INGENIERIA INDUSTRIAL
UNIVERSIDAD DE CHILE

**IN4522 – WebMining**
Profesor: Juan D. Velásquez
Auxiliares: Felipe Bravo – Gabriel Oberreuter

# Tarea 1: Text Mining

El objetivo de esta tarea es realizar minería de datos en páginas web. Se realizará *crawling* de páginas de internet, para proceder a caracterizarlas y finalmente hacer clustering según su contenido. Para ello, se le pide utilizar el software libre *RAPIDMINER versión 5*.

Los conceptos a tener en cuenta para la realización de la tarea son: parser del contenido de la página, tokenizar, stopwords, TF-IDF y clustering (mediante k-Means).

Requerimientos:

- Elegir un tema (por ejemplo, deportes)
- Proveer al software una página web inicial donde comenzar
- Instruir al software a crawlear ~100 páginas
- Segmentación de las páginas de acuerdo a su contenido.

Al finalizar la tarea debe elaborar un informe, el cual debe contener al menos:

- los parámetros iniciales escogidos (tema, url inicial, etc)
- Analizar los resultados de clasificación para diferentes *k* (por ejemplo k=2,…,5)
- Escoger un *k* que según ud. sea adecuado y explicar por qué.
- Especificar los *centroides* y el vector más cercano a cada uno, describiendo cada tema, obteniendo así las posibles clasificaciones.

Debe entregar el informe + los archivos del proyecto de *rapidminer* el día jueves 9 de septiembre. Cada día de atraso restará 1 punto de la nota máxima. Se puede utilizar algún programa o código adicional para el trabajo; en tal caso deberá justificar su uso y documentar acorde el informe. Es recomendable apoyar el análisis mediante inspección visual, por ejemplo, de los valores de las tablas en Excel (se puede exportar a .xls en el software, para "ir viendo lo que se va realizando").

**Fecha de entrega**: 9 de septiembre 2010.
Grupos de a 2 personas.