

Tabla de contenido

1. Introducción	1
1.1. Antecedentes	1
1.1.1. WIC	2
1.1.2. OpinionZoom	2
1.2. Justificación	4
1.3. Objetivos	5
1.3.1. Objetivo general	5
1.3.2. Objetivos específicos	5
1.4. Hipótesis de investigación	6
1.5. Resultados esperados	6
1.6. Alcances	6
1.7. Metodología	7
1.8. Estructura del informe	7
2. Marco Conceptual	9
2.1. La Web	9
2.1.1. Web 2.0	9
2.2. Redes sociales	11
2.2.1. Twitter	11
2.3. KDD y Minería de datos	12
2.3.1. Descubrimiento de Conocimiento en bases de datos	12
2.3.2. Minería de Datos	14
2.4. Minería de Texto	15
2.4.1. Preprocesamiento de texto	15
2.4.2. Ponderación de términos	16
2.4.3. Preprocesamiento lingüístico	17
2.5. El usuario	18
2.6. User Modeling	18
2.6.1. Recolección de información de usuarios	19
2.6.2. Dinamismo del perfil de usuario	20
2.6.3. Representaciones de perfiles de usuario	21
2.6.4. Construcción de perfiles de usuario	24
2.6.5. Método Who-Likes-What	26
2.7. APIs	28
2.8. Métricas de evaluación	30
3. Diseño del sistema	31

3.1.	Arquitectura general	31
3.2.	Módulo de Extracción de Datos de Twitter	33
3.3.	Módulo de Procesamiento de Listas	34
3.3.1.	Módulo de Procesamiento de Texto	34
3.4.	Módulo de Identificación de Tópicos de Influencia	38
3.5.	Módulo de Identificación de Tópicos de Interés	39
3.5.1.	Agregador	39
3.5.2.	Filtro	40
3.5.3.	TF-IDF	40
3.6.	API	41
3.7.	Módulo de Visualización	42
4.	Desarrollo del sistema	45
4.1.	Herramientas tecnológicas	45
4.1.1.	Java	45
4.1.2.	PHP	46
4.1.3.	MariaDB	46
4.1.4.	REST API de Twitter	46
4.1.5.	Twitter4J	47
4.1.6.	Wordcloud2.js	47
4.1.7.	Freeling	47
4.1.8.	Netbeans	48
4.1.9.	Maven	48
4.1.10.	Spring	49
4.2.	Extracción de Datos de Twitter	49
4.2.1.	Obtención de amigos	49
4.2.2.	Obtención de Listas	51
4.3.	Procesamiento de listas	53
4.3.1.	Procesamiento de texto	54
4.4.	Identificación de Tópicos de Influencia	57
4.5.	Identificación de Tópicos de Interés	59
4.6.	API	60
4.6.1.	Agregar Usuario	60
4.6.2.	Obtener Tópicos de Interés	62
4.7.	Visualización	62
5.	Resultados	64
5.1.	Validación	64
5.1.1.	Metodología	64
5.1.2.	Resultados	65
5.2.	Discusión	70
5.2.1.	Comparación con Klout	70
5.2.2.	Discusión General	73
6.	Conclusiones	75
6.1.	Conclusiones generales	75
6.2.	Trabajo futuro	76

Bibliografía	79
A. Encuesta de validación	84
B. Listado de tópicos	85

Índice de tablas

3.1. Lista Negra	41
4.1. Límites REST API de Twitter	47
4.2. Expresiones regulares utilizadas en tokenización previa	56
4.3. Resultado Consulta 4.11	58
5.1. Probabilidad de identificación de tópicos de interés	68
5.2. Probabilidad de importancia relativa de tópicos de interés	68
5.3. Tiempos de ejecución del sistema	69
5.4. Ejemplo resultados de Klout	70
5.5. Tópicos con más frecuencia de Klout	72
5.6. Tópicos más frecuentes del nuevo sistema	72
B.1. Frecuencia de tópicos de Klout	85
B.2. Resultados Klout 34 usuarios	86
B.3. Frecuencia de tópicos de nuevo sistema - Primera parte	87
B.4. Frecuencia de tópicos de nuevo sistema - Segunda parte	88

Índice de figuras

2.1. Proceso KDD	13
2.2. Perfil de usuario basado en keywords	22
2.3. Perfil de usuario basado en redes semánticas	23
2.4. Perfil de usuario basado en conceptos	24
2.5. Creación de perfil de usuario en OBIWAN	27
2.6. Resultado sistema <i>who-is-who</i> para el usuario @BarackObama	28
2.7. Resultado sistema <i>Who Likes What</i>	29
2.8. 2 aplicaciones comunicadas por medio de una API	29
3.1. Arquitectura General	32
3.2. Arquitectura de Módulo de Extracción de Datos de Twitter	34
3.3. Módulo de Procesamiento de Listas	35
3.4. Etapas de procesamiento de texto	36
3.5. Etapas del Módulo de Identificación de Influencia	39
3.6. Etapas del Módulo de Identificación de Tópicos de Interés	40
3.7. Módulo de Visualización	42
3.8. Landing page de Módulo de Visualización	43
3.9. Página de resultado de Módulo de Visualización	44
4.1. Cola de credenciales de Twitter	50
4.2. JSON de resultado	61
5.1. Nubes de palabras de usuarios encuestados	66
5.2. Representatividad de los tópicos de interés	67
5.3. Histograma de tópicos correctamente identificados	67
5.4. Representatividad del tamaño de los tópicos de interés	69
5.5. Frecuencia de tópicos en 34 usuarios de Klout	71
5.6. Frecuencia de tópicos en 34 usuarios del nuevo sistema	71