

Tabla de contenidos

1	ANALÍTICA TEXTUAL	25
1.1	INTRODUCCIÓN	25
1.2	MINERÍA DE TEXTOS Y ANALÍTICA TEXTUAL	28
1.3	TAREAS Y APLICACIONES	30
1.4	EL PROCESO DE LA ANALÍTICA TEXTUAL.....	33
1.5	RESUMEN.....	36
1.6	PREGUNTAS	37
2	PROCESAMIENTO DEL LENGUAJE NATURAL	39
2.1	INTRODUCCIÓN	39
2.2	PROCESAMIENTO DEL LENGUAJE NATURAL.....	41
2.3	NIVELES Y TAREAS EN NLP	42
2.3.1	Fonología	43
2.3.2	Morfología	44
2.3.3	Léxico.....	45
2.3.4	Sintaxis.....	51
2.3.5	Semántica	55
2.3.6	Razonamiento y pragmática	60
2.4	RESUMEN.....	60
2.5	EJERCICIOS.....	62
2.5.1	Análisis morfológico	62
2.5.2	Análisis léxico.....	66
2.5.3	Análisis sintáctico	68
3	EXTRACCIÓN DE INFORMACIÓN.....	71
3.1	INTRODUCCIÓN	71
3.2	EXTRACCIÓN DE INFORMACIÓN BASADA EN REGLAS	75
3.3	EXTRACCIÓN DE ENTIDADES NOMBRADAS	76
3.3.1	Modelos de N-gramas	78
3.4	EXTRACCIÓN DE RELACIONES.....	81
3.5	EVALUACIÓN.....	86
3.6	RESUMEN.....	88
3.7	EJERCICIOS.....	90

3.7.1	Extracción vía expresiones regulares	90
3.7.2	Reconocimiento de entidades nombradas (NER).....	94
4	REPRESENTACIÓN DE DOCUMENTOS	97
4.1	INTRODUCCIÓN	97
4.2	INDEXACIÓN DE DOCUMENTOS	99
4.3	MODELOS DE ESPACIO VECTORIAL	101
4.3.1	Modelo de representación booleana	102
4.3.2	Modelo de frecuencia de términos	103
4.3.3	Modelo de frecuencia inversa de documentos	104
4.4	RESUMEN	106
4.5	EJERCICIOS.....	107
4.5.1	Modelo de representación TFxIDF	107
5	ANÁLISIS DE REGLAS DE ASOCIACIÓN	115
5.1	INTRODUCCIÓN	115
5.2	PATRONES DE ASOCIACIÓN.....	116
5.3	EVALUACIÓN	118
5.3.1	Support.....	118
5.3.2	Confidence	119
5.3.3	Lift.....	119
5.4	GENERACIÓN DE REGLAS DE ASOCIACIÓN	120
5.5	RESUMEN	124
5.6	EJERCICIOS.....	126
5.6.1	Extracción de reglas de asociación	126
6	ANÁLISIS SEMÁNTICO BASADO EN CORPUS	131
6.1	INTRODUCCIÓN	131
6.2	ANÁLISIS BASADO EN CORPUS	133
6.3	ANÁLISIS SEMÁNTICO LATENTE	135
6.3.1	Generación de vectores con LSA	136
6.4	WORD2VEC.....	140
6.4.1	Aprendizaje de embeddings en CBOW.....	143
6.4.2	Predicción e interpretación de embeddings	146
6.5	RESUMEN	148
6.6	EJERCICIOS.....	149
6.6.1	Análisis semántico latente (LSA).....	149
6.6.2	Modelo de Word embedding del tipo Word2Vec	156
7	AGRUPACIÓN DE DOCUMENTOS	161
7.1	INTRODUCCIÓN	161

7.2	CLUSTERING DE DOCUMENTOS.....	163
7.3	CLUSTERING K-MEANS.....	169
7.4	MAPAS AUTOORGANIZATIVOS.....	172
7.4.1	Aprendizaje de mapas topológicos.....	174
7.5	RESUMEN.....	178
7.6	EJERCICIOS.....	179
7.6.1	Clustering via K-means	179
7.6.2	Clustering vía mapas autoorganizativos	185
8	MODELAMIENTO DE TÓPICOS.....	188
8.1	INTRODUCCIÓN	189
8.2	MODELAMIENTO DE TÓPICOS.....	191
8.3	LATENT DIRICHLET ALLOCATION	193
8.4	EVALUACIÓN.....	200
8.5	RESUMEN.....	202
8.6	EJERCICIOS.....	203
8.6.1	Modelamiento de tópicos con LDA	203
9	CATEGORIZACIÓN DE DOCUMENTOS.....	209
9.1	INTRODUCCIÓN	209
9.2	MODELOS DE CATEGORIZACIÓN	211
9.3	CLASIFICACIÓN BAYESIANA	214
9.4	CATEGORIZACIÓN POR MÁXIMA ENTROPÍA	218
9.5	EVALUACIÓN.....	223
9.6	RESUMEN.....	225
9.7	EJERCICIOS.....	227
9.7.1	Categorización con Naïve Bayes	227
9.7.2	Categorización con Máxima Entropía.....	232
10	CONCLUSIONES	239
	Bibliografía	244
	Glosario	244
	Índice onomástico	244

PLATAFORMA DE CONTENIDOS INTERACTIVOS

Para tener acceso al material de la plataforma de contenidos interactivos de *Análítica textual*, siga los siguientes pasos:

1. Ir a la página: https://libroweb.alfaomega.com.mx/book/analitica_textual
2. En la sección *Materiales de apoyo* podrá descargar gratis el contenido adicional, complemento imprescindible de este libro, el cual podrá descomprimir con la clave: **TEXT23**

Figura 1.1: Búsqueda versus descubrimiento en datos.....	26
Figura 1.2: El ámbito del text mining	30
Figura 1.3: Agrupación de información textual.....	31
Figura 1.4: Extracción de información.....	32
Figura 1.5: Categorización de textos	32
Figura 1.6: Inferencia de relaciones	33
Figura 1.7: El proceso del text mining	33
Figura 2.1: Niveles, tareas y recursos lingüísticos en el NLP	43
Figura 2.2: Un modelo de Márkov simple	48
Figura 2.3: Una HMM con probabilidades de transición y emisión	50
Figura 2.4: Tarea de análisis sintáctico.....	52
Figura 2.5: Reglas de una CFG	53
Figura 2.6: Árbol sintáctico para la frase «El vuelo despegó».....	53
Figura 2.7: Gramática de dependencias para «el vuelo despegó sin problemas» ...	54
Figura 2.8: Grafo semántico para «El vuelo despegó sin problemas»	56
Figura 2.9: Desambiguación del sentido de las palabras (WSD)	57
Figura 2.10: Estructura de relaciones retóricas de un texto de ejemplo	60
Figura 3.1: Un texto de queja.....	71
Figura 3.2: Extracción de relaciones específicas para alimentar una tabla.....	72
Figura 3.3: Pasos en la extracción de información.....	73
Figura 3.4: Extracción de información simple y relacional.....	74
Figura 3.5: Ejemplo de extracción basado en reglas en cascada.....	75
Figura 3.6: Asociación y búsqueda de relaciones específicas	82
Figura 3.7: Extracción de relaciones de interacción proteína-proteína	82
Figura 4.1: La tarea de indexación o generación de características.....	99
Figura 4.2: Representación vectorial de textos de ejemplo.....	102
Figura 5.1: Transacciones vistas como canastas de compra	117
Figura 5.2: Generación de itemsets frecuentes con método APRIORI.....	122
Figura 5.3: Espacio de búsqueda de reglas de asociación.....	123
Figura 6.1: Representación de word embeddings.....	134
Figura 6.2: Descomposición vía SVD	137
Figura 6.3: Elección del mejor número de dimensiones basado en importancia de valores singulares.....	138
Figura 6.4: Tipos de modelos Word2Vec	141
Figura 6.5: Ejemplos de ventanas de contexto para entrenamiento	142
Figura 6.6: Arquitectura de un modelo CBOW.....	143
Figura 6.7: Actualización de valores de la capa oculta	144

Figura 6.8: Cálculo para neuronas de salida	144
Figura 6.9: Predicción de palabras de salida según clasificador SoftMax.....	145
Figura 6.10: Proceso de entrenamiento en CBOW para generar predicciones	146
Figura 6.11: Estructura general de la red para predicción de contextos.....	146
Figura 7.1: Agrupación simple de noticias.....	162
Figura 7.2: Combinaciones de grupos de documentos	163
Figura 7.3: Ejemplo de tres clusters considerando dos dimensiones en los datos	164
Figura 7.4: Clustering de documentos.....	164
Figura 7.5: Distribución de clusters según sus centroides.....	166
Figura 7.6: Distribución espacial de vectores de documentos	168
Figura 7.7: Generación de $K = 3$ clusters para un cierto conjunto de datos.....	170
Figura 7.8: Diferentes <i>clusterings</i> según los centroides iniciales	171
Figura 7.9: Selección del mejor número de clusters	172
Figura 7.10: Arquitectura general de un SOM.....	173
Figura 7.11: Algunas formas de topologías de SOM: (a) rectángulo (2D), (b) octaedro (2D) y (c) lineal (1D)	174
Figura 7.12: Representación geométrica de clusters en SOM.....	178
Figura 8.1: Distribución de tópicos, palabras y documentos.....	191
Figura 8.2: Una mezcla de tres distribuciones.....	192
Figura 8.3: Interpretación geométrica de tópicos	195
Figura 8.4: Modelamiento de dependencia entre documentos, tópicos y palabras	196
Figura 8.5: Modelamiento de tópicos con distribución de Dirichlet	197
Figura 8.6: Distribucion Dirichlet sobre un 2-simplex para diferentes valores de α	199
Figura 8.7: Visualización de un modelo de tópicos para un corpus de entrada	200
Figura 8.8: Evaluación de un modelo de tópicos según perplejidad	201
Figura 8.9: Evaluación de modelos de tópicos según coherencia	202
Figura 9.1: Entrenamiento de un modelo de clasificación de textos	212
Figura 9.2: Clasificación de textos no vistos	212
Figura 9.3: Entropía para una distribución de una variable X	219
Figura 9.4: Cálculo de funciones indicadoras y probabilidad de clases MaxEnt.....	223
Figura 9.5: Un ejemplo de curva ROC.....	225

Tabla 3.1: Matriz de confusión.....	86
Tabla 3.2: Ejemplo de una matriz de confusión	88
Tabla 4.1: Muestra de documentos	101
Tabla 4.2: Representación vectorial TF normalizada	103
Tabla 4.3: Correlación entre vectores de documentos: (a) modelo TF versus y (b) modelo TFxIDF	105
Tabla 6.1: Matriz inicial de términos por documentos	137