**Especificación de Requisitos de Software (SRS) Sistema de Web Scraping para X (antes Twitter) RascaCazuelas S.A.**

### Administración de Proyectos I

**Profesora:** Alma Delia De Jesús Islao

**Ciclo escolar:** 2025-I

### Elaborado por:

Leonardo Valdes Palafox - 18-011-0249

Andrea Marlene Ortega Almendares - 17-003-1026

**Fecha de entrega:** 01 de marzo de 2025

**Versión:** 1.0

# Historial de Revisiones

|  |  |  |  |
| --- | --- | --- | --- |
| **Fecha** | **Versión** | **Descripción** | **Autor** |
| 01/03/2025 | 1.0 | Versión inicial del documento | Leonardo Valdes, Andrea Ortega |
| C C | | | |

**Tabla de Contenidos**

1. Introducción
   1. Propósito
   2. Alcance
   3. Definiciones, Acrónimos y Abreviaturas
   4. Referencias
   5. Visión General del Documento
2. Descripción General
   1. Perspectiva del Producto
   2. Funciones del Producto
   3. Características de los Usuarios
   4. Restricciones
   5. Suposiciones y Dependencias
3. Requisitos Específicos
   1. Requisitos Funcionales
   2. Requisitos No Funcionales
4. Apéndices

# Introducción

## Propósito

Este documento especifica los requisitos funcionales y no funcionales para el desarrollo de una herramienta de web scraping destinada a extraer información de la red social X (anteriormente Twitter). El documento está dirigido al equipo de desarrollo, stakeholders y usuarios finales de RascaCazuelas S.A.

## Alcance

El sistema de web scraping, denominado "XScraper", permitirá automatizar la extracción de datos públicos de X (antes Twitter), procesar esta información y almacenarla en una base de datos. El sistema no incluirá funcionalidades para extraer datos de otras redes sociales ni accederá a información privada o protegida.

## Definiciones, Acrónimos y Abreviaturas

 **API**: Application Programming Interface (Interfaz de Programación de Aplicaciones)

 **BD**: Base de Datos

 **GUI**: Graphical User Interface (Interfaz Gráfica de Usuario)

 **HTTPS**: Hypertext Transfer Protocol Secure (Protocolo Seguro de Transferencia de Hipertexto)

 **RF**: Requisito Funcional

 **RNF**: Requisito No Funcional

 **SRS**: Software Requirements Specification (Especificación de Requisitos de Software)

 **Web Scraping**: Técnica de extracción automatizada de información de sitios web

## Referencias

 Documento de Visión y Alcance, RascaCazuelas S.A., Versión 1.0, 17/02/2025

 IEEE Std 830-1998, IEEE Recommended Practice for Software Requirements Specifications  Términos de Servicio de X (Twitter): <https://twitter.com/en/tos>

 Laínez Fuentes, J. R. (2024). Gestión de proyectos de software: Metodologías, herramientas y casos prácticos (2ª ed.). RA-MA Editorial.

## Visión General del Documento

Este documento está organizado en tres secciones principales. La primera sección proporciona una introducción al documento y al proyecto. La segunda sección ofrece una descripción general del sistema, incluyendo su contexto y funcionalidades principales. La tercera sección detalla los requisitos específicos, tanto funcionales como no funcionales, que el sistema debe cumplir.

# Descripción General

## Perspectiva del Producto

XScraper será un componente nuevo dentro de la infraestructura de análisis de datos de RascaCazuelas

S.A. El sistema interactuará con la red social X (antes Twitter) para extraer datos y con la base de datos corporativa para almacenar la información procesada. Los analistas de datos accederán a esta información a través de las herramientas de análisis existentes en la empresa.

## Funciones del Producto

Las principales funciones del sistema incluyen:

* + 1. Configuración de parámetros y reglas de extracción para X (antes Twitter)
    2. Programación de tareas de extracción automática
    3. Extracción y recopilación de datos de perfiles y tweets públicos
    4. Validación y limpieza de datos recopilados
    5. Almacenamiento estructurado de datos en la base de datos corporativa
    6. Generación de registros de actividad (logs)

## Características de los Usuarios

|  |  |  |
| --- | --- | --- |
| **Tipo de Usuario** | **Nivel de**  **Experiencia** | **Responsabilidades** |
| Administrador del  Sistema | Técnico | Configuración inicial del sistema, gestión de parámetros y monitoreo  de operaciones |
| Analista de Datos | Medio | Definición de reglas de extracción, programación de tareas y revisión  de datos extraídos |
| Usuario Final | Básico | Consulta de datos procesados a través de la plataforma de análisis |
| C C | | |

* 1. **Restricciones**

 El sistema debe respetar los términos de servicio de X (antes Twitter) y las leyes de protección de datos

 El desarrollo debe completarse en un plazo máximo de 4 meses

 El sistema debe operar en equipos domésticos con recursos limitados

 La solución debe integrarse con la infraestructura existente de RascaCazuelas S.A.

 El sistema debe ser compatible con los navegadores web más utilizados (Chrome, Firefox, Edge, Brave)

## Suposiciones y Dependencias

 Se asume que los usuarios tendrán conexión a Internet estable

 Se asume que la estructura de la plataforma X (antes Twitter) permanecerá relativamente estable durante el desarrollo

 El sistema depende de las políticas de acceso a datos de X (antes Twitter), que pueden cambiar

 El sistema depende de la disponibilidad de la base de datos corporativa para el almacenamiento de información

# Requisitos Específicos

## Requisitos Funcionales

### RF-01: Configuración de Reglas de Extracción

 **Descripción**: El sistema debe permitir al usuario definir reglas de extracción para perfiles, hashtags y términos de búsqueda en X (antes Twitter).

 **Entradas**: Criterios de búsqueda, elementos a extraer, frecuencia de extracción.

 **Proceso**: El sistema almacena las reglas definidas por el usuario.

 **Salidas**: Confirmación de reglas guardadas.

 **Prioridad**: Alta

 **Dependencias**: Ninguna

### RF-02: Programación de Extracciones

 **Descripción**: El sistema debe permitir programar extracciones automáticas según horarios definidos.

 **Entradas**: Fecha y hora de inicio, periodicidad, duración máxima.

 **Proceso**: El sistema registra la programación y ejecuta las tareas en los momentos definidos.

 **Salidas**: Confirmación de programación y notificaciones de ejecución.

 **Prioridad**: Alta

 **Dependencias**: RF-01

### RF-03: Extracción de Datos

 **Descripción**: El sistema debe extraer datos de perfiles, tweets y tendencias públicas en X (antes Twitter) según las reglas configuradas.

 **Entradas**: Reglas de extracción.

 **Proceso**: El sistema navega por las páginas de Facebook y extrae la información solicitada.

 **Salidas**: Datos en formato estructurado.

 **Prioridad**: Alta

 **Dependencias**: RF-01, RF-02

### RF-04: Validación de Datos

 **Descripción**: El sistema debe validar y limpiar los datos extraídos para garantizar su calidad.

 **Entradas**: Datos extraídos sin procesar.

 **Proceso**: El sistema aplica reglas de validación y transformación.

 **Salidas**: Datos validados y estructurados.

 **Prioridad**: Alta

 **Dependencias**: RF-03

### RF-05: Almacenamiento de Datos

 **Descripción**: El sistema debe almacenar los datos procesados en la base de datos corporativa.

 **Entradas**: Datos validados y estructurados.

 **Proceso**: El sistema realiza operaciones de inserción/actualización en la base de datos.

 **Salidas**: Confirmación de almacenamiento exitoso.

 **Prioridad**: Media

 **Dependencias**: RF-04

### RF-06: Generación de Registros de Actividad

 **Descripción**: El sistema debe generar registros detallados de todas las operaciones realizadas.

 **Entradas**: Eventos del sistema.

 **Proceso**: El sistema registra cada actividad con timestamp y detalles relevantes.

 **Salidas**: Archivos de log.

 **Prioridad**: Media

 **Dependencias**: RF-01, RF-02, RF-03, RF-04, RF-05

### RF-07: Gestión de Errores

 **Descripción**: El sistema debe detectar y gestionar errores durante la extracción y procesamiento de datos.

 **Entradas**: Errores detectados durante la operación.

 **Proceso**: El sistema registra el error y ejecuta procedimientos de recuperación.

 **Salidas**: Notificaciones de error y registros.

 **Prioridad**: Media

 **Dependencias**: RF-03, RF-04, RF-05

### RF-08: Interfaz de Usuario

 **Descripción**: El sistema debe proporcionar una interfaz gráfica para la configuración y monitoreo.

 **Entradas**: Interacciones del usuario.

 **Proceso**: El sistema presenta opciones y recibe comandos del usuario.

 **Salidas**: Visualización de datos y opciones de configuración.

 **Prioridad**: Baja

 **Dependencias**: Ninguna

## Requisitos No Funcionales

### RNF-01: Rendimiento

 **Descripción**: El sistema debe realizar extracciones con un uso de recursos limitado.

 **Métrica**: El proceso de extracción no debe consumir más del 30% de CPU y 2GB de RAM.

 **Prioridad**: Alta

### RNF-02: Seguridad

 **Descripción**: El sistema debe implementar mecanismos de seguridad para proteger datos sensibles.

 **Métrica**: Uso de conexiones HTTPS, cifrado de credenciales y control de acceso basado en roles.

 **Prioridad**: Alta

### RNF-03: Usabilidad

 **Descripción**: La interfaz de usuario debe ser intuitiva para usuarios con conocimientos básicos.

 **Métrica**: Usuarios nuevos deben poder configurar una extracción básica en menos de 10 minutos.

 **Prioridad**: Media

### RNF-04: Fiabilidad

 **Descripción**: El sistema debe ser capaz de recuperarse automáticamente de errores comunes.

 **Métrica**: 95% de los errores deben ser gestionados sin intervención manual.

 **Prioridad**: Media

### RNF-05: Compatibilidad

 **Descripción**: El sistema debe ser compatible con diferentes navegadores y sistemas operativos.

 **Métrica**: Funcionamiento verificado en Chrome, Firefox, Edge y Brave en Windows 10/11 y macOS.

 **Prioridad**: Media

### RNF-06: Mantenibilidad

 **Descripción**: El código debe estar bien documentado y seguir estándares de desarrollo.

 **Métrica**: Cobertura de documentación del 90% y cumplimiento de estándares PEP 8 (Python).

 **Prioridad**: Media

### RNF-07: Escalabilidad

 **Descripción**: El sistema debe poder escalar para manejar un mayor volumen de extracción.

 **Métrica**: Capacidad para gestionar hasta 100 extracciones diarias sin degradación del rendimiento.

 **Prioridad**: Baja

### RNF-08: Disponibilidad

 **Descripción**: El sistema debe estar disponible durante las horas de operación establecidas.

 **Métrica**: Disponibilidad del 99% durante el horario laboral.

 **Prioridad**: Baja

# Apéndices

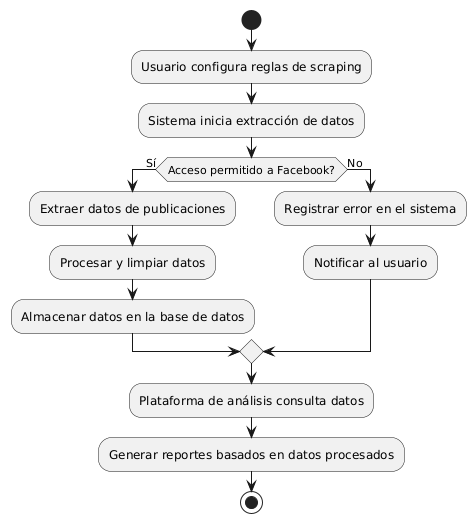
## Modelo de Datos Preliminar

El sistema utilizará el siguiente modelo de datos simplificado para almacenar la información extraída:

Entidades principales:

* Configuración (reglas, programación)
* Extracción (metadatos de cada extracción realizada)
* DatosX (datos extraídos)
* Logs (registros de actividad)

## Diagrama de Flujo de Proceso



## Glosario

**Web Scraping**: Técnica de extracción automatizada de datos de sitios web mediante la simulación de navegación humana o llamadas programáticas.

**Datos Estructurados**: Información organizada en un formato predefinido y consistente que facilita su análisis y procesamiento.

**API**: Conjunto de definiciones y protocolos que permite la comunicación entre componentes de software.