MINERÍA DE DATOS. MINERÍA DE TEXTO:

El objetivo de la minería de texto es el descubrimiento de información nueva a partir de colecciones de documentos de texto no estructurado. Esto quiere decir que la información que contienen dichas colecciones de documentos son en lenguaje natural aunque también se podría utilizar otro tipo de información textual.

La minería de textos se dedica principalmente a la categorización, clasificación y agrupamiento de textos.

- La categorización se encarga de identificar las categorías, temas, materias o conceptos presentes en los textos
- La clasificación se encarga de asignar una clase o categorización de textos

Según diversos autores, categorización y clasificación se utilizan como sinónimos; aunque otros autores piensan que la categorización forma parte de la clasificación.

Es de gran importancia decir que la minería de textos (realizada de forma automática) juega un papel muy interesante en la manipulación de información tanto dinámica como personalizada, porque, por ejemplo, cataloga artículos de novedosa aparición o paginas web.

Lo primero que se realiza en la minería de textos es representar el texto en un formato determinado como los algoritmos de aprendizaje. Para ello, lo inicial es usar la representación más abstracta:

- "Bolsas de Palabras", es decir, la representación de información se basa en vectores (cada documentos se representa como un vector de dimensión según el número de palabras y de acuerdo cono unas determinadas características, esto es, si aparece o no en el documento y la frecuencia con la que aparece)
- Frases: el documento es considerado como un conjunto de frases sintácticas, al igual que se realiza en el procesamiento del lenguaje natural
- N-Gramas: la información se utiliza según la posición en el texto. Se suele utilizar para tratar la información del texto expresada en frases negativas
- Representación Relacional: se utiliza para detectar patrones complejos

- Categorías de Conceptos o Indexación Semántica Latente: sirve para reducir la aparición de las palabras según su raíz morfológica, es decir, las palabras se clasifican según la raíz semántica de los vocablos

El segundo paso consiste en reducir el conjunto de características originales, esto quiere decir que tenemos que eliminar las palabras con poca semántica o palabras vacías, como son los artículos, proposiciones, conjunciones…