Universidad Nacional de Luján
Departamento de
Ciencias Básicas

DISPOSICION CONSEJO DIRECTIVO DEPARTAMENTAL DE CIENCIAS BÁSICAS DISPCD-CB : 443 / 2025

LUJAN, 13 DE NOVIEMBRE DE 2025

VISTO: El programa de la asignatura Bases de Datos Textuales (14033) para la carrera Licenciatura en Sistemas de Información presentado por la División Computación; y

CONSIDERANDO:

Que la Comisión Plan de Estudio ha tomado intervención en el trámite.

Que se ha tratado y aprobado por el Consejo Directivo Departamental de Ciencias Básicas en su Sesión Ordinaria del día 6 de noviembre de 2025.

Por ello,

EL CONSEJO DIRECTIVO DEPARTAMENTAL

DE CIENCIAS BÁSICAS

D I S P O N E :

ARTÍCULO 1°.- Aprobar el programa de la asignatura Bases de Datos Textuales (14033) para la carrera Licenciatura en Sistemas de Información presentado por la División Computación que como anexo I forma parte de la presente Disposición.-

ARTICULO 2°.- Establecer que el mismo tendrá vigencia para los años 2025-2026.-

ARTÍCULO 3°.- Regístrese, comuníquese, cumplido, archívese.-

Lic. Ariel H. REAL - Secretario Académico - Departamento de Ciencias Básicas

Lic. Emma L. FERRERO - Directora Decana - Departamento de Ciencias Básicas

UNIVERSIDAD NACIONAL DE LUJÁN
DEPARTAMENTO DE CIENCIAS BÁSICAS

PROGRAMA OFICIAL

DENOMINACIÓN DE LA ACTIVIDAD: **14033 – Bases de Datos Textuales**

TIPO DE ACTIVIDAD ACADÉMICA: **Asignatura**

CARRERA: **Licenciatura en Sistemas de Información**

PLAN DE ESTUDIOS: **17.14 (Resolución H.C.S. N° 260/24, Resolución H.C.S. N° 836/24 y Disposición S.A. N° 395/24).**

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 Gonzalez, Ayudante de Primera**

ACTIVIDADES CORRELATIVAS PRECEDENTES:

PARA CURSAR: **11288 (Gestión de Datos Masivos)**
PARA APROBAR: **11288 (Gestión de Datos Masivos)**

CARGA HORARIA TOTAL
HORAS SEMANALES: **4**
HORAS TOTALES: **64**

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

PERÍODO DE VIGENCIA DEL PRESENTE PROGRAMA: **2025-2026**

CONTENIDOS MÍNIMOS O DESCRIPTORES

Concepto de Base de Datos textual. Procesamiento de datos no estructurados. Análisis de textos y representación de textos. Modelos de recuperación de información. Evaluación. Estructuras de datos asociadas. Indexación y recuperación de gran escala. Introducción al Procesamiento del Lenguaje Natural para recuperación de información.

FUNDAMENTACIÓN, OBJETIVOS, COMPETENCIAS

Las bases de datos textuales se basan en modelos que permiten organizar, almacenar y soportar búsquedas eficientes sobre sobre datos no estructurados (como documentos de textos) o semi-estructurados (como páginas HTML). En las Ciencia de la Computación pertenecen al área de la Recuperación de Información, la cual hoy en día se interseca en varios puntos con el Procesamiento del Lenguaje Natural moderno.
Desde la aparición de la web como plataforma que habilita compartir cualquier tipo de recurso digital la cantidad de información (no estructurada) que se genera y distribuye 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. La aplicación típica es un motor de búsqueda, los cuales tratan con grandes volúmenes de información, millones de usuarios y la heterogeneidad propia del ambiente web. En general, estas aplicaciones se basan en representaciones basadas en palabras, frecuencias y enlaces. Sin embargo, en los últimos años, aparecieron nuevas técnicas que permiten plantear representación de documentos y consultas en espacios que intentan modelar la semántica del lenguaje, habilitando nuevas maneras de organizar, almacenar y soportar búsquedas, aportando a la eficacia y desafiando la eficiencia de los sistemas.

Esta asignatura brinda los fundamentos de la recuperación de información para construir bases de datos textuales, incluyendo los modelos que permiten comprender cómo se tratan documentos y consultas. El objetivo final es satisfacer la necesidad de información de un usuario. Este problema por un lado se complejiza con la posibilidad de la multimodalidad en los datos y por el otro se beneficia de los nuevos avances. En ambos casos, el estudio de las formas de representación, sus alcances y limitaciones abren el panorama de soluciones posibles a los diferentes problemas. Además, es un espacio donde convergen y se integran conocimientos de otras asignaturas (estructuras de datos, programación, sistemas operativos, redes, gestión de datos masivos) en el diseño de soluciones.

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 bases de datos textuales (y su integración con un sistema de información)
- Cuenten con los fundamentos teóricos sobre los modelos de representación de información textual y métodos de recuperación, las estructuras de datos necesarias y las implicancias en eficacia y eficiencia en cada caso.
- Adquieran criterios de evaluación (y dominen las métricas más relevantes) 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.
- Adquieran manejo conceptual de las nuevas técnicas de procesamiento del lenguaje natural y su aplicación en una base de datos textual.
- 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 pares y/o docentes en presentaciones públicas.

CONTENIDOS

Unidad 1 – Bases de datos textuales
El problema de la recuperación de información. Conceptos sobre bases de datos textuales. Arquitectura de un Sistema de Recuperación de Información. Necesidades de información y expresiones de consultas. Procesamiento de datos no estructurados. Análisis de texto y representación de documentos y consultas. Modelos sobre las distribuciones de ítems en los textos.

Unidad 2 – Modelos de Recuperación de Información

Taxonomía de los modelos de recuperación. Relación con las representaciones de documentos. Conceptos sobre similitud y matching. Ponderación de términos. Medidas de similitud. Modelos clásicos y extendidos. Modelos de Lenguaje. Conceptos y técnicas asociadas aplicadas a la Recuperación de Información

Unidad 3 – Evaluación de la Recuperación
Conceptos sobre evaluación de la recuperación (efectividad y eficiencia). Métricas clásicas basadas en conjuntos. Métricas para rankings y medidas complementarias. Colecciones de prueba y evaluación de sistemas. Las conferencias TREC y su importancia en la metodología.

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. Estructuras de datos para vectores densos.

Unidad 5 – 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 la performance.

Unidad 6 – 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 7 – Introducción al Procesamiento del Lenguaje Natural para RI
Representación del lenguaje natural. Problemas asociados. Representaciones multidimensionales para términos y documentos (vectores densos). Algoritmos de recuperación exactos y aproximados. Estado del arte.

METODOLOGÍA

El desarrollo del curso es de carácter teórico/práctico, con aplicación de los conceptos a las actividades prácticas. En las clases teóricas se plantean los conceptos, modelos, ejemplos y aplicaciones del área con ejemplos sobre el uso de bases de datos textuales en sistemas de información. Se proponen problemas y se discuten ideas sobre potenciales soluciones hasta converger en los conceptos que corresponden a cada tema. Por cada uno el equipo docente prepara una guía de clases que ayuda y sugiere a los estudiantes cómo abordarlo, qué recursos complementarios se proponen (diapositivas, videos, preguntas, tutoriales) junto con la bibliografía (no se sigue un solo texto). En cada clase teórica, se destina un espacio de tiempo para resolver dudas derivadas de la clase previa y/o del material de estudio.

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 y práctica de la disciplina.

Complementariamente, los estudiantes deben preparar una exposición sobre la base de la lectura de un artículo de 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.

Finalmente, el proyecto final para aprobar el curso consiste en un trabajo de investigación que incluye lectura, elaboración de una propuesta de mejora o reproducción, ejecución de experimentos y análisis de resultados. Su desarrollo implica investigación aplicada, aplicación de criterios, diseño, prototipado y pruebas empíricas. Para este trabajo los estudiantes elaboran una propuesta que es evaluada primero por el equipo docente en cuanto a la pertinencia con los temas de la asignatura, los desafíos a resolver, la aplicación de la solución a desarrollar o pregunta a responder y la metodología para el abordaje. La presentación de este trabajo se debe abordar como un artículo de investigación (*paper*[1]). Debe considerar las mismas preguntas que se formularon en la propuesta pero con un nivel de profundidad

mayor. El informe debe responder a la estructura y estilo clásico de un artículo de la disciplina (Por ejemplo: Introducción, Trabajos Relacionados, Metodología, Experimentos y Resultados, Conclusiones, Bibliografía). Durante este proceso, el equipo docente monitorea los avances y sugiere acciones (como si fueran revisores del artículo).

Respecto de la modalidad de dictado, la asignatura se propone presencial (60%) y mediada sincrónica (40%). Semana tras semana se intercalan en la versión mediada teoría y práctica. Por tratarse de una asignatura de años superiores, esta modalidad facilita a los estudiantes un mejor aprovechamiento de su tiempo. La alternancia de la opción mediada se propone para mantener un registro cercano tanto en teoría como en práctica. En general, cuando la práctica es mediada, los estudiantes deben hacer presentaciones parciales de la aproximación utilizada para resolver los ejercicios y plantear inconvenientes.

ACTIVIDADES PRÁCTICAS

En las actividades prácticas se considera tanto la resolución de problemas y el desarrollo de prototipos o pruebas de concepto con los que se pueda ejemplificar una solución o explorar un concepto. Aquí 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 (por ejemplo, Terrier o PISA). 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 (discutido en el apartado anterior), los estudiantes deben programar los módulos para ejecutar los experimentos propuestos y/o analizar los resultados.

REQUISITOS DE APROBACION Y CRITERIOS DE CALIFICACIÓN:

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

DE ACUERDO AL ART.23 DEL RÉGIMEN GENERAL DE ESTUDIOS RES.HCS 261-21 y su ANEXO PARA CARRERAS CON MODALIDAD PEDAGÓGICA A DISTANCIA

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 RES.HCS 261-21 y su ANEXO PARA CARRERAS CON MODALIDAD PEDAGÓGICA A DISTANCIA.

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 sólo podrá recuperarse en una oportunidad.

Antes de presentarse a un examen, el estudiante debe tener **aprobado** el proyecto final de curso.

EXÁMENES PARA ESTUDIANTES EN CONDICIÓN DE LIBRES

1. Para aquellos estudiantes que, habiéndose inscripto oportunamente en la presente actividad hayan quedado en condición de “libre”, podrán rendir el final en tal condición cumpliendo con los mismos requisitos de la cursada (tener entregados y aprobados todos los trabajos prácticos, incluyendo el proyecto final de curso). El final incluirá los contenidos teóricos prácticos del programa vigente.

2. Para aquellos estudiantes que no cursaron la asignatura y se presenten en condición de “libres” en la Carrera, sólo podrán rendir en tal condición la presente actividad después de haber cumplido con los mismos requisitos de la cursada respecto de los trabajos prácticos y el proyecto final de curso. El final incluirá los contenidos teóricos prácticos del programa vigente.

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.
- D. Jurafsky, Martin James. Speech and Language Processing, 2nd Ed. Prentice Hall, 2008 (draft 3rd Ed. online, 2023)

COMPLEMENTARIA

- L. Tunstall, L. von Werra, T. Wolf. Natural Language Processing with Transformers. O'Reilly Media, Inc., 2022.
- C. Manning, P. Raghavan, H. Schutze. Introduction to Information Retrieval. Cambridge University Press. 2008.
- 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. UNLu.

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

[1]Se sugiere que el artículo se escriba en Latex, usando una plantilla para Ciencias de la Computación. Puede usar OverLeaf para editar su documento: http://www.overleaf.com/

Hoja de firmas