
DENOMINACIÓN DE LA ACTIVIDAD: **11090– Recuperación de Información**

TIPO DE ACTIVIDAD ACADÉMICA: **Asignatura**

CARRERA: **Licenciatura en Sistemas de Información**

PLAN DE ESTUDIOS: **17.13**

DOCENTE RESPONSABLE: **Dr. Gabriel H. Tolosa, Profesor Asociado**

EQUIPO DOCENTE:

Lic. Pablo J. Lavallén, Ayudante de Primera
Lic. Esteban Ríssola, Ayudante de Primera
Lic. Francisco Tonin Monzón, Ayudante de Primera
A.S. Agustín Marrone, Ayudante de Segunda
A.S. Agustín Gonzalez, Ayudante de Segunda

ACTIVIDADES CORRELATIVAS PRECEDENTES:

PARA CURSAR: **11078 (Bases de Datos II)**
11086 (Programación en Ambiente Web)

PARA APROBAR: **11078 (Bases de Datos II)**
11086 (Programación en Ambiente Web)

CARGA HORARIA TOTAL
HORAS SEMANALES: **6**
HORAS TOTALES: **96**

DISTRIBUCIÓN INTERNA DE LA CARGA HORARIA:
CLASES TEÓRICAS: **50%**
CLASES PRÁCTICAS: **50%**

PERÍODO DE VIGENCIA DEL PRESENTE PROGRAMA: **2024-2025**

CONTENIDOS MÍNIMOS O DESCRIPTORES (Según RES.HCS. Nº478/12)

Concepto de Base de Datos textual. Procesamiento de datos no estructurados. Modelos clásicos de recuperación: booleano, vectorial, probabilístico. Conceptos sobre similitud y matching. Medidas de similitud. Modelos de Lenguaje. Análisis de textos y representación de documentos. Estructuras de datos para Recuperación de Información. Evaluación. Métricas. Recuperación de Información en la Web. Arquitectura de los motores de búsqueda. Recolección (crawling), indexación y recuperación a gran escala. Modelos de la Web. Algoritmos de ranking basados en el análisis de enlaces. Aplicaciones.

FUNDAMENTACIÓN, OBJETIVOS, COMPETENCIAS

La recuperación de información trata de la organización, el almacenamiento y búsqueda eficiente sobre datos no estructurados (como documentos de textos) o semi-estructurados (como páginas HTML). Es la disciplina que estudia las bases de datos textuales o documentales. En la actualidad, la cantidad de información no estructurada que se genera y distribuye (especialmente en redes globales como Internet) supera ampliamente las posibilidades de los usuarios para su procesamiento y uso eficiente. Por ello, se requieren de modelos, algoritmos y técnicas que permitan su gestión eficaz y eficiente. Entre las aplicaciones típicas se incluyen las bibliotecas digitales y los motores de búsquedas web. Estos últimos imponen múltiples desafíos ya que tratan con grandes volúmenes de información, millones de usuarios y la heterogeneidad propia del ambiente web.

Esta asignatura brinda los fundamentos de la recuperación de información junto a los modelos clásicos que permiten comprender cómo se tratan documentos y consultas a los efectos de determinar cómo satisfacer la necesidad de información de un usuario. Se estudian las características estadísticas del texto escrito, los modelos de representación y las aplicaciones clásicas, extendiendo los temas al ámbito de la web y los motores de búsqueda.

OBJETIVOS

Se espera que al completar la asignatura los estudiantes:

- Comprendan los alcances de la disciplina, junto con criterios que les permitan determinar sus ámbitos de aplicación y entiendan la problemática de la construcción de sistemas de información basado en RI.
- Cuenten con los fundamentos teóricos sobre los modelos clásicos de recuperación de información y las estructuras de datos necesarias de almacenamiento y recuperación de datos masivos no estructurados o débilmente estructurados.
- Adquieran criterios de evaluación basados tanto en los sistemas como en los usuarios de los mismos.
- Comprendan la estructura del espacio Web y sean capaces de plantear aplicaciones de recuperación de información basadas en éste.

- Aumenten sus capacidades para la implementación de módulos de software, en particular a partir de implementar técnicas de recuperación de información.

Complementariamente, se propone que también incrementen sus habilidades para:

- Redactar informes de desarrollo, reportes técnicos o trabajos de investigación siguiendo objetivos y metodología concreta.
- Comunicar sus conocimientos, resultados de investigación a pes y/o superiores en presentaciones públicas.

CONTENIDOS

Unidad 1 – Introducción a la Recuperación de Información
El problema de la recuperación de información. Diferencias con el concepto de recuperación de datos. Conceptos sobre bases de datos textuales. Arquitectura de un Sistema de Recuperación de Información. Necesidades de Información y expresiones de consultas (queries). Procesamiento de datos no estructurados. Introducción a los modelos de recuperación a partir de ejemplos.

Unidad 2 – Modelos de Recuperación de Información
Taxonomía de los modelos clásicos: booleano, vectorial y probabilístico. Conceptos sobre similitud y matching. Medidas de similitud. Modelos extendidos. Introducción a los Modelos de Lenguaje para Recuperación de Información.

Unidad 3 – Análisis de Textos y Representación de Documentos
Representación de documentos a partir de su contenido. Análisis estadístico de las propiedades del texto. Leyes de Zipf y Heaps y su aplicación. Ponderación de términos a partir de su frecuencia. Indexación manual y automática. Extracción de términos a partir de sus pesos.

Unidad 4 – Estructuras de Datos
Estructuras de datos y algoritmos para soportar los modelos de recuperación. Archivos invertidos y listas de posteo. Archivos invertidos posicionales. Soporte para frases y operadores de proximidad. Recuperación por evaluación completa. Compresión del índice.

Unidad 5 – Evaluación de la Recuperación
Conceptos sobre evaluación de la recuperación y relevancia. Definiciones de las métricas de Exhaustividad (Recall) y Precisión (Precision). Diagramas de Exhaustividad/Precisión. F-Measure y medidas complementarias. Colecciones de prueba y evaluación de sistemas. Las conferencias TREC y su importancia en la metodología.

Unidad 6 – Recuperación desde el Índice
Algoritmos básicos de recuperación desde el índice: DAAT y TAAT. Algoritmos de poda dinámica. Recuperación eficiente sobre índices por bloques. Evaluación de performance.

Unidad 7 – Recuperación de Información en la Web
Características del espacio Web y los lenguajes de marcado. Arquitectura de los motores de búsqueda. Recolección (crawling), indexación y recuperación a gran escala. Modelos de la Web. Algoritmos de ranking basados en el análisis de enlaces. Arquitectura de un Motor de Búsqueda de escala Web.

Unidad 8 – Introducción al Procesamiento del Lenguaje Natural para RI
Representación del lenguaje natural. Problemas asociados. Representaciones multidimensionales para términos y documentos. Operaciones básicas de recuperación. Estado del arte.

METODOLOGÍA

El desarrollo del curso es de carácter teórico/práctico, con aplicación de los conceptos a las actividades propiamente de laboratorio. En las clases teóricas se plantean los conceptos, modelos, ejemplos y aplicaciones del área de recuperación de información y demás temas propuestos en este programa.

En las clases prácticas se realizan implementaciones de los modelos desarrollados como así también de experimentos de recuperación y evaluación. Se trabaja tanto con software propio como con toolkits existentes y ampliamente utilizados para la enseñanza de la disciplina.

Complementariamente, los estudiantes deben preparar una exposición sobre la base de la lectura e investigación de un tema propuesto por el equipo docente. Esta actividad introduce en la lectura de literatura netamente de investigación y se propone como motivadora para la discusión en clase con todo el grupo. La última actividad de evaluación consiste en un trabajo final de curso sobre algún tema del programa. Éste puede ser de carácter teórico/práctico o de un desarrollo concreto.

ACTIVIDADES PRÁCTICAS

En las actividades prácticas se considera tanto la resolución de problemas como la ejercitación de laboratorio. Con las mismas se pretende reforzar los conceptos planteados en clase ya que permiten la exploración y aplicación concreta de la mayoría de los temas.

En las tareas de laboratorio se deben realizar pequeñas aplicaciones orientadas a diferentes problemas del área como análisis de textos, indexación, recuperación y presentación de resultados. Complementariamente, se utilizan herramientas libres existentes a modo demostrativo o cuando el tema lo requiere.

Las aplicaciones se pueden programar en lenguaje C, C++, Python, Perl o Java y los estudiantes deben demostrar sus habilidades en la programación como así también en el análisis de la situación propuesta previo a la construcción de la solución. En ambos casos, cuentan con el soporte del equipo docente.

Para el trabajo final, los estudiantes deben presentar su propio proyecto, el cual se discute con los docentes. En éste deben realizar una investigación relacionada a

algunos de los tópicos desarrollados en la asignatura o bien una propuesta con estudio experimental de algún enfoque alternativo a las técnicas existentes. En cualquiera de los casos, se debe elaborar un documento con formato de artículo de investigación (paper) donde se expongan los objetivos, antecedentes, la propuesta, la metodología utilizada y los resultados obtenidos.

REQUISITOS DE APROBACION Y CRITERIOS DE CALIFICACIÓN:

La evaluación consta de 1 (un) examen parcial y un trabajo final integrador (descripto en el apartado anterior) obligatorio. El examen parcial se aprueba con nota 4 cuatro) o superior mientras que el integrador con 7 (siete) o superior.

CONDICIONES PARA PROMOVER (SIN EL REQUISITO DE EXAMEN FINAL), DE ACUERDO AL ART.23 DEL RÉGIMEN GENERAL DE ESTUDIOS RESHCS-LUJ:0000996-15

a) Tener aprobadas las actividades correlativas al finalizar el turno de examen extraordinario de ese cuatrimestre.
b) Cumplir con un mínimo del 80% de asistencia para todas las las actividades.
c) Aprobar todos los *trabajos prácticos* previstos en este programa, pudiendo recuperarse hasta un 25% del total por ausencias o aplazos.
d) Aprobar el 100% de las evaluaciones previstas con un promedio no inferior a seis (6) puntos sin recuperar ninguna.
d) Aprobar una evaluación integradora de la asignatura con calificación no inferior a siete (7) puntos.

CONDICIONES PARA APROBAR COMO REGULAR (CON REQUISITO DE EXAMEN FINAL) DE ACUERDO AL ART.24 DEL RÉGIMEN GENERAL DE ESTUDIOS RESHCS-LUJ:0000996-15

a) Estar en condición de regular en las actividades correlativas al momento de su inscripción al cursado de la asignatura.
b) Cumplir con un mínimo del 70% de asistencia para todas las actividades.
c) Aprobar todos los trabajos prácticos previstos en este programa, pudiendo recuperarse hasta un 40% del total por ausencias o aplazos.
d) Aprobar el 100% de las evaluaciones previstas con un promedio no inferior a cuatro (4) puntos, pudiendo recuperar el 50% de las mismas. Cada evaluación solo podrá recuperarse en una oportunidad.

Antes de presentarse a un examen, el estudiante debe tener aprobado el trabajo práctico integrador.

EXAMENES PARA ESTUDIANTES EN CONDICIÓN DE LIBRES

1. Para aquellos estudiantes que, habiéndose inscripto oportunamente en la presente actividad hayan quedado en condición de libres por aplicación de los

artículos 22, 25, 27, 29 o 32 del Régimen General de Estudios, podrán rendir en tal condición la presente actividad.

2. Para aquellos estudiantes que no cursaron la asignatura y se presenten en condición de estudiantes libres en la Carrera, por aplicación de los artículos 10 o 19 del Régimen General de Estudios, podrán rendir en tal condición la presente actividad. Además, quince días antes de la fecha de sustanciación de mesa, el alumno deberá entregar la resolución de todas las actividades prácticas vigentes en la última cursada.

BIBLIOGRAFÍA

SUGERIDA

- S. Büttcher, C.L.A. Clarke, G.V. Cormack. Information Retrieval: Implementing and Evaluating Search Engines. The MIT Press, 2016
- B. Croft; D. Meltzer, T. Strohman. Search Engines: Information Retrieval in Practice. Pearson Education. 2009.
- R. Baeza-Yates, B. Ribeiro-Neto. Modern Information Retrieval: The concepts and technology behind search. 2nd Ed. Addison-Wesley, 2011.
- C. Manning, P. Raghavan, H. Schutze. Introduction to Information Retrieval. Cambridge University Press. 2008.
- D. Jurafsky, Martin James. Speech and Language Processing, 2nd Ed. Prentice Hall, 2008 (draft 3rd Ed. online, 2023)

- Material provisto por el equipo docente:
    - G.H. Tolosa y F.R.A. Bordignon. Introducción a la Recuperación de Información. Conceptos, modelos y algoritmos básicos. Laboratorio de Redes de Datos. Universidad Nacional de Luján.

DE CONSULTA

- I.H. Witten, A. Moffat, T.C. Bell. Edit. Managing Gigabytes: Compressing and Indexing Documents and Images. 2ª Edition. Morgan Kaufmann, 1999.
- W. B. Frakes, R. Baeza-Yates. Edit. Information Retrieval. Data Structures & Algorithms. Prentice-Hall, 1992.
- S. Chakrabarti. Mining the Web. Discovering Knowledge from Hypertext Data. Morgan-Kaufmann Publishers. 2003.
- S. Bird, E. Klein, E. Loper. Natural Language Processing with Python. O'Reilly Media, Inc., 2009.
- L. Tunstall, L. von Werra, T. Wolf. Natural Language Processing with Transformers. O'Reilly Media, Inc., 2022.

RECURSOS ADICIONALES

El equipo docente mantiene un sitio web de la asignatura (http://www.labredes.unlu.edu.ar/) en el cual se publica el cronograma, guías de clase, material regular y las novedades. Todos los años se actualiza una lista de artículos de

investigación, tutoriales y white papers que se utilizan durante la cursada. Además, se atienden durante todo el año consultas por correo electrónico y/o sesiones de chat.

CONFERENCIAS/JOURNALS RELACIONADOS A LA DISCIPLINA

- SIGIR - Special Interest Group in Information Retrieval, http://www.sigir.org/
- Conference on Information and Knowledge Management, http://www.cikmconference.org/
- Web Search and Data Mining - https://www.wsdm-conference.org/
- WWW – International World Wide Web Conference
- ECIR - European Conference on Information Retrieval
- TREC - Text REtrieval Conference , http://trec.nist.gov/
- Information Processing & Management, https://www.journals.elsevier.com/information-processing-and-management
- International Journal on Digital Libraries, http://www.dljournal.org/

DISPOSICIÓN DE APROBACIÓN: CD