

Tabla de Contenido

Introducción	1
1. Objetivos e Hipótesis	4
1.1. Objetivo General	4
1.2. Objetivos Específicos	4
1.3. Hipótesis	5
1.4. Preguntas de Investigación	5
1.5. Metodología	5
2. Preliminares	7
2.1. La Web Semántica	7
2.2. RDF	7
2.2.1. Vocabulario RDF	8
2.2.2. Esquema RDF	9
2.3. Wikidata	9
2.4. Apache Spark	11
3. Trabajos Relacionados	13
3.1. Medidas Globales de Similitud de Vértices	13
3.2. Comparaciones de Medidas de Similitud	16
3.3. Aplicaciones de medidas de similitud	17
3.4. Investigaciones en Wikidata	18
3.5. Aporte de esta Investigación	19
4. Algoritmos de Similitud Propuestos	20
4.1. Preliminares	20
4.2. Conteo de Vecinos Comunes	21
4.3. Suma Probabilística de Vecinos Comunes	22
4.4. Conteo de Intersección	24
4.5. SimRank	25
4.6. Comparaciones de las Medidas Propuestas	26
5. Implementación de las Medidas	27
5.1. Extractor de Tuplas	27
5.2. Conteo de Vecinos Comunes	28
5.3. Suma Probabilística de Vecinos Comunes	29
5.4. Conteo de Intersección	29

5.5. Costo de las Implementaciones	31
6. Evaluación	33
6.1. Descripción de Datos	33
6.2. Tiempos en Calcular las Medidas	34
6.3. Efecto del valor de k en las medidas.	38
6.3.1. Coeficiente de Correlación tau-b de Kendall	38
6.3.2. Valor de p y la significación estadística	39
6.3.3. Resultados de las Pruebas	40
6.3.4. Análisis del efecto de k	42
6.4. Análisis de la Calidad de los Resultados	43
6.5. Comparación de los Resultados con Medida Base	46
6.6. Aplicación de las medidas en aristas inversas	49
Conclusión	54
Bibliografía	56