COMP. 2007
(2P)

Universidad de Buenos Aires
Facultad de Ciencias Exactas y Naturales

Planilla a completar para presentación de Cursos de Posgrado

1.- DEPARTAMENTO de COMPUTACION.....

2.- NOMBRE DEL CURSO: Recuperación de Información y "Web Mining"

3.- DOCENTES:

RESPONSABLE/S: **Dr. José Castaño**
COLABORADORES:.....
AUXILIARES:.....

4.- CARRERA de DOCTORADO

5.- AÑO: 2007..... CUATRIMESTRE/S: 2° 2007

6.- PUNTAJE PROPUESTO PARA CARRERA DE DOCTORADO: 2 (dos) puntos

7.- DURACIÓN (anual, cuatrimestral, bimestral u otra): un cuatrimestre

8.- CARGA HORARIA SEMANAL:

Teóricas:.....
Problemas:.....
Laboratorio:.....
Seminarios:.....
Teórico – Práctico: 3 horas.....
Salida a Campo:.....

9.- CARGA HORARIA TOTAL: 48 hs.....

10.- FORMA DE EVALUACIÓN: Aprobación de parciales y examen final.

11.- PROGRAMA ANALÍTICO (adjuntarlo).

12.- BIBLIOGRAFÍA (indicar título del libro, autor, Editorial y año de publicación)(adjuntada)

UNIVERSIDAD DE BUENOS AIRES
FACULTAD DE CIENCIAS EXACTAS Y NATURALES
ENTRO | SALIO
7 AGO 2007

11. PROGRAMA

- Introducción a Recuperación de Información. Indices invertidos y consultas boolianas. Optimización de las consultas. La naturaleza del texto no-estructurado y del semi-estructurado.
- Codificación del Texto: Segmentación en 'tokens', extracción de lemas, 'stop words', y frases. Optimización de índices para el procesamiento de las consultas. Proximidad de frases y de consultas. Indices posicionales.
- Recuperación tolerante. Corrección ortográfica, sinónimos. Consultas con símbolos comodines ('wild cards'), permutación de índices. Indices de n-gramas. Distancia de Edición, "Soundex", detección del lenguaje.
- Construcción de Indices. Estimación del tamaño de los 'postings'. Indexación de los n-gramas. Cuestiones prácticas.
- Compresión de Indices. Compresión del léxico y de las listas de 'postings'. Codificación de 'gaps', códigos gamma, Ley de Zipf, Bloqueo, compresión extrema.
- Búsqueda parametrizada o por dominios. Zonas de documentos. El modelo del Espacio Vectorial ('Vector Space Model'). Esquemas de asignación de pesos. El esquema 'tf-idf'. Asignación de un valor ('score') a los documentos.
- Valoración en el modelo del Espacio Vectorial. La medida del Coseno. Consideraciones de eficiencia. Técnicas del vecino próximo. Aproximaciones de dimensionalidad reducida, proyecciones al azar.

- Presentación de resultados. Resumenes estáticos y dinámicos. Evaluación de resultados. Satisfacción del usuario. Medidas de Precisión, Recuperación y "F". Creación de colecciones. Medida kappa, concardancia de anotación. Relevancia, aproximación a la recuperación vectorial.

- Feedback de la relevancia. Pseudo-feedback. Expansión de las queries. Generación automatica de un thesuaurus.
Recuperación basada en los sentidos de los términos.
- Agrupamiento ("Clustering") y sus Métodos.
- Clasificación de textos y Métodos.

- Recuperación de información de la Web. Protocolos. Robots de navegación de la "web" ("Arañas"). Ranking basado en conducta y en 'links'. Analisis de 'links'.

- Recuperación y Extracción de Información. Sistemas de Respuesta a preguntas.

12.- BIBLIOGRAFÍA (indicar título del libro, autor, Editorial y año de publicación)(adjuntada)
Texto principal:
Introduction to Information Retrieval, by C. Manning, P. Raghavan, and H. Schütze.
Bibliografía adicional:
Managing Gigabytes. I. Witten, A. Moffat, and T. Bell.
Modern Information Retrieval. Baeza-Yates and B. Ribeiro-Neto.
Mining the Web, by S. Chakrabarti.
Foundations of Statistical Natural Language Processing, by C. Manning and H. Schütze.
Information Retrieval: Algorithms and Heuristics by D. Grossman and O. Frieder.

Dr. Alejandro Ríos
Subcomisión de Doctorado