

Tabla de contenido

1. Introducción	1
1.1. Motivación	1
1.2. Objetivos de la memoria	3
2. Marco teórico	4
2.1. <i>Tag Clouds</i>	4
2.2. Técnicas de extracción de <i>keywords</i>	5
2.2.1. Tf-idf	6
2.2.2. <i>Rapid Automatic Keyword Extraction</i> (RAKE)	7
2.2.3. KEA	8
2.3. Aprendizaje automático	9
2.3.1. Clasificación	10
2.3.2. Regresión	11
2.3.3. Modelos de aprendizaje supervisado	11
2.4. <i>Learning to rank</i>	13
2.4.1. Métricas de evaluación	14
2.5. Resumen	15
3. Solución propuesta	17
3.1. Arquitectura de la solución	17
3.2. Extracción de las publicaciones	18
3.2.1. Fuente de información	18
3.2.2. Acceso a DBLP	19
3.3. Extracción de PDFs	21
3.4. Extracción de <i>keywords</i>	22
3.4.1. Extracción del texto y pre-procesamiento	23
3.4.2. Extracción de <i>keywords</i> según su tf-idf	23
3.4.3. Extracción de <i>keywords</i> con RAKE	24
3.4.4. Extracción de <i>keywords</i> con KEA	24
3.4.5. Enfoque escogido	25
3.5. Selección de <i>keywords</i>	25
3.5.1. Características	25
3.5.2. Clasificación	26
3.5.3. <i>Learning to Rank</i>	27
3.6. Creación de <i>tag clouds</i>	28
3.7. Sistema final	28

3.7.1. API	29
3.7.2. Interfaz de usuario	30
4. Evaluación de la solución	32
4.1. Conjunto de datos	32
4.1.1. Extracción de <i>keywords</i>	32
4.1.2. Etiquetado de las <i>keywords</i>	33
4.1.3. Características	34
4.2. Evaluación de los modelos de <i>learning to rank</i>	35
4.2.1. Validación cruzada	36
4.2.2. Validación final	40
4.3. Discusión	41
5. Conclusiones y trabajo a futuro	44
Bibliografía	45
A. Evaluaciones	48
A.1. Tau de Kendall	48

Índice de tablas

4.1. Resumen de recuperación y tiempos del proceso de extracción de <i>keywords</i>	33
4.2. Tabla completa de valores p.	40
A.1. Tabla completa de los índices de correlación de Kendall.	48