

TABLA DE CONTENIDOS

	página
Dedicatoria	I
Agradecimientos	II
Tabla de Contenidos	III
Índice de Figuras	VI
Índice de Tablas	VII
Resumen	VIII
Abstract	IX
1. Introducción	1
1.1. Descripción del problema	3
1.2. Objetivos	4
1.2.1. Objetivo general	4
1.2.2. Objetivos específicos	4
1.3. Alcance	4
1.4. Limitaciones	5
2. Antecedentes	6
2.1. Traducción automática	6
2.2. ¿Cómo es la Web?	7
2.3. Máquinas de búsqueda	8
2.4. Web Crawlers	10
2.4.1. Principales <i>Web Crawler</i> creados	11
2.4.2. Aspectos a considerar en la construcción de un <i>Web Crawler</i> .	11
2.4.3. Arquitectura de un <i>Web Crawler</i>	16
2.4.4. Análisis de páginas Web	16
2.4.5. Crawler paralelo	17
2.5. Computación paralela y distribuida	20

2.5.1.	Arquitecturas paralelas	21
2.5.2.	Modelos de programación paralela	22
2.5.3.	Invocación remota de métodos	24
2.5.4.	Métricas de desempeño	25
3.	Desarrollo	27
3.1.	Consideraciones metodológicas	27
3.1.1.	Etapa 1: Construcción del Crawler secuencial	28
3.1.2.	Etapa 2: Construcción de Crawler paralelo	29
3.1.3.	Etapa 3: Evaluación del Crawler	29
3.2.	Construcción del Crawler secuencial	30
3.2.1.	Diseño	30
3.2.2.	Implementación	31
3.3.	Construcción del Crawler paralelo	32
3.3.1.	Diseño	32
3.3.2.	Implementación	37
3.4.	Otros aspectos relevantes	38
3.4.1.	Planificación de URLs	38
3.4.2.	Detección de idioma	40
4.	Evaluación experimental	42
4.1.	Evaluación del Crawler secuencial	42
4.1.1.	Diseño de prueba	42
4.1.2.	Resultados obtenidos	43
4.1.3.	Discusión de resultados	43
4.2.	Evaluación del Crawler paralelo	43
4.2.1.	Diseño de prueba	45
4.2.2.	Definición del espacio de prueba	46
4.2.3.	Resultados obtenidos	46
4.2.4.	Discusión de resultados	46
4.2.5.	Determinación del número de <i>Escritores</i> y carga de trabajo de <i>Lectores</i>	49
4.3.	Evaluación de la planificación de URLs	51
4.3.1.	Diseño de prueba	51

4.3.2. Resultados obtenidos	51
4.3.3. Discusión de resultados	52
4.4. Determinación del porcentaje promedio de <i>Stopwords</i> utilizados en idioma Español	52
4.4.1. Diseño de prueba	52
4.4.2. Resultados obtenidos	53
4.4.3. Discusión de resultados	53
4.5. Evaluación de la detección de idioma	54
4.5.1. Diseño de prueba	54
4.5.2. Resultados obtenidos	54
4.5.3. Discusión de resultados	61
5. Conclusiones	62
5.1. Trabajo Futuro	63
Bibliografía	64

ÍNDICE DE FIGURAS

	página
1.1. Arquitectura proyecto de <i>Refinamiento de Traducciones Automáticas</i>	2
2.1. Ejemplos ilustrativos de una red aleatoria y una red libre de escala.	7
2.2. Funcionamiento de un motor de búsqueda.	9
2.3. La vista del motor de búsqueda.	12
2.4. Evolución de la frescura y la edad en el tiempo	14
2.5. Arquitectura de alto nivel de un <i>Web Crawler</i>	17
2.6. Flujo de un Crawler secuencial.	18
2.7. Modelo de Crawler multi-hilos.	19
2.8. Sistema Paralelo.	20
2.9. Sistema Distribuido.	21
2.10. Interfaz remota.	25
3.1. Etapas para el desarrollo del Crawler.	28
3.2. Arquitectura de alto nivel propuesta para el desarrollo del Crawler.	30
3.3. Diagrama de Clases simplificado.	33
3.4. Arquitectura de alto nivel propuesta para el desarrollo del Crawler paralelo.	36
4.1. Aceleración obtenida de acuerdo al número de procesadores activos.	47
4.2. Eficiencia obtenida de acuerdo a la cantidad de procesadores activos.	48
4.3. Comparación del Crawler con planificador v/s Crawler sin planificador.	51
4.4. Porcentaje de <i>Stopwords</i> contenidos en idioma Español.	53
4.5. Comparación entre Crawler con y sin detección de idioma.	55

ÍNDICE DE TABLAS

	página
3.1. Descripción del diagrama de clases.	34
4.1. Semillas utilizadas en la prueba del Crawler secuencial.	43
4.2. Cantidad de frases obtenidas por el Crawler secuencial en 20 minutos.	44
4.3. Tiempo utilizado en descargar y analizar 5.000 páginas Web.	47
4.4. Estadística del número de Escritores utilizados.	50
4.5. Dominios más populares con detección de idioma.	56
4.6. Dominios más populares sin detección de idioma.	57
4.7. Dominios menos populares con detección de idioma	58
4.8. Dominios menos populares sin detección de idioma	59
4.9. Resumen del análisis de idioma realizado a frases en Italiano.	60