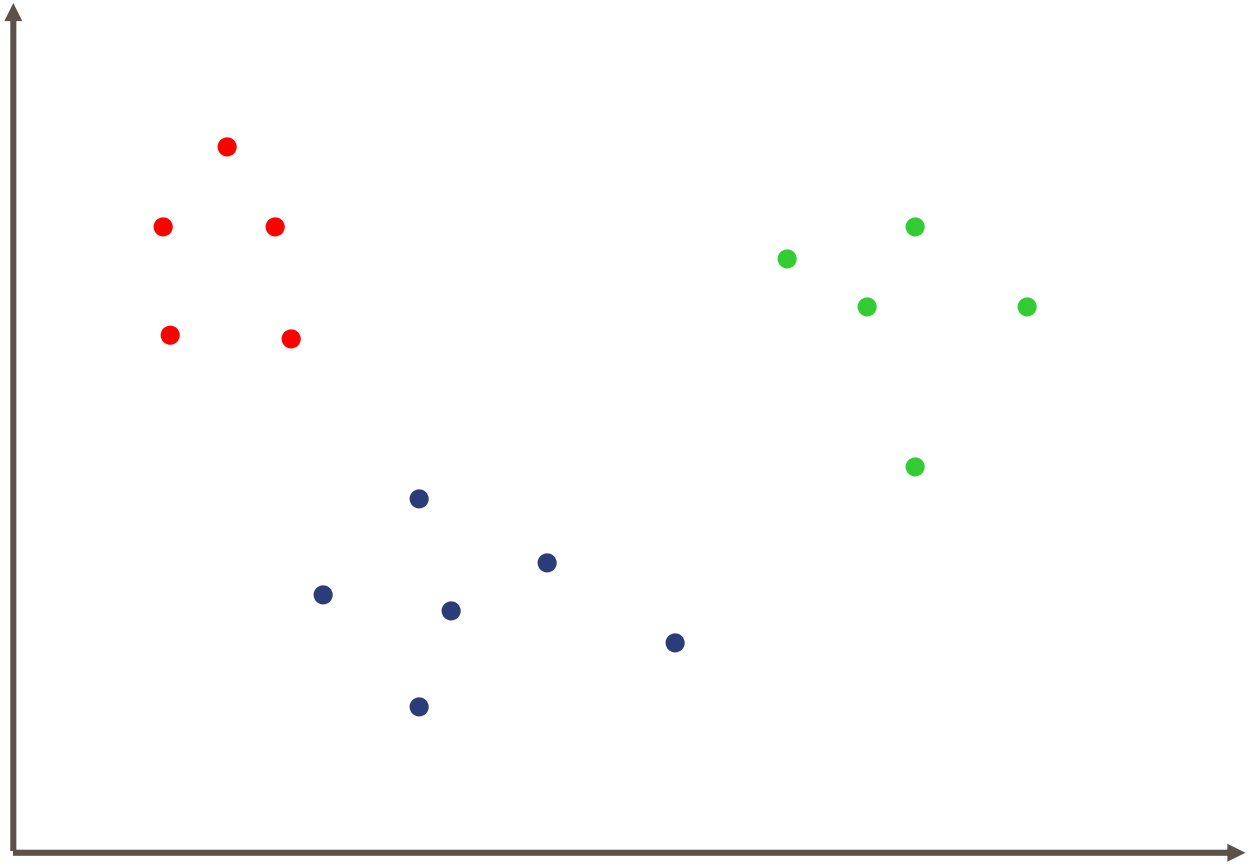
Tesauros (cont.)

- La detección automática de estos tipos de relaciones continua siendo un reto:
  - Sin embargo, algunas metodologías automáticas han emergido recientemente.
  - Al igual que la mayoría de los subcampos de la RI estos métodos están influenciados fuertemente por la estadística.
- La construcción manual de tesauros es una tarea altamente conceptual y de conocimiento intensivo, y por lo tanto un proceso muy laborioso.

Clustering

- Es la división de los datos en grupos significantes llamados *clusters*. Ayuda a realizar una agrupación natural o estructural de un conjunto de datos.
  - Es un método alternativo que permite organizar los resultados obtenidos a través de grupos de términos (cluster base) tomando en cuenta algún tópico en especial.
  - Es una técnica para presentar los términos después de haberlos recuperado en grupos pequeños que están.
  - Es una técnica estadística que se usa para generar una estructura de categorías y para agrupar un conjunto de términos.

Clustering (cont.)



Compresión

- La compresión de texto es para buscar maneras de representar el texto en pocos bits o bytes.
- La cantidad de espacio requerido para almacenar el texto en la computadora puede ser reducido significantemente usando técnicas de compresión:
  - Los métodos de compresión reducen la representación al identificar y usar estructuras que existen en el texto.
  - Desde una versión comprimida, el texto original puede ser reconstruido exactamente.
  - La comprensión es considerada muy importante en el extenso uso de librerías digitales, sistemas automáticos, bases de datos textuales.

Compresión (cont.)

- Ventajas:
  - Reduce los costos asociados con los requerimientos de espacio en disco, *overhead* de I/O y retrasos de comunicación.
  - Toma menos tiempo en la búsqueda.
- Desventajas:
  - El precio del tiempo necesario para codificar y descodificar el texto.
  - El mayor obstáculo para almacenar texto comprimido es la necesidad de los SRI de acceder aletoriamente al texto, ya que para accesar una palabra en el texto comprimido es usualmente necesario decodificar por entero el texto del principio hasta que se alcanza la palabra deseada.
    - Un texto grande se podría dividir en bloques que se comprimen independientemente, así permite el acceso aleatorio y rápido a cada bloque.
    - Sin embargo, los métodos eficientes de la compresión necesitan procesar un poco de texto antes de hacer la compresión eficaz (generalmente más de 10 kilobytes).
    - Cuanto más pequeños son los bloques, menos eficiente se espera la compresión.
    - La velocidad de descompresión es más importante que velocidad de compresión.

Compresión (cont.)

- Otra característica importante de un método de compresión es la posibilidad de realizar la concordancia con el texto comprimido, realizar la tarea de búsqueda con el modelo en un texto comprimido sin descomprimirlo.
- En este caso, la búsqueda secuencial puede ser acelerada comprimiendo la llave de búsqueda, más bien que descifrando el texto comprimido que es buscado.

Compresión (cont.)

- Dos aproximaciones de compresión de texto son:
  - **Métodos estadísticos:** Confía en la generación de buenas estimaciones de la probabilidad (del aspecto en el texto) para cada símbolo. Cuanto más exactas son las estimaciones, mejor es la compresión obtenida. Dos estrategias estadísticas bien conocidas de la codificación son:
    - Código de Huffman.
    - Código aritmético.
  - **Métodos de diccionario:** Substituyen una secuencia de símbolos por un indicador a una ocurrencia anterior de esa secuencia. Las representaciones del indicador son referencias a entradas en un diccionario integrado por una lista de los símbolos (a menudo llamados frases) que se espera que ocurran con frecuencia. Los métodos de diccionario más conocidos son los de la familia de *Ziv-Lempel*.

Referencias Bibliográficas

- La información fue tomada de:
  - Libro de texto del curso.
  - http://www.gedlc.ulpgc.es/docencia/seminarios/rit/.
  - http://www.cse.unl.edu/~lksoh/Classes/CSCE410_810_Fall03/sup6.html.