

Tabla de Contenido

1. Introducción	1
1.1. Contexto	1
1.2. Problema	2
1.3. Situación Actual	3
1.4. Objetivos	4
1.4.1. Objetivo General	4
1.4.2. Objetivos Específicos	4
1.5. Descripción General de la Solución	5
2. Marco Teórico	7
2.1. Linked Data	7
2.1.1. RDF	8
2.1.2. SPARQL	10
2.2. Grafo de Conocimiento	10
2.3. Wikidata	11
2.3.1. Modelo de Datos	11
2.3.2. Wikidata en la Web Semántica	15
2.4. Web Crawling	16
2.5. Índices Invertidos	18
2.6. TF-IDF	18
2.6.1. Algoritmo de ranking de Apache Lucene	19
3. Problema	21
3.1. Descripción	21
3.2. Situación Actual	23
3.3. Desafíos	25
4. Solución	27
4.1. Descripción	27
4.2. Obtener Datos de Elementos	27
4.3. Referencias del Índice	28
4.3.1. Opción offline	28
4.3.2. Opción online	28
4.3.3. Comparación de opciones: offline vs online	29
4.3.4. Índice de documentos	29
4.4. Rango de Búsqueda	30
4.4.1. Comparación de Opciones de Rangos de Búsqueda	31

4.5. Términos de Búsqueda	31
4.6. Documentos Retornados	35
4.7. Ejemplo de uso	35
5. Validación	39
5.1. Datos	39
5.1.1. Crawl Completo	39
5.1.2. Referencias Indexadas	40
5.1.3. Muestra de entidades	41
5.2. Referencias conocidas de Wikidata	43
5.3. Gold Standards	44
5.3.1. nDCG	45
5.3.2. Any at k: 1-5	46
5.3.3. F_1	46
5.4. Validación de extractos	48
5.5. Rendimiento offline	48
6. Conclusión	50
Glosario	52
Bibliografía	54
Anexo A. Datos usados en los gráficos de la sección 5	56
A.1. Referencias descargadas en top 10 dominios del total de referencias.	56
A.2. Resultados referencias existentes	57
A.3. Resultados validación nDCG	57
A.4. Resultados validación Any at k	57
A.5. Resultados validación F_1	58
A.6. Resultados rendimiento offline	58