

Índice general

1. Introducción	1
1.1. Conceptos generales	3
1.1.1. Término	3
1.1.2. Documento	4
1.1.3. Consultas y preguntas	6
1.1.4. Indexación	7
1.1.5. Ficheros invertidos	8
1.1.6. Palabras comunes o <i>stopwords</i>	9
1.1.7. Términos relevantes o pivotes	10
1.1.8. <i>Stemming</i> y lematización	11
1.1.9. Pasaje	12
1.1.10. Expansión de la pregunta	13
1.2. Áreas de la Recuperación de Información	16
1.2.1. Recuperación de la Información <i>Ad hoc</i>	17
1.2.2. Búsqueda de Respuestas	20
1.2.3. Extracción de la Información	20
1.3. Motivación y objetivos de la tesis	21
1.4. Organización de la tesis	24
2. Estado del arte	27
2.1. Sistemas de Recuperación de Información	27
2.1.1. Modelo booleano	27
2.1.2. Modelo Booleano Difuso	29
2.1.3. Modelo Booleano Extendido	29
2.1.4. Modelo Espacio Vectorial	31
2.2. Sistemas de Recuperación de Pasajes	33
2.3. El tamaño del pasaje	34

2.3.1.	Modelos de ventanas	35
2.3.2.	Modelos de discurso	36
2.3.3.	Modelos semánticos	37
2.3.4.	Comparativa entre modelos de división de documentos	38
2.3.5.	Solapamiento	40
2.4.	Sistemas de Búsqueda de Respuestas	41
2.5.	Análisis de la pregunta	44
2.6.	Sistemas de Recuperación de pasajes	44
2.7.	Extracción de la respuesta	47
2.8.	Campañas de evaluación de sistemas de Recuperación de la Información y Búsqueda de Respuestas	48
2.8.1.	Text REtrieval Conference	48
2.8.2.	Cross-Language Evaluation Forum	52
3.	Modelos de n-gramas	55
3.1.	Arquitectura del sistema	57
3.2.	Modelo de N -gramas Simple	59
3.3.	El modelo de n -gramas Term Weight	64
3.4.	El modelo de Densidad de Distancias de N -gramas	66
3.5.	Reformulaciones de la pregunta	71
3.6.	Filtros de pasajes	73
4.	Descripción del sistema	77
4.1.	Arquitectura del sistema	77
4.2.	Java Process Manager	79
4.3.	Archivo de Configuración de Ejecuciones	80
4.3.1.	Acciones	81
4.3.2.	Ejecuciones	82
4.3.3.	Procesos y subprocessos	83
4.3.4.	Métodos	87
4.4.	Parámetros	90
4.4.1.	Parámetros estáticos	90
4.4.2.	Parámetros dinámicos	92
4.4.3.	Referencia entre parámetros	94
4.4.4.	Array de parámetros	94
4.5.	Ámbito de las acciones y parámetros	95

4.6.	JAVA Database Information Retrieval	96
4.7.	Acciones y parámetros condicionados	100
4.8.	JAVA Information Retrieval System	103
4.8.1.	Indexación de la colección de documentos	103
4.8.2.	Búsqueda de pasajes local	104
4.8.3.	Búsqueda de pasajes remota	105
4.8.4.	Búsqueda de pasajes distribuida	107
4.8.5.	Recuperación de pasajes sobre colección de preguntas	108
4.8.6.	Creación de tablas de cobertura	109
4.8.7.	Aplicar modelos de n -gramas a otros sistemas . . .	110
5.	Evaluación de JIRS	111
5.1.	Medidas de evaluación	112
5.1.1.	Cobertura	112
5.1.2.	Mean Reciprocal Rank	112
5.1.3.	Redundancia	113
5.1.4.	Precisión	114
5.2.	El corpus	114
5.2.1.	Clasificación de las preguntas	115
5.3.	Ajustes de parámetros	119
5.3.1.	Modelo Simple	120
5.3.2.	Modelo Term Weight	130
5.3.3.	Modelo Densidad de Distancias de N -gramas . . .	138
5.3.4.	Comparación de los modelos de n -gramas	149
5.3.5.	Tamaño del pasaje	157
5.4.	Comparación con otros sistemas de RP	162
5.4.1.	Descripción de los sistemas de RP	162
5.4.2.	Evaluación de los sistemas de RP	163
5.5.	Resultados en Búsqueda de Respuestas	170
6.	Conclusiones	177
	Bibliografía	182
	A. Glosario de abreviaturas	201

B. Manual de usuario	203
B.1. Instalación	203
B.2. Modo de uso	204
B.3. Parámetros globales	205
B.4. Indexar documentos	208
B.5. Comprobar si la indexación se ha realizado con éxito . .	213
B.6. Lanzar el motor de búsqueda de pasajes	214
B.7. Realizar una búsqueda de pasajes remota	218
B.8. Realizar una búsqueda de forma local	218
B.9. Buscar una colección de preguntas automáticamente . . .	219
B.9.1. Formato de preguntas del CLEF	219
B.9.2. Formato XML de JIRS (JIRS Questions)	220
B.9.3. Búsqueda en modo cliente/servidor	220
B.9.4. Búsqueda en modo local	222
B.10. Crear una tabla de cobertura	222
B.11. Adaptar JIRS a nuevos idiomas no soportados	225
B.12. Errores comunes	228
C. GNU General Public License	229
C.1. Preamble	229
C.2. TERMS AND CONDITIONS FOR COPYING, DISTRI- BUTION AND MODIFICATION	231
