

Tabla de Contenido

1. Introducción	1
1.1. Web Intelligence Centre	2
1.2. Descripción y justificación	3
1.3. Hipótesis de investigación	5
1.4. Objetivos	5
1.4.1. Objetivo General	5
1.4.2. Objetivos Específicos	5
1.5. Metodología	6
1.6. Resultados esperados	7
1.7. Estructura de la memoria	7
2. Marco teórico y conceptual	9
2.1. World Wide Web	9
2.2. Medios sociales	10
2.2.1. Twitter	10
2.3. Application Programming Interfaces (API)	11
2.3.1. APIs de <i>Twitter</i>	12
2.4. Extracción de datos	13
2.4.1. <i>Crawler</i>	13
2.5. <i>Ciencia de datos</i>	14
2.5.1. Minería de datos	14
2.5.2. Minería de texto	14
2.5.3. <i>Machine learning</i>	15
2.6. Algoritmos de clasificación	16
2.6.1. Naïve Bayes	16
2.6.2. Support Vector Machine	17
2.6.3. Random Forest	19
2.7. Procesamiento de texto	19
2.7.1. Matriz Tf-Idf	19
2.8. Evaluación de resultados de modelos de clasificación	20
2.8.1. Matriz de confusión	20
2.8.2. Métricas de desempeño	21
2.8.3. Error de entrenamiento	22
2.9. Predicción de ubicación geográfica en Twitter	22
2.9.1. Text mining para determinar <i>Home Location</i>	23
2.9.2. Métricas de evaluación de desempeño para geolocalización en <i>Twitter</i>	24

2.10. Datos personales	24
3. Construcción set de datos	26
3.1. Definición de modelos	26
3.1.1. Modelo de clasificación: País	26
3.1.2. Modelo de clasificación: Región	27
3.2. Obtención de datos	27
3.2.1. Recopilación de usuarios	27
3.2.2. Recopilación de los <i>tweets</i> de los usuarios obtenidos	29
3.2.3. Etiquetado	30
3.3. Selección bases de entrenamiento	35
4. Modelamiento	40
4.1. Recursos utilizados	40
4.2. Pre-procesamiento de datos	40
4.2.1. Atributos de texto	42
4.2.2. Atributos categóricos	44
4.3. Modelo país	45
4.3.1. Modelos entrenados y sus variaciones	47
4.4. Modelo región	51
4.4.1. Modelos entrenados y sus variaciones	52
5. Análisis de resultados	56
5.1. Evaluación de desempeño de modelo país	56
5.1.1. Análisis MNB	56
5.1.2. Análisis SGD	57
5.1.3. Análisis RF	58
5.1.4. Análisis General	59
5.2. Evaluación de desempeño heurística actual de clasificación de usuarios chilenos	61
5.2.1. Descripción de heurística de clasificación	61
5.3. Evaluación de desempeño de modelo región	63
5.3.1. Análisis MNB	63
5.3.2. Análisis SGD	64
5.3.3. Análisis RF	64
5.3.4. Análisis General	64
6. Conclusiones	68
6.1. Conclusiones	68
6.2. Trabajo futuro	70
Bibliografía	73
Anexos	76
A.	76
B.	77

