

Tabla de Contenido

1	Introducción	1
1.1	Objetivo general	4
1.2	Objetivos específicos.....	4
2	Antecedentes	5
2.1	Agregadores de noticias actuales y sus limitaciones	5
2.1.1	Agregadores de Noticias.....	5
2.1.2	Galean	5
2.1.3	Knowhere News.....	6
2.1.4	Swipe News	6
2.1.5	Event Registry	6
2.1.6	Disponibilidad de contenido periodístico.....	7
2.2	Procesamiento del lenguaje natural	7
2.2.1	<i>Transformers</i>	9
2.2.2	Uso práctico.....	14
2.3	<i>Clustering</i>	14
2.3.1	<i>Embeddings</i> de texto	14
2.3.2	Métricas y algoritmos	15
2.3.3	Criterios de evaluación de resultados de <i>clustering</i>	16
2.4	Base de datos de texto	17
2.4.1	Tecnologías existentes	17
2.4.2	Escalabilidad.....	18
2.5	Arquitecturas de software	19
2.6	Oportunidad de desarrollo	20
3	Diseño de solución	21
3.1	Selección de medios de comunicación.....	21
3.2	Recopilación de noticias	22
3.2.1	Aspecto técnico	22
3.2.2	Aspecto legal.....	23
3.3	Sesgos periodísticos y otras funcionalidades	24
3.3.1	Polaridad	24
3.3.2	Subjetividad	25
3.3.3	Resumidor	26
3.3.4	Limitaciones y consideraciones	27
3.4	<i>Clustering</i>	29
3.4.1	Problema de fondo y limitaciones	29
3.4.2	Método con <i>embeddings</i>	29

3.4.3	Método con heurística	32
3.5	Base de datos	33
3.5.1	Modelo de datos	35
3.6	Buscador avanzado	39
3.7	Arquitectura del sistema	40
3.7.1	Componentes y microservicios	40
3.7.2	Acotamientos	43
3.8	Visualización	44
3.8.1	Metodología	44
3.8.2	<i>Mockups</i>	45
3.8.3	Comentarios y observaciones.....	50
4	Implementación	51
4.1	Metodologías de desarrollo	51
4.2	Selección de medios	51
4.3	<i>Scrappers</i>	54
4.3.1	Extracción de <i>urls</i> desde Twitter	54
4.3.2	Portales digitales	54
4.4	Procesamiento del lenguaje natural	57
4.4.1	Clases y componentes	57
4.4.2	<i>Machine translation</i>	58
4.4.3	Polaridad	60
4.4.4	Subjetividad	62
4.4.5	Resumidor	63
4.4.6	<i>Preprocess service</i>	65
4.4.7	<i>Summary service</i>	65
4.5	<i>Clustering</i>	66
4.5.1	Método con <i>embeddings</i>	66
4.5.2	Método con heurística	84
4.5.3	<i>Clustering service</i>	90
4.6	Base de datos	91
4.6.1	Clase <i>QueryMaker</i>	91
4.6.2	<i>Query service</i>	93
4.7	Buscador avanzado	93
4.7.1	Método principal	93
4.7.2	<i>Search service</i>	94
4.8	Visualización	95
4.8.1	Herramientas y metodología	95
4.9	Automatización de procesos.....	103
4.9.1	<i>Control service</i>	103
4.9.2	Otros procesos	103
5	Evaluación y validación	104
5.1	Evaluación de funcionalidades	104
5.1.1	Buscador	105
5.1.2	Agrupador de noticias	111
5.1.3	Análisis de polaridad y subjetividad de titulares generados	132

5.2	Validación con usuarios del <i>Proof of Concept</i>	133
5.2.1	Metodología	133
5.2.2	Resultados	135
6	Conclusiones	137
6.1	Conclusiones generales	137
6.2	Trabajo futuro	139
	Bibliografía	142
	Anexos	150
A	Software similar	151
A.1	Event Registry	151
A.2	Google News	152
B	Mappings para Elasticsearch	153
B.1	Mapping de artículos	153
B.2	Mapping de eventos.....	157
B.3	Mapping de agrupaciones	161
C	Experimentos previos	162
C.1	Experimentación previa con heurística	162
C.2	Experimentación con modelos de traducción	164
C.2.1	Experimento con texto sobre inauguración de Convención Constitucional .	164
C.2.2	Experimento de noticia de CNN sobre inauguración de Convención Cons- titucional	167
C.3	Experimentación con resumidores	170
C.3.1	Generación resúmenes por evento de inauguración de Convención Cons- titucional	170
D	Resultados de agrupación del 23 al 25 de febrero de 2022 mediante heurística	172
D.1	Eventos sobre el conflicto ruso-ucraniano	172
E	Entrevistas a usuarios	175
E.1	Perfiles de los entrevistados	175
E.2	Entrevistas de la 1 a la 10	177
E.3	Entrevistas de la 11 a la 20	181
E.4	Entrevistas de la 21 a la 24	183