Formulario de Aprobación Curso de Posgrado

Asignatura: "Minería de textos"

Profesor de la asignatura [1] : Dr. Adrien Guille, Maître de Conférences, (Université de Lyon, Lumière 2, ERIC EA 3083).

(título, nombre, grado o cargo, Instituto o Institución)

Profesor Responsable Local [1] : Dr. Mathias Bourel, Grado 3, DT, IMERL.
(título, nombre, grado, Instituto)

Otros docentes de la Facultad:
(título, nombre, grado, Instituto)

Docentes fuera de Facultad:
(título, nombre, cargo, Institución, país)

Instituto ó Unidad: Instituto de Matemática y Estadística "Prof. Ing. Rafael Laguardia".
Departamento ó Area: Laboratorio de Probabilidad y Estadística.

[1] Agregar CV si el curso se dicta por primera vez.
(Si el profesor de la asignatura no es docente de la Facultad se deberá designar un responsable local)

Horas Presenciales: 24
(se deberán discriminar las mismas en el ítem Metodología de enseñanza)

Nº de Créditos: 5
(de acuerdo a la definición de la UdelaR, un crédito equivale a 15 horas de dedicación del estudiante según se detalla en el ítem metodología de la enseñanza)

Público objetivo y Cupos:
Estudiantes de la Licenciatura en Estadística, Maestría en Ingeniería Matemática y otros estudiantes de posgrado interesados.

Objetivos: Introducir los aspectos metodológicos de minería de texto así como algunas contribuciones modernas. Aplicar dichas técnicas a conjuntos de datos reales e interpretar los resultados obtenidos. Acercar el estudiante al empleo de los paquetes para el análisis estadístico de datos disponibles en el ambiente de desarrollo de software libre R (http://www.r-project.org/).

Conocimientos previos exigidos: Un curso de Análisis Multivariado.

Conocimientos previos recomendados: Uso del software estadístico R.

Metodología de enseñanza:
(comprende una descripción de la metodología de enseñanza y de las horas dedicadas por el estudiante a la asignatura, distribuidas en horas presenciales -de clase práctica, teórico, laboratorio, consulta, etc.- y no presenciales de trabajo personal del estudiante)

- Horas clase (teórico): 12
- Horas clase (práctico):
- Horas clase (laboratorio): 6
- Horas consulta:5

- Horas evaluación:2
    - Subtotal horas presenciales:25
- Horas estudio: 15
- Horas resolución ejercicios/prácticos:10
- Horas proyecto final/monografía:25
    - Total de horas de dedicación del estudiante: 75

Forma de evaluación: carpeta de ejercicios

Temario:

1. Vectorisación de texto y test de hipótesis sobre la frecuencia de palabras.
2. Clasificación supervisada de documentos (clasificador bayesiano ingenuo, regresión logística)
3. Modelado Temático (NMF, LDA)
4. Clasificación supervisada de documentos (kmeans)
5. Aprendizaje Profundo : word embedding & LSTM

Bibliografía:

- **Turney & Pantel**. From frequency to meaning: vector space models of semantics. Journal of artificial intelligence research (37), pp. 141-188, 2010.
- **Lijffijt et al**. Significance testing of word frequencies in corpora. Literary and linguistic computing (31), pp. 374-397, 2016.
- **Bishop**. Pattern recognition and machine learning. Springer. 2006.
- **Blei et al**. Latent Dirichlet allocation. Journal of machine learning research (3), pp. 993-1022, 2003.
- **Mikolov et al**. Distributed representations of words and phrases and their compositionality. Proceedings of the international conference on neural information processing systems, pp. 3111-3119, 2013.
- **Kim**. Convolutional neural networks for sentence classification. Proceedings of the international conference on empirical methods in natural language processing, pp. 1746-1751, 2014.

Datos del curso

Fecha de inicio y finalización: 01/10/18 al 12/10/18 (2 semanas)

Horario y Salón: Lunes, Miércoles y Jueves de 17 a 19. Salón de seminarios IMERL