

TABLA DE CONTENIDO

Introducción	1
Objetivo.....	2
Estructura de la tesis	2
1 Procesamiento de Lenguaje Natural	4
1.1 Recursos y herramientas empleadas en PLN	5
1.1.1 Corpus lingüísticos.....	5
1.1.2 Tokenizadores	7
1.1.3 N-gramas.....	9
1.1.4 Etiquetadores de partes de la oración.....	10
1.1.5 Lematizadores	12
1.1.6 Palabras funcionales.....	14
1.2 Recuperación de información.....	15
1.2.1 Term frequency – Inverse document frequency (TF-IDF)	16
1.2.2 Normalización de la longitud del documento	20
1.2.2.1 Normalización de coseno.....	21
1.2.2.2 Normalización por pivote	24
1.2.2.3 Normalización por máximo TF	26
1.2.3 Evaluación de sistemas de recuperación de información	27
1.2.4 Extracción y recuperación de información	29
2 Terminología.....	31

2.1	Terminología y terminografía	31
2.1.1	Los términos.....	32
2.1.2	La terminografía.....	33
2.1.3	Extracción de información terminológica.....	35
2.2	Sistemas actuales de extracción terminológica	36
2.2.1	Sistemas basados en conocimiento lingüístico	37
2.2.1.1	LEXTER.....	38
2.2.1.2	HEID.....	41
2.2.2	Sistemas basados en conocimiento estadístico	42
2.2.2.1	ANA.....	43
2.2.2.2	Extractor de términos estadístico basado en corpus	45
2.2.3	Sistemas basados en conocimiento híbrido.....	46
2.2.3.1	Termext.....	46
2.2.3.2	YATE.....	47
2.3	Evaluación de los extractores terminológicos.....	49
2.3.1	Lista de referencia.....	49
2.3.2	Validación	50
2.4	Recursos electrónicos para la validación	50
2.4.1	WordNet y EuroWordNet	51
2.4.2	Lexicón Specialist UMLS.....	52
2.4.3	Wikipedia.....	52

3	Obtención automática de términos y su validación	54
3.1	Corpus de textos científicos en español de México (COCIAM)	54
3.1.1	Estructura del COCIEM.....	55
3.2	Preprocesamiento del COCIEM.....	56
3.2.1	Revisión, limpieza y adecuación de los documentos.....	57
3.2.2	Lematización usando FreeLing.....	57
3.2.3	Tokenización del COCIEM	61
3.2.4	Creación de n-gramas	62
3.3	Extracción de candidatos a término	64
3.3.1	Cálculo de TF.....	65
3.3.2	Limpieza de los n-gramas generados.....	66
3.3.3	Cálculo de IDF, TF-IDF y su normalización.....	68
3.4	Validación de los candidatos a término.....	71
3.4.1	Wikipedia para la validación	72
3.4.1.1	Conversión a una base de datos.....	75
3.4.1.2	Lematización de Wikipedia	78
3.4.2	Cálculo del coeficiente de dominio.....	79
3.5	Arquitectura del sistema.....	84
4	Resultados y evaluación.....	87
4.1	Extracción de candidatos a término del COCIEM.....	87
4.2	Selección de los candidatos a término a validar.....	89

4.3	Obtención de términos validados por Wikipedia	93
4.4	Evaluación de resultados	101
4.4.1	Observaciones de la evaluación de resultados	105
5	Conclusión y trabajo a futuro.....	110
	Bibliografía	113
	Anexos	119
	Anexo A: Lista de palabras funcionales	120
	Anexo B: Gráficas de precisión contra cobertura de matemáticas de bachillerato.....	125
	Anexo C: Gráficas de precisión contra cobertura de ecología de bachillerato	131
	Anexo D: Gráficas de precisión contra cobertura de matemáticas de primaria.....	137
	Anexo E: Gráficas de precisión contra cobertura de matemáticas de bachillerato evaluada con la segunda lista de términos	143
	Anexo F: Lista de términos validados de matemáticas de bachillerato	145
	Anexo G: Lista de términos validados de ecología de bachillerato.....	157
	Anexo H: Lista de términos validados de matemáticas de primaria.....	160

