Universidad Mayor de San Simón
Facultad de Ciencias y Tecnología
Carreras de Informática y Sistemas
Recuperación de la Información
Mayo de 2013

Trabajo semestral
Recuperación de la Información

El trabajo semestral consiste en construir un sistema recuperador de información, las presentaciones serán en dos partes, los puntos importantes a considerar se describe a continuación:

Se tomará sobre una cantidad mínima de 50 documentos electrónicos sobre tópicos diversos, y con un mínimo de 10 paginas por documento (en formatos html o txt).

Se debe preprocesar los documentos considerando:

- Stopwords
- Stemming (sufijos)
- Determinación de términos índice
- Construcción de estructuras de índices o thesauros

Adicionalmente, también se debe contemplar:

- La implementación del modelo booleano
- La implementación del modelo vectorial
  - Implementar el ranking de los documentos encontrados
- El sistema debe permitir añadir nuevos documentos e indexarlos automáticamente.

NOTA: Se deben considerar todos los aspectos relevantes expuestos en clase tanto en los modelos como en las estructuras de datos.

Por ejemplo: Dados los documentos:

Nro doc	Documentos
1	.....Las bases de datos constituyen una de las ramas mas importantes de las ciencias de la computación.....
2	.....Las Bases de Datos contienen inmensas cantidades de datos.....
3	.....los datos son elementales o complejos, y son importantes para las bases de datos.....
4	.....constutuyen las bases importantes.....

Preprocesamiento

stopwords
Las
De
Una
La
O
Y
.....
.....
.....
.....
.....
.....

Sufijos
S
Os
En
As
Es
Ción
.....
.....
.....
.....
.....
.....

stemmning	Docs. donde aparecen
Bas	1, 2, 3,4
Dat	1,2,3
Constituy	1,4
Ram	1
Cienci	1
Computa	1
Contien	2
Inmens	2
Cantidad	2
Elemental	3
Complej	3
Important	1, 3, 4

Un ejemplo de la consulta:

Datos complejos e inmensos

Insertar consulta

a) .....los **datos** son elementales o **complejos**, y son importantes para las bases de **datos**.....
b) .....Las Bases de **Datos** contienen **inmensas** cantidades de **datos.....**
c) .....Las bases de **datos** constituyen una de las ramas mas importantes de las ciencias de la computación.....

Respuesta

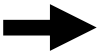
El proceso de recuperación fue:

Nota: no quitar sufijos a palabras de tamaño menor a cuatro letras

Consulta 1 inicial	Consulta 2 despues de stopwords	Consulta 3 Despues de auxiliar
Datos	Datos	Dat
Complejos	Complejos	Complej

stemming	Docs. donde aparecen (posteo)
Bas	1, 2, 3
Dat	***1,2,3***

E	Inmensos	Inmens
Inmensos		



Luego buscar

Constituy	1
Ram	1
Cienci	1
Computa	1
Contien	2
Inmens	***2***
Cantidad	2
Elemental	3
Complej	***3***
Important	1, 3

COMO COMPLEMENTO AL TRABAJO:

Al modelo anterior se debe añadir funcionalidad adicional:

- búsqueda por proximidad con distancia de máximo tres palabras (puede ser 0,1), es decir palabras que juntas tienen significado y separadas no.

Herramientas de desarrollo de software:

- **PHP**
- **mySQL o PostgreSQL**
- **O a elección del grupo**

La nota:

- Se distribuirá en dos partes: 60% sobre la implementación y defensa y 40% el documento

Entrega:

- CD con fuente y manual de instalación
- Informe del sistema en UP, XP, SCRUM ,UP agile
  - Poner énfasis en descripción de los modelos de recuperación vectorial
  - Índices usados y estructuras de datos
- Informe del Recall y precisión (respecto de su sistema)

Fecha de entrega:

- martes 25 de junio, para todos los grupos en horas de clase

- defensas los días 27 y 28 de junio bajo rol por programar bajo rol a publicar en el departamento