Pregunta 1:

En el contexto de Recuperación de Información (*Information Retrieval*), definir el concepto relevancia de un documento ¿Qué medidas y procedimientos se suelen emplear para evaluar la relevancia de un documento en el modelo vectorial?

Pregunta 2:

¿Qué tipo de medidas se emplean para evaluar la eficacia de un sistema de recuperación de información?

Pregunta 3:

¿Qué es un índice invertido? Nombra algún contexto en el que se emplee.

Pregunta 4:

En el contexto de los sistemas de recuperación de información define Precisión y *Recall*.

Pregunta 5:

En el contexto de los sistemas de recuperación de información en que consiste la

eliminación de *stopwords*.

Pregunta 6:

En el contexto de sistemas de recuperación de información, define el proceso de lematización o *stemming*.

Pregunta 7:

Ventajas e inconvenientes de los modelos booleanos y vectoriales. Diferencias entre ellos.

Pregunta 8:
Diferencias entre eficiencia y eficacia en el contexto de sistemas de recuperación de información en corpus documentales cerrados y en la Web.

Pregunta 9:
¿Cómo se mide el rendimiento de un sistema de recuperación de información?

Pregunta 10:
¿Cuáles son los principales componentes de un sistema de recuperación de información? Descríbelos brevemente.

Pregunta 12:

En el contexto de recuperación de información qué se entiende por nombre compuesto (*compound noun*) y qué problemas puede acarrear su uso.

Pregunta 13:

En el contexto de recuperación de información qué se entiende por nombre de entidad (*entity name*) y qué técnicas se suelen emplear para tratar con este tipo de nombres.

Pregunta 14:

En la construcción de metabuscadores no es necesario disponer de corpus documentales ni construir y actualizar índices. ¿Cuáles son los componentes de un meta-buscador donde se invierten los mayores esfuerzos de desarrollo e implementación?

Pregunta 15:

Principales diferencias entre el algoritmo de ranking HITS y el *Page-Rank.*

Curso T1: Information Retrieval. Examen Final

Dr. Hugo Zaragoza. Julio 2004

****

Una empresa cuenta con una colección de documentos internos, y muchos de esos documentos están etiquetados indicando el autor del documento y su departamento.

Considerad por ejemplo que hay un millón de documentos, y 60% de ellos están etiquetados con autor y departamento, y un 99.5% de las etiquetas son correctas. Un 10% de los documentos etiquetados tienen más de un autor. El número de autores distintos está alrededor de 500 y el número de departamentos está alrededor de 10.

Disponéis de un sistema de IR que indexa los documentos y puede implementar cualquier tipo de búsqueda. Pero hoy en día solo utiliza el contenido (las palabras) de los documentos, y no hace uso de las etiquetas de autores o departamentos.

Vuestro jefe, deprimido, os viene un día rogando...

- Toda esa información..., no podrías hacer algo con ella?

El ejercicio consiste en:

1) Describir una manera en la que se podría utilizar la información de autor y departamento. Podéis intentar mejorar el sistema de búsqueda existente, o proponer un nuevo tipo de búsqueda.

Existen muchas cosas que se pueden hacer con esta información; es inútil intentar abarcarlo todo. Concéntrense en desarrollar hasta el final una sola idea, utilizando el mayor número de conceptos tratados durante el curso.

El tamaño del trabajo debe estar entre 900 y 1800 palabras.