Formulario de aprobación de curso de posgrado/educación permanente

Asignatura: Ciencia de Datos y Lenguaje Natural
(Si el nombre contiene siglas deberán ser aclaradas)

Modalidad: (posgrado, educación permanente o ambas)	**Posgrado**	**X**
	Educación permanente	

Profesor de la asignatura [1]: Profesora Titular Dra. Ing. Dina Wonsever, Instituto de Computación
(título, nombre, grado o cargo, instituto o institución)

Profesor Responsable Local [1]:
(título, nombre, grado, instituto)

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

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

[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)

[Si es curso de posgrado]
Programa(s) de posgrado: Maestría / Doctorado en Informática Pedeciba-Udelar, Maestría en Ciencia de Datos y Aprendizaje Automático (CDAA)

Instituto o unidad: Instituto de Computación

Departamento o área: Grupo Procesamiento de Lenguaje Natural

Horas Presenciales: 55
(se deberán discriminar las horas en el ítem Metodología de enseñanza)

Nº de Créditos: 10
[Exclusivamente para curso de posgrado]
(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 enseñanza)

Público objetivo: Estudiantes de posgrado en Informática o Ciencias de Datos.

Cupos: El curso requiere de al menos 2 docentes con un cupo de 30 estudiantes. El cupo podrá extenderse con mayor dotación docente.
(si corresponde, se indicará el número de plazas, mínimo y máximo y los criterios de selección. Asimismo, se adjuntará en nota aparte los fundamentos de los cupos propuestos. Si no existe indicación particular para el cupo máximo, el criterio general será el orden de inscripción, hasta completar el cupo asignado)

Objetivos: Los objetivos del curso son brindar formación en fundamentos, técnicas y aplicaciones de procesamiento automático de lenguaje natural con metodologías basadas en datos. El curso es teórico-práctico e incluye experimentación a través de trabajos de taller.

Conocimientos previos exigidos: Formación básica en Matemática, Estadística, Programación.

Conocimientos previos recomendados: Además de los conocimientos exigidos son de interés conocimientos en Lógica, en Lingüística y en Procesamiento de Lenguaje Natural

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)

Descripción de la metodología:
[Obligatorio]
El curso está organizado en unidades temáticas. Por cada unidad se prevé un conjunto de clases téoricas expositivas previamente filmadas, sesiones interactivas de discusión y trabajos de taller. Las sesiones interactivas de discusión serán de asistencia obligatoria y otorgarán la posibilidad de generar puntuación para el resultado final. Los estudiantes realizarán además presentaciones de artículos científicos, con posterior respuesta a preguntas y discusión. Se constituirán equipos para la realización de los talleres y la presentación de artículos. La prueba escrita será individual.

Detalle de horas:

- Horas de clase (teórico): 20
- Horas de clase (práctico): 10
- Horas de clase (laboratorio):10
- Horas de consulta: 12
- Horas de evaluación: 3
    - Subtotal de horas presenciales: 55
- Horas de estudio: 40
- Horas de resolución de ejercicios/prácticos: 30
- Horas proyecto final/monografía: 30
    - Total de horas de dedicación del estudiante: 155

Forma de evaluación:
El curso se evalúa en base a tareas de taller y proyecto (contribuye con 40% al puntaje total), la presentación de un artículo en clase (contribuye con 12% al puntaje total), intervenciones orales en las clases interactivas (contribuye con 8% al puntaje total) y una prueba individual escrita (contribuye con 40% al puntaje total). La prueba final y los talleres tienen cada uno un mínimo de suficiencia de un 60%.

Temario:

1- Conceptos básicos de gramática y de procesamiento automático del lenguaje

En esta unidad el estudiante adquiere conceptos básicos relativos al lenguaje humano y a algunas herramientas informáticas que se han desarrollado para procesarlo. Se usan herramientas ya hechas de distintos tipos, analizando fundamentalmente cómo y para qué se utilizan y cómo se mide la performance y se visualizan los resultados.

2- Corpus, experimentos, modelos, medidas de evaluación.

Los enfoques empíricos en procesamiento automático de lenguaje se basan usualmente en repositorios de datos lingüísticos y en experimentos y modelos que se diseñan según el problema particular del que se trate.

En esta unidad se verán conceptos de estadística y de teoría de la información que proporcionan medidas para evaluar modelos y experimentos sobre datos lingüísticos y se experimentará con casos simples.

3. Recursos para el Procesamiento de Lenguaje

Los métodos actuales de procesamiento de lenguaje se apoyan fuertemente en datos. Además de los datos propios a cada aplicación existen recursos de uso general, tales como corpus , diccionarios o bases de datos léxicas, o repositorios generales de conocimiento del mundo, que son aprovechables por diversas aplicaciones. Algunas de estos repositorios fueron construidos manualmente, mientras que otros contienen datos extraídos de modo automático de grandes corpus, en los que se incluye la web.
El objetivo de esta unidad es conocer los recursos de uso más extendido y experimentar con herramientas que los soportan.

4. Representaciones semánticas

Una de las innovaciones recientes más significativas para la semántica léxica es la propuesta de representaciones vectoriales para las palabras. El objetivo de esta unidad es adquirir conceptos básicos del análisis del significado en el lenguaje humano y experimentar con modelos de semántica léxica distribuida.

5. Aplicaciones

El objetivo de esta unidad es tratar, en todos sus aspectos, algún tipo de problema de interés práctico, adquiriendo conocimientos del marco teórico y experimentando en un taller. Se seleccionará un problema de interés práctico actual, y se propondrá un análisis teórico y trabajos prácticos. El problema seleccionado puede variar en los distintos dictados del curso.

Bibliografía:
(título del libro-nombre del autor-editorial-ISBN-fecha de edición)

Speech and Language Processing (Third ed.). Jurafsky, D. and J. H. Martin (2021).
https://web.stanford.edu/~jurafsky/slp3/

Introduction to Natural Language Processing. Eisenstein, J. , MIT Press. ISBN: 9780262042840 (2018).

Para cada unidad temática se indicará bibliografía específica adicional.

Datos del curso

Fecha de inicio y finalización: Inicio: 5/8/2022, fin: 15/11/2022

Horario y Salón: A determinar

Arancel: No corresponde
[Si la modalidad no corresponde indique “no corresponde”. Si el curso contempla otorgar becas, indíquelo]

Arancel para estudiantes inscriptos en la modalidad posgrado:
Arancel para estudiantes inscriptos en la modalidad educación permanente: