Universidad Nacional de Luján
República Argentina

Ruta 5 y Av. Constitución
C.C. 221 - 6700 – LUJÁN (Bs. As.)

N° DISPOSICIÓN: .....

DEPARTAMENTO DE: **Ciencias Básicas**

CARRERA/S: **Licenciatura en Sistemas de Información**

(RES.HCS. N°238/04)

PROGRAMA DE LA ASIGNATURA: **Taller Libre I (11421)**

RESPONSABLE:	**HORAS DE CLASE**
Mg. Gabriel H. Tolosa, Profesor Adjunto	SEMANALES: **8**
EQUIPO DOCENTE:	TEÓRICAS: **4**
A.S. Pablo J. Lavallén, Ayudante de Primera	PRÁCTICAS: **4**
	HS.TOTALES:**128**

ASIGNATURAS CORRELATIVAS

CURSADAS	**APROBADAS**
11413, 10040 (Regular para cursar)	11413, 10040 (Para aprobar)

VIGENCIA:2012

CONTENIDOS MÍNIMOS: Según RES.HCS. N°238/04

Esta asignatura tiene tema libre. Su objetivo es el de actualizar a los alumnos en temas de reciente desarrollo. Por esta razón, la Coordinación de la carrera definirá para cada año los temas a tratar.

FUNDAMENTACIÓN

La recuperación de información trata del almacenamiento y búsqueda eficiente sobre datos no estructurados (como documentos de textos) o 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 del texto escrito, los modelos de representación y las aplicaciones clásicas como recuperación, filtrado, clasificación y clustering, extendiendo éstos al ámbito de la web.

OBJETIVOS GENERALES:

Se espera que al completar el taller los alumnos:

- 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 pares 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 documentos y colecciones. Arquitectura de un Sistema de Recuperación de Información. Necesidades de Información y expresiones de consultas (*queries*). Introducción a los modelos de recuperación a partir de ejemplos.

Unidad 2 – Modelos Clásicos de Recuperación de Información

Taxonomía de los modelos clásicos. El modelo booleano. Conceptos sobre similitud y *matching*. El modelo Booleano extendido y el modelo vectorial. Medidas de similitud. 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. Ley de Zipf, ley de Heaps y su aplicación. Ponderación de términos a partir de su frecuencia. Indexación manual y automática. Extracción de términos a partir de sus pesos. Construcción automática de un tesauro.

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. Archivos de firmas.

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 – Tratamiento de Consultas y Documentos

Conceptos sobre retroalimentación por relevancia: Pseudo y directa. Retroalimentación en el modelo vectorial. Expansión de consultas con tesauros. Clasificación de documentos basado en el teorema de Bayes. Clustering. Métodos jerárquicos y no jerárquicos. Aplicaciones.

Unidad 7 – Introducción a la Recuperación de Información Distribuida

La problemática de la distribución de contenidos. Representación de repositorios textuales. Algoritmos para selección de recursos y fusión de resultados. Introducción a los sistemas peer-to-peer para recuperación de información.

Unidad 8 – Recuperación de Información en la Web

Características del espacio Web y los lenguajes de marcado. Arquitectura de los motores de búsqueda. Directorios y metabuscadores. Recolección (*crawling*), indexación y recuperación a gran escala. Modelos de la Web. Algoritmos de ranking basados en el análisis de enlaces. Aplicaciones y tendencias.

METODOLOGÍA:

El desarrollo del curso es de carácter teórico – práctico, con un alta preponderancia de actividades propiamente de laboratorio (dada la modalidad taller). En las clases teóricas se plantearán los conceptos, modelos, ejemplos y aplicaciones del área de recuperación de información y demás temas propuestos en este programa.

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

Complementariamente, los alumnos deberán 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 espera que sea motivadora para la discusión en clase con todo el grupo. La última actividad consiste en la realización de 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 consideran tanto la resolución de problemas como la ejercitación de laboratorio. Con las mismas se pretende reforzar los conceptos planteados en clase ya que permitirán la exploración en profundidad de los temas.

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

Las aplicaciones podrán estar escritas en lenguaje C, Perl, Python o Java y los alumnos deberán 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, contarán con el soporte del equipo docente.

Para el trabajo final, los alumnos presentarán su propio proyecto, el cual discutirán con los docentes. En éste deben realizar una aplicación 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 elaborará un informe (en un formato que se especificará de acuerdo a la naturaleza del proyecto) donde se exponga los objetivos, antecedentes, la propuesta, la metodología utilizada y los resultados obtenidos.

EVALUACIÓN

La evaluación consta de 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.

Al finalizar, existe una instancia de recuperatorio para quien no haya aprobado el examen parcial. Las condiciones luego de cursar la asignatura son las siguientes:

PROMOVIDO:

- Aprobar todos los trabajos prácticos y/o actividades académicas especiales previstas.
- Aprobar el cien por ciento (100%) de las evaluaciones previstas (2 parciales) con un promedio final no inferior a seis (6) puntos, sin haber recuperado ninguna.
- Aprobar la evaluación integradora (TP Final Integrador) de la asignatura con calificación no inferior a siete (7) puntos.

REGULAR:

- Aprobar todos los trabajos prácticos y/o actividades académicas especiales previstas.

- Aprobar el cien por ciento (100%) de las evaluaciones previstas (2 parciales) con una calificación no inferior a cuatro (4) puntos. Se podrá recuperar una de las instancias.
- Aprobar la evaluación integradora (TP Final Integrador) de la asignatura con calificación no inferior a cuatro (4) puntos.

LIBRE:
- Es aquel que habiendo participado en al menos una (1) de las evaluaciones establecidas como obligatorias en el programa oficial de la asignatura, o de las instancias de recuperación de la misma, no hubiera alcanzado el rendimiento exigido para ser considerado regular.
- Para el examen libre: 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.

AUSENTE:
- Es aquél que habiéndose inscripto en la asignatura, no ha cumplido con ninguna de las actividades evaluables establecidas por el programa oficial de la misma.

Las condiciones de asistencia se rigen según el Capítulo II y V del Régimen General de Estudios de la Universidad.

BIBLIOGRAFÍA

SUGERIDA

- Introduction to Information Retrieval. C. Manning, P. Raghavan, H. Schutze. Cambridge University Press. 2008.
- Search Engines: Information Retrieval in Practice. B. Croft; D. Meltzer, T. Strohman. Pearson Education. 2009.
- Modern Information Retrieval. R. Baeza-Yates, B. Ribeiro-Neto. 2nd Ed. Addison-Wesley, 2011.

- Information Retrieval. Algorithms and Heuristics. D. A. Grossman, O. Frieder. Kluwer, 1998.
- Material provisto por el equipo docente. Libro: "*Introducción a la Recuperación de Información. Conceptos, modelos y algoritmos básicos*". Gabriel H. Tolosa y Fernando R.A. Bordignon. Laboratorio de Redes de Datos. Universidad Nacional de Luján.

DE CONSULTA

- Van Rijsbergen, C. J. Information Retrieval. Butterworth. 1979. Recurso disponible en linea: http://www.dcs.gla.ac.uk/Keith/Preface.html
- Managing Gigabytes: Compressing and Indexing Documents and Images. 2ª Edition. I.H. Witten, A. Moffat, T.C. Bell. Edit. Morgan Kaufmann, 1999.
- Gestón Digital de la Información. De Bits a Bibliotecas Digitales y la Web. Peña, Rosalía; Baeza-Yates, Ricardo y Rodriguez, José. Alfaomega-Rama. 2002.
- Information Retrieval Interaction. Peter Ingwersen. London: Taylor Graham, 1992. Recurso disponible en linea: http://www.db.dk/pi/iri/
- Information Retrieval. Data Structures & Algorithms. W. B. Frakes, R. Baeza-Yates. Edit. Prentice-Hall, 1992.
- Advances in Information Retrieval. 27th European Conference on IR Research, ECIR 2005, Santiago de Compostela, Spain, March 21-23, 2005 (Proceedings).
- Mining the Web. Discovering Knowledge from Hypertext Data. Soumen Chakrabarti. Morgan-Kaufmann Publishers. 2003.
- Readings in Information Retrieval, First Edition. Karen Spark Jones & Peter Willett (editors). Morgan Kaufmann Series in Multimedia Information and Systems. 1997.

OTROS RECURSOS

Artículos de investigación, *surveys*, tutorialess y *white papers* provistos por el equipo docente.

CONFERENCIAS RELACIONADAS

- SIGIR (Special Interest Group in Information Retrieval). http://www.sigir.org/
- ECIR (European Conference on Information Retrieval). http://ecir2006.soi.city.ac.uk/
- TREC (Text REtrieval Conference). http://trec.nist.gov/
- International Journal on Digital Libraries. http://www.dljournal.org/

RECURSOS ADICIONALES

El equipo docente mantiene un sitio web de la asignatura (http://www.tyr.unlu.edu.ar/TallerIR/) con un blog en el cual se publica el material regular y las novedades. Además, se atienden durante todo el año consultas por correo electrónico y/o sesiones de chat.