Índice general

1. Introducción **2**

1.1. Descripción del problema 4

1.2. Objetivos 5

1.2.1. Objetivo general 5

1.2.2. Objetivos específicos 5

1.3. Alcances 6

1.4. Limitaciones 7

1.5. Organización del documento 9

2. Marco teórico **11**

2.1. Formulaicidad lingüística 11

2.1.1. Teoría formulaica 12

2.1.2. Terminología 13

2.1.3. Dominios formulaicos 16

2.1.4. Detección de formulaicidad 16

2.2. Lingüística computacional 20

2.3. Recuperación de información 21

2.4. Minería de textos 21

2.4.1. Categorización 22

2.4.2. Sumarización 22
2.4.3. Agrupamiento (*Clustering*) 23
2.4.4. Reglas de asociación 23
2.5. Algoritmos 23
2.5.1. Algoritmo de reducción a raíces léxicas (*stemming*) 23
2.5.2. Algoritmo de colocación 24
2.5.3. *N*-gramas 24
2.5.4. Algoritmo de distancias 24
2.5.5. Vectorización 25
2.5.6. Redes neuronales 25
2.5.7. Indizado 26
2.5.8. Método bayesiano 26
2.5.9. SUBDUE (Substructure discovery system) 26
2.6. Representación 28
2.6.1. Grafos 28
2.6.2. Estructuras *trie* 28
2.6.3. Estructuras *two-trie* 28
2.6.4. Redes semánticas 28
2.6.5. Vectores 29
2.6.6. Árboles gramaticales 29
2.7. Trabajos relacionados 29
2.7.1. Proyectos en lingüística y lingüística computacional 29
2.7.2. La Web semántica 31
2.7.3. Minería de textos 33

3. Metodología **35**

3.1. Fase de entrenamiento 37

3.1.1. Entradas 37

3.1.2. Salidas 39

3.1.3. Los experimentos 40

3.2. Fase de evaluación 43

3.2.1. Entradas de clasificación 43

3.2.2. Salidas de clasificación 44

3.2.3. Experimentos de clasificación 44

3.2.4. Entradas de análisis 47

3.2.5. Salidas de análisis 47

3.2.6. Experimentos de análisis 48

3.3. Refinamiento de fórmulas 49

3.3.1. Entradas 49

3.3.2. Salidas 49

3.3.3. Experimentos 50

3.4. Fase de pruebas 53

3.4.1. Entradas de clasificación 53

3.4.2. Salidas de clasificación 54

3.4.3. Experimentos de clasificación 55

3.4.4. Entradas de análisis 55

3.4.5. Salida de análisis 55

3.4.6. Experimentos 56

4. Diseño **58**

4.1. Arquitectura del sistema 59

4.2. Preprocesamiento 60
4.2.1. Eliminar palabras vacías 60
4.2.2. Aplicar reducción 62
4.2.3. Puntuación 62
4.2.4. *Parser* 65
4.3. Módulo de entrenamiento: procesamiento 65
4.3.1. SUBDUE 66
4.3.2. Algoritmo AD-HOC 70
4.3.3. Elección del algoritmo 72
4.4. Módulo de evaluación 74
4.4.1. Clasificación 75
4.4.2. Análisis 76
4.5. Módulo de pruebas 77
4.6. Modelo de datos 77

5. Implementación **80**
5.1. Consideraciones 80
5.1.1. Algoritmo de Porter 81
5.2. Modelo de datos 81
5.3. Módulos del sistema 82
5.3.1. Módulo de limpieza 82
5.3.2. Módulo de reconocimiento de patrones 82
5.3.3. Módulo de clasificación 84
5.3.4. Módulo de análisis 86
5.3.5. Módulo de interfaces 86
5.4. Funciones principales del sistema 89

5.4.1. Parser 90
5.4.2. Sustituir abreviaturas 91
5.4.3. Eliminar palabras vacías 92
5.4.4. Obtener relaciones 94
5.4.5. Construir fórmulas 96
5.4.6. Calcular índice de similitud 97
5.4.7. Elegir dominio 97
5.4.8. Calcular porcentaje de éxito 98
5.4.9. Iniciar el sistema 98
5.5. Características de la implementación 100
5.5.1. Cohesión 100
5.5.2. Acoplamiento 100
5.5.3. Modularidad 100
5.5.4. Extensibilidad 100

6. Pruebas y resultados **103**

6.1. Recolección de las muestras 103
6.1.1. Córpora de entrenamiento 103
6.1.2. Corpus de prueba 104
6.2. Fase de entrenamiento 105
6.2.1. Preprocesamiento 105
6.2.2. Lista de fórmulas 107
6.3. Fase de evaluación 107
6.3.1. *Clustering* 107
6.3.2. Análisis 108
6.4. Fase de pruebas 108

6.4.1. *Clustering* 108
6.4.2. Análisis 109
6.5. Dominios 109
6.6. Complejidad del algoritmo 110

7. Trabajo a futuro **113**
7.1. Detección de formulaicidad 113
7.2. Tamaño de las fórmulas 114
7.3. Extensión a otros idiomas 114
7.4. Definición de términos 115
7.5. Más intentos con SUBDUE 115
7.6. Pruebas con otros córpora 116
7.7. Comparaciones con otros trabajos 116

8. Conclusiones **118**

A. Diccionario de datos **121**

B. Notas de instalación **123**
B.1. Herramientas requeridas 123
B.2. Descargar 123
B.3. Ejecución 123

C. Diagramas de clases **125**

D. Corpora **131**
D.1. Corpus de entrenamiento: cartas de amor 131
D.2. Corpus de entrenamiento: cartas empresariales 140
D.3. Corpus de pruebas 144

E. Preprocesamiento **148**

F. Fórmulas **154**

F.1. Cartas de amor 154

F.2. Cartas empresariales 154

G. Resultados **155**

G.1. Evaluación 155

G.2. Pruebas 155

Bibliografía **158**