TUTORIAL : Minería de datos textuales utilizando Treetagger y Weka.

Presentado por : Juan Antonio Lossio Ventura y Hugo Alatrista Salas

juan.lossio@lirmm.fr , hugo.alatristasalas@lirmm.fr

Descargar el software de acuerdo al sistema operativo que se tiene:

TreeTagger: http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

Weka: http://www.cs.waikato.ac.nz/ml/weka/downloading.html

La minería de textos o la extracción de conocimientos a partir de datos textuales es una especialización de la minería de datos. Ella permite, a partir de un conjunto de documentos textuales denominado corpus, extraer conocimientos según un criterio de novedad o de similaridad. Este tutorial se inscribe justamente en este marco. Para esto, utilizaremos el método denominado "bag of words", así como diferentes algoritmos de clasificación sobre un corpus otorgado por los expositores.

El proyecto se desarrollará en 4 etapas :

Etapa 1: Estudio del corpus.

Los participantes deberán conocer el corpus que utilizarán para realizar las experimentaciones. Así mismo, los participantes deben identificar las "clases" de documentos y en numero de documentos por clases.

Etapa 2: Utilización del software TreeTagger.

TreeTagger es una herramienta para la anotación de texto. Con esto podremos realizar el análisis morfo sintáctico de datos textuales.

Etapa 3: Utilización del software Weka.

Weka nos ofrece una gran variedad de algoritmos y herramientas. En vista que utilizaremos el método "saco de palabras" (bag of words), necesitamos representar nuestro corpus según tres modelos : Booleano, de Frecuencias y TD-IDF. Los participantes representarán el corpus en uno de los tres modelos antes citados utilizando las herramientas propuestas por Weka.

Etapa 4: Utilización de métodos de clasificación.

En esta etapa, los participantes utilizaran un protocolo experimental a fin de ponen en marcha tres algoritmos de clasificación propuestos por Weka: K-means, Bayes nativo y Arboles de decisión. En esta etapa, los participantes utilizarán su capacidad de análisis para evaluar e interpretar los resultados.

Etapa 5: Extracción de reglas de asociación.

Utilizando el corpus completo, los participantes están invitados a extraer las 20 primeras reglas de asociación (clasificados en función a la confianza). Existen algunos algoritmos propuestos por Weka, sin embargo, utilizaremos en esta etapa el algoritmo

“A priori”. Posteriormente, los participantes están invitados a realizar el mismo proceso (independientemente), pero para cada una de las clases identificada en la Etapa 1.

Finalmente, los participantes compararán “a la mano” los dos conjuntos de reglas de asociación extraídos en la etapa anterior (por cada clase). Nuevamente, la capacidad de análisis de los participantes será decisiva para analizar y comparar los resultados.