Universidad Nacional de San Luis
Rectorado

ES COPIA
OSCAR GUILLERMO SEGURA
Director de Despacho
UNSL

SAN LUIS, 21 NOV. 2016

VISTO:

El Expediente EXP-USL: 13052/2016 mediante el cual se solicita la protocolización del Curso de Posgrado: **INDEXACIÓN Y COMPRESIÓN DE TEXTO**; y

CONSIDERANDO:

Que el Curso de Posgrado se dictará en el ámbito de la Facultad de Ciencias Físico Matemáticas y Naturales del 21 de noviembre al 16 de diciembre de 2016, con un crédito horario de 40 horas presenciales y bajo la coordinación del Lic. Dario **RUANO**.

Que la Comisión Asesora de Posgrado de la Facultad de Ciencias Físico Matemáticas y Naturales recomienda aprobar el curso de referencia.

Que el Consejo de Posgrado de la Universidad Nacional de San Luis en su reunión del 8 de noviembre de 2016, analizó la propuesta y observa que el programa del curso, bibliografía, metodología de evaluación y docentes a cargo, constituyen una propuesta de formación de posgrado de calidad en su campo específico de estudio.

Que, por lo expuesto, el Consejo de Posgrado aprueba la propuesta como Curso de Posgrado, según lo establecido en Ordenanza CS N° 35/16.

Que corresponde su protocolización.

Dr. José Roberto Saad
Vicerrector - UNSL
A/C Rectorado RR N
1183/16

Por ello y en uso de sus atribuciones

EL RECTOR DE LA UNIVERSIDAD NACIONAL DE SAN LUIS

RESUELVE:

ARTÍCULO 1°.- Protocolizar el dictado del Curso de Posgrado: **INDEXACIÓN Y COMPRESIÓN DE TEXTO**, en el ámbito de la Facultad de Ciencias Físico Matemáticas y Naturales del 21 de noviembre al 16 de diciembre de 2016, con un crédito horario de 40 horas presenciales.

Dra. Alicia Marcela PRATTICA
A/C Secretaria de Posgrado
U.N.S.L.

ARTÍCULO 2°.- Protocolizar como docente responsable del curso a la Mg. Norma Edith **HERRERA** (DNI N° 18.469.411) de esta Casa de Estudios Superiores.

Cpde RESOLUCIÓN R N° 2036

Universidad Nacional de San Luis
Rectorado

ES COPIA
OSCAR GUILLERMO SEGURA
Director de Despacho
UNSL

ARTÍCULO 3º.- Aprobar el programa del Curso de referencia, de acuerdo al **ANEXO** de la presente disposición.-

ARTÍCULO 4º.- Comuníquese, insértese en el Libro de Resoluciones, publíquese en el Digesto Electrónico de la UNSL y archívese.-

RESOLUCIÓN R Nº 2036
mav

Dra. Alicia Marcela PRINTISTA
A/C Secretaria de Posgrado
U.N.S.L.

Dr. José Roberto Saad
Vicerrector - UNSL
A/C Rectorado RR N 1983/16

Universidad Nacional de San Luis
Rectorado

ES COPIA
OSCAR GUILLERMO SEGURA
Director de Despacho
UNSL

ANEXO

DENOMINACIÓN DEL CURSO: INDEXACIÓN Y COMPRESIÓN DE TEXTO

UNIDAD ACADÉMICA RESPONSABLE: Facultad de Ciencias Físico Matemáticas y Naturales

CATEGORIZACIÓN: Perfeccionamiento

RESPONSABLE: Mg. Norma Edith **HERRERA**

COORDINADOR: Dr. Dario **RUANO**

CRÉDITO HORARIO: 40 horas

MODALIDAD DE DICTADO: Presencial

FECHA DE DICTADO DEL CURSO: 21 de noviembre al 16 de diciembre de 2016

FECHA PREVISTA PARA ELEVAR LA NÓMINA DE ALUMNOS APROBADOS: Abril de 2017

DESTINATARIOS: Egresados con título de grado universitario de 4 años o más en Ciencias Informáticas y en disciplinas afines a la temática del curso.

LUGAR DE DICTADO: Departamento de Informática – UNSL – San Luis.

CUPO: 15 personas.

Dr. José Roberto Saez
Vicerrector - UNSL
A/C Rectorado RRN
1983/16

FUNDAMENTACIÓN: La información disponible en formato digital aumenta día a día su tamaño de manera exponencial y gran parte de esta información se representa en forma de texto. Por esta razón, los temas de investigación relacionados a la gestión de cantidades masivas de texto se han convertido en los últimos años en un foco de interés en Ciencias de la Computación.
Una base de datos de texto es un sistema que mantiene una colección grande de texto y que provee acceso rápido y seguro al mismo. Uno de los tópicos de interés en base de datos de texto son las técnicas de compresión que explotan las repeticiones existentes en los textos a representar con el fin de reducir el espacio a usar en el almacenamiento de los mismos. Si esta compresión es realizada con una técnica que además permita buscar directamente sobre el texto comprimido, no sólo se beneficia el espacio ocupado sino también el tiempo insumido durante una consulta a la base de datos.

Dra. Alicia Marcela PRINTISTA
A/C Secretaria de Posgrado
U.N.S.L.

OBJETIVOS:

- Introducir al alumno en las temáticas de indexación y compresión de texto, con el fin de que adquiera los conocimientos y las destrezas necesarias sobre la temática de estudio.
- Proveer al alumno de los criterios necesarios para evaluar el desempeño de las técnicas estudiadas en casos concretos de aplicación.

CONTENIDOS MÍNIMOS:

- Introducción a compresión e indexación en bases de datos de texto.
- Compresores semiestáticos.

Cpde RESOLUCIÓN R Nº 2036

ES COPIA
OSCAR GUILLERMO SEGURA
Director de Despacho
UNSL

- Compresores dinámicos.
- Indexación: árboles y arreglos de sufijos. Indexación +compresión: autoíndices.

PROGRAMA:

UNIDAD 1: INTRODUCCIÓN.

Bases de datos de texto. El problema de pattern matching. Indexación. Operaciones count y locate. Compresión y bases de datos de texto. Por qué y para qué comprimir. Clasificación de compresores. Medidas de eficiencia. El concepto de entropía.

UNIDAD 2: COMPRESORES SEMIESTÁTICOS.

Huffman Clásico. Huffman canónico. Huffman orientado a palabras. End Tagged Dense Code y y (*s*-*c*)-Dense Code. Byte Pair Encoding. Burrows – Wheeler Transform.

UNIDAD 3: COMPRESORES DINÁMICOS.

Introducción. Estadísticos: Huffman dinámico o adaptativo. Basados en diccionarios: LZ77, LZ78, LZW. End-Tagged Dense Code dinámico y (*s*-*c*)-Dense Code dinámico.

UNIDAD 4: INDEXACIÓN.

Árboles y arreglos de sufijos. Indexación + compresión: autoíndices. Arreglos de sufijos comprimidos. Arboles de sufijos comprimidos. Wavelet tree. Directly Addressable Codes (DACs).

Dr. José Roberto Saad
Vicerrector - UNSL
A/C Rectorado RR N
1983/16

SISTEMA DE EVALUACIÓN:

Para aprobar el curso el alumno deberá realizar y entregar los prácticos que se le soliciten y además deberá aprobar la evaluación final que será de carácter individual.

BIBLIOGRAFÍA:

- R. Baeza Yates, B. Ribeiro-Neto. *Modern Information Retrieval: The Concepts and Technology behind Search* (3nd edition). Addison Wesley. ACM Press Books. ISBN-10: 0321416910. Dic. 2010.
- N. Brisaboa, A. Fariña, G. Navarro, J. Paramá. *New adaptive compressors for natural language text. Software, Practice and Experience,* 38(13), pp. 1429-1450. Sussex, United Kingdom, 2008.
- T. Bell, J. Cleary and I. Witten. *Text Compression.* Prentice Hall, 1990.
- A. Fariña. *New compression codes for text databases*. Ph.D. thesis, Database Laboratory, University of A Coruña, España, 2005.
- P. Ferragina, G. Manzini, V. Makinen, and G. Navarro. *Compressed representations of sequences and full-text indexes*. ACM Transactions on Algorithms (TALG), 3(2): article 20, 2007.
- Ling Liu and M. Tamer Özsu (editores). *Encyclopedia of Database Systems*. Springer Verlag, 2009.
- Gonzalo Navarro and Mathieu Raffinot, *Flexible Pattern Matching in Strings*. Cambridge University Press; 1 edition (July 30, 2007).
- David Salomon, *A Concise Introduction to Data Compression*, Springer; 2008 edition (January 14, 2008).

Dra. Alicia Marcela PRINTISTA
A/C Secretaria de Posgrado
U.N.S.L.

ES COPIA
OSCAR GUILLERMO SEGURA
Director de Despacho
UNSL

- Donald Adjeroh, Timothy Bell, Amar Mukherjee *The Burrows-Wheeler Transform: Data Compression, Suffix Arrays, and Pattern Matching*. Springer; 2008 edition.
- Gonzalo Navarro. *Wavelet Trees for All*. Journal of Discrete Algorithms. 25:2-20, 2014.
- Nieves Brisaboa, Susana Ladra, and Gonzalo Navarro. *Directly Addressable Variable-Length Codes. Proc. SPIRE'09*, pages 122-130. LNCS 5721.
- R. Grossi, A. Gupta, and J. Vitter. *High-order entropy compressed text indexes*. In proceedings of 14th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA 03), pages 841–850, 2003.

ARANCEL: $300 (pesos trescientos).
–Docentes y Alumnos de Posgrado de la UNSL: Gratuito.

COSTOS Y FUENTE DE FINANCIAMIENTO: Se realiza como parte de las tareas docentes del profesor responsable.

Cpde RESOLUCIÓN R Nº 2036
mav

Dra. Alicia Marcela PRINTISTA
A/C Secretaria de Posgrado
U.N.S.L.

Dr. José Roberto Saad
Vicerrector - UNSL
A/C Rectorado RR Nº 1983/16