

Información general del curso

| | |
|---|-----------------------------|
| Nombre de la asignatura | Text y Web analytics |
| Créditos académicos | 3 |
| Horas presenciales por semana | 3 horas |
| Horas de trabajo no presencial por semana | 6 horas |

TEXT Y WEB ANALYTICS

Ante el problema de los grandes volúmenes de información se plantea la necesidad de nuevas técnicas y herramientas para un análisis inteligente y automático de la información. La extracción de conocimiento a partir de datos se basa en técnicas inductivas y de aprendizaje automático y su uso se ha generalizado a diferentes áreas, desde la tradicional extracción de modelos a partir de bases de datos (minería de datos), hasta ámbitos relativamente más novedosos como la combinación de modelos y decisiones, la minería web, los sistemas recomendadores, y otras áreas.

Objetivos de aprendizaje

Objetivo General

El objetivo principal de este curso es presentar las técnicas y métodos para la extracción de conocimiento, e información, desde textos en particular y la web en general

Objetivos Específicos

Los principales objetivos de aprendizaje del curso serán:

- Conocer las características especiales de la extracción automática de conocimiento desde otras fuentes de información no estructurada (textos y web) y semiestructurada (XML).
- Ver las técnicas de aprendizaje automático más apropiadas y su adaptación a estos problemas.
- Conocer herramientas para intercambiar conocimiento.
- Conocer técnicas para extraer patrones de navegación y asistentes Web para personalización/recomendación.
- Conocer las herramientas computacionales R y Python para Text y Web analytics.

Evaluación

Aspectos a evaluar:

- Identificación y apropiación conceptual del aprendizaje automático
- Conocimiento y apropiación de los métodos clasificación, regresión y agrupamiento de datos
- Adquisición y aplicación de métodos
- Capacidad de desarrollar proyectos de aplicación

CALIFICACIONES

- 50 % Tareas y Talleres
- 50 % Proyecto

Dinámica de clase/ Metodología

El curso tiene un carácter crítico-constructivo, con clases magistrales que propicien el debate. Existirá material de apoyo como transparencias o copias. Al mismo tiempo, se trabaja sobre talleres con aplicaciones reales, los cuales deben ser solucionados con base en la fundamentación teórica impartida durante el curso y donde el estudiante confronte sus resultados. El trabajo se organizará mediante:

- Clases presenciales
- Prácticas en laboratorio
- Lecturas asignadas por el profesor
- Proyecto sobre un caso práctico
- Talleres

Trabajo del estudiante

- Reforzar los temas vistos en clase realizando lectura de textos y artículos relacionados
- Completar los ejercicios/talleres propuestos
- Realizar un laboratorio sobre Conocimiento de documentos No estructurados
- Realizar un laboratorio sobre Conocimiento de la Estructura
- Realizar un laboratorio sobre Conocimiento de Patrones de Uso
- Preparar un documento y presentación sobre los temas de proyectos/casos prácticos asignados

Contenido del curso

- Módulo 1. Conceptos Básicos de Recuperación de la Información:
- Taller 0 -> web scraping/adquirir redes sociales
- Módulo 2. Extracción de Conocimiento a partir de Documentos No Estructurados (Text Content Mining)
- Preprocesamiento de texto
 - Leyendo texto de diversas fuentes
 - Expresiones regulares
 - Tokenización
 - Lematización
 - Tagging: modelos de Markov
 - Analítica cuantitativa de texto
 - Matrices de documentos y texto
 - Frecuencias: frecuencias e importancia, Ley de Zip
 - Análisis de sentimientos
 - Co-ocurrencias, asociaciones, n-grams
 - Clusterización
 - Modelos de Tópicos latentes: LDA y CTM
 - Clasificación
- Módulo 3. Analítica de redes Sociales
- Caso de estudio de inteligencia competitiva con redes sociales
 - Librerías para adquisición de datos y análisis de datos
 - Taller. Análisis de datos de twitter
- Módulo 4. Extracción de Conocimiento a partir de Patrones de Uso (Web Usage Mining).
- Extracción de conocimiento a partir de herramientas en la nube. Caso de estudio: Google Cloud Platform.
 - Adquisición, Comportamiento y Conversión de usuarios. Caso de estudio: Google Analytics.
 - Taller: Analizando la información de Google Analytics con R.
- Módulo 5. Personalización, Recomendación y Asistentes (Web) 'Inteligentes'
- Herramientas para la creación de Bots conversacionales. Caso de estudio: Dialog Flow.
 - Taller dirigido: Aplicaciones de Procesamiento de Lenguaje Natural para el desarrollo de un ChatBot.
 - Detección de Intenciones
 - Reconocimiento de Entidades
 - Generación de texto basado en intenciones, entidades y reglas.

Bibliografía

1. Mastering Text Mining with R: Master text-taming techniques and build effective text-processing applications with R. Ashish Kumar, Avinash
2. Fundamentals of Predictive Text Mining. Paul Sholom, M. Weiss, Nitin Indurkha, Tong Zhang
3. Mining Text Data. Charu C. Aggarwal, ChengXiang Zhai (Eds)
4. Practical Text Mining With Perl. Roger Bilisoly
5. Natural Language Processing and Text Mining. Anne Kao and Stephen R. Potet (Eds)
6. THE TEXT MINING HANDBOOK: Advanced Approaches in Analyzing Unstructured Data. Ronen Feldman, James Sanger¹
7. Aggarwal, Charu C., Han, Jiawei (Eds.) Frequent Pattern Mining. Springer Verlag. 2014.
8. Manning, D., Raghavan, P. and Schütze, H. Introduction to Information Retrieval, Cambridge University Press. 2008.
9. Lliu, Bing. Web Data Mining Exploring Hyperlinks, Contents, and Usage Data. Springer Verlag. 2011.
10. Chakrabarti, S. Mining the Web: Discovering Knowledge from Hypertext Data 1st Edition. Morgan Kauffman Publishers. 2003.
11. Sakurai S. Theory and Applications for Advanced Text Mining. InTech. 2011. <http://www.intechopen.com/books/theory-and-applications-for-advanced-text-mining>
12. Perkins, J. Python Text Processing with NLTK 2.0 Cookbook. Cookbook. 2010.

Enlaces de Interés

<https://datawarrior.wordpress.com/2018/01/22/document-term-matrix-text-mining-in-r-and-python/>

https://www.analyticsvidhya.com/blog/2018/10/mining-online-reviews-topicmodelinglda/?utm_source=feedburner&utm_medium=email&utm_campaign=Feed%3A+AnalyticsVidhya+%28Analytics+Vidhya%29

<http://www.rtexttools.com/>

<https://cran.r-project.org/web/packages/syuzhet/vignettes/syuzhet-vignette.html>

<http://jorpela.fi/perl/regexp.html>

<https://www.tidytextmining.com/>