

Curso Avanzado de Análisis Predictivo Minería de Textos (DMO42)



SUMILLA

Este curso expone las técnicas y tecnologías de la minería de textos. La **minería de textos** es el proceso para descubrir conocimiento almacenado en documentos (datos no estructurados), el conocimiento se puede representar como tendencias, promedios, desviaciones, dependencias, grupos, etc. Se entiende como la extensión de los métodos estándar de predicción y descripción de la minería de datos. La minería de textos comprende las siguientes actividades fundamentales: Clasificación de documentos para la asignación automática a clases pre-definidas. Agrupamiento de documentos para la identificación de documentos similares. Recuperación de información (similar a un buscador). Extracción de la información incluida en esos textos (hechos) y Extracción de asociaciones entre los hechos extraídos. Por otro lado, dado que el vector de características obtenido es por lo general muy grande se requiere el uso de técnicas para la reducción de la dimensionalidad.

OBJETIVOS

Al final del curso los alumnos estarán en capacidad de:

- Conocer y entender los fundamentos y problemáticas actuales de Minería de Textos en comparación a Data Mining tradicional.
- Comprender y usar las técnicas para el análisis y la preparación de documentos.
- Aplicar y evaluar técnicas de agrupamiento y clasificación de documentos.
- Aplicar y evaluar técnicas para la recuperación de documentos.
- Aplicar y evaluar técnicas para la extracción de información.
- Aprender a utilizar las herramientas disponibles para la minería de textos.

CONTENIDO

El curso está conformado por los siguientes temas.





TEMARIO

Nombre de los temas a tratar	Duración
INTRODUCCIÓN A LA MINERÍA DE TEXTOS (TEXT MINING) ¿Qué es la minería de textos?. Colecciones de documentos corpus. Datos semi-estructurados y datos no estructurados. Técnicas de minería de textos. El proceso de la minería de textos. Herramientas	2 h
PREPARACION DE DATOS Lingüística, Morfología, Sintaxis, Semántica. Conceptos básicos, Gramática, Léxico, Sintaxis. Tokenización. Stop words. Stemming.	4 h
MODELOS PREDICTIVOS PARA TEXTOS - CLASIFICACION Clasificación de documento. Similaridad de documentos y el vecino más cercano. Reglas de decisión. Red Bayesiana, Máquina de Vector Soporte. Evaluación del desempeño. Aplicaciones.	4 h
AGRUPAMIENTO (CLUSTERING) DE TEXTOS Medida de similitud para la recuperación. Búsqueda de documentos basados en la Web y análisis de links. Matching de Documentos. Agrupamiento por similitud. Agrupamiento K-means. Agrupamiento jerárquico. Evaluación del agrupamiento. Aplicaciones.	4 h
RECUPERACIÓN DE INFORMACIÓN (BINARIO Y VECTORIAL) Modelos de recuperación de información, Relevancia. Modelos clásicos, índice de términos, importancia, medidas de similitud. Modelo Booleano. Modelo Vectorial, pesos, similitud.	4 h
EXTRACCION DE INFORMACION DESDE TEXTOS Objetivos de extracción de información. Búsqueda de patrones y entidades. Expresiones regulares. Extracción de entidades y el método de máxima entropía. Plantillas de llenado. Aplicaciones. Tagging.	4 h
SELECCIÓN DE CARACTERÍSTICAS EN TEXTOS (REDUCCIÓN DE DIMENSIONALIDAD) Métodos de muestreo, métodos de selección de características, búsqueda.	2 h
Total de horas a dictar	24 h

DOMINIOS DE APLICACIÓN

- Análisis de similaridad de documentos
- Búsqueda e indexación de documentos
- Análisis de mensajes en redes sociales
- Análisis de encuestas abiertas.
- Análisis de post en blogs.
- Análisis de correos electrónicos (spam).
- Estructuración de base de datos.

METODOLOGÍA

- Desarrollo de clases teóricas para explicar los conceptos necesarios.
- Desarrollo de talleres, donde se aplica lo aprendido en clase.
- Uso de medios audiovisuales (proyectores)
- Materiales de clase impresos y en CD.

REQUISITOS

Es deseable que los alumnos tengan experiencia en (no indispensable):

- Estadística y probabilidades.
- Métodos de clasificación, agrupamiento y asociación

Disponer de una computadora para el desarrollo de las clases

HERRAMIENTAS

Para el desarrollo del curso se hace uso de las siguientes herramientas de software



QUIENES PUEDEN ASISTIR

Profesionales en TI. Consultores en minería de datos. Analistas de marketing en la web. Investigadores de mercado que desean analizar encuestas abiertas. Profesionales en estadística interesados en analizar el contenido de textos no estructurados (formularios, encuestas, etc.).

INSTRUCTOR

Ing. Samuel Oporto Díaz. Magíster en Inteligencia Artificial – ITESM-México. Especialización en robótica aplicada-CNAD-México DF. Ingeniero de Sistemas – UNI-Perú. Jefe del Proyectos en el CTIC-UNI. Investigador Principal del Instituto de Investigación de la FIIS (IIFIIS). Especialista en Visión Artificial, Reconocimiento de Patrones y Redes Neuronales. Docente del curso de Inteligencia Artificial en la UNI, UPAO, USMP y UPC. Docente del Curso de Minería de Datos en el IIFIIS, CTIC-UNI y la UPC. Investigador en Ciencias de Computación con publicaciones en: IJCNN2007, ICAIPR2007, ICIAR2005, LNCS2005, CLEI2004, CLEI2006. Consultor en Sistemas Inteligentes y Sistemas Autónomos. Consultor del programa de Modernización del Estado Peruano.





KASPeru

Calle Germán Schreiber 291, Piso 2, oficina 201

San Isidro

Lima – Perú

(51-1) 697-8227 (51-1) 725-7209

www.kasperu.com informes@kasperu.com

Todos los derechos reservados.

Todos los nombres de empresas y/o productos mencionados tienen propósitos de identificación únicamente, ellos son registrados por sus respectivos dueños.