

# Índice general

Índice de figuras	VIII
Índice de tablas	IX
<b>1. Introducción</b>	<b>1</b>
1.1. Planteamiento del problema . . . . .	1
1.2. Objetivos . . . . .	2
1.3. Estado del arte . . . . .	3
1.4. Descripción de la tesis . . . . .	6
<b>2. Conceptos de la tesis</b>	<b>8</b>
2.1. Minería de datos . . . . .	8
2.2. Minería de textos . . . . .	9
2.2.1. Etapas de un sistema de minería de textos . . . . .	11
2.2.1.1. Etapa I: Recopilación de documentos . . . . .	11
2.2.1.2. Etapa II: Tareas de preprocesamiento . . . . .	11
2.2.1.2.1. Estandarización de los documentos . . . . .	12
2.2.1.2.2. Segmentación . . . . .	13
2.2.1.2.3. Lematización . . . . .	14
2.2.1.2.4. Generación del vector de rasgos . . . . .	14
2.2.1.2.5. N-gramas . . . . .	15
2.2.1.2.6. Etiquetado PoS ( <i>Part of Speech tagging</i> ) . . . . .	15
2.2.1.3. Etapa III: Operaciones principales de minería . . . . .	17
2.2.1.4. Etapa IV: Presentación . . . . .	17

2.3. Aprendizaje de Máquinas ( <i>Machine Learning</i> ) . . . . .	17
2.3.1. Clasificación . . . . .	18
2.3.1.1. Método de Bayes ingenuo ( <i>naïve Bayes</i> ) . . . . .	18
2.3.2. Evaluación del clasificador . . . . .	22
2.3.2.1. Remuestreo ( <i>resampling</i> ) . . . . .	22
2.3.2.1.1. Validación cruzada con $k$ pliegues . . . . .	23
2.3.3. Medición del desempeño del clasificador . . . . .	24
2.3.3.1. Medidas de desempeño . . . . .	24
2.3.3.1.1. Espacio ROC . . . . .	26
2.3.4. Agrupamiento . . . . .	28
2.3.4.1. Tipos de algoritmos de agrupamiento . . . . .	30
2.3.4.2. Descomposición de matrices . . . . .	33
2.3.4.2.1. Factorización no negativa de matrices . . . . .	35
2.4. Minería de Opiniones . . . . .	39
2.4.1. Introducción . . . . .	39
2.4.2. Aplicaciones . . . . .	41
<b>3. Metodología . . . . .</b>	<b>42</b>
3.1. Herramientas de programación utilizadas . . . . .	42
3.1.1. Python . . . . .	42
3.1.2. Eclipse IDE (Integrated Development Environment) . . . . .	44
3.1.3. Pydev . . . . .	45
3.2. Procesos del sistema . . . . .	45
3.2.1. Obtención de artículos sobre películas desde Wikipedia . . . . .	45
3.2.2. Extracción de los títulos de las películas . . . . .	48
3.2.3. Extracción y almacenamiento de datos generales y reseñas . . . . .	48
3.2.4. Separación de reseñas por orientación . . . . .	53
3.2.5. Selección de rasgos y generación de matrices de datos . . . . .	54
3.2.6. Agrupamiento con factorización no negativa de matrices (FNM) . . . . .	57
3.2.7. Detección de oraciones subjetivas . . . . .	59
3.2.7.1. Oraciones con adjetivos o adverbios . . . . .	59
3.2.7.2. Oraciones con disparadores de presuposición . . . . .	60

3.2.7.3. Oraciones con disparadores o adjetivos o adverbios	60
3.2.8. Validación cruzada con 10 pliegues . . . . .	60
<b>4. Resultados</b>	<b>64</b>
4.1. Resultados generales . . . . .	64
4.2. Enunciados con adjetivos o adverbios . . . . .	69
4.3. Enunciados agrupados con FNM . . . . .	69
4.4. Enunciados con disparadores de presuposición o adjetivos o adverbios	70
4.5. Todos los enunciados . . . . .	70
4.6. Enunciados con disparadores de presuposición . . . . .	70
<b>5. Conclusiones</b>	<b>72</b>
<b>Referencias</b>	<b>77</b>
<b>Apéndices</b>	<b>83</b>
Apéndice A: Matrices de confusión y resultados de precisión, exhausti- vidad y medida F . . . . .	84
Apéndice B: Descripción de los módulos del sistema . . . . .	87
Apéndice C: Descripción del módulo de agrupamiento automático con FNM . . . . .	90
Apéndice D: Descripción del módulo de clasificación binaria mediante Bayes ingenuo . . . . .	91

# Índice de figuras

2.1. Etapas de un sistema de minería de textos. . . . .	11
2.2. Diagrama de un etiquetador PoS . . . . .	16
2.3. Espacio ROC . . . . .	29
2.4. Ejemplo de un agrupamiento jerárquico . . . . .	31
2.5. Ejemplo de un agrupamiento particional . . . . .	32
2.6. Descomposición de la matriz de datos $A$ en las matrices $W$ , $C$ y $H$	34
2.7. Matrices obtenidas después de la factorización con FNM. . . . .	35
3.1. Procesos que comprenden al sistema de clasificación . . . . .	46
3.2. Archivo <code>2009_films.xml</code> . . . . .	47
3.3. Forma en la que se encuentran los datos en la página original de IMDb . . . . .	50
3.4. Archivo <code>peliculas.xml</code> . . . . .	51
3.5. Archivo <code>reseñas.xml</code> . . . . .	52
3.6. Matriz dispersa . . . . .	56
3.7. Archivo <code>NMFpositivos.txt</code> . . . . .	58
3.8. Validación cruzada con 10 pliegues . . . . .	62
4.1. Ubicación de los clasificadores en el espacio ROC . . . . .	66
4.2. Curva ROC para cada método utilizado . . . . .	67

# Índice de tablas

2.1. Matriz de confusión . . . . .	24
2.2. Medidas de desempeño para clasificadores . . . . .	25
4.1. Resultados promediados de la validación con 10 pliegues del clasificador creado. . . . .	65
4.2. Resultados promediados de la validación cruzada con 3 pliegues del clasificador creado para cada experimento. . . . .	68