Minería de textos

Salvador López Mendoza

Mayo de 2018

Introducción

Contexto:

- La cantidad de datos almacenados en soportes digitales ha crecido mucho en los últimos años.
  Necesidad de procesar esa *información*
  . . . para extraer *conocimiento*.
  Minería de Datos.
- La cantidad de *datos textuales*, en lenguaje natural, es cada vez mayor.
  Fuentes de datos textuales:
  Páginas web, noticias, literatura científica, correo electrónico, blogs, tweets, mensajes, foros, opiniones de productos, etc.

Se requiere de *herramientas de software* poderosas para ayudar a la gente a *analizar y manejar* grandes cantidades de textos en forma **eficiente y efectiva**.

PEUVI

Datos textuales

Los datos textuales son generados, en la mayoría de las ocasiones, por personas. *Son escritos por personas para ser leídos por personas.*

Requieren de un contexto para determinar con precisión su significado.

Hay muchas clases de conocimiento que se pueden codifcar en el texto, pero es muy valioso lo que se puede descubrir acerca de las *opiniones* y *preferencias* de la gente.

PEUVI

Datos textuales (II)

Problemas:

- El texto que corresponde a lenguaje natural carece de estructura.
  Son datos no-estructurados.
  La tecnología actual para el procesamiento de lenguaje natural no permite que las computadoras entiendan con precisión el texto en lenguaje natural.

Solución:

Se utilizan enfoques estadísticos y heurísticos para analizar y minar datos textuales.

Son métodos robustos y se aplican en prácticamente cualquier lenguaje natural.

PEUVI