UNIVERSIDAD NACIONAL DEL ALTIPLANO

ESCUELA DE POSGRADO

MAESTRÍA EN INFORMÁTICA

TESIS

INDEXACIÓN DE SITIOS WEB PARA OPTIMIZAR LA BÚSQUEDA DE PAQUETES TURÍSTICOS DE LA REGIÓN DE PUNO BASADO EN WEB SCRAPING

PRESENTADA POR:

GERARDINO JUVENAL CAUNA HUANCA

PARA OPTAR EL GRADO ACADÉMICO DE:

MAGISTER SCIENTIAE EN INFORMÁTICA

MENCIÓN EN GERENCIA DE TECNOLOGÍAS DE INFORMACIÓN Y COMUNICACIONES.

PUNO, PERÚ

2021

DEDICATORIA

Con todo mi cariño e infinita gratitud a mis padres Juan pablo y Epifania quienes hicieron todo en la vida para que yo pudiera lograr mis sueños, por motivarme y darme la mano cuando sentía que el camino se terminaba, a ustedes por siempre mi corazón y mi agradecimiento.

A mi hijo Leonardo por qué se convirtió en la razón de cada acción que tomó en esta vida

Mi Pareja Celia, por su comprensión y aliento, por haber soportado a mi lado tantas vicisitudes y, haberme dado el mejor regalo que nadie me dio en la vida, mi hijo.

AGRADECIMIENTOS

A la Universidad Nacional del Altiplano, escuela de Postgrado, la Maestría en Informática, en la mención de Gerencia de Tecnologías de la Información y Comunicaciones por haberme formado y así poder alcanzar una más de mis metas.

A mis jurados, asesor y docentes de la escuela de Postgrado: Maestría en Informática, por sus estimadas sugerencias y por haberme transmitido sus conocimientos, y apoyarme en la culminación de mi trabajo de investigación.

ÍNDICE GENERAL

Pág.

DEDICATORIA i

AGRADECIMIENTOS ii

ÍNDICE GENERAL iii

ÍNDICE DE TABLAS vi

ÍNDICE DE FIGURAS vii

ÍNDICE DE ANEXOS ix

RESUMEN x

ABSTRACT xi

INTRODUCCIÓN 1

CAPÍTULO I

REVISIÓN DE LITERATURA

1.1 Marco teórico 3

1.1.1 Web Scraping 3

1.1.2 Indexación web 10

1.1.3 Calidad del producto de software 11

1.1.4 Calidad del producto software modelos y definiciones 13

1.1.5 Serie ISO/IEC 25000 14

1.1.6 Complejidad algorítmica 18

1.2 Antecedentes 21

CAPÍTULO II

PLANTEAMIENTO DEL PROBLEMA

2.1 Identificación del problema 27

2.2 Enunciados del problema 28

2.3 Justificación 28

2.4 Objetivos 29

2.4.1 Objetivo general 29

2.4.2 Objetivos específicos 29

2.5 Hipótesis 29

2.5.1 Hipótesis general 29

2.5.2 Hipótesis específicas 29

CAPÍTULO III

MATERIALES Y MÉTODOS

3.1 Lugar de estudio 30

3.2 Población 30

3.3 Muestra 30

3.4 Método de investigación 31

3.4.1 Tipo de investigación 31

3.4.2 Diseño de investigación 31

3.4.3 Método de tratamiento de datos 32

3.5 Descripción detallada de métodos por objetivos específicos 33

3.5.1 Metodología para desarrollar el algoritmo de extracción 33

3.5.2 Determinación del rendimiento del sitio web gopuno 35

3.5.3 Determinación del grado en que el producto de software satisface a un usuario final basado en la norma ISO/IEC 25000. 36

CAPÍTULO IV

RESULTADOS Y DISCUSIÓN

4.1 Resultado conforme al objetivo específico 1 37

4.1.1 Extracción de datos 39

4.1.2 Análisis del sistema 39

4.1.3 Diseño 41

4.1.4 Desarrollo del web Scraping 44

4.2 Resultados conforme al objetivo específico 2 56

4.3 Resultados conforme al objetivo específico 3 63

4.4 Hipótesis para prueba de los rangos con signo de Wilcoxon 66

CONCLUSIONES 70

RECOMENDACIONES 71

BIBLIOGRAFÍA 72

ANEXOS 79

Puno, 21 de julio de 2021

ÁREA: Informática
TEMA: Indexación de sitios web
LÍNEA: Tecnologías de Información

ÍNDICE DE TABLAS

Pág.

1. Especificación formal de las métricas de calidad en uso 17
2. Diseño pre prueba y post prueba con un solo grupo 32
3. Comparación de tiempo y los datos extraídos 50
4. Puntuación FCP 59
5. Puntuación del Speed index 60
6. Puntuación Largest Contentful Paint 60
7. Puntuación Time to Interactive 61
8. Puntuación Total Blocking Time 61
9. Puntuación de rendimiento 63
10. Métricas de calidad de uso 63
11. Ponderación de las características de calidad en uso 64
12. Niveles de Puntuación final 65
13. Valor total obtenido de la calidad en uso 66
14. Búsqueda de paquetes (en minutos) turísticos antes y después 67
15. Prueba de la normalidad 67
16. Prueba de rangos Wilcoxon 68

ÍNDICE DE FIGURAS

Pág.

1. Representación de un contenedor 4
2. Arquitectura de un sistema de extracción de datos web 8
3. Estructura usada por el modelo de calidad 11
4. Calidad en el ciclo de vida 12
5. Modelo de ciclo de vida de calidad del sistema / software 13
6. Modelo para calidad interna y externa del producto software 14
7. Modelo para calidad en uso del producto software. 14
8. Tabla de complejidad Big-O 18
9. Complejidad del tiempo 20
10. Diagrama de flujo para la implementación del algoritmo 34
11. Distribución de paquetes turísticos en la web 37
12. Información relevante a extraer 38
13. Código fuente de la portada principal 38
14. Código fuente del paquete turístico 39
15. Diagrama de caso de uso de consultar paquete turístico 41
16. Diagrama de casos de uso para la consulta de paquetes 42
17. Diagrama de secuencia para la extracción de datos 42
18. Diagrama de secuencia para la transformación de datos 43
19. Diagrama de secuencia para la exportación de datos 43
20. Diagrama de secuencia para la búsqueda de paquetes 44
21. Estructura de las páginas web a scrapear 45
22. Código para realzar solicitudes y manipular las etiquetas HTML 46
23. Código para extraer hipervínculos de los paquetes turísticos 47
24. Acopio de la información relevante del paquete turístico 47
25. Código para la extracción de los paquetes turísticos 48
26. Datos extraídos de las agencias de viaje 49
27. Extensión de Google Chrome denominado Webscraper 49
28. Extracción de datos con wenscraper.io 50
29. Complejidad algorítmica para la obtención de hipervínculos 51
30. Complejidad algorítmica para la extracción de los datos 52
31. Limpieza y transformación de datos 54

32. CSV con datos transformados 54
33. Configuración básica para exportar a SQLite 55
34. Configuración de las columnas de la base de datos 56
35. Ventana principal de sitio web 56
36. Paquetes turísticos extraídos 57
37. Descripción detallada del paquete turístico 58
38. Directorio de agencias 58
39. Puntuación global del rendimiento para ordenadores 62
40. Puntuación global del rendimiento para dispositivos móviles 62
41. Matriz de calidad en uso 66

ÍNDICE DE ANEXOS

Pág.

1. Archivo Robot.txt de las páginas a Scrapear 79

2. Encuesta de satisfacción 81

3. Ficha de validación 82

4. Directorio de agencias IPERI 2019 85

5. Matriz de consistencia 87

6. Archivo main.py 89

7. Archivo page.py 92

RESUMEN

La técnica del Web Scraping permite la extracción de contenido de varios sitios web, recabando información de interés para el usuario; a fin de ser presentada de forma ordenada y estructurada para su posterior utilización. El presente proyecto tiene como finalidad desarrollar un sitio web en la cual pueda almacenar información de los diferentes paquetes turísticos que son ofertados por las agencias de viaje que operan en la región de Puno utilizando la técnica del web Scraping. La población está conformada por 38 páginas web según inscritas en IPERÚ Puno. Para la elaboración del algoritmo de extracción se utilizó la metodología de desarrollo de software XP y para el contaste de la hipótesis se utilizó prueba de rangos con signo de Wilcoxon. Como resultado, el análisis de la estructura DOM permitió el desarrollo del algoritmo de extracción, haciendo uso de Python como lenguaje de programación, también se puso a prueba la eficiencia del algoritmo, el cual demostró ser eficiente en comparación con la el programa webscraper. Se determinó que la complejidad algorítmica es lineal O(n). Del desempeño de nuestro sitio web según la puntuación global de PageSpeed Insights está en la categoría rápida (97 puntos). La evaluación del sitio web basado en la norma ISO 25000 proporcionó una valoración de 6.96/10 puntos como calidad total, considerado como nivel aceptable y grado satisfactorio. Se concluye que la implementación del sitio web facilita la búsqueda de diferentes paquetes turísticos, reduciendo el tiempo empleado de forma significativa p-valor (0.015)< α(0.05)

Palabras clave: Paquetes, recuperación de información, sistema web, Scraping, turismo

ABSTRACT

The Web Scraping technique allows the extraction of content from various websites, collecting information of interest to the user; in order to be presented in an orderly and structured way for later use. The purpose of this project is to develop a website in which you can store information on the different tourist packages that are offered by travel agencies that operate in the Puno region using the web Scraping technique. The population is made up of 38 web pages as registered in IPERÚ Puno. For the elaboration of the extraction algorithm, the XP software development methodology was used and the Wilcoxon signed rank test was used to test the hypothesis. As a result, the analysis of the DOM structure allowed the development of the extraction algorithm, making use of Python as the programming language, the efficiency of the algorithm was also tested, which proved to be efficient compared to the webscraper program. The algorithmic complexity was determined to be linear O (n). The performance of our website according to the global PageSpeed Insights score is in the fast category (97 points). The evaluation of the website based on the ISO 25000 standard gave a rating of 6.96 / 10 points as total quality, considered as acceptable level and satisfactory grade. It is concluded that the implementation of the website facilitates the search for different tourist packages, reducing the time spent significantly p-value (0.015) <α (0.05)

Keywords: Information retrieval, packages, Scraping, tourism, web system

INTRODUCCIÓN

La sobreabundancia de información en Internet, es uno de los principales componentes de su éxito; sin embargo, el tratamiento de esta, exige una enorme cantidad de tiempo y energía a fin de seleccionar la calidad de los datos dentro de un enorme repositorio. Según Villarroel (2015) refiere que en la actualidad la sobrecarga de información que recibe un usuario, en especial de Internet en todas sus formas, puede causarle la sensación de no poder abarcarla ni gestionarla y, por tanto, llegar a generarle una gran angustia, para Toffler (1974) menciona que hay demasiados conocimientos para tomar una decisión o para mantenerse constantemente informado sobre algún asunto.

La información que se muestra en las páginas web tiene como fin ser entendida y procesada por personas, por ello, resulta muy difícil que una máquina sea capaz de entender un texto que no esté estructurado. Actualmente las páginas web cuentan con metadatos, semántica que describen el contenido y la relación entre los datos, de forma que es posible evaluarlas automáticamente por máquinas. Además, la web 3.0 intenta conseguir que los datos sean identificables dentro de la estructura de internet, es decir, que las búsquedas sean mucho más concretas y fiables. Este motivo es una de las principales razones por las que emerge el concepto de web scraping que se va a aplicar durante el trabajo investigación.

Web Scraping (raspado web) es el proceso de extracción de datos específicos de sitios web (Julian & Natalia, 2015), usando un programa que simula la exploración humana mediante el envío de peticiones HTTP (Muñoz *et al.*, 2018) o emulando un navegador web completo, además Web Scraping está muy relacionado con la indexación de la web, la cual indexa información utilizando un agente web automatizado y es una técnica global adoptada por la mayoría de los motores de búsqueda. En este sentido, cada vez que hemos realizado la acción de copiar y pegar por diferentes páginas de la web para obtener datos, y que posteriormente hemos utilizado para otra actividad diferente, lo que se ha realizado ha sido un raspado de datos en la web. Pero esta práctica resulta muy costosa al tener que emplear mucho tiempo en el caso de querer obtener muchos datos de diferentes páginas, luego organizar, estructurar, almacenar y analizar en una base de datos central, en una hoja de cálculo o en alguna otra fuente de almacenamiento (Hanretty, 2013).

Por ello, lo que pretende simular la técnica del web scraping, es realizar la acción de copiar y pegar de un modo más eficiente, mediante el uso de la programación. De modo

que, programando se pueda obtener y organizar la información residente en las diferentes páginas, en concreto, para nuestro caso de estudio, páginas web de agencias de viajes que operan en la región de Puno, dado que el turismo es uno de los sectores con enorme potencial de desarrollo, y cuenta con importantes recursos turísticos reconocidos y otros que recién están tomando auge, además de poseer una cultura tradicional y proveer una gran cantidad de posibilidades para los turistas que nos visitan. El sector del turismo es un ambiente muy dinámico, los productos (paquetes turísticos) cambian continuamente, como así también los intereses de los usuarios y precisamente estos, deben pasar bastante tiempo en el ordenador o su dispositivo móvil a fin de poder encontrar un paquete acorde a sus necesidades.

El trabajo de investigación se estructuro de la siguiente manera:

Capítulo I: Revisión literaria que hace referencia al estado que sostiene la investigación y los antecedentes.

Capítulo II: Se plantea el problema de la investigación

Capítulo III: Enfoca sobre los materiales y métodos empleada durante la investigación, el diseño, tipo, población, tamaño de la muestra y tratamiento estadístico.

Capítulo IV: Comprende la interpretación de los resultados con la finalidad de evaluar si los objetivos propuestos han sido alcanzados.

Finalmente se presentó las conclusiones y las recomendaciones de la investigación, seguida por la referencia

CAPÍTULO I

REVISIÓN DE LITERATURA

1.1 Marco teórico

1.1.1 Web Scraping

Para Julian & Natalia (2015) Web Scraping (raspado web o extracción de datos web) es el proceso de extracción de datos específicos de sitios web, usando un programa que simula la exploración humana mediante el envío de peticiones HTTP simples según Muñoz *et al.* (2018) o emulando un navegador web completo. Web Scraping, Tambien es conocido como Content Scraping, Screen Scraping, Web Harvesting o Web Data Extraction. En general, cualquier cosa que se puede ver en Internet, puede ser extraído y este proceso puede ser automatizado. Web Scraping está muy relacionado con la indexación de la web, la cual indexa información de la web utilizando un agente web automatizado y es una técnica global adoptada por la mayoría de los motores de búsqueda.

Mientras que Sanchez (2020) manifiesta que Web scraping es la encargada de obtener los datos mediante un procesamiento del código HTML que forma la página. En este sentido, cada vez que hemos realizado la acción de copiar y pegar por diferentes páginas de la web para obtener datos, y que posteriormente hemos utilizado para otra actividad diferente, lo que se ha realizado ha sido un raspado de datos en la web. Pero esta práctica puede resultar muy costosa al tener que emplear mucho tiempo en el caso de querer obtener muchos datos de diferentes páginas, y posteriormente, organizar, estructurar, almacenarlos y analizados en una base de datos central, en una hoja de cálculo o en alguna otra fuente de almacenamiento dichos datos para poder ser utilizados con otra finalidad (Hanretty, 2013).

Por ello, lo que pretende la técnica del web scraping, es realizar la acción de copiar y pegar de un modo más eficiente, mediante el uso de la programación. De modo que, programando se pueda obtener y organizar la información residente en las

diferentes páginas (Diab *et al.*, 2018). Existen diversos procedimientos para utilizar web scraping y mucho va depender del objetivo que se quiera lograr, para Zhao, (2017) indica que se puede dividir en dos pasos secuenciales; donde el primero consta de adquirir recursos web y luego extraer la información deseada de los datos adquiridos. Por otro lado Baumgartner *et al.(*2009), indica que el proceso de extracción de datos orientado a una página Web puede ser dividido en 5 pasos :

- Paso 1: Interacción con el sitio Web. El programa extractor interactúa con el sitio Web mediante una dirección URL.
- Paso 2: Soporte para la generación y ejecución del contenedor. El programa extractor necesita la estructura de la sección HTML como parámetro de entrada, dicha sección es conocida como contenedor o wrapper e incluye detalladamente la estructura de etiquetas que contiene la información a ser extraída.

Figura 1. Representación de un contenedor

La Figura 1 muestra un ejemplo de cómo puede estar estructurado un contenedor en una agencia de viajes, en ese caso se cuenta con una etiqueta <html> la cual hace el papel de contenedor principal, seguida de una etiqueta que contiene los datos de un destino turístico, dentro de ella se aprecian 3 etiquetas <img>, <p> y <label> las que contienen la imagen del destino, la descripción y el precio respectivamente.

- Paso 3: Planificación. Se refiere al proceso repetitivo de la extracción. Un programa informático puede ejecutarse permanentemente sobre un sitio

Web, intentando extraer más información o esperando alguna actualización de contenido.

- Paso 4: Transformación. Una vez obtenida la información, se procede al filtrado o refinamiento de los datos extraídos.
- Paso 5: Provisión. Finalmente, la información extraída es enviada a aplicaciones externas.

1.1.1.1 Uso del web Scraping

La tecnología y los procesos que lo permiten se han explorado ampliamente en el campo de la ciencia de datos, donde los investigadores aplican esta información a múltiples dominios de contenido (Landers *et al.*, 2016). Según lo descrito por Marres & Weltevrede (2013), en el periodismo el raspado web se ha utilizado para evaluar la importancia de las noticias internacionales contando la cantidad de veces que los usuarios de las redes sociales mencionaron esas historias. En marketing, se ha aplicado para recopilar datos de información del cliente a través de perfiles en línea y participación en el tablero de mensajes. En la investigación de políticas (Hernández *et al.*, 2015), se ha utilizado para medir el respaldo u oposición de los esfuerzos políticos a medida que cambian las percepciones públicas, en tiempo real, mientras que Medrano, (2020) lo utilizo para analizar la oferta de inmuebles a partir de avisos clasificados

Web Scraping se utiliza normalmente por diferentes motivos, como la detección de cambio, la investigación de mercado, seguimiento de datos y en algunos casos hasta el robo de datos.

1.1.1.2 Técnicas usadas para el web Scraping

Las técnicas para Web Scraping varían ampliamente en el esfuerzo y la complejidad. Estos métodos son principalmente técnicas especiales utilizadas para encontrar y aislar los elementos de datos dentro del código HTML de una página web, entre ellos, los selectores CSS y XPath son los métodos que se utilizan comúnmente en el proceso de rastreo (Rizaldi *et al.*, 2017).

Algunas de estas técnicas son las siguientes:

- Copiar y Pegar. Esta técnica es la solución más simple para extraer datos de un sitio web, los que hacen este trabajo pueden seleccionar las secciones de

interés de la página web, copiarlos usando atajo de teclado (Ctrl + C) o por medio de un ratón y pegue simultáneamente (Ctrl + V) todo el material necesario en las hojas de datos externos (hojas de cálculo, etc.).

- Expresiones regulares: son una secuencia de caracteres que forman un patrón de búsqueda. Normalmente es usado para la búsqueda o reconocimiento de cadenas de caracteres. De modo que, utilizando motores de búsqueda se puede conseguir la búsqueda de expresiones regulares en una página web, este método resulta sencillo y potente, aunque laborioso y lento. No se recomienda su uso para procesar formatos HTML.
- Protocolo HTTP: se obtienen páginas web mediante peticiones HTTP a un servidor remoto mediante sockets. De modo que, se consigue la página web completa con todos los datos que ésta contiene, y que posteriormente deberán ser almacenados y clasificados.
- Algoritmos de minería de datos: la minería de datos se encarga de extraer información de un conjunto de datos, transformándola en una estructura comprensible para su posterior uso.
- Muchos sitios web tienen páginas web similares, ya que para agilizar la creación hacen uso de plantillas o scripts. Entonces, la minería de datos se encarga de detectar la plantilla y extraer los datos.
- Parsers de HTML: mediante el uso de lenguajes de programación se procesan documentos HTML, recuperando y transformando el contenido. Algunos ejemplos de lenguajes que permiten realizar este procesamiento son Xquery o HTQL (HyperText Query Language).
- Reconocimiento de web semántica: algunas páginas contienen metadatos o información semántica, como anotaciones o comentarios. Esta 'metainformación' puede ser usada para extraer la información deseada. Las anotaciones pueden contenerse en la misma página, siendo de utilidad cuando procesamos el DOM (Document Object Model) del documento; o pueden estar en una capa semántica, que se encuentra almacenada de forma separada, pudiendo obtener estos esquemas antes de analizar los documentos.
- Aplicaciones para web scraping: existe una gran variedad de aplicaciones disponibles que permiten obtener información. Para conseguirlo, deben ser

capaces de conocer la estructura de una página, o permitir al usuario seleccionar los campos que son de su interés en un documento. Ésta es la solución más recomendable, ya que dispone de diferentes algoritmos que permiten la recolección de la información.

1.1.1.3 Herramientas para el web Scraping

Existen multitud de herramientas destinadas a la obtención de datos, cada herramienta se encuentra escrita en un lenguaje de programación, como puede ser Python, Ruby, PHP, entre otros.

Usuarios sin conocimientos de programación: las aplicaciones destinadas a este tipo de usuarios se caracterizan por tratarse de interfaces sencillas, que mediante diferentes clics el usuario puede obtener la información y en las cuales se evita tener que programar. También es cierto, que están bastante más limitadas en capacidad de obtener datos, que aquellas que están destinadas a usuarios con conocimientos de programación (Michael, 2014).

La empresa BBVA (2016) recomienda el lengua de programación Python para la extracción de datos no estructurados, debido a que los usuarios con conocimientos de programación le permitir programar los datos que desea obtener de las páginas web.

- BeautifulSoup: es una librería en Python que sirve para la extracción sencilla de datos concretos de una página web en HTML sin excesiva programación. Es lo que técnicamente recibe el nombre de parsear HTML. Una de las ventajas de esta biblioteca en Python es que todos los documentos salientes de la extracción de datos lo hacen en UTF-8, lo cual es bastante interesante porque el problema típico de las codificaciones queda totalmente resuelto.
- Python Mechanize: es un navegador virtual que consigue rastrear página web con lenguaje de programación Python. Está basado en el módulo urllib.
- Scrapy: es un marco de desarrollo de código abierto para la extracción de datos con Python. Este framework permite a los desarrolladores la programación de arañas que sirven para rastrear y extraer información concreta de una o varias páginas web a la vez. El mecanismo que utiliza

recibe el nombre de selectores, aunque también se pueden utilizar librerías en Python como BeautifulSoup o lxml.

En la figura 2 se observa la arquitectura típica de un sistema de extracción de información Web planteada por Baumgartner *et al.* (2009). Este sistema comprende varios componentes estrechamente conectados e interactúa con tres entidades externas: (1) la Web, que contiene páginas con información de interés; (2) una aplicación de destino, a la que finalmente se entregarán los datos extraídos y refinados; y (3) el usuario, que diseña interactivamente el envoltorio.

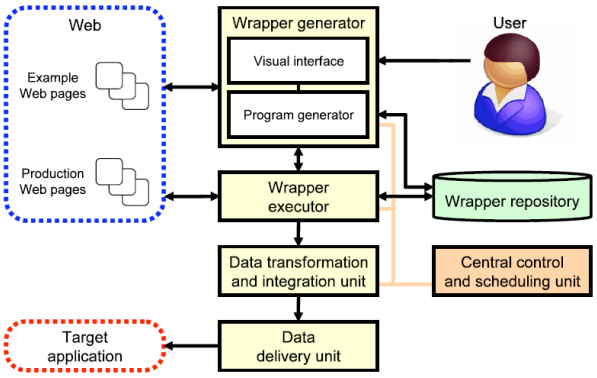


Figura 2. Arquitectura de un sistema de extracción de datos web

Fuente: (Baumgartner *et al.*, 2009)

1.1.1.4 Dificultades para realizar web Scraping

Las mayores limitaciones para hacer uso de la técnica Web Scraping se dividen en dos grupos, factores que dificultan el proceso de Web Scraping y barreras legales desarrolladas a través de los años (Hernández *et al.*, 2015).

a) Factores que dificultan el uso de aplicaciones para el web scraping

Parte de los factores que dificultan el proceso de Web Scraping son aquellos que hacen más confuso o evitan que se realice los métodos que se aplican, estos son:

- Códigos HTML mal estructurados, cuando el sitio web al que accedemos tiene un buen tiempo de vida, y por tal no se ha actualizado en su estructura, por ser tan sencilla puede ser no posible extraer información de su etiquetado HTML. Siendo en este caso que la información que adquiramos será confusa o con contenido de datos innecesario, o el caso

contrario. No hay manera de que la información sea obtenida vaya a ser precisa.

- Sistemas de autenticación, las más conocidas son captcha y paywalls, estos sistemas permiten diferenciar un humano de una máquina. Directamente al tener que pasar por una prueba de autenticación, es seguro determinar que será una situación inesperada para las aplicaciones que desean obtener datos.
- Bloqueo al acceso masivo, cuando se realiza una operación idéntica (en este caso una solicitud) el servidor reacciona de manera desfavorable, ya que la detecta como una intrusión, por lo cual nuestra dirección IP puede resultar perjudicada.
- Sistemas que usan cookies, lo cual realiza un seguimiento al usuario y determina que acciones realiza.

b) Barreras legales que limitan el Web Scraping

De acuerdo con las leyes de cada país, existen restricciones legales, las cuáles no brindan suficientes privilegios para el uso de datos o información. Considerando este hecho, podemos apreciar que pocas páginas están protegiendo su contenido, con términos legales, agregando usos comerciales permitidos y no permitidos en sus sitios web, ya sean con fines lucrativos o no. Así podemos decir que no existen leyes que exijan alguna legitimidad para el control de procesos de Web Scraping (Hernández, 2014).

Algunos casos judiciales que indican que el web sacaping es legal

En (*FEIST PUBLICATIONS, INC. V. TEL. RURAL SERVICIO CO. | FindLaw*, 1991), la Corte Suprema de los Estados Unidos decidió descartar y volver a publicar hechos, como listados telefónicos, permitidos. Un caso similar en Australia, (*Telstra Corporation Limited v Directorios telefónicos Company Pty Ltd [2010] FCA 44*, 2010), demostró que solo los datos con un autor identificable pueden tener derechos de autor.

Según Jarmul & Lawson (2017), afirma casos un caso de la Unión Europea en Dinamarca, ofir.dk vs home.dk, concluyó que el rastreo regular y los enlaces profundos son permisibles. También QVC v., como resultado, dictaminó que, a menos que el raspado resultara en daños a la propiedad privada, no podría

considerarse un daño intencional, a pesar de la actividad del rastreador que conduce a algunos problemas de estabilidad del sitio.

Según Hernández (2014), existen casos donde se ha presenciado el abuso en la cantidad de información que puede ser utilizada por terceros, como, por ejemplo: American Airlines, la cual emitió una queja contra FareChase, que se dedicaba a vender un software que permitía comparar las tarifas de los vuelos. El resultado fue que la sentencia dio la razón a American Airlines, impidiendo a la empresa FareChase vender el software si el sitio de American Airlines estaba incluido, al considerar que entraba en los servidores de la compañía sin permiso de ésta.

1.1.1.5 ¿Es legal el Web Scraping en el Perú?

No existe una legislación clara y precisa respecto el Web Scraping en sí, cabe mencionar que el congreso de la Republica implemento (*Proyectos de ley emitidos por el Congreso de la República del Perú*, s. f.), la cual es una aplicación web creada para mostrar forma ordenada y accesible los proyectos de ley presentados en el Congreso peruano, para recolectar la información que se muestra en el sitio web aplicaron Web Scraping.

También podemos observar que diversos sitios web que ofrecen bolsas de trabajo realizan esta práctica, recolectando información de los sitios Web confiables donde las empresas publican información sobre sus bolsas de trabajo.

1.1.2 Indexación web

La tarea principal de un buscador web es encontrar documentos que contengan los términos que se ingresan en la consulta de información por parte del usuario (Martínez, 2012). Dada una consulta, una opción es revisar secuencialmente la base de datos de documentos para encontrar aquellos que contienen dichos términos de búsqueda. Sin embargo, la técnica anterior se vuelve impráctica cuando se maneja una gran cantidad de documentos, como en la Web. La otra alternativa es la creación de índices para acelerar el proceso de búsqueda. También conocido como indizado o indexado, el cual es responsable de la extracción de contenido textual de las páginas almacenadas y de la construcción de un índice para facilitar el procesamiento de consultas de usuario.

1.1.3 Calidad del producto de software

Para Baldeón (2015), el software es el elemento más importante en las actividades cotidianas y, con frecuencia, su correcta operación es vital para el éxito del negocio y/o la seguridad de las personas; por lo tanto, el desarrollo de productos software de calidad es de suma importancia, mientras que para Valenciano, (2015) la calidad del producto software se puede interpretar como el grado en que dicho producto satisface los requisitos de sus usuarios aportando de esta manera un valor.

Una evaluación de la calidad del producto software es un factor clave para asegurar la calidad adecuada, ello se puede lograr al definir de manera apropiada un proceso para su evaluación, que debe considerar en su contenido la verificación de las características relevantes de la calidad del producto utilizando métricas validadas o de amplia aceptación. Para poder comprender la calidad del producto software, es necesario recurrir a un modelo de calidad. La norma ISO/IEC 25010, (2013) lo define como un conjunto de características y la relación entre las mismas, que conforman la base para la evaluación de calidad. La Figura 3 representa un modelo de calidad que clasifican la calidad del producto en características, que en algunos casos se subdividen en sub características. Esta descomposición jerárquica proporciona un desglose conveniente de la calidad del producto. Las propiedades medibles relacionadas con la calidad de un sistema se denominan propiedades de calidad, con medidas de calidad asociadas.

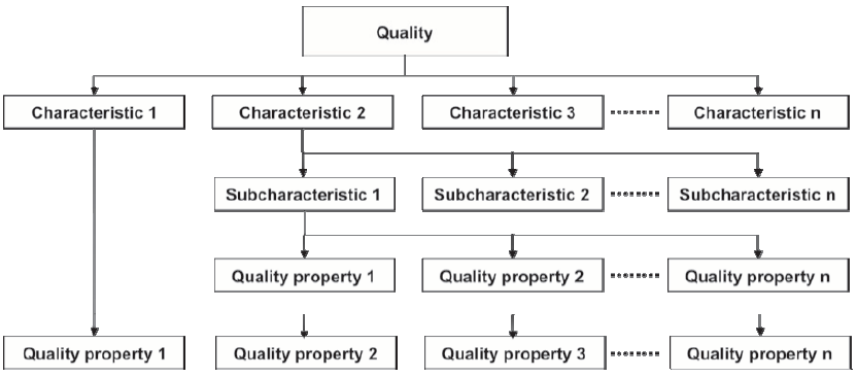


Figura 3. Estructura usada por el modelo de calidad

Fuente: (ISO/IEC 25010, 2013)

El modelo de calidad de la serie ISO/IEC 25000 presenta el concepto de calidad en uso, calidad externa y calidad interna. También señala que “la calidad del proceso

contribuye a mejorar la calidad del producto, y la calidad del producto contribuye a mejorar la calidad en uso. Por lo tanto, evaluar y mejorar un proceso es una manera de mejorar la calidad del producto, y evaluar y mejorar la calidad del producto es una manera de mejorar la calidad en uso. De igual manera, evaluar la calidad en uso puede proporcionar una retro alimentación para mejorar el producto, y evaluando un producto puede proporcionar una retroalimentación para mejorar un proceso"

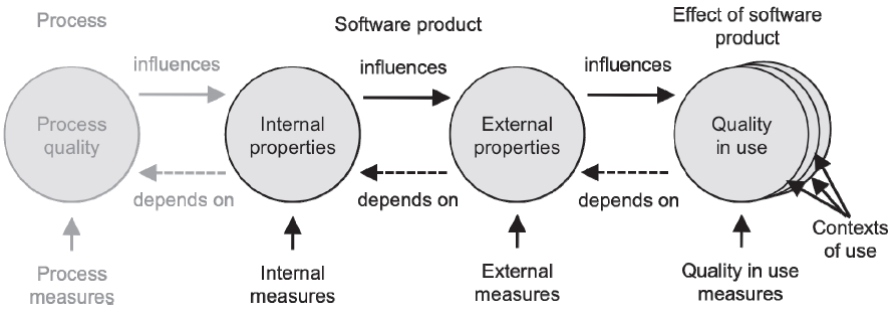


Figura 4. Calidad en el ciclo de vida

Fuente: (ISO/IEC 25010, 2013)

La figura 4 representa el ciclo de vida de la calidad que muestra la dependencia entre los distintos enfoques de calidad (interna, externa y en uso) y en la figura 5 se puede apreciar que las necesidades de calidad en uso contribuyen a especificar los requerimientos de calidad externa y estos a su vez los requerimientos de calidad interna. El cumplimiento de los requisitos de calidad interna se comprobará en un proceso de verificación que permitirá medirlo, el cumplimiento de los requisitos de calidad externa se comprobará en un proceso de validación que permitirá medirlo y finalmente la satisfacción de las necesidades de la calidad del producto se comprobarán en un proceso de evaluación que permitirá medir la calidad en uso.

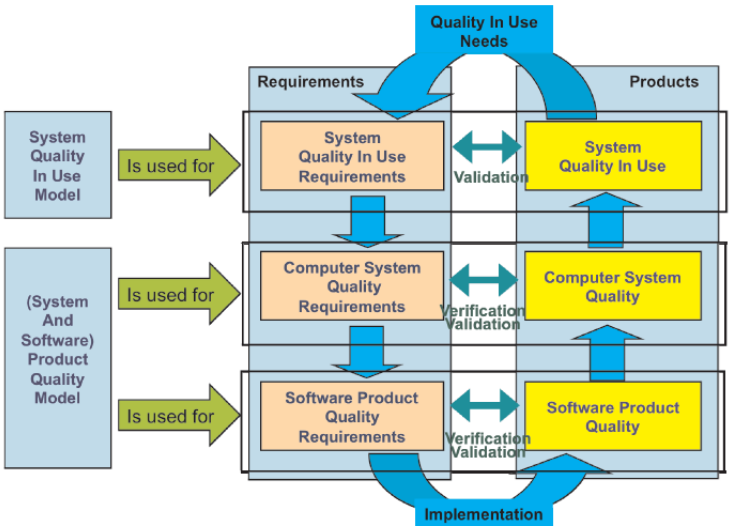


Figura 5. Modelo de ciclo de vida de calidad del sistema / software

Fuente: (ISO/IEC 25010, 2013)

1.1.4 Calidad del producto software modelos y definiciones

La serie ISO/IEC 25000 presenta dos modelos de calidad, la primera referida a la calidad interna y externa y el segundo modelo referido a la calidad en uso.

1.1.4.1 Calidad externa e interna

La norma ISO/IEC 25010 (2013) indica que la calidad interna es medida y evaluada en base a los requerimientos de calidad interna. Los detalles de la calidad del producto software pueden ser mejorados durante la implementación, revisión y prueba del código software, pero la naturaleza fundamental de la calidad del producto software representada por la calidad interna permanece sin cambios a menos que sea rediseñado, además la norma también define a la calidad externa como la totalidad de las características del producto software desde una perspectiva externa, cuando el software es ejecutado, también es medida y evaluada mientras se prueba en un ambiente simulado con datos simulados y usando métricas externas. Durante las pruebas, muchas fallas serán descubiertas y eliminadas. Sin embargo, algunas fallas todavía pueden permanecer después de las pruebas. La Figura 6 representa el modelo de calidad interna o externa, que muestra un conjunto de 8 características.

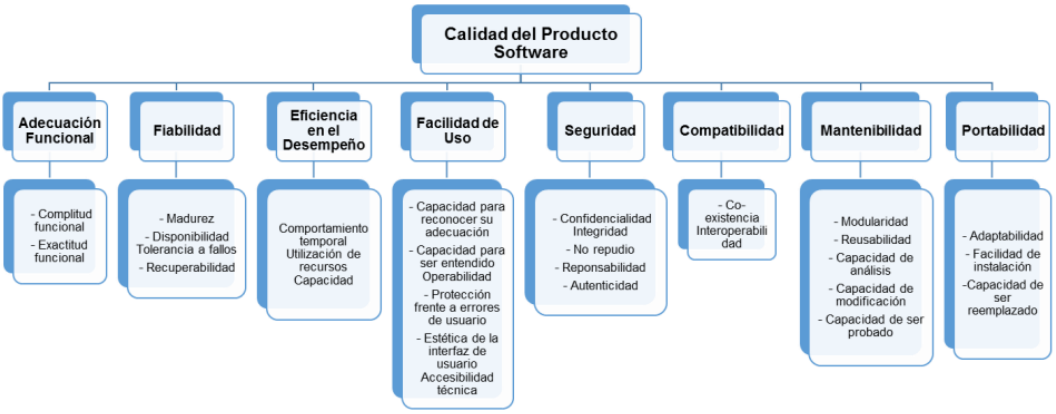


Figura 6. Modelo para calidad interna y externa del producto software

Fuente: (ISO/IEC 25010, 2013)

1.1.4.2 Calidad en uso

La norma ISO/IEC 25010 (2013) define la calidad en uso como la perspectiva del usuario de la calidad del producto software cuando este es usado en un ambiente específico y un contexto de uso específico. Esta mide la extensión para la cual los usuarios pueden conseguir sus metas en un ambiente particular, en vez de medir las propiedades del software en sí mismo. En la Figura 5 se presenta el modelo de calidad en uso que muestra un conjunto de cinco características.

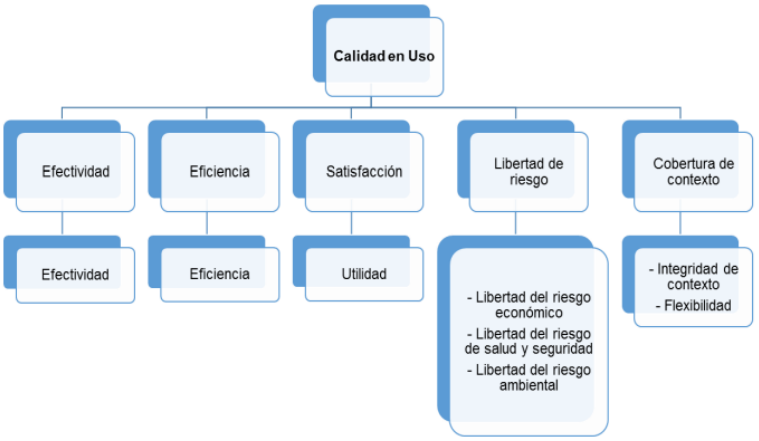


Figura 7. Modelo para calidad en uso del producto software.

Fuente: (ISO/IEC 25010, 2013)

1.1.5 Serie ISO/IEC 25000

Es una familia de normas que tiene por objetivo la creación de un marco de trabajo común para evaluar la calidad del producto software.

La serie ISO/IEC 25000 es el resultado de la evolución de otras normas anteriores, especialmente de la serie de normas ISO/IEC 9126, que describe las particularidades de un modelo de calidad del producto software, y la serie ISO/IEC 14598, que abordaba el proceso de evaluación de productos software. Esta familia de normas ISO/IEC 25000 se encuentra compuesta por cinco divisiones.

1.1.5.1 ISO/IEC 2500n - División de Gestión de Calidad

Las normas que forman este apartado definen todos los modelos, términos y definiciones comunes referenciados por todas las otras normas de la familia 25000. Actualmente, esta división se encuentra formada por:

- **ISO/IEC 25000 – Guía para SQuaRE.** contiene el modelo de la arquitectura de SQuaRE, la terminología de la familia, un resumen de las partes, los usuarios previstos y las partes asociadas, así como los modelos de referencia.
- **ISO/IEC 25001 – Planeamiento y Gestión.** Establece los requisitos y orientaciones para gestionar la evaluación y especificación de los requisitos del producto software.

1.1.5.2 ISO/IEC 2501n - División de Modelo de Calidad

Las normas de este apartado presentan modelos de calidad detallados incluyendo características para calidad interna, externa y en uso del producto software. Actualmente, esta división se encuentra formada por:

- **ISO/IEC 25010 (2010) –Modelos de Calidad de Software y Sistemas.** Describe el modelo de calidad para el producto software y para la calidad en uso.
- **ISO/IEC 25012 – Modelo de Calidad de Datos.** Define un modelo general para la calidad de los datos, aplicable a aquellos que se encuentran almacenados de manera estructurada y forman parte de un Sistema de Información.

1.1.5.3 ISO/IEC 2502n - División de Medición de Calidad

Estas normas incluyen un modelo de referencia de la medición de la calidad del producto, definiciones de medidas de calidad (interna, externa

y en uso) y guías prácticas para su aplicación. Actualmente esta división se encuentra formada por:

- **ISO/IEC 25020 - Guía y Modelo de Referencia para la medición.** Presenta una explicación introductoria y un modelo de referencia común a los elementos de medición de la calidad. También proporciona una guía para que los usuarios seleccionen o desarrollen y apliquen medidas propuestas por normas ISO.
- **ISO/IEC 25021 (2011) - Elementos de Medida de Calidad.** Define y especifica un conjunto recomendado de métricas base y derivadas que puedan ser usadas a lo largo de todo el ciclo de vida del desarrollo software.
- **ISO/IEC 25022 - Medición de Calidad en Uso.** Define específicamente las métricas para realizar la medición de la calidad en uso del producto.
- **ISO/IEC 25023 - Medición de la Calidad del Producto Software y Sistemas.** Define específicamente las métricas para realizar la medición de la calidad de productos y sistemas software.
- **ISO/IEC 25024 - Medición de la Calidad de Datos.** Define específicamente las métricas para realizar la medición de la calidad de datos.

1.1.5.4 ISO/IEC 2503n - División de Requisitos de Calidad

Las normas que forman este apartado ayudan a especificar requisitos de calidad que pueden ser utilizados como entrada del proceso de evaluación. Para ello, este apartado se compone de:

- **ISO/IEC 25030 – Requisitos de Calidad**. Provee de un conjunto de recomendaciones para realizar la especificación de los requisitos de calidad del producto software.

1.1.5.5 ISO/IEC 2504n - División de Evaluación de Calidad

Este apartado incluye normas que proporcionan requisitos, recomendaciones y guías para llevar a cabo el proceso de evaluación del producto software. Se encuentra formada por:

- **ISO/IEC 25040 - Guía y Modelo de Referencia de la Evaluación.** Propone un modelo de referencia general para la evaluación, que considera las entradas al proceso de evaluación, las restricciones y los recursos necesarios para obtener las correspondientes salidas.
- **ISO/IEC 25041 - Guía de Evaluación para Desarrolladores, Adquirentes y Evaluadores Independientes.** Describe los requisitos y recomendaciones para la implementación práctica de la evaluación del producto software desde el punto de vista de los desarrolladores, de los adquirentes y de los evaluadores independientes.
- **ISO/IEC 25042 - Módulos de Evaluación.** Define lo que la Norma considera un módulo de evaluación y la documentación, estructura y contenido que se debe utilizar a la hora de definir uno de estos módulos.
- **ISO/IEC 25045 – Módulo de Evaluación para Recuperabilidad**. Define un módulo para la evaluación de la sub característica Recuperabilidad.

Tabla 1

Especificación formal de las métricas de calidad en uso

Sub-característica	**Nombre de la métrica**	**Fase del ciclo de vida de calidad del producto**	**Propósito**	**Método de aplicación**	**Fórmula**	**Valor deseado**
Utilidad	Uso discrecional de las funciones	En Uso	¿Qué porcentaje de los usuarios optan por utilizar las funciones del sistema?	Observación de manejo	X=A/B Donde B>0	0<=X<=1 Cuanto más cercano a 1, mejor.

Fuente (ISO/IEC 25022)

1.1.6 Complejidad algorítmica

La complejidad algorítmica es una medida que permite conocer la cantidad de recursos que necesita un algoritmo para resolver un problema en función de su tamaño (Minguillón, 2015). Para ello se utiliza la notación Big O, lo cual es una forma matemática básica de expresar cuanto tarda un algoritmo en ejecutarse y es útil para evaluar cómo es la curva de crecimiento de su tiempo de ejecución conforme aumentan el tamaño del input de entrada(Cianes, 2020)

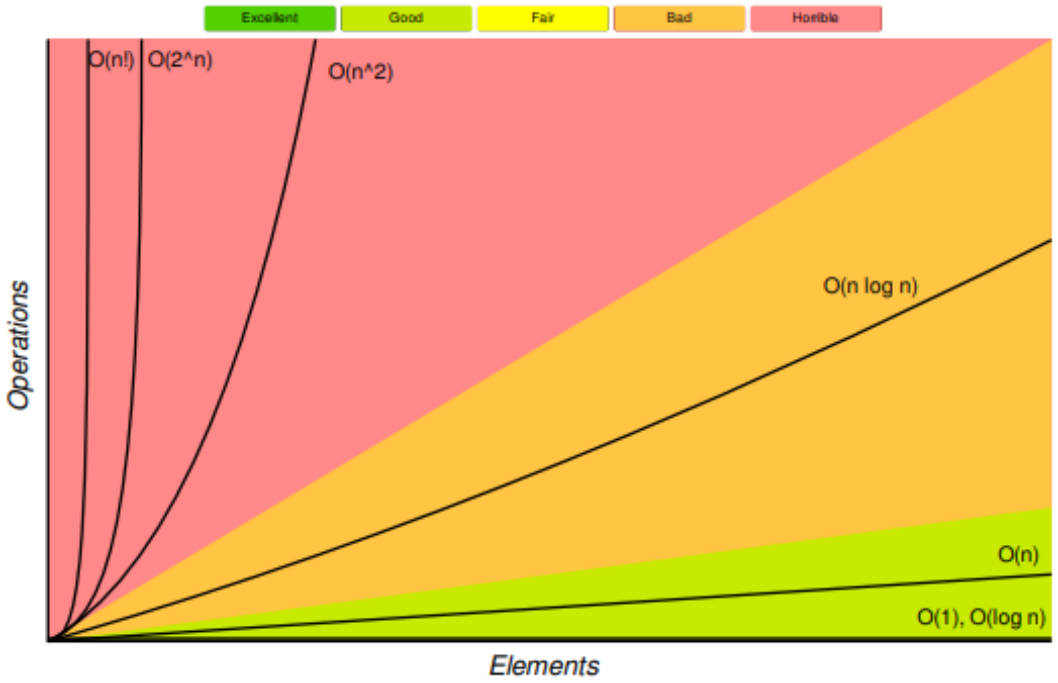


Figura 8. Tabla de complejidad Big-O

Fuente: (Sourav, 2016)

a) Big O Notation

La O viene de clasificar el ORDEN de magnitud de la función a analizar, resumiéndose en que nos sirve para determinar cómo un algoritmo escala en complejidad tanto en tiempo como en espacio necesario, en función de su tamaño de entrada.

Según Minguillón (2015), Se dice que un algoritmo F tiene una complejidad O(G(N)) si existen dos constantes C y N_0 para las que se cumpla $|F(N)| < C \times G(N)$ para todo $N > N_0$.; Es decir, un algoritmo F tiene complejidad O(G) si la cantidad de operaciones necesarias queda fijado por el comportamiento de G para valores

grandes de N. Normalmente, las funciones G (órdenes de complejidad) son sencillas, sin constantes, es decir, equivalentes con independencia de factores externos (hardware, sistema operativo, lenguaje de programación, etc.). Respecto a la constante C, su presencia en la definición de la notación O hace que, en una función G con complejidad O(G), todas las funciones de la forma C x G (en las que C es una constante) pertenecen también a O(G). Por lo tanto, se habla de complejidad O(N), por ejemplo, pero no de complejidad O(2 x N), ya que asintóticamente son equivalentes según la definición. Así, por ejemplo, N pertenece a O(N) y 2 x N también pertenece a O(N).

b) Clases de complejidad

Según Laaksonen (2018) muestra las complejidades algorítmicas más importantes

- O(1) - Tiempo constante: es el mejor resultado, y quiere decir que el tiempo de ejecución no varía conforme aumenta el tamaño de los datos de entrada, y la respuesta siempre tarda lo mismo sin importar la magnitud de entrada.
- O(n) - Tiempo lineal: el crecimiento es lineal en tanto el tiempo de ejecución es cada vez mayor de modo proporcional a cómo se incrementa el tamaño de la entrada. Por lo que, si tenemos el doble de elementos de entrada, tardará el doble, aunque despreciamos realmente la pendiente de la misma y sólo nos quedamos con que aumenta de forma lineal.
- O(log n) - tiempo logarítmico: una forma de crecimiento que crece al inicio pero tiende a estabilizarse conforme aumentan el tamaño de entrada, por lo que es una buena nota para un algoritmo ya que no tiende a resentirse.
- O(n2) - tiempo cuadrático: el crecimiento es de forma exponencial por lo que será un algoritmo a evitar ya que para valores pequeños de entrada el tiempo será asumible, pero conforme aumente el tamaño de los datos de

entrada el tiempo tenderá a ser muy elevado y es probable que el procesador se quede inoperativo.

- O(n!) - tiempo factorial: el crecimiento es factorial, por lo que rápidamente tiende a valores imposibles de tratar, en lo que sería una recta vertical.

Operation	Average Case	Amortized Worst Case
Copy	O(n)	O(n)
Append[1]	O(1)	O(1)
Pop last	O(1)	O(1)
Pop intermediate[2]	O(n)	O(n)
Insert	O(n)	O(n)
Get Item	O(1)	O(1)
Set Item	O(1)	O(1)
Delete Item	O(n)	O(n)
Iteration	O(n)	O(n)
Get Slice	O(k)	O(k)
Del Slice	O(n)	O(n)
Set Slice	O(k+n)	O(k+n)
Extend[1]	O(k)	O(k)
Sort	O(n log n)	O(n log n)
Multiply	O(nk)	O(nk)
x in s	O(n)	
min(s), max(s)	O(n)	
Get Length	O(1)	O(1)

Figura 9. Complejidad del tiempo

Fuente: (*TimeComplexity - Python Wiki*, s. f.)

1.2 Antecedentes

Los estudios previos que guardan relación con este trabajo de investigación son los siguientes:

Muehlethaler *et al.* (2021), utilizaron una interfaz de búsqueda fragmentada para rastrear a un importador minorista de ropa en línea y afirma que en menos de 24 h, lograron extraer 68 campos basados en texto que describen un total de 24,701 prendas para ayudar a proporcionar estimaciones precisas de los tipos de fibras y las frecuencias de color, además proporcionar datos del algodón, el poliéster, la viscosa y el elastano son los 4 tipos principales de fibras que se utilizan en la industria textil.

Kiran *et al.* (2021), afirma que la compleja recuperación y análisis de información es un área crucial de estudio en tiempo real que incluye estudios de información, investigaciones policiales, análisis forense digital, minería de impresiones, estudios predictivos y varias otras aplicaciones. El Scraping es el procedimiento que se utiliza comúnmente en la minería, y garantizar que el conocimiento veraz sobre la persona u objeto se perciba desde Internet.

Uriarte *et al.* (2020), a través de esta contribución, se afirma que el raspado web es una herramienta de empoderamiento. La motivación de índice de precios del consumidor Online, permite la verificación cruzada de una variable muy importante, como los precios con diferentes ámbitos geográficos, a la vez se ha demostrado cómo los índices urbanos basados en datos de desguace web se estimaron y contrastaron con los índices tradicionales que otorgan mediciones altamente eficientes a una milésima de costo.

Awangga *et al.* (2019), implementaron web Scraping para el monitoreo integrado con GitHub llegando a la conclusión que la aplicación construida ha sido capaz de responder a los problemas discutidos en los capítulos anteriores y afirma que su trabajo muestra con el diseño del sistema facilitan las tareas de recopilación de datos utilizando los medios de comunicación de redes sociales GitHub, la documentación y la recopilación de tareas más estructuradas.

Moskalenko *et al.* (2019), realizaron un estudio dedicado al proceso de raspado web y el problema de bloquear los raspadores web en Internet, además consideraron los principios y conceptos básicos del proceso de raspado web y la clasificación de los raspadores web, investigaron las técnicas para evitar los bloqueos del raspador web y su impacto en el

proceso de raspado web. Se propone un programa desarrollado en el lenguaje de programación Python que utiliza técnicas para evitar los bloqueos de la web scrapper.

Januzaj *et al.* (2019), aplicaron técnicas para comparar conjuntos de datos que contienen información sobre la demanda del mercado y los planes de estudio universitarios. Durante su trabajo indican cómo se extraen los datos de sitios web específicos. Este fue un paso necesario, mientras que la información para las ofertas de trabajo y los programas de estudio se extrajeron de los sitios web. Después de extraer la información, indican los procedimientos de procesamiento de datos para aplicar técnicas de aprendizaje automático.

Dewi *et al.* (2019), utilizaron el método de Web Scraping para la búsqueda de información, combinarla y presentarla de una mejor manera según las preferencias del usuario. Se desarrolla un sistema para implementar el método propuesto mediante el uso de una API que proporcionaron los desarrolladores de Facebook y Twitter.

Huaman *et al.* (2019), investigaron las tecnologías de asistente virtual tipo Chatbot y extracción de datos con Web Scraping, donde se pudo determinar que el Chatbot ofrece un servicio al cliente más personalizado, brindándoles información en todo momento, también se determinó que el Chatbot debe estar en constante entrenamiento para brindar respuestas asertivas. En cuanto el Web Scraping se pudo determinar que es una técnica que facilita la extracción de información de forma masiva y automatizada, para que pueda ser usada de acuerdo a las necesidades del usuario.

Ullah *et al.* (2018), presentaron un estudio donde propusieron un sistema de filtrado de precios web que explota las técnicas de raspado web para extraer tendencias y sugerir el mejor precio de un producto objetivo desde sitios web comerciales de primera línea como amazon.com, alibaba.com y daraz.pk . El marco diseñado incorpora el marco Scrapy para el rastreo y el raspado web.

Gheorghe *et al.* (2018), afirma, en comparación con el análisis convencional de las respuestas HTTP, la arquitectura y las técnicas de raspado propuestas tienen la ventaja de tratar con los sitios web modernos de una manera más práctica, en particular para los usuarios que no dominan la administración de servidores y los protocolos de Internet. Sin embargo, esto conlleva el aumento de los tiempos de procesamiento, la baja portabilidad (es improbable que un raspador creado para un sitio web dinámico en particular sea

operativo para otro diferente) y el gasto indebido de los recursos de red (para imitar un comportamiento humano normal, el navegador carga elementos como imágenes, contenido multimedia y secuencias de comandos de estilo que no se pueden utilizar para lograr el objetivo deseado.

Muñoz *et al.* (2018), afirma que almacenar datos climatológicos permitirá tener un compilatorio en forma de serie de tiempo, algo muy importante para poder analizar fenómenos como lluvias atemporales, sequías o inundaciones mediante cálculos estadísticos o empleando diversos algoritmos. Asimismo, esta serie de tiempo puede ser empleada para generar pronósticos mediante modelos estadísticos, así como entrenamiento para inteligencias artificiales que permitan generar pronósticos, o realizar análisis de las tormentas sucedidas.

Marques *et al.* (2018), presentaron resultados sobre el uso de diversas herramientas de monitoreo para la detección de actividad maliciosa de raspado web, además llevaron a cabo un análisis de un conjunto de datos real de registros de Apache HTTP Access para una aplicación de comercio electrónico proporcionada por un gran proveedor multinacional de TI para la industria global de viajes y turismo, también se utilizaron dos herramientas para detectar actividades de raspado basadas en las solicitudes HTTP: una herramienta comercial y una herramienta interna llamada Arcane. Mostramos los beneficios que se pueden lograr mediante el uso de ambos sistemas, en términos de sensibilidad y especificidad generales, y discutimos las posibles fuentes de diversidad entre los patrones de alerta de la herramienta.

Diab *et al.* (2018), realizaron un estudio donde utilizaron el modelo Markup Randomizer para detener los scrappers cambiando las marcas CSS y HTML periódicamente, lo que detiene eficientemente los scrapers para siempre. Por lo tanto, el modelo se prueba en un conjunto de datos que se recopiló al azar de diferentes categorías. La prueba se realizó en tres etapas raspar, aleatorizar, volver a raspar y los resultados fueron muy buenos. Finalmente, el modelo está deteniendo completamente los scrappers basados en CSS y los experimentos muestran que el tiempo requerido es muy bueno para ese montón de resultados y es aceptado en absoluto.

Baskaran *et al.* (2018), propuso un framework de extracción automática de registros de datos del Foro de discusión sobre salud. El contenido de los sitios del Foro Médico sirve como una fuente complementaria de información para muchas aplicaciones de minería de

datos, como la predicción de reacciones a medicamentos, la predicción de enfermedades basada en síntomas, la sugerencia de pruebas clínicas, etc.

Torres *et al.* (2017), manifiesta que la herramienta desarrollada, los participantes redujeron el tiempo de costeo en dos terceras partes del tiempo actual invertido con un proceso de estandarización de recetas para obtener calidad y el tamaño estándar de porciones. Los resultados comenzaron a ser notorios conforme hubo mayor interacción con la herramienta, debido a que entre más familiarizado se encuentre el usuario con la tecnología, más fácil será su uso y el tiempo estimado del proceso se reduce.

Murillo *et al.* (2017), realizaron un estudio con objetivo de utilizar la técnica Web Scraping para extraer datos de Google Scholar a través de diferentes métodos utilizando el lenguaje R, a la vez manifiesta que el Web Scripting resulta ser una alternativa funcional para extraer datos de un sitio web, sin embargo, con los métodos web online y de escritorio realizados a través de aplicaciones no logramos obtener el objetivo deseado, en tiempo y datos estructurados.

Rizaldi *et al.* (2017), manifiesta que el uso del método del selector XPATH para los sitios de noticias de raspado web produce artículos más completos que el uso del método del selector CSS. Esto se indica por la cantidad de elementos y el tamaño del archivo obtenido es mayor que el método del selector de CSS. Pero esto también deja más trabajo, ya que requiere otro proceso para eliminar el código HTML no deseado del artículo generado mediante el método del selector XPATH.

Vàllez (2017), Afirma que el uso del marcado semántico, a través de las etiquetas HTML, y de los vocabularios controlados resulta una combinación ventajosa para la indexación de contenidos web, además, la indexación semiautomática puede utilizarse como una herramienta complementaria para el indexador humano y el uso de la información obtenida de la analítica web debe tenerse en cuenta para mejorar los procesos de recuperación de información.

Khalil *et al.* (2017), presentaron RCrawler, un webcrawler potente, flexible y con subprocesos avanzados basado en R que proporciona un conjunto de funciones útiles para el rastreo web, el web Scraping y también el análisis potencial de enlaces.

Landers *et al.* (2016), sostenemos que el raspado web ofrece un gran potencial para la psicología al aumentar el acceso a los datos de comportamiento sin una intrusión

significativa del investigador, aumentando drásticamente el tamaño de las muestras, disminuyendo la cantidad de tiempo dedicado durante los datos fase de recopilación, aumentando el acceso a investigadores en países recientemente industrializados y subdesarrollados que no pueden pagar proyectos de investigación tradicionales a gran escala, y mejorando la aplicación interdisciplinaria de nuestra vasta literatura de investigación sobre psicometría. El raspado web, en esencia, se trata de encontrar significado en los patrones de comportamiento humano, el objetivo fundamental de toda investigación psicológica.

Almeida *et al.* (2016), manifiestan que las extracciones realizadas permitieron analizar información relevante sobre las atracciones turísticas de Minas Gerais, como la oferta de atracciones por categorías y municipios, la calificación promedio de cada atracción, el perfil de los visitantes y el momento de mayor visita. Esta información puede ser utilizada por los gestores públicos y también para cada atracción, controlar el nivel de satisfacción de los visitantes, ayudar en la creación de proyectos para mejorar la calidad del servicio prestado, la creación de itinerarios o la difusión de atracciones por segmentos y motivaciones para viajar. Por lo tanto, se cree que la extracción de información de los usuarios en el sitio de TripAdvisor puede considerarse como una buena fuente de datos para la planificación del sector.

Vernier *et al.* (2016), realizaron un estudio donde propusieron una metodología basada en minería de datos web, para ello utilizaron técnicas de rastreo y extracción de flujos de noticias de 37 medios de comunicación chilenos que presentan una vida activa en Twitter y proponemos varios indicadores para compararlos. Analizamos los volúmenes de producción, sus audiencias potenciales y, usando técnicas de procesamiento natural del lenguaje, exploramos el contenido de la producción informativa, sus tendencias editoriales y cobertura geográfica.

Hernández *et al.* (2015), realizaron una revisión del estado del arte de diferentes metodologías que se han propuesto para realizar análisis político en las redes sociales y en el internet en general. Esta es la primera etapa de un proyecto de investigación en el que se pretende obtener información de diferentes fuentes en internet mediante técnicas de web scraping y analizar lo obtenido mediante técnicas de text mining. Esta revisión nos será de gran utilidad para definir los indicadores necesarios, así como los resulta-dos esperados de nuestro proyecto.

Chu *et al.* (2015), realizo un estudio en la cual presenta un enfoque novedoso para la extracción de datos en páginas web. El enfoque propuesto puede extraer de manera efectiva los registros de datos mediante la coincidencia de la ruta de datos e identificar los elementos de datos mediante la alineación del código de la ruta de datos. Los procesos son compatibles con información visual para filtrar regiones de datos inesperadas con características espaciales. Los resultados experimentales revelan que la tasa de precisión promedio es tan alta como el 96% y la tasa de recuperación es superior al 93% para el enfoque propuesto, que supera al otro método conocido.

CAPÍTULO II

PLANTEAMIENTO DEL PROBLEMA

2.1 Identificación del problema

Las diversas agencias de viaje de la ciudad de Puno, en la gran mayoría, cuentan con sus respectivas páginas web, las cuales ofrecen información de los atractivos y diversos paquetes turísticos (destino, precio, estadía, incluye, etc.). Un usuario para poder elegir cuál de los paquetes le conviene, debe de navegar por cada una de las páginas webs, volviéndose un proceso bastante engorroso y demanda mucho tiempo, además los usuarios deberán conocer los dominios web de todas las agencias de viaje para realizar la comparación de precios, estadía, que servicios incluye el paquete, promociones, etc.; ahora no todas las páginas web están bien posicionadas dentro de los buscadores, lo que hace que algunos sitios estén inubicables; además, la información que se muestra son datos que no tienen estructura interna identificable, es un conglomerado masivo y desorganizado de varios objetos que no tienen valor hasta que se logre identificar y almacenar de manera organizada. Una vez que se organizan, los elementos que conforman su contenido pueden ser buscados y categorizados para obtener información.

En la actualidad no se cuenta con un sitio web que pueda albergue la información relevante, actualizada de las agencias de viajes, además no se puede realizar el análisis y la comparación de lugares, precios y promociones para agilizar la selección del paquete más conveniente.

El propósito de este estudio es desarrollar un sitio web en la cual pueda almacenar información de los diferentes paquetes turísticos que son ofertados por las diversas agencias de viaje que operan en la región de Puno utilizando la técnica del web Scraping para la extracción de la data, y posteriormente ser almacenado en una base de datos.

2.2 Enunciados del problema

De lo expuesto anteriormente nos planteamos la siguiente premisa:

¿El uso del software para indexar sitios web optimiza la búsqueda de paquetes turísticos de la región de Puno utilizando la tecnología web Scraping?

Y las siguientes interrogantes especificas:

- ¿En qué medida el análisis de la estructura DOM de las páginas web de las agencias de viaje contribuye al desarrollo del algoritmo para la extracción de la información relevante?
- ¿En qué medida el rendimiento de un sitio web basado en la experiencia de usuario facilita la búsqueda de paquetes turísticos?
- ¿La norma ISO/IEC 25000 permite establecer el grado de satisfacción de un producto de software?

2.3 Justificación

La sobreabundancia de información en Internet, es uno de los principales componentes de su éxito; sin embargo, el tratamiento de esta, exige una enorme cantidad de tiempo y energía a fin de desgranar la calidad de los datos sumergidos en tan enorme repositorio. Según Guzñay, (2016), refiere que actualmente la sobrecarga de información que recibe un usuario, en especial de Internet en todas sus formas, puede causarle la sensación de no poder abarcarla ni gestionarla y, por tanto, llegar a generarle una gran angustia

En la región de Puno, el turismo es uno de los sectores con enorme potencial de desarrollo, ya que cuenta con importantes recursos turísticos reconocidos y otros que recién están tomando auge, además de poseer una cultura tradicional y proveer una gran cantidad de posibilidades para los turistas que nos visitan.

El sector del turismo es un ambiente muy dinámico, los productos (paquetes turísticos) cambian continuamente, como así también los intereses de los usuarios y precisamente estos, deben pasar bastante tiempo en el ordenador o su dispositivo móvil a fin de poder encontrar un paquete acorde a sus necesidades.

2.4 Objetivos

2.4.1 Objetivo general

Desarrollar un software para la indexación de sitios web y optimizar la búsqueda de paquetes turísticos de la región de Puno basado en web Scraping.

2.4.2 Objetivos específicos

- Desarrollar un algoritmo para extraer la información relevante de las páginas web de las agencias de viaje que operan en la región de Puno.
- Implementar y determinar el rendimiento del sitio web basado en la experiencia de usuario que facilite la búsqueda de los paquetes turísticos
- Establecer el grado en que el producto satisface a un usuario final basado en la norma ISO/IEC 25000

2.5 Hipótesis

2.5.1 Hipótesis general

La implementación del software para la indexación de sitios web basado en web scraping optimiza la búsqueda de paquetes turísticos de la región de Puno

2.5.2 Hipótesis específicas

- El análisis de la estructura DOM facilita la implementación del algoritmo para la extracción de la información relevante de las páginas web de las agencias de viaje de la región Puno.
- El rendimiento de un sitio web basado en la experiencia de usuario facilita la búsqueda de paquetes turísticos
- La norma ISO/IEC 25000 permite establecer el grado de satisfacción de un producto de software.

CAPÍTULO III

MATERIALES Y MÉTODOS

3.1 Lugar de estudio

La región de Puno se encuentra en la parte sur del Perú. Está ubicada a orillas del lago Titicaca y sobre los 3,827 metros s.n.m. según el estudio realizado por Laurente & Machaca (2020) Puno es la cuarta región más visitada por los turistas internacionales. Una de las grandes festividades del Perú es la Fiesta de la Virgen Candelaria en Puno, que se desarrolla todos los años en el mes de febrero, llegando turistas nacionales y extranjeros.

La ciudad de Puno es considerada la capital del Folclore peruano. Se desarrolló una de las culturas más importantes del antiguo Perú, la Cultura Tiahuanaco, máxima expresión del antiguo pueblo Aymara

3.2 Población

La población está conformada 38 páginas web de las agencias de viaje que operan en la ciudad de Puno («IPERÚ Puno», 2019)

3.3 Muestra

El diseño de muestra que se utilizó es el muestreo por conveniencia, el cual es un método de muestreo no probabilístico donde los sujetos son seleccionados dada la conveniente accesibilidad y proximidad de los sujetos para el investigador, para la indexación de los sitios web se seleccionó a trece (13) página web bajo los siguientes criterios.

Criterios de Inclusión

- Páginas web activas (cuentan con un domino habilitado)
- Permisos en el archivo robots.txt
- Páginas web que promociones destinos turísticos de la región de Puno
- Páginas web con descripción de los paquetes turísticos

Criterios de Exclusión

- Páginas web inactivas (no cuentan con dominio habilitado)
- Páginas web denegadas por medio del archivo robots.txt
- Páginas web que no promocionen destinos turísticos de la región de puno
- Páginas web sin contenido y/o en mantenimiento

Criterios de Eliminación

- Agencias de viaje que no cuentas con un sitio web

Para recabar opiniones y establecer el grado en que en que el producto de software satisface a un usuario final basado en la norma ISO/IEC 25000 se seleccionó a diez participantes

3.4 Método de investigación

3.4.1 Tipo de investigación

- Según su finalidad: Es una Investigación Aplicada (Arias, 2012), porque busca mejorar la problemática existente a través del software para la indexación de sitios web y optimizar la búsqueda de paquetes turísticos de la región de Puno
- Según el enfoque de la investigación: Es una Investigación Explicativa porque describe las causas de los fenómenos que están en estudio (Hernández *et al.*, 2014)

3.4.2 Diseño de investigación

El diseño de investigación que se va a utilizar, es el diseño experimental, pre experimento (pre prueba y post prueba con un solo grupo), en tal sentido:

- Se va a seleccionar y se va a manipular la variable independiente con la solución software (indexación de sitios web).
- Se aplicará un Pre test para medir a la variable dependiente después de la aplicación del producto se aplicará un Pos test para evaluar los beneficios de la solución.

Tabla 2

Diseño pre prueba y post prueba con un solo grupo

Aplicación de pre-test o medición inicial	Aplicación del estímulo o tratamiento	Aplicación del post-test o medición final
G O1	X	O2

Fuente: (Arias, 2012)

- G: Personas seleccionadas para determinar la optimización de la búsqueda de paquetes turísticos.
- O1: Búsqueda de paquetes turísticos de la región de Puno antes de la indexación de sitios web basado en web Scraping.
- X: Software para la indexación de sitios web basado en web Scraping.
- O2: Búsqueda de paquetes turísticos de la región de Puno después de la indexación de sitios web basado en web Scraping.

Al finalizar se establecen las diferencias entre O1 y O2 para determinar si se logró optimizar la búsqueda de paquetes turísticos

3.4.3 Método de tratamiento de datos

La prueba de los rangos con signo de Wilcoxon es un estadístico no paramétrico que se utiliza para comparar la media de dos muestras relacionadas y determinar si existen diferencias entre ellas. Se utiliza como alternativa a la prueba t de Student cuando no se puede suponer la normalidad de dichas muestras.

i. Planteamiento de hipótesis

$\mathbf{H_0}$: $\boldsymbol{\mu_x \geq \mu_y}$ (Con la implementación del sitio web, no se reduce el tiempo en la búsqueda de paquetes turísticos de la región de Puno basado en web Scraping).

$\mathbf{H_a}$: $\boldsymbol{\mu_x < \mu_y}$ **(**Con la implementación del sitio web, si se reduce el tiempo en la búsqueda de paquetes turísticos de la región de Puno basado en web Scraping).

ii. Nivel de significancia

Se usará un nivel de significancia del 5%, es decir $\alpha = 0.05$

iii. Regla de decisión

Si $Z_c < Z_t$, entonces se rechaza la H_O y se acepta la H_a.

iv. Conclusión

Dependiendo del resultado de la regla de decisión, se dará una interpretación acerca de los datos analizados.

3.5 Descripción detallada de métodos por objetivos específicos

Para el cumplimiento de los objetivos específicos recurrimos a los siguientes métodos:

3.5.1 Metodología para desarrollar el algoritmo de extracción

En la actualidad gran parte de los sitios web hacen uso de Sistemas Gestores de Contenido (CMS) para la publicación y la administración de la información, promueven las buenas prácticas del desarrollo web, utiliza estándares de desarrollo, entre otras ventajas el cual facilito en el análisis de la estructura de las pagina web de cada agencia de viaje.

Lo primero que se realizó fue el establecer la URL padre de cada página web, la siguiente tarea fue analizar la estructura DOM y encontrar los contenedores HTML repetitivos con ayuda del inspector de elementos del navegador; la tarea es ubicar el enlace que redirige a la información detallada del paquete turístico ofertado, finalmente establecer el selector de CSS para extraer la información relevante.

La otra etapa es automatizar la extracción de los datos identificados en cada página web de las agencias de viaje sometidas al presente estudio. Se opto por hacer uso del lenguaje de programación de Python, el cual posee librerías para manipular contenido de la web. Para implementar el algoritmo de extracción se utilizó la metodología de desarrollo XP por ser muy rápida y ligera de desarrollo de software que se basa en la simplicidad, la comunicación y la realimentación o reutilización del código desarrollado

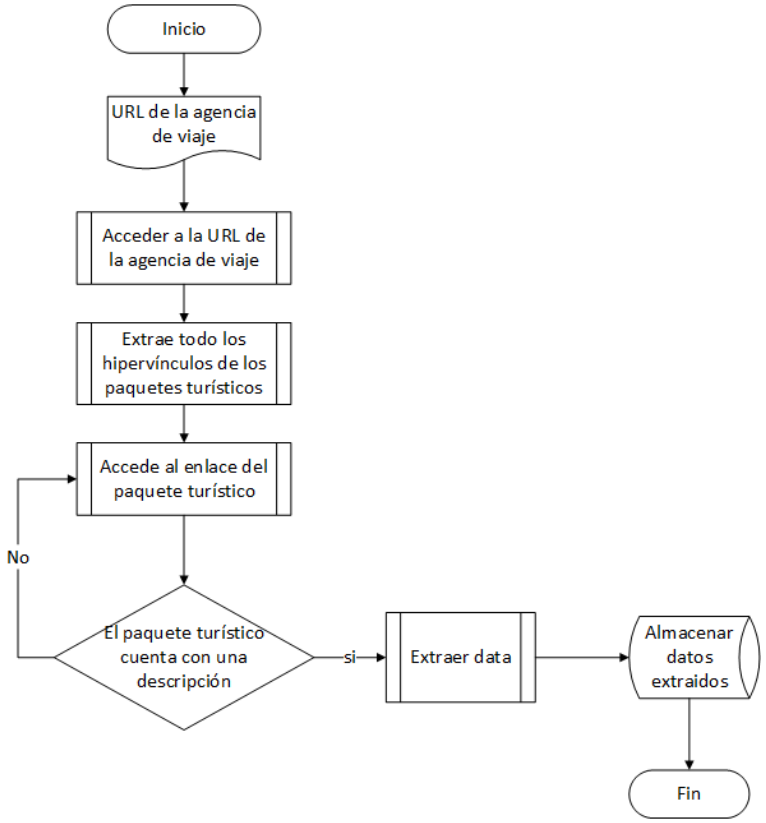


Figura 10. Diagrama de flujo para la implementación del algoritmo

A. Análisis

- Se utiliza historias de usuarios: se describe las necesidades y las funcionalidades que debe poseer el algoritmo, con la ayuda de los diseños.
- Se crean los planes de entrega, los cuales estiman el tiempo de desarrollo de las historias de usuario.
- Se llevan a cabo la planificación de iteración: identificar las historias de usuarios que se van a desarrollar en una iteración especifica.

B. Diseño

- Se proponen soluciones a problemas técnicos o de diseño.
- Se ignoran las funcionalidades extra que podrían incorporarse al proyecto, centrar en lo principal.
- Se remueve la redundancia, se eliminan las funcionalidades no necesarias y se renuevan los diseños obsoletos.

C. Desarrollo

- Se utilizan estándares para escribir el código.
- Se crean las pruebas antes de empezar a codificar, lo cual hará más sencillas y efectivas las pruebas.
- Se deja la optimización para el final. Una vez que el código requerido este completo

D. Pruebas

La evaluación de la calidad en uso del software basado en la ISO/IEC 2500

3.5.2 Determinación del rendimiento del sitio web gopuno

PageSpeed Insights es una herramienta gratuita creada por Google que informa sobre el rendimiento de las páginas tanto en dispositivos móviles como en ordenadores, pero además de esto, es capaz de ofrecer una serie de sugerencias y herramientas asociadas para mejorar los resultados.

PSI facilita datos de experimentos y datos de campo sobre las páginas. Los datos de experimentos son útiles para depurar problemas de rendimiento, ya que se recogen en un entorno controlado, pero es posible que con ellos no se detecten problemas de capacidad producidos por volúmenes reales de tráfico. En cuanto a los datos de campo, resultan útiles para saber qué pasa con las experiencias de usuario auténticas y reales.

Para realizar este análisis, lo que PSI hace es comprobar que el portal que está siendo analizado cumple algunas de las buenas prácticas que Google considera necesarias para que la web sea mostrada lo más rápido posible a los usuarios. Estas buenas prácticas abarcan tanto a nivel front-end (imágenes, archivos, carga de JavaScript, hojas de estilos...), así como la configuración del servidor donde la web está hospedada.

Los objetivos planteados por esta herramienta para mejorar la velocidad de carga del sitio se pueden clasificar en tres aspectos fundamentales.

- Reducir al máximo el número de llamadas HTTP realizadas
- Reducir a su mínima expresión el tamaño de las respuestas tras una petición HTTP

- Optimizar el renderizado de la página en el navegador del usuario

PSI de Google ofrece seis métricas que miden diversos aspectos del rendimiento relevantes para los usuarios

- First Contentful Paint (FCP): mide cuánto tiempo le toma al navegador procesar la primera parte del contenido DOM después de que un usuario navega a la página web.
- Speed index: mide la rapidez con la que se muestra visualmente el contenido durante la carga de la página.
- Largest Contentful Paint(LCP): informa el tiempo de renderizado de la imagen o el bloque de texto más grande visible dentro de la ventana gráfica.
- Time to Interactive (TTI): mide el tiempo que tarda una página en volverse completamente interactiva.
- Total Blocking Time: mide la cantidad total de tiempo que una página está bloqueada para que no responda a la entrada del usuario, como los clics del mouse, los toques de la pantalla o las pulsaciones del teclado.
- Cumulative Layout Shift (CLS): mide la suma total de todas las puntuaciones de cambio de diseño individuales para cada cambio de diseño inesperado que se produce durante toda la vida útil de la página.

3.5.3 Determinación del grado en que el producto de software satisface a un usuario final basado en la norma ISO/IEC 25000.

La nueva familia de normas ISO/IEC 25000 que se desarrolla en el proyecto SQuaRE, es el esfuerzo que hace la ISO para cubrir más temas relacionados a la calidad de producto software y toma como base y reemplaza las series ISO/IEC 9126 y 14598. El modelo de calidad de uso según la ISO/IEC 25010 la define como el grado en el que un producto o sistema puede ser utilizado por usuarios específicos para satisfacer sus necesidades y alcanzar sus objetivos específicos con eficacia, eficiencia, libertad de riesgo y satisfacción en un contexto específico de uso, y el modelo ISO/IEC 25022 define específicamente las métricas para realizar la medición de la calidad en uso del producto.

CAPÍTULO IV

RESULTADOS Y DISCUSIÓN

En este capítulo presento los resultados y discusión concerniente a cada objetivo específico

4.1 Resultado conforme al objetivo específico 1

Para poder obtener la información deseada (nombre del paquete, descripción, itinerario, que incluye el paquete, logo, imagen, URL del Dominio, URL del paquete) del paquete turístico, primero se realizó el análisis de cada página web e identificar su estructura HTML. Por un lado, tenemos la página web que contiene el listado de todos paquetes turísticos, un ejemplo de esta página web se muestra en la figura11

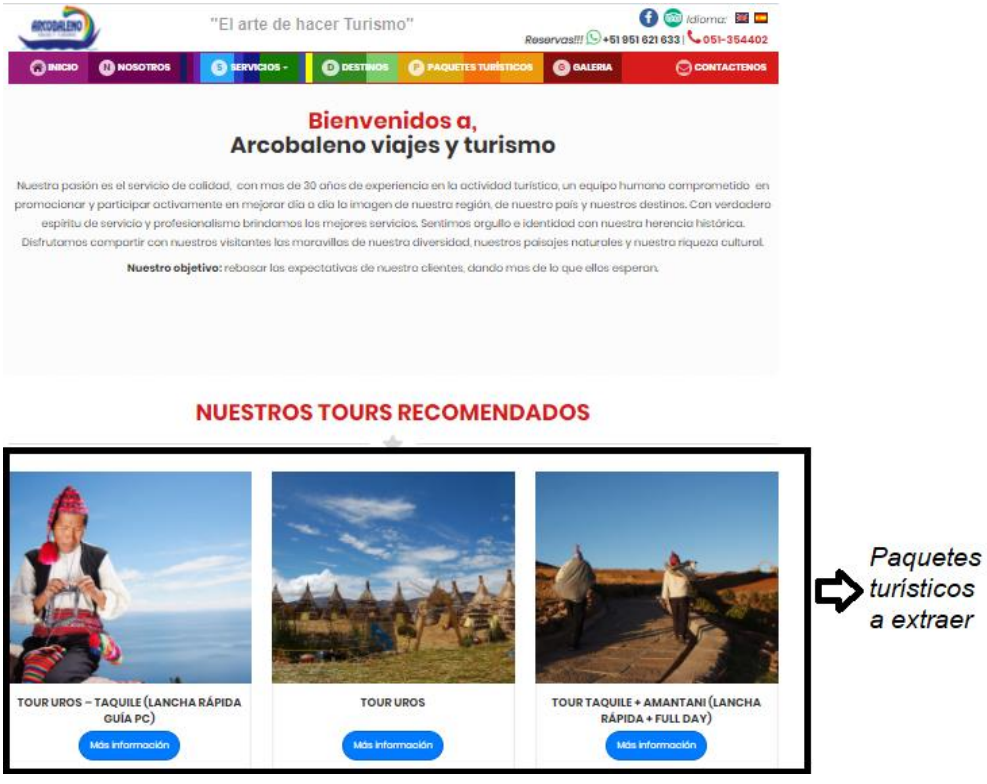


Figura 11. Distribución de paquetes turísticos en la web

Al acceder al paquete turísticos identificamos toda la información que se va extraer con el scraper ver figura 12

Figura 12. Información relevante a extraer

De la figura 11 se analizaron el código fuente (ver figura 13) de la web para determinar el enlace principal que será visitado por el sistema de forma automática. Una vez que el enlace se ha lanzado con éxito, se identificó el texto o imagen que contiene el hipervínculo, los que nos redirigen dentro del sitio donde se obtuvo la información del paquete turístico. Una vez extraídos los enlaces del texto o la imagen, se almacenarán en la memoria para que sean visitados posteriormente. Una vez que los enlaces estén almacenados en la memoria, serán visitados y se extraerá información específica del contenido de ese enlace.

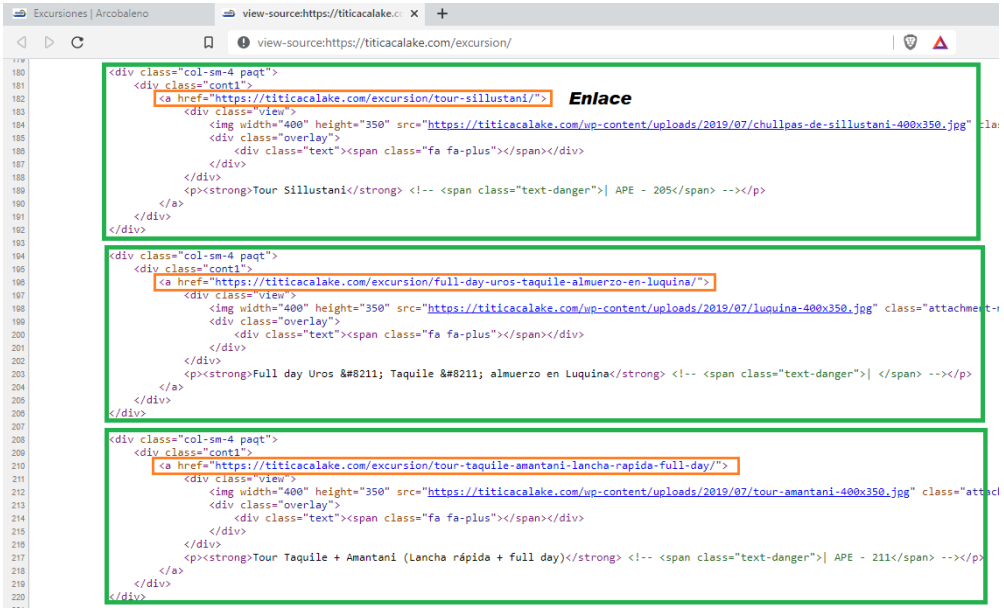


Figura 13. Código fuente de la portada principal

Una vez que se almacene el enlace y se hace clic en él, el sistema va redirigir al sitio web que incluye información a ser extraída ver figura 14.

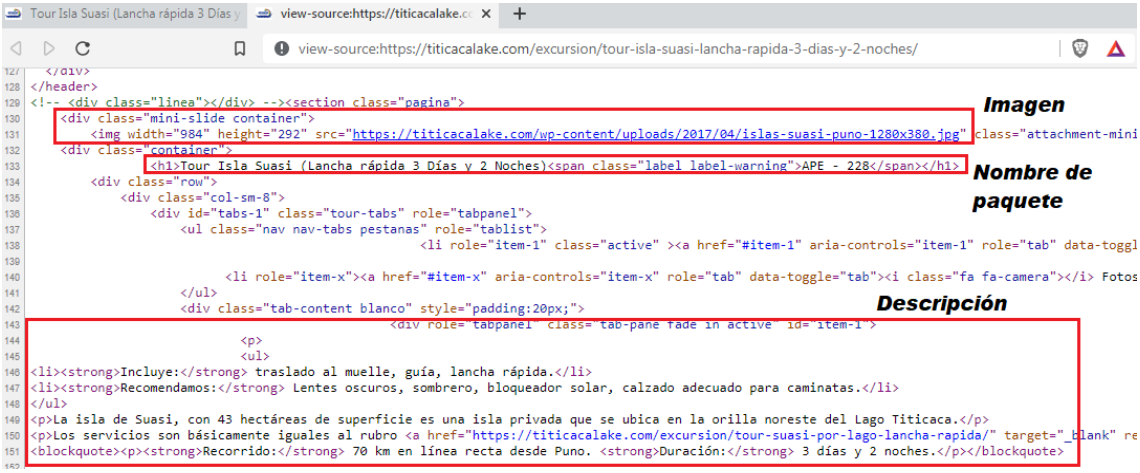


Figura 14. Código fuente del paquete turístico

En la figura 14 vemos cómo se utiliza la identificación de los elementos que se utilizan para definir la imagen de portada, nombre y descripción del paquete turístico. Según la plantilla del diseño web, dentro de la etiqueta div cuya clase es “mini-slide", se encuentra la imagen de portada, el elemento que se usó para definir el nombre del paquete es la etiqueta "h1", mientras que la etiqueta que se usa para definir la descripción del paquete se utiliza la clase “tap-pane" así como otros elementos como "p" y "ul".

4.1.1 Extracción de datos

Una vez realizado el análisis de la estructura de cada una de las páginas web, el siguiente paso es extraer los datos relevantes, para el desarrollo del software se utilizó la metodología de XP.

La programación extrema como metodología de desarrollo muy rápida y ligera, cuya base basa es la simplicidad, la comunicación y la realimentación o reutilización del código desarrollado

4.1.2 Análisis del sistema

Se tomo en cuenta las historias de usuarios y los requerimientos funcionales, para ello se elabora los diagramas de casos de uso que representa gráficamente el comportamiento y los procesos que el web Scraping ejecuta y sus relaciones que hace el sistema.

a) Requerimientos funcionales

- Crear un mapa de sitios web para realizar las consultas
- Acceder a la página web en consulta e identificar los vínculos que nos llevara a los paquetes turísticos
- Una vez obtenido la lista de vínculos se debe visitar cada vinculo y extraer el nombre del paquete, descripción y otros datos relevantes para construir el data set
- Almacenar los datos obtenidos en formato CSV
- Realizar la limpieza (datos redundantes, datos faltantes, caracteres especiales) del data set extraído y almacenarlo en formato CSV
- Exportación de datos en formato SQLite

b) Requerimientos no funcionales

- Recorrer cada vinculo y determinar si el paquete cuenta con descripción o detalle del producto
- Aplicación multiplataforma.
- Interfaz agradable para fácil entendimiento del software.
- Disponibilidad del sistema de encontrarse disponible todos los días.
- Portabilidad estará diseñado en un lenguaje multiplataforma.
- Mantenimiento y escalabilidad diseñado pensando en el crecimiento del sistema.

c) Definición de roles

Dada la coyuntura de la investigación, y la disponibilidad de recursos humamos, el investigador ha asumido los roles de directa relación con el desarrollo del sistema. Solo se han tomado en consideración los roles más importantes según el desarrollo de la presente investigación.

- Programador: El investigador asume el rol de programador por tal motivo es el encargado de escribir el código fuente necesario para la implementación del Web Scraping.
- Tester: Este rol es asumido por el asesor de tesis por el desarrollador con el fin de apoyar al programador en la preparación y realización de pruebas también está encargado de explicar los resultados al equipo.
- Tracker: El investigador analiza la información sobre la marcha del proyecto sin afectar demasiado el proceso.

- Entrenador: El investigador, es el responsable global del proyecto también es el encargado de verificar que se estén aplicando correctamente las guías XP.

4.1.3 Diseño

El propósito del diseño es de crear la solución propuesta que se planteó en la primera parte, se ha tomado en consideración el especial énfasis que hace XP en los diseños simples y claros. Para un mejor entendimiento de las propiedades del sistema y aplicando los diagramas UML.

A. Diagrama de caso de uso para la extracción, transformación y exportación de datos

En la figura 15 se muestra el caso de uso para generar el data set, los cuales realizan las siguientes acciones

- Extraer datos: el Scraping previamente programado deberá extraer los datos relevantes del sitio web en consulta y posteriormente almacenarlos en formato CVS
- Transformación de datos: una vez que tengamos los datos almacenados en formatos CVS, se debe de limpiar los datos, esto quiere decir que el Scraper debe eliminar datos repetidos, eliminar registros en blanco y corregir caracteres especiales, finalmente deberá almacenar en otro archivo CSV
- Exportar Datos: de poseer datos transformados, se debe de exportar a un formato que ser utilizado por gestores de base de datos, para nuestra investigación haremos uso de SQLite

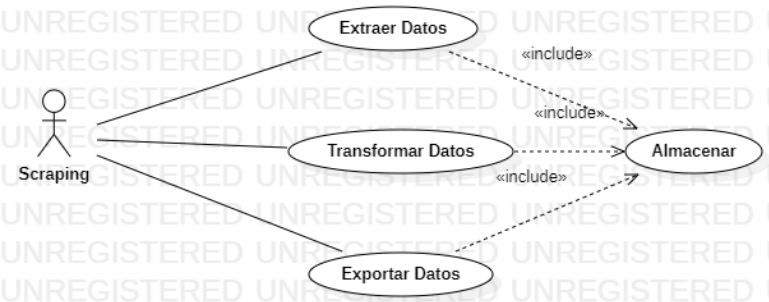


Figura 15. Diagrama de caso de uso de consultar paquete turístico

En la figura 16 muestra el caso de uso donde el actor usuario puede realizar la consulta de paquetes turísticos ya sea por el nombre de la agencia de viajes o por

el nombre del destino; por otro lado, el Sitio Web deberá procesar la consulta y luego enviar una respuesta.

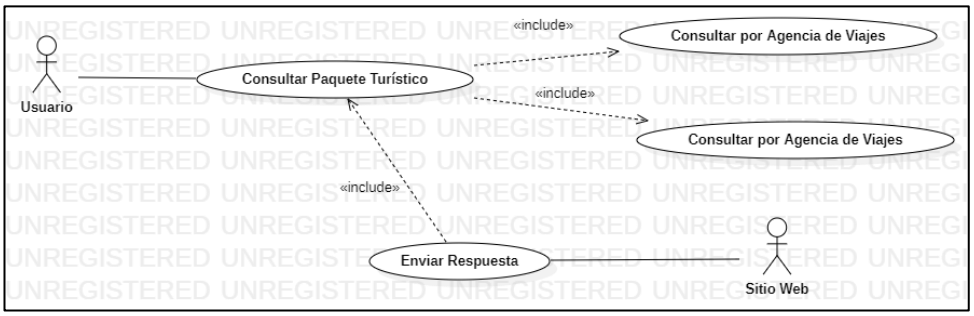


Figura 16. Diagrama de casos de uso para la consulta de paquetes

B. Diagrama de interacción para la extracción de datos

En la figura 17 se muestra la interacción con los diferentes objetos y las secuencias que debe realizar para la extracción de los datos de las páginas web en consulta

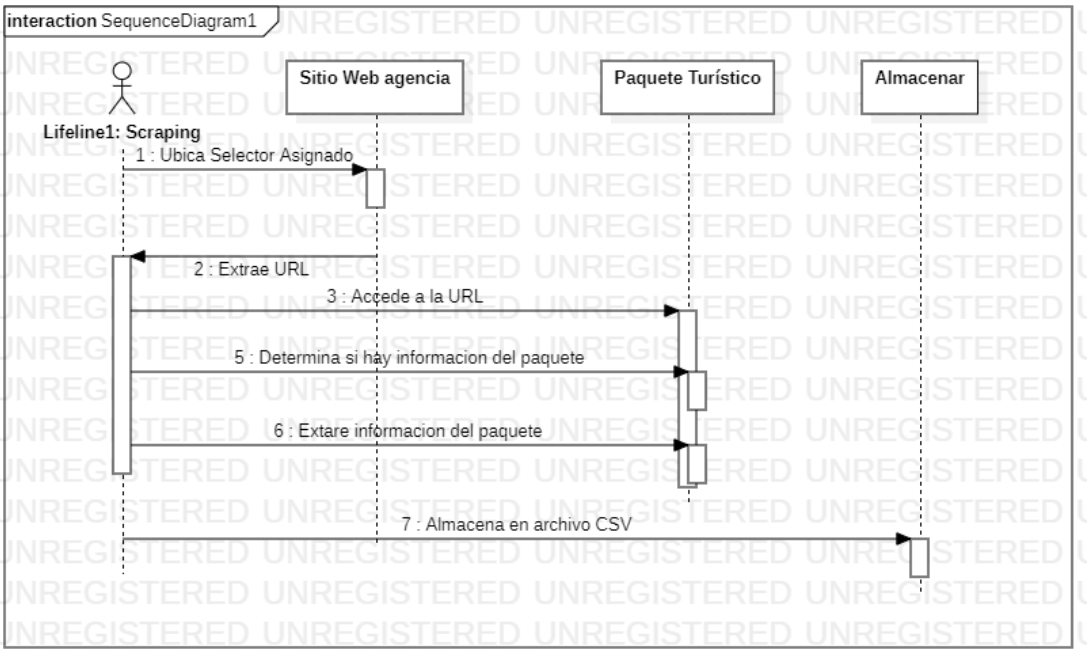


Figura 17. Diagrama de secuencia para la extracción de datos

C. Diagrama de interacción para la transformación de los datos

En la figura 18 se observa la interacción con los diferentes objetos y las secuencias que debe realizar para la transformación de los datos, los cuales fueron almacenados previamente en formatos CSV.

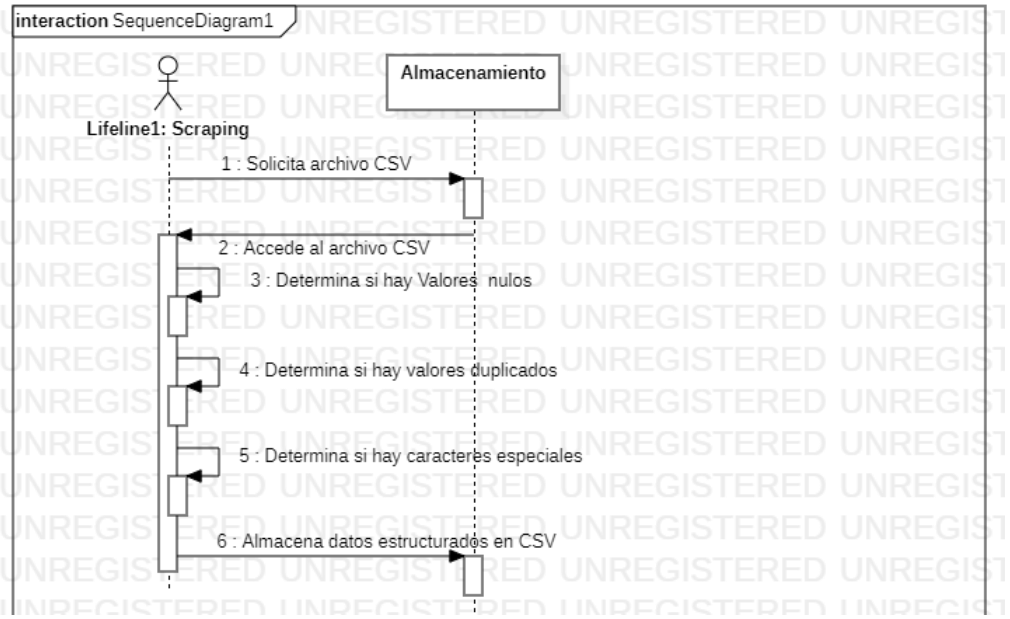


Figura 18. Diagrama de secuencia para la transformación de datos

D. Diagrama de interacción para la exportación de datos

En la figura 19 se observa la interacción con los diferentes objetos y las secuencias que debe realizar para la exportación de los datos para que los sistemas gestores de base de datos pueda manipularlos.

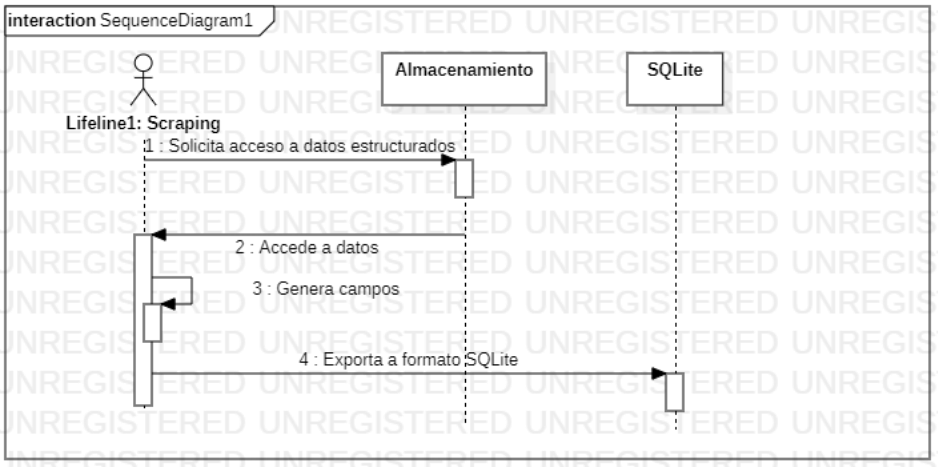


Figura 19. Diagrama de secuencia para la exportación de datos

E. Diagrama de interacción para la búsqueda de paquetes turísticos en nuestro sitio web

En la figura 20 se observa la interacción con los diferentes objetos y las secuencias que debe realizar para la búsqueda de paquetes turísticos haciendo uso del

aplicativo web, que tiene la indexación de los diferentes paquetes turísticos ofertados por las agencias de viaje que operan en la ciudad de puno.

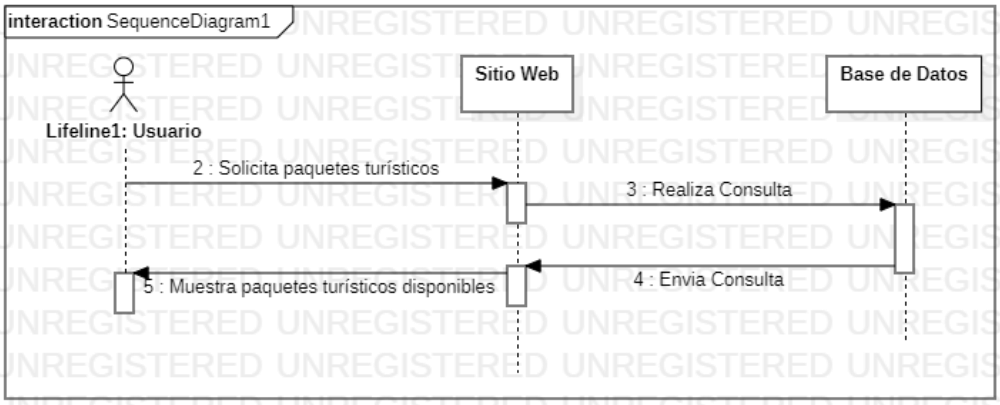


Figura 20. Diagrama de secuencia para la búsqueda de paquetes

4.1.4 Desarrollo del web Scraping

A. Algoritmo para la extracción de datos

Para la extracción de los datos se utilizó Python por ser es un lenguaje de programación interpretado; además, Python cuenta con librerías que permite analizar el contenido HTML de una página web. Para el desarrollo del web Scraping se utilizó el patrón de diseño POM (Page Object Model) una herramienta esencial para encontrar errores rápidamente durante las fases iniciales de los ciclos de desarrollo de software.

Iniciamos le algoritmo definiendo la estructura de cada página web a ser extraída, indicando las variables asociadas a las etiquetas HTML y/o selectores CSS en un archivo denominado Config.yaml, el cual actúa como un mapa para realizar la extracción de los datos. YAML es un formato para guardar objetos de datos con estructura de árbol, este lenguaje es muy legible para las personas, más legible que un JSON y XML (Contreras, 2016).

Figura 21. Estructura de las páginas web a scrapear

Dado que la web es un lugar dinámico del cual no se tiene el control y puede ser modificado en cualquier instante, se debe de separar los selectores de la lógica de nuestro algoritmo, de tener alguna modificación en la web, basta con actualizar los selectores en el archivo YAML. El programa utilizara la variable news_sites para acceder a todas las url de los sitios web a ser scrapeado; cuando acceda a una página web, realizara las siguientes consultas:

- Homepage_article_links: Obtiene el hipervinculo de los paquetes turísticos ofertados por la agencia de viaje
- Article_title: Obtiene el nombre del paquete turístico
- Article_body: Obtiene la descripción del paquete turístico
- Itinerario: Obtiene el itinerario del paquete turístico
- Incluye: Obtiene la descripción de los servicios que incluye el paquete turístico
- Email: Obtiene los datos de contacto de la agencia de viajes
- Img: Obtiene el link del logo de la agencia de viaje
- Imagen: Obtiene el link de la imagen de portada del paquete turístico

En el archivo news_page_objects(ver figura 22) iniciamos importando la librerías BeautifulSoup, Requests e importamos la configuración de nuestra archivo YAML. Para manipular documentos HTML en Python utilizamos BeautifulSoup, que permite analizar gramaticalmente dicho documento, y luego mediante código

poder realizar las consultas. Requests me permite realizar solicitudes HTTP a la web sin necesidad de tanta labor manual, haciendo que la integración con los servicios web sea mucho más fácil. No es necesario agregar manualmente consultas a las URLs o de convertir información a formularios para realizar una solicitud POST, GET, PUT, ETC. Para el caso de estudio se utilizó el verbo GET y como parámetro enviamos la URL, que devuelve un objeto de tipo respuesta.

La función _visit() recibe como parámetro la URL en consulta, su función es comunicarnos con la web, determinar el estado de la respuesta del servidor y enviarle como parámetro a la librería BeautifulSoup y extraer los datos del archivo HTML. La función _select recibe como parámetro una cadena de caracteres que contiene un selector CSS y ayuda a obtener información del árbol de nodos que BeautifulSoup almaceno en la variable _html

news_page_objects.py 1 ×

agencia > extract > news_page_objects.py > HomePage > __init__

```python
import bs4
import requests
from common import config


class NewsPage:
    def __init__(self, news_site_uid,url):
        self._config = config()['news_sites'][news_site_uid]
        self._queries = self._config['queries']
        self._html = None
        self._visit(url)
        self._url=url

    def _select(self, query_string): # funcion auxilia
        return self._html.select(query_string)


    def _visit(self, url):
        response = requests.get(url)
        response.raise_for_status()
        self._html = bs4.BeautifulSoup(response.text, 'html.parser')
```

Figura 22. Código para realzar solicitudes y manipular las etiquetas HTML

La clase HomePage(NewsPage) de la figura 23 es una instancia de la clase NewsPage, el cual tiene una función denominado article_links cuyo objetivo es

determinar si existe un enlace y almacenar en una lista (link_list) de todos los que tengan la propiedad href, el análisis y almacenamiento de los enlaces lo obtiene consultando a la variable homepage_article_link por cada URL del archivo YAML. Finalmente retornamos un conjunto de elementos únicos a través de la clase SET en Python.

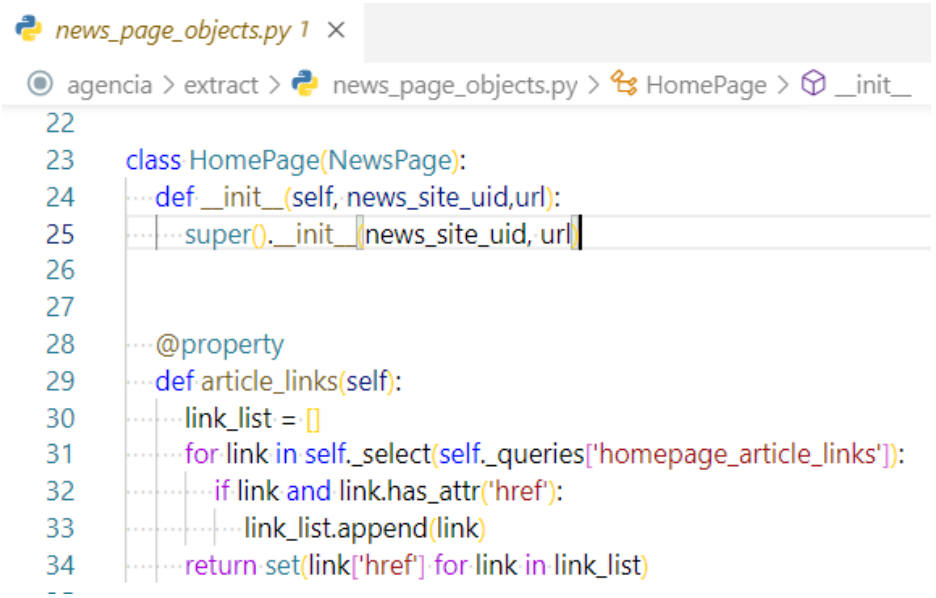


Figura 23. Código para extraer hipervínculos de los paquetes turísticos

La clase ArticlePage(NewsPage) ver figura 24 es también una instancia de la clase NewsPage, su finalidad es visitar cada hipervínculo y extraer, el nombre, la descripción, el itinerario, que servicios incluye, el correo electrónico, logo de la web y la imagen de portada por cada paquete turístico que visita el scraper.

Figura 24. Acopio de la información relevante del paquete turístico

El archivo principal denominado main.py es el encargado de realizar el raspado de la web(scraper) y de almacenar los datos en formato CSV, se inicia haciendo referencia al archivo news_page_objects y a nuestro archivo de configuración YAML

La función _news_scraper recibe como parámetro el identificador de cada sitio web, además, se inicializa la variable host con la URL de cada sitio web a ser analizado; la variable homepage obtiene todos los enlaces de todo los paquetes turísticos a los que se hicieron la consulta a través del archivo de configuración, homepage recibe como parámetro el identificador del sitio web y la URL, posteriormente itera en todos los vínculos encontrados en el homepage, accede y se determina si posee los campos requeridos en la clase ArticlePage

main.py 2 ×

agencia > extract > main.py > _save_articles > out_file_name

```
12  import news_page_objects as news
13
14  from common import config
15
16  logger = logging.getLogger(__name__)
17  is_well_formed_link = re.compile(r'^https?://.+/.+$')
18  is_well_formed_link2 = re.compile(r'^https?://.+/.+/.+$')
19  is_root_path = re.compile(r'^/.+$')
20
21  def _news_scraper(news_site_uid):
22      host = config()['news_sites'][news_site_uid]['url']
23      logging.info('Iniciando el scraper para {}'.format(host))
24      homepage = news.HomePage(news_site_uid, host)
25      articles = []
26      for link in homepage.article_links:
27          article = _fetch_article(news_site_uid,host,link)
28          if article:
29              logger.info('Articulo encontrado!!!')
30              articles.append(article)
31
32      _save_articles(news_site_uid, articles)
```

Figura 25. Código para la extracción de los paquetes turísticos

Todos los datos obtenidos se almacenan en una lista denominado article, y son enviados a la función _save_article los cuales almacenan en archivos CSV

Con el resultado de la extracción, ya contamos con datos que en su gran mayoría posee caracteres especiales, espaciados en blancos, tabulaciones, etc., los cuales se muestran en la figura 26

	body	incluye	itinerario	title	url
2	Incluye: Transporte por tierra y	lago, alimentaciÃ³n com	Incluye: Tran	Tour Isla Anapia OvernightAF	https://titicacalake.com/excursion/tour-isla-anapia-overnight/
3	Incluye: Transporte, guÃ-a, entradas al Templo de Santia		Incluye: Tran	Tour Pukara & LampaAPE - 22	https://titicacalake.com/excursion/tour-pukara-lampa/
4	La isla Tikonata se ubica en el e	traslado al muelleguÃ-a	La isla Tikona	Tour Tikonata OvernightAPE	https://titicacalake.com/excursion/tour-tikonata-overnight/
5	Incluye: almuerzo y cena del primer dÃ-a, desayuno y a		Incluye: alm	Tour Taquile overnight + Ama	https://titicacalake.com/excursion/tour-taquile-overnight-amantani-o-v-v/
6	La excursiÃ³n a Taquile o Aman	Traslado hacia y del mue	La excursiÃ³	Tour Taquile o Amantani o Ll	https://titicacalake.com/excursion/tour-taquile-o-amantani-o-llachon-lancha-rapida/
7	UBICACIÃ"N:A ORILLAS DEL LAG	Actualmente en la comu	UBICACIÃ"N	Luquina chico 2D/1N	https://titicacalake.com/excursion/luquina-chico-2d-1noche/
8	Incluye: entradas, guÃ-a, transporte hacia y del muelle		Incluye: entr	Tour LlachÃ³n + Uros Kayak (L	https://titicacalake.com/excursion/tour-llachon-uros-kayak-lancha-normal/
9	Esta NecrÃ³polis de origen Pre-	entradasguÃ-atransport	Esta NecrÃ³p	Tour SillustaniAPE - 205	https://titicacalake.com/excursion/tour-sillustani/
10	En esta excursiÃ³n visitamos: E	entradasguÃ-atransport	En esta excu	City tour PunoAPE - 201	https://titicacalake.com/excursion/city-tour-puno/
11	La misma excursiÃ³n que la anterior, con la diferencia q		La misma ex	Puno â€" Isla del Sol â€" La Pa	https://titicacalake.com/excursion/puno-isla-del-sol-la-paz/
12	Saliendo de la ciudad de Puno,	TransporteguÃ-aentrada	Saliendo de	Tour ChucuitoAPE - 212	https://titicacalake.com/excursion/tour-chucuito/
13	Este es un cementerio Inca, ub	TransporteguÃ-aentrada	Este es un ce	Tour CutimboAPE - 215	https://titicacalake.com/excursion/tour-cutimbo/
14	A 75 kilÃ³metros al norte de la	TransporteguÃ-aentrada	A 75 kilÃ³me	Tour PukaraAPE - 217	https://titicacalake.com/excursion/tour-pukara/
15	Tiahuanaco estÃ¡ ubicado a 15	TransporteguÃ-aalmuer	Tiahuanaco e	Tour TiahuanacoAPE - 227	https://titicacalake.com/excursion/tour-tiahuanaco/
16	A orillas del Lago Titikaka y 50 k	transporteguÃ-a.	A orillas del	Titilaca o CharcasAPE - 216	https://titicacalake.com/excursion/titilaca-o-charcas/
17	Incluye: Lancha rÃ¡pida, guÃ-a, traslados al muelle, ent		Incluye: Lanc	Tour Taquile + Amantani (Lar	https://titicacalake.com/excursion/tour-taquile-amantani-lancha-rapida-full-day/
18	Esta visita tiene la particularida	almuerzo y cena del prir	Esta visita tie	Tour Taquile o Amantani â€"	https://titicacalake.com/excursion/tour-taquile-o-amantani-overnight/
19	Las Islas Flotantes de Los Uros,	guÃ-atransporte hacia y	Las Islas Flot	Tour UrosAPE - 202	https://titicacalake.com/excursion/tour-uros/
20	Incluye: Traslado hacia y del muelle de Puno, guÃ-a, lan		Incluye: Tras	Tour Uros â€" Taquile (Lancha	https://titicacalake.com/excursion/tour-uros-taquile-lancha-rapida-guia-pc/

Figura 26. Datos extraídos de las agencias de viaje

Se puso a prueba la eficiencia del algoritmo para la extracción de los datos, comparándolo con un software gratuito desarrollado para scrapear paginas web, se utilizó la extensión de Google Chrome denominado Webscraper.io, el cual posee una interface web gráfica, se configura los selectores de la web a ser extraídos, en la figura 27 se muestra el instalador de la extensión en la tienda de aplicaciones de Google Chrome

Figura 27. Extensión de Google Chrome denominado Webscraper

Una vez instalado y configurado la extensión se procedió a registrar y analizar el tiempo promedio (en minutos) que se toma para extraer los datos y la cantidad de paquetes, en la figura 28 muestra el resultado de la extracción de los datos haciendo uso Webscraper.io.

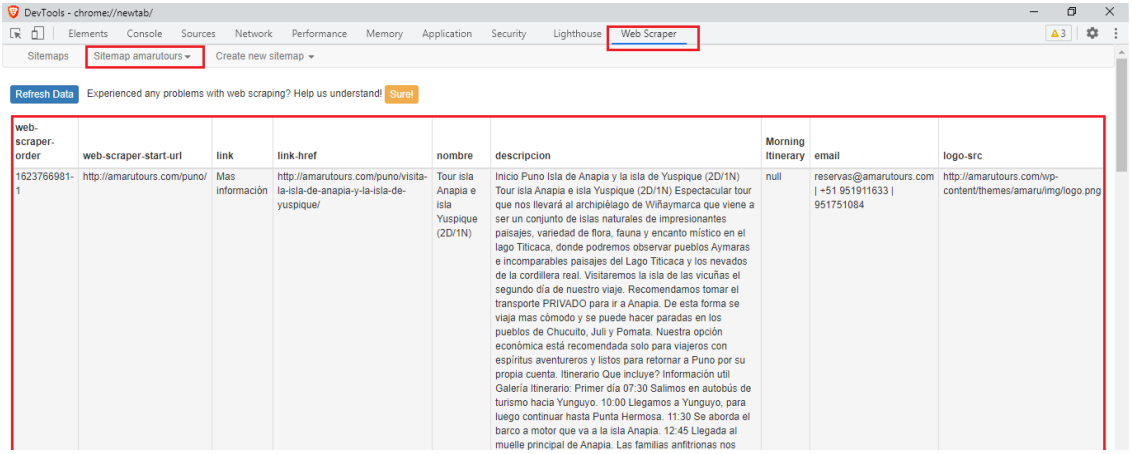


Figura 28. Extracción de datos con wenscraper.io

Para efectos de prueba se configuro tres agencias de viajes dentro de Webscraper.io y se hizo el raspado de los datos, en la Tabla 3, se puede visualizar que Webscraper.io le tomas más tiempo, además la cantidad de paquetes turísticos extraídos es menor en comparación en el algoritmo de extracción propuesto en la presente investigación.

Tabla 3

Comparación de tiempo y los datos extraídos

Agencia	Extensión Web Scraper		Extracción por el algoritmo	
	Tiempo promedio en min.	Cantidad de paquetes extraídos	Tiempo promedio en min.	Cantidad de paquetes extraídos
Arcobaleno	01:43:02	27	00:29:45	31
Andespuno	00:36:29	10	00:17:45	12
Amarutours	00:16:19	06	00:11:74	09

B. Análisis de la complejidad algorítmica

Se realizo el análisis de la complejidad algorítmica con la notación Big O del código para la extracción de todos los enlaces de los paquetes turísticos encontrados en una página web (ver figura 29).

news_page_objects.py 1

agencia > extract > news_page_objects.py > ...

```
23  class HomePage(NewsPage):
24      def __init__(self, news_site_uid,url):
25          super().__init__(news_site_uid, url)
26
27
28      @property
29      def article_links(self):
30          link_list = []      #O(1)
31          for link in self._select(self._queries['homepage_article_links']):  #O(n)
32              if link and link.has_attr('href'):  #O(2)
33                  link_list.append(link)  # O(1)
34          return set(link['href'] for link in link_list)  # O(1)
35
```

Figura 29. Complejidad algorítmica para la obtención de hipervínculos

- La línea de código número 30 se crea una lista vacía, el cual tiene una valoración de O(1) por que actúa como una constante y siempre va inicializar en una lista vacía cada que se invoque a la función article_links
- El ciclo for de la línea de código numero 31 itera dependiendo de cuantos enlaces va recorrer y eso equivale a un O(n)
- La línea de código numero 31 realiza dos comparaciones las cuales va devolver un verdadero o un falso, en el peor de los casos de que ambos fueran verdaderos realiza dos acciones eso equivale a un O(2)
- La línea de código numero 33 agrega elementos a la lista vacía, eso equivale a O(1)
- La línea de código numero 34 va a retornar un valor, lo cual equivale O(1)

Realizamos la sumatoria

$$O(G) = O(1) + O(n) + O(2) + O(1) + O(1)$$
$$O(G) = O(n+ 5)$$
$$O(G) = O(n)$$

Por lo tanto, podemos determinar que la complejidad algorítmica para la extracción de los hipervínculos de los paquetes turísticos es lineal O(n) y el tiempo de ejecución es cada vez mayor de modo proporcional a cómo se incrementa el tamaño de los enlaces encontrados en cada página web.

También se realizó el análisis de la complejidad algorítmica con la notación Big O del código para la extracción de todos los datos obtenidos por agencia de viaje (ver figura 30).

main.py 2 ×

agencia > extract > main.py > ...

```
20
21  def _news_scraper(news_site_uid):
22      host = config()['news_sites'][news_site_uid]['url'] #O(1)
23      logging.info('Iniciando el scraper para {}'.format(host))
24      homepage = news.HomePage(news_site_uid, host) #O(1)
25      articles = [] #O(1)
26      for link in homepage.article_links: #O(n)
27          article = _fetch_article(news_site_uid,host,link) #O(1)
28          if article: #O(1)
29              logger.info('Articulo encontrado!!!')
30              articles.append(article) #O(1)
31
32      _save_articles(news_site_uid, articles) #O(1)
```

Figura 30. Complejidad algorítmica para la extracción de los datos

- La línea de código 22 se asigna un valor a la variable host cada que se invoque a la función _news_scraper, por lo que equivale a un O(1)
- La línea de código 24 se asigna un valor a la variable homepage cada que se invoque a la función _news_scraper, por lo que equivale a un O(1)
- La línea de código número 25 se creando una lista vacía, el cual tiene una valoración de O(1) siempre va inicializar en una lista vacía cada que se invoque a la función _news_scraper
- La línea de código número 26 existe un ciclo for el cual va iterar dependiendo de cuantos enlaces va recorrer y eso equivale a un O(n)
- La línea de código número 28 realiza una comparación el cual devolverá un verdadero o un falso, en el peor de los casos de que sea siempre sea verdadero realiza una acciones eso equivale a un O(1)
- La línea de código numero 30 agrega elementos a la lista vacía, eso equivale a O(1)

- La línea de código numero 32 envía un elemento a la función _save_articles, eso equivale a O(1)

Realizando la sumatoria

O(G) = O(1) + O(1) + O(1) + O(n) + O(1) + O(1) + O(1)

O(G) = O(n+ 6)

O(G) = O(n)

Por lo tanto, podemos determinar que la complejidad algorítmica para la extracción de los datos de las páginas web de las agencias de viaje es lineal O(n), el tiempo de ejecución de la función _news_scraper va ser proporcional a la cantidad de páginas web a extraer

C. Transformación de datos

Luego de obtener nuestro data set, se debe realizar la limpieza y transformación de los datos, esto implica eliminar filas con valores nulos, buscar valores redundantes, corregir palabras con caracteres especiales, eliminar tabulaciones y saltos de línea.

Para ello utilizamos la librería de pandas en Python; este paquete otorga facilidades para realizar un Data Wranglin o un domado de datos, esto quiere decir que proporciona herramientas que permiten:

- Leer y escribir datos en diferentes formatos: CSV, Microsoft Excel, bases SQL y formato HDF5
- Seleccionar y filtrar de manera sencilla tablas de datos en función de posición, valor o etiquetas
- Fusionar y unir datos, entro otros

Para el caso de estudio se implementó función para:

- Leer archivos CSV
- Generar nuevas columnas con los datos obtenidos (host, ID)
- Eliminar filas con valores vacíos
- Eliminar saltos de línea (\n)
- Eliminar valores duplicados
- Almacenar datos en otro archivo CSV

En la figura 31 muestra algunas funciones para realizar la limpieza y transformación

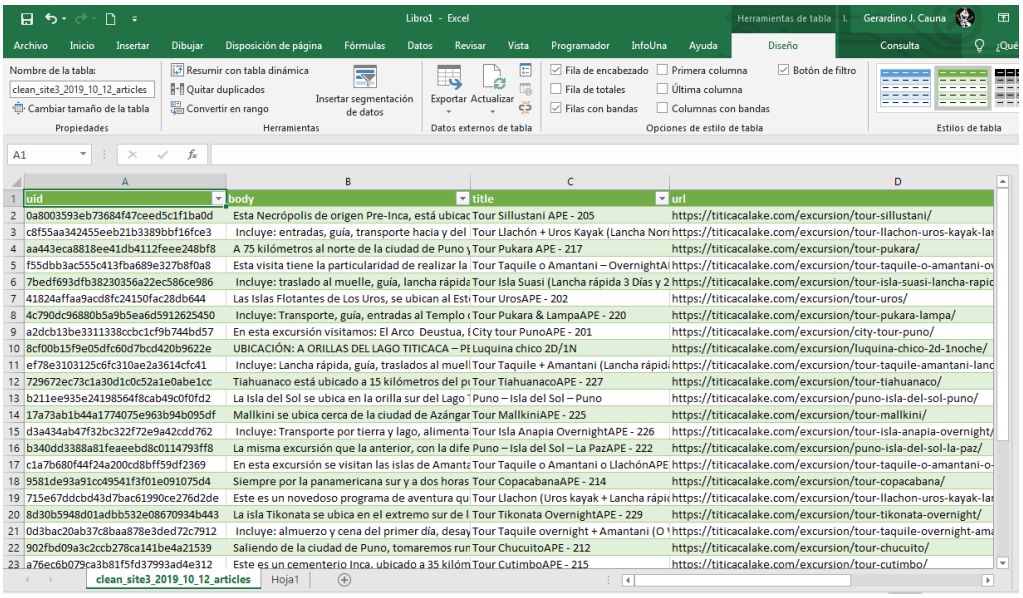
main.py

agencia > transform > main.py > ...

```python
def _read_data(filename):
    logger.info('Leendo archivo {}'.format(filename))
    return pd.read_csv(filename)

def _extract_agency_uid(filename):
    logger.info('Extraendo el Sitio por uid')
    agency_uid = filename.split('_')[0]

    logger.info('Sitio uid detectado: {}'.format(agency_uid))
    return agency_uid

def _extract_host(df):
    logger.info('Extraendo el host de las urls')
    df['host'] = df['url'].apply(lambda url: urlparse(url).netloc)
    return df

def _drop_rows_with_missing_values(df):
    logger.info('Eliminando filas con valores vacios')
    return df.dropna()
```

Figura 31. Limpieza y transformación de datos

Al terminar el proceso de limpieza y transformación de datos, el software genera otro archivo CSV con los valores deseados.



uid	body	title	url
0a8003593eb73684f47ceed5c1f1ba0d	Esta Necrópolis de origen Pre-Inca, está ubicac	Tour Sillustani APE - 205	https://titicacalake.com/excursion/tour-sillustani/
c8f55aa342455eeb21b3389bbf16fce3	Incluye: entradas, guía, transporte hacia y del	Tour Llachón + Uros Kayak (Lancha Nor	https://titicacalake.com/excursion/tour-llachon-uros-kayak-la
aa443eca8818ee41db4112feee248bf8	A 75 kilómetros al norte de la ciudad de Puno y	Tour Pukara APE - 217	https://titicacalake.com/excursion/tour-pukara/
f55dbb3ac555c413fba689e327b8f0a8	Esta visita tiene la particularidad de realizar la	Tour Taquile o Amantani – OvernightA	https://titicacalake.com/excursion/tour-taquile-o-amantani-o
7bedf693dfb38230356a22ec586ce986	Incluye: traslado al muelle, guía, lancha rápida	Tour Isla Suasi (Lancha rápida 3 Días y 2	https://titicacalake.com/excursion/tour-isla-suasi-lancha-rapic
41824affaa9acd8fc24150fac28db644	Las Islas Flotantes de Los Uros, se ubican al Est	Tour UrosAPE - 202	https://titicacalake.com/excursion/tour-uros/
4c790dc96880b5a9b5ea6d5912625450	Incluye: Transporte, guía, entradas al Templo (	Tour Pukara & LampaAPE - 220	https://titicacalake.com/excursion/tour-pukara-lampa/
a2dcb13be3311338ccbc1cf9b744bd57	En esta excursión visitamos: El Arco Deustua, (	City tour PunoAPE - 201	https://titicacalake.com/excursion/city-tour-puno/
8cf00b15f9e05dfc60d7bcd420b9622e	UBICACIÓN: A ORILLAS DEL LAGO TITICACA – PE	Luquina chico 2D/1N	https://titicacalake.com/excursion/luquina-chico-2d-1noche/
ef78e3103125c6fc310ae2a3614cfc41	Incluye: Lancha rápida, guía, traslados al muell	Tour Taquile + Amantani (Lancha rápid	https://titicacalake.com/excursion/tour-taquile-amantani-lanc
729672ec73c1a30d1c0c52a1e0abe1cc	Tiahuanaco está ubicado a 15 kilómetros del pu	Tour TiahuanacoAPE - 227	https://titicacalake.com/excursion/tour-tiahuanaco/
b211ee935e24198564f8cab49c0f0fd2	La Isla del Sol se ubica en la orilla sur del Lago	Puno – Isla del Sol – Puno	https://titicacalake.com/excursion/puno-isla-del-sol-puno/
17a73ab1b44a1774075e963b94b095df	Mallkini se ubica cerca de la ciudad de Azángar	Tour MallkiniAPE - 225	https://titicacalake.com/excursion/tour-mallkini/
d3a434ab47f32bc322f72e9a42cdd762	Incluye: Transporte por tierra y lago, alimenta	Tour Isla Anapia OvernightAPE - 226	https://titicacalake.com/excursion/tour-isla-anapia-overnight/
b340dd3388a81feaeebd8c0114793ff8	La misma excursión que la anterior, con la dife	Puno – Isla del Sol – La PazAPE - 222	https://titicacalake.com/excursion/puno-isla-del-sol-la-paz/
c1a7b680f44f24a200cd8bff59df2369	En esta excursión se visitan las islas de Amanta	Tour Taquile o Amantani o LlachónAPE	https://titicacalake.com/excursion/tour-taquile-o-amantani-o-
9581de93a91cc49541f3f01e091075d4	Siempre por la panamericana sur y a dos horas	Tour CopacabanaAPE - 214	https://titicacalake.com/excursion/tour-copacabana/
715e67dcbcd43d7bac61990ce276d2de	Este es un novedoso programa de aventura qu	Tour Llachon (Uros kayak + Lancha rápi	https://titicacalake.com/excursion/tour-llachon-uros-kayak-la
8d30b5948d01adbb532e08670934b443	La isla Tikonata se ubica en el extremo sur de l	Tour Tikonata OvernightAPE - 229	https://titicacalake.com/excursion/tour-tikonata-overnight/
0d3bac20ab37c8baa878e3ded72c7912	Incluye: almuerzo y cena del primer día, desay	Tour Taquile overnight + Amantani (O	https://titicacalake.com/excursion/tour-taquile-overnight-ama
902fbd09a3c2ccb278ca141be4a21539	Saliendo de la ciudad de Puno, tomaremos run	Tour ChucuitoAPE - 212	https://titicacalake.com/excursion/tour-chucuito/
a76ec6b079ca3b81f5fd37993ad4e312	Este es un cementerio Inca, ubicado a 35 kilóm	Tour CutimboAPE - 215	https://titicacalake.com/excursion/tour-cutimbo/

Figura 32. CSV con datos transformados

D. Exportación de datos a SQLite

Para generar la indexación de los sitios web de las agencias de viaje que ofertar paquetes turísticos de la región de Puno, debemos tener nuestro data set previamente transformado, en un formato que puede ser gestionado por un sistema gestor de base de datos, para el presente trabajo de investigación se utilizo SQLite, que implementa un motor de base de datos SQL pequeño , rápido , autónomo , de alta confiabilidad y con todas las funciones, además es de dominio público y libre para cualquier uso, ya sea comercial o privado.

En la figura 33 muestra la importación de los módulos en este caso SQLAlchemy, que es el conjunto de herramientas Python SQL y Object Relational Mapper (Mapeo Objeto-relacional) que brinda las bondades y flexibilidad de SQL.

base.py

agencia > load > base.py > ...

```
1  from sqlalchemy import create_engine
2  from sqlalchemy.ext.declarative import declarative_base
3  from sqlalchemy.orm import sessionmaker
4
5  engine = create_engine('sqlite:///agencia.db')
6
7  Session = sessionmaker(bind=engine)
8
9  Base = declarative_base()
```

Figura 33. Configuración básica para exportar a SQLite

Una vez realizada la conexión a la base de datos, en este caso corresponde a SQLite, por medio de un archivo, mapeamos las tablas como clases, utilizando los tipos de datos de sqlalchemy para expresar los diferentes tipos de datos de SQLite: Integer, String ver figura 34.

article.py 1 ×

agencia > load > article.py > Article

```python
from sqlalchemy import Column, String, Integer
from base import Base

class Article(Base):
    __tablename__ = 'articles'

    id = Column(String, primary_key=True)
    body = Column(String)
    incluye = Column(String)
    itinerario = Column(String)
    host = Column(String)
    title = Column(String)
    imagen = Column(String)
    newspaper_uid = Column(String)
    url = Column(String,unique=True)
```

Figura 34. Configuración de las columnas de la base de datos

4.2 Resultados conforme al objetivo específico 2

Para la implementación de la página web, se utilizó CodeIgniter que es un entorno de desarrollo web escrito en PHP, cuya base es el patrón Modelo-Vista-Controlador (MVC) y permite diseñar software de forma flexible, ya que se pueden substituir, editar y reutilizar los módulos individuales de programación muy fácilmente.

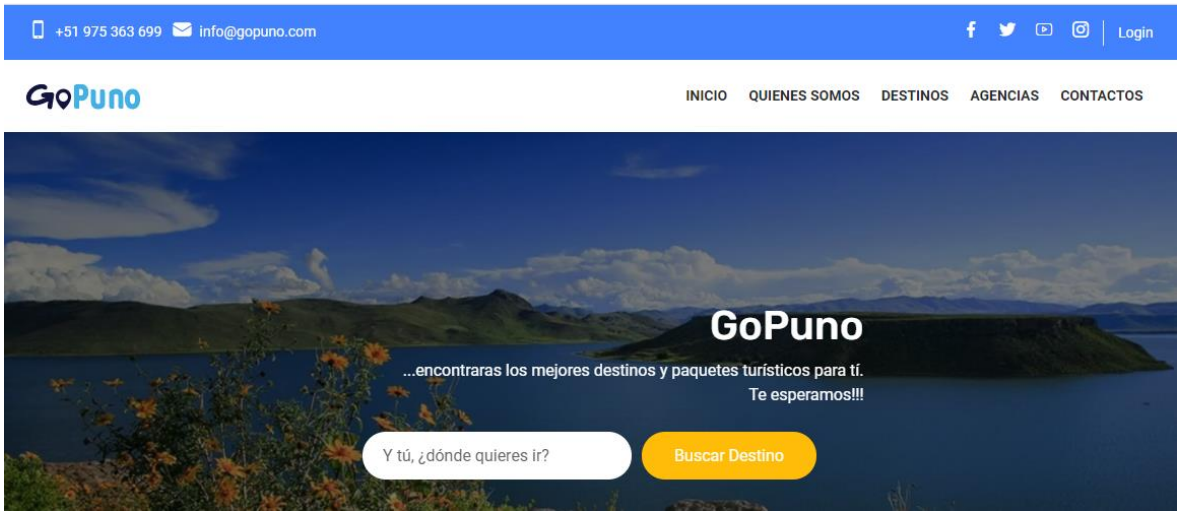


Figura 35. Ventana principal de sitio web

En la figura 35 muestra la ventana principal de la página web, el cual posee un menú principal, un buscador, categorías de los destinos más populares de la región de Puno y la lista de agencias con las que se trabajó.

En la opción destinos se muestra toda la información extraída de las diferentes páginas web de las agencias de viaje, cada una de ellas con el enlace a la web oficial para ampliar la información.

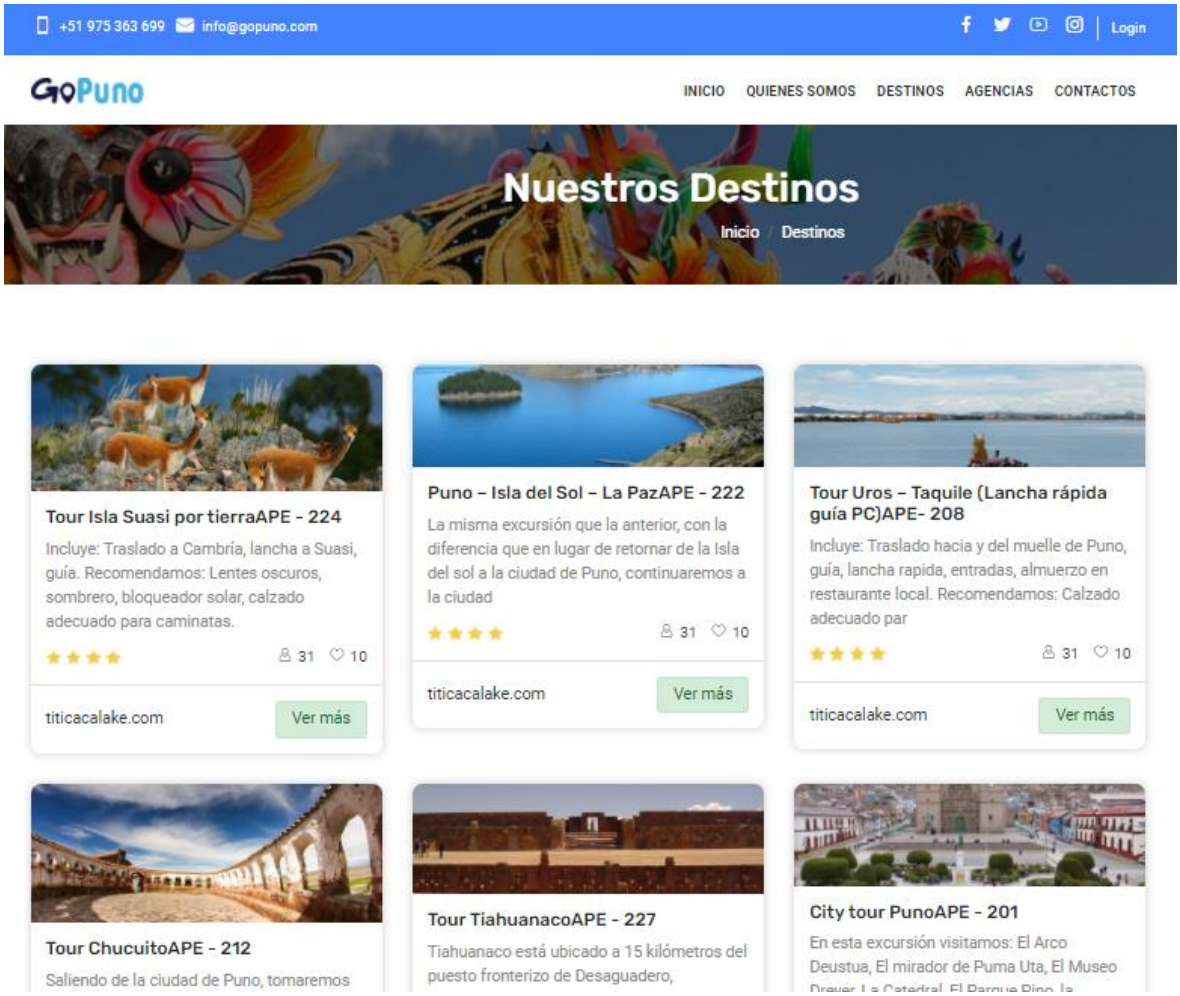


Figura 36. Paquetes turísticos extraídos

Primeramente, muestra la página web oficial, seguida del nombre del paquete con una pequeña descripción y al darle clic en el botón ver más muestra toda la información extraída, entre ellas: descripción, itinerario y que incluye el paquete turístico, además muestra un enlace a la web oficial, y datos de contacto

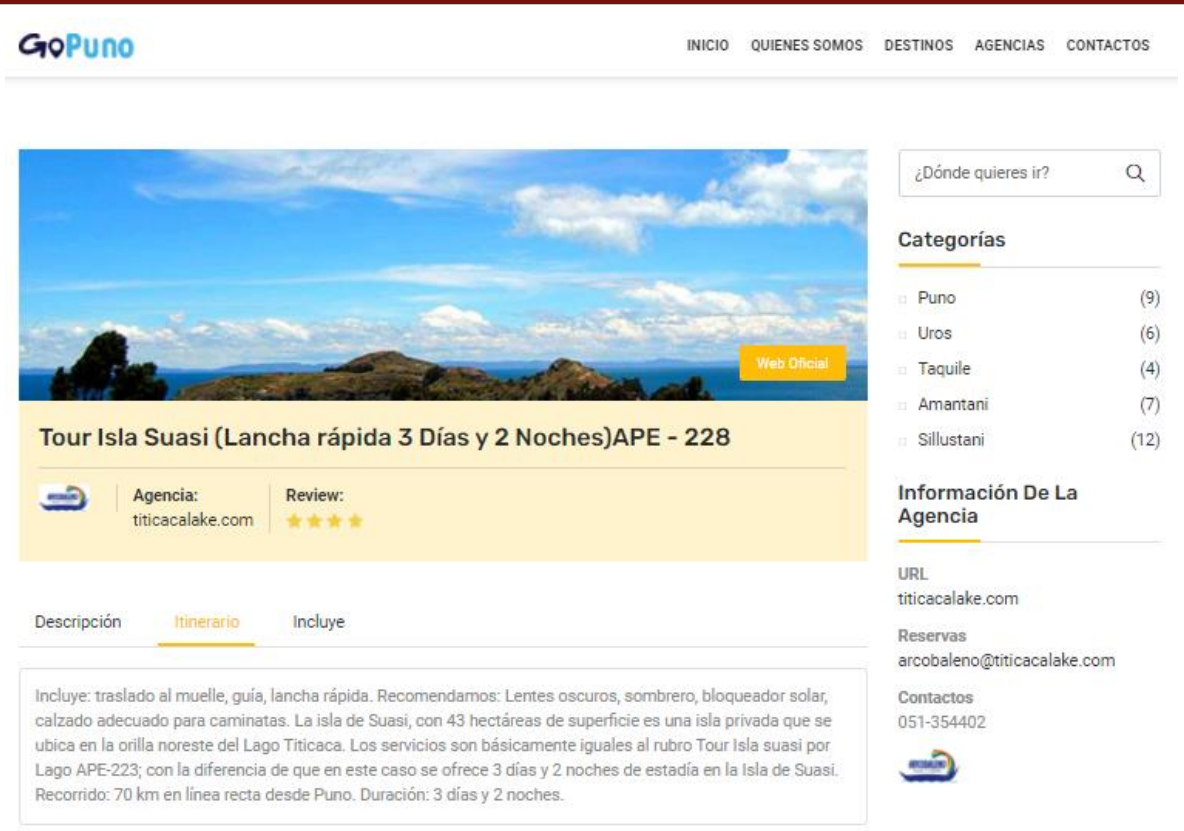
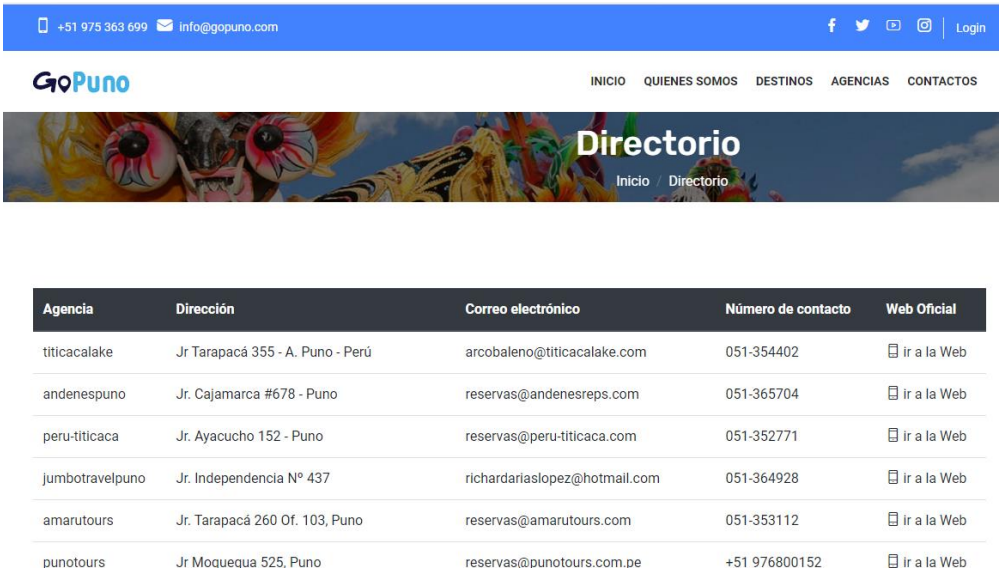


Figura 37. Descripción detallada del paquete turístico

La opción agencias en menú principal, muestra el directorio de cada una de ellas, al darle clic en el enlace muestra todos los paquetes turísticos asociados a la agencia.



Agencia	Dirección	Correo electrónico	Número de contacto	Web Oficial
titicacalake	Jr Tarapacá 355 - A. Puno - Perú	arcobaleno@titicacalake.com	051-354402	ir a la Web
andenespuno	Jr. Cajamarca #678 - Puno	reservas@andenesreps.com	051-365704	ir a la Web
peru-titicaca	Jr. Ayacucho 152 - Puno	reservas@peru-titicaca.com	051-352771	ir a la Web
jumbotravelpuno	Jr. Independencia Nº 437	richardariaslopez@hotmail.com	051-364928	ir a la Web
amarutours	Jr. Tarapacá 260 Of. 103, Puno	reservas@amarutours.com	051-353112	ir a la Web
punotours	Jr Moquegua 525, Puno	reservas@punotours.com.pe	+51 976800152	ir a la Web

Figura 38. Directorio de agencias

Para determinar el rendimiento del sitio web en dispositivos móviles como en ordenadores utilizaremos la API de PageSpeed Insights (PSI) de Google el cual nos ofrece

seis métricas que miden muchos de los diversos aspectos del rendimiento relevantes para los usuarios.

- **First Contentful Paint (FCP):** mide cuánto tiempo le toma al navegador procesar la primera parte del contenido DOM después de que un usuario navega a la página web.

Tabla 4

Puntuación FCP

Tiempo FCP (en segundos)	**Código de colores**	**Puntuación FCP (percentil de archivo HTTP)**
0-2	Verde (rápido)	75-100
2-4	Naranja (moderado)	50-74
Más de 4	Rojo (lento)	0-49

Fuente: (PageSpeed Insights, 2021)

Luego de realizar el análisis se obtuvo un resultado de 1.8 segundos para dispositivos móviles y 0.5 para ordenadores, en ambos casos está dentro de la categoría rápido (color verde), el cual indica que no le toma mucho tiempo al navegador procesar el contenido DOM desde que comenzó la navegación hasta que el contenido principal de la página se muestre en la pantalla.

- **Speed index:** mide la rapidez con la que se muestra visualmente el contenido durante la carga de la página.

Tabla 5

Puntuación del Speed index

Tiempo FCP (en segundos)	Código de colores	Puntuación del índice de velocidad
0-4.3	Verde (rápido)	75-100
4.4-5.8	Naranja (moderado)	50-74
Más de 5.8	Rojo (lento)	0-49

Fuente: (PageSpeed Insights, 2021)

Luego de realizar el análisis se obtuvo un resultado de 2.9 segundos para dispositivos móviles y 1.1 para ordenadores, en ambos casos está dentro de la categoría rápido (color verde), lo cual indica el tiempo en mostrar el contenido del sitio web

- **Largest Contentful Paint(LCP):** informa el tiempo de renderizado de la imagen o el bloque de texto más grande visible dentro de la ventana gráfica.

Tabla 6

Puntuación Largest Contentful Paint

Tiempo FCP (en segundos)	Código de colores	Puntuación del índice de velocidad
0-4.3	Verde (rápido)	75-100
4.4-5.8	Naranja (moderado)	50-74
Más de 5.8	Rojo (lento)	0-49

Fuente: (PageSpeed Insights, 2021)

Como resultado se obtuvo 2.3 segundos en dispositivos móviles y 0.8 segundos para ordenadores. tiempo desde que la página comienza a cargarse hasta que el bloque de texto o elemento de imagen más grande se representa en la pantalla, en ambos casos están en la categoría rápido (color verde)

- **Time to Interactive (TTI):** mide el tiempo que tarda una página en volverse completamente interactiva.

Tabla 7

Puntuación Time to Interactive

Métrica TTI (en segundos)	Código de colores
0 – 3.8	Verde (rápido)
3.9 – 7.3	Naranja (moderado)
Más de 7.3	Rojo (lento)

Fuente: (PageSpeed Insights, 2021)

Como resultado se obtuvo 2.3 segundos en dispositivos móviles y 0.6 segundos para ordenadores, ambos casos están en la categoría rápido (color verde), lo que indica que la página muestra contenido útil, los controladores de eventos están registrados para la mayoría de los elementos visibles y la página responde a las interacciones del usuario en 50 milisegundo

- **Total Blocking Time:** mide la cantidad total de tiempo que una página está bloqueada para que no responda a la entrada del usuario, como los clics del mouse, los toques de la pantalla o las pulsaciones del teclado.

Tabla 8

Puntuación Total Blocking Time

Tiempo TBT (en milisegundos)	Código de colores
0 – 300	Verde (rápido)
300 – 600	Naranja (moderado)
Más de 600	Rojo (lento)

Fuente: (PageSpeed Insights, 2021)

Como resultado se obtuvo 90ms para dispositivos móviles y 0ms para ordenadores, ambos están dentro de la categoría rápido (color verde)

- **Cumulative Layout Shift (CLS):** mide la suma total de todas las puntuaciones de cambio de diseño individuales para cada cambio de diseño inesperado que se produce durante toda la vida útil de la página. Para proporcionar una buena experiencia de usuario, los sitios deben esforzarse por tener una puntuación CLS de menos de 0.1, para nuestro caso de estudio se obtuvo 0.001 para dispositivos móviles y 0.0 para ordenadores.

El informe que generó PSI muestra una puntuación global que resume el rendimiento del sitio web GoPuno, la figura 39 muestra la puntuación para ordenadores y la figura 40 muestra la puntuación en dispositivos móviles.

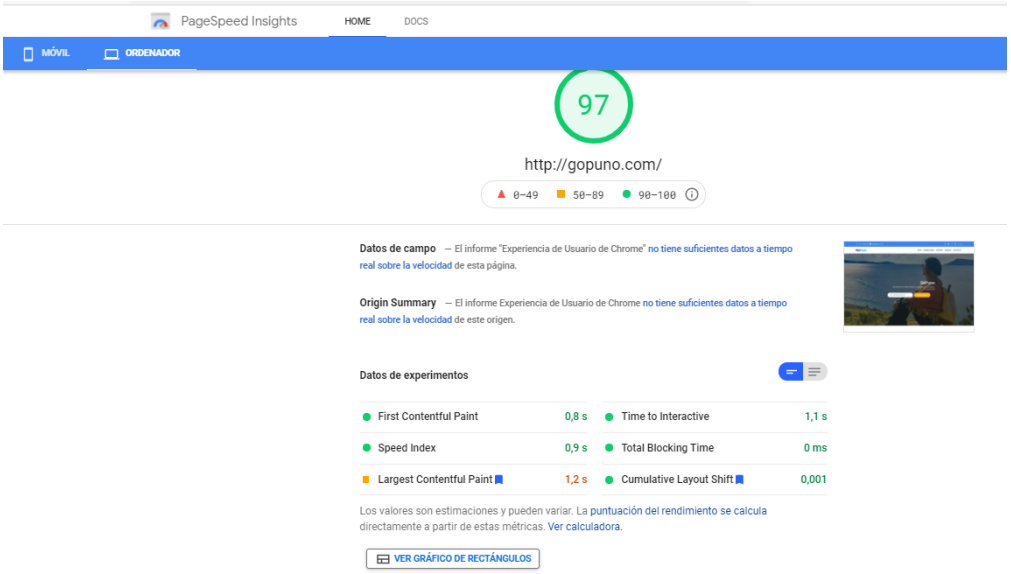


Figura 39. Puntuación global del rendimiento para ordenadores

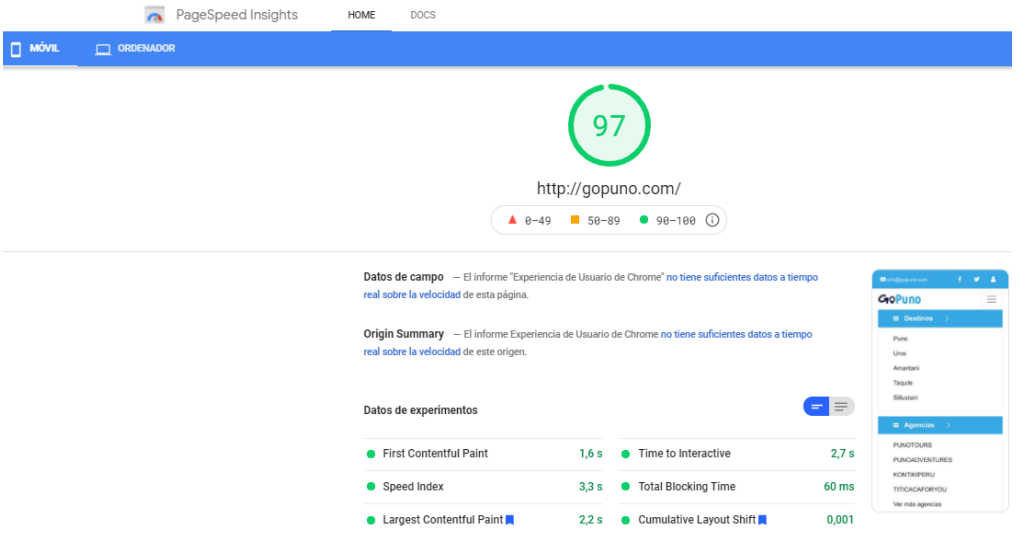


Figura 40. Puntuación global del rendimiento para dispositivos móviles

Del desempeño de nuestro sitio web para dispositivos móviles y ordenadores según la puntuación global está en la categoría rápida (color verde), el cual garantiza una buena experiencia de usuario.

Tabla 9

Puntuación de rendimiento

Puntuación de rendimiento	**Código de colores**
90 – 100	Verde (rápido)
50 – 89	Naranja (normal)
0 – 49	Rojo (lento)

Fuente: (PageSpeed Insights, 2021)

4.3 Resultados conforme al objetivo específico 3

Para la obtención de los resultados se selecciona a 10 participantes, entre ellos Figuran:

- 01 gerente de una agencia de viajes
- 02 profesionales en turismo
- 07 personas naturales

Para la selección de los participantes se utilizó la técnica del muestreo por conveniencia para recabar opiniones y establecer el grado en que en que el producto de software satisface a un usuario final basado en la norma ISO/IEC 25000.

Tabla 10

Métricas de calidad de uso

Característica	**Sub característica**	**Métricas**	**Significado**
Efectividad	Efectividad	Completitud de tarea	Cantidad de tareas de forma completa y satisfactoria.
		Efectividad de la tarea	Cantidad de objetivos completados

		Tiempo de la tarea	Tiempo real en que completa una tarea
Eficiencia	Eficiencia	Tiempo relativo de la tarea	Tiempo que necesita un usuario nuevo para completar una tarea
		Eficiencia de la tarea	Eficiencia de los usuarios
		Nivel de satisfacción	Satisfacción del usuario
Satisfacción	Utilidad	Uso discrecional de las funciones	Número de veces que el usuario hace uso de las funcionalidades
		% de quejas de los clientes	% de quejas realizadas por los clientes

En la Tabla 11, se detalla la ponderación en porcentaje que las características de calidad en uso para evaluar el sitio web.

Tabla 11

Ponderación de las características de calidad en uso

Características	Nivel de importancia	Ponderación	Motivación de ponderación
Efectividad	A	30%	Se pondera con 30% porque es necesario evaluar si la página web permite al usuario alcanzar sus objetivos o necesidades

Eficiencia	A	30%	Se pondera con 30% porque es necesario evaluar si la página web permite al usuario alcanzar sus objetivos o necesidades haciendo uso de recursos mínimos.
Satisfacción	A	40%	Se pondera con 40% porque es necesario evaluar si la página web al ser utilizado satisface las necesidades del usuario.
Libertad de riesgo	B	0%	Se pondera con 0% porque no es necesario evaluar
Cobertura de contexto	B	0%	Se pondera con 0% porque no es necesario evaluar

Para fines de este estudio se estableció una escala numérica propia que se asignó a los cuatro niveles de puntuación y tres grados de satisfacción propuestos en la ISO/IEC 25040 como niveles de puntuación final.

Tabla 12

Niveles de Puntuación final

Escala de medición	Niveles de puntuación	Grado de satisfacción
8.76 – 10.00	Cumple con los requisitos	Muy satisfactorio
5.10 – 8.75	Aceptable	Satisfactorio
2.76 – 5.00	Mínimamente aceptable	Insatisfactorio
0.00 – 2.75	Inaceptable	

Fuente: (ISO/IEC 25040, 2011)

Con la finalidad de disponer de un instrumento para el registro, cálculo y análisis de calidad del producto software, se elaboró una matriz en formato Excel, que reúne los pasos anteriores descritos en una hoja de cálculo. La figura 41 se muestra la hoja de cálculo diseñada para evaluación de la Calidad de Uso.

Característica	Subcaracterística	Métrica	Fórmula / Variable	Valor deseado	Aplica	Valor Obtenido	Ponderación	Valor parcial total (/10)	Nivel de importancia	Porcentaje de Importancia	Valor Final
Efectividad	Efectividad	Completitud de la tarea	X = A/B A = Número de tareas completadas. B = Número total de tareas intentadas. Dónde B > 0	1	SI	A = 6 B = 6 X = 1	10	10	A	30%	3
		Efectividad de la tarea	X = A/B A = Cantidad de objetivos completados por la tarea. B = Cantidad de objetivos planteados por la tarea. Dónde B > 0	1	SI	A = 1 B = 1 X = 1	10				
		Tiempo de busqueda de	X = A/B A = Timepo planteado (Min).	1	SI	A = 2 B = 3	6.7				

Figura 41. Matriz de calidad en uso

De acuerdo a los valores obtenidos de las características de calidad en uso evaluadas, se tuvo un resultado satisfactorio, lo que indica que el nivel de uso del sitio web el usuario se encuentra satisfecho con la utilización.

Tabla 13

Valor total obtenido de la calidad en uso.

Característica	Valor parcial total (/10)	Nivel de importancia	Porcentaje de importancia	Valor final	Calidad del Sistema
Efectividad	10	A	30%	3	
Eficiencia	5.57	A	30%	1.67	6.67
Satisfacción	5	A	40%	2	

4.4 Hipótesis para prueba de los rangos con signo de Wilcoxon

Realizando prueba de hipótesis para determinar la optimización de la busque da paquetes turísticos de la región de Puno, dos medias muéstrales y poder así sustentar uno de los

objetivos específicos planteado: Analizar el tiempo antes y después de la implementación del sitio web.

Tabla 14

Búsqueda de paquetes (en minutos) turísticos antes y después

Individuo	Antes (y)	Después (x)	Diferencia (y -x)
1	2.6	1.08	1.52
2	2.09	1.26	0.83
3	2.44	1.5	0.94
4	3.01	1.32	1.69
5	5.75	2.01	3.74
6	2.87	1.21	1.66
7	3.1	2.67	0.43
8	4.55	2.9	1.65
9	2.63	1.08	1.55
10	3.2	1.5	1.7
Totales	32.24	16.53	15.71

Tabla 15

Prueba de la normalidad

Medidas	**Búsqueda de paquetes turísticos**		**Diferencia**
	Antes	Después	
Media	3.22	1.65	1.57
Desviación estándar	1.10	0.66	0.88
Varianza	1.22	0.43	0.78
Shapiro-Wilk (P-valor)	0.021	0.020	**0.015**

Como P-valor (0.015) < 0.05, la diferencia entre la búsqueda de paquetes turísticos antes y después de implementar el sitio web no tiene una distribución normal, por lo tanto, se utilizará un estadístico no paramétrico

i. Planteamiento de hipótesis

$\mathbf{H_0}$: $\boldsymbol{\mu_x \geq \mu_y}$ (Con la implementación del sitio web. no se reduce el tiempo en la búsqueda de paquetes turísticos de la región de Puno basado en web Scraping).

$\mathbf{H_a}$: $\boldsymbol{\mu_x < \mu_y}$ **(**Con la implementación del sitio web. si se reduce el tiempo en la búsqueda de paquetes turísticos de la región de Puno basado en web Scraping).

ii. Nivel de significancia

Se usará un nivel de significancia del 5%. es decir $\alpha = 0.05$

iii. Prueba estadística

Prueba de rangos con signo de Wilcoxon

Tabla 16

Prueba de rangos Wilcoxon

Individuo	Prueba 1	Prueba 2	Diferencia	|Diferencia|	Rangos
1	3.1	2.67	0.43	0.43	1
2	2.09	1.26	0.83	0.83	2
3	2.44	1.5	0.94	0.94	3
4	2.6	1.08	1.52	1.52	4
5	2.63	1.08	1.55	1.55	5
6	4.55	2.9	1.65	1.65	6
7	2.87	1.21	1.66	1.66	7
8	3.01	1.32	1.69	1.69	8
9	3.2	1.5	1.7	1.7	9
10	5.75	2.01	3.74	3.74	10

W=min(55,0)

W=0

$$Z_{cal.}=\frac{W-\frac{n(n-1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{12}}}$$

Donde:

$Z_{cal.}$: Z calculado

W:Valor de Wilcoxon

n:Tamaño de la muestra

$$Z_{cal.}=\frac{0-\frac{10(10-1)}{4}}{\sqrt{\frac{10(10+1)(2*10+1)}{12}}}$$

$Z_{cal.}$=-1.98

Z_{tabla}=1.96

p-valor=0.024

iv. Regla de decisión

Si p-valor(0.024) < α(0.05) . entonces se rechaza la H_O y se acepta la H_a.

v. Decisión

Hay una diferencia en las medias de los tiempos del antes y después de la implementación del sitio web. Por lo cual se concluye que la implementación redujo el tiempo en la búsqueda de paquetes turísticos.

CONCLUSIONES

Se ha logrado la implementación del sitio web que facilita la búsqueda de diferentes paquetes turísticos ofertados por diversas agencias de viajes que operan en la región de Puno. reduciendo el tiempo empleado en la búsqueda antes y después de forma significativa $p\ (0.005) < \alpha(0.05)$

El análisis de la estructura DOM de cada uno de los sitios web de las agencias de viaje que operan en la región de Puno, permitió el desarrollo del algoritmo de extracción, limpieza y almacenamiento de los datos haciendo uso de Python como lenguaje de programación, también se puso a prueba la eficiencia del algoritmo, el cual demostró ser mucho más rápido y la cantidad de datos extraídos fue mayor en comparación con la extensión de Google denominado webscraper.io, además se determinó que la complejidad algorítmica es lineal O(n) para la extracción de los datos, que el tiempo de ejecución es proporcional a la cantidad de páginas a extraer.

La API de PageSpeed Insights (PSI) de Google reporto el rendimiento de la página web en base a seis (06) métricas centradas en proporcionar una buena experiencia de usuario tanto en dispositivos móviles como en ordenadores. en ambos casos se obtuvo una puntuación superior a 90. y se considera que la velocidad de la página GoPuno es rápida.

La evaluación del sitio web basado en la norma ISO 25000 proporcionó una valoración de 6.96 sobre 10 puntos como calidad total. considerado como nivel “Aceptable” y grado “Satisfactorio”; resultado que evidenció la importancia de tomar en cuenta métricas de calidad en uso según el modelo de calidad de software aplicado.

RECOMENDACIONES

La técnica del web Scraping es ampliamente utilizada en varias industrias para la obtención de datos. pero existen páginas que posiblemente no quieran que se descargue masiva mente información de su sitio. por ello se recomienda revisar el archivo Robot.txt y tener en cuenta los derechos de propiedad intelectual de los sitios web.

Implementar al algoritmo la extracción de datos con la búsqueda de enlaces dentro de la página web de forma inteligente ampliando el análisis de datos más complejos y dotándole de autonomía para la extracción.

Ampliar la cobertura de páginas web, ya no solo limitarlo a la región de Puno, dado que con la técnica del web Scraping se rompe la barrera del espacio

Las métricas centradas en el usuario otorgadas por la API de Google proporcionan una buena línea de base. pero en muchos casos es necesario medir más que estas métricas para capturar la experiencia completa de un sitio en particular. por ejemplo. determinar el tiempo que tarda una página en mostrar los datos obtenidos de una base de datos para los usuarios que han iniciado sesión; por ende. se recomienda utilizar métricas según sus necesidades y objetivos.

Se recomienda utilizar el modelo de calidad ISO/IEC 25000 para evaluar productos software. ya que presenta una amplia información sobre las características de calidad del producto software y a la vez es integrado con el proceso de evaluación

Se recomienda guardar históricos de la base de datos del Web Scraping para realizar Minería de Datos.

BIBLIOGRAFÍA

Almeida de Oliveira, R., & Arantes Baracho Porto, R. M. (2016). Extração de dados do site tripadvisor como suporte na elaboração de indicadores do turismo de minas gerais: Uma iniciativa em big data. *Pesq. Bras. em Ci. da Inf. e Bib.*, *11*(2), 026-037.

Arias, F. G. (2012). *El proyecto de investigación Introduccion a la metodología científica* (Sexta). Editorial Episteme.

Awangga, R. M., Pane, S. F., & Dwi Astuti, R. (2019). Implementation of web scraping on GitHub task monitoring system. *Telkomnika*, *17*(1), 275-281. https://doi.org/10.12928/TELKOMNIKA.v17i1.11613

Baldeón VIllanes, E. J. (2015). *Método para la evaluación de calidad de software basado en ISO/IEC 2500 (Tesis de Maestria)* [Tesis de Maestria]. Universidad de San Martin de Porres.

Baskaran, U., & Ramanujam, K. (2018). Automated scraping of structured data records from health discussion forums using semantic analysis. *Informatics in Medicine Unlocked*, *10*, 149-158. https://doi.org/DOI: 10.1016/j.imu.2018.01.003

Baumgartner, R., Gatterbauer, W., & Gottlob, G. (2009). Web data extraction system. *Encyclopedia of Database Systems*, 3465-3471.

BBVA. (2016, enero 11). *Herramientas de extracción de datos: Para principiantes y profesionales*. BBVAOpen4U. Recuperado de https://bbvaopen4u.com/es/actualidad/herramientas-de-extraccion-de-datos-para-principiantes-y-profesionales

Chu, Y.-C., Hsu, C.-C., Lee, C.-J., & Tsai, Y.-T. (2015). Automatic data extraction of websites using data path matching and alignment. *2015 Fifth International*

Conference on Digital Information Processing and Communications (ICDIPC), 60-64. https://doi.org/10.1109/ICDIPC.2015.7323006

Cianes, P. (2020). *Notación big O*. Recuperado de https://pablocianes.com/notacion-big-o/

Contreras, F. (2016, septiembre 27). *Conoce que es un YAML - fercontreras*. Recuperado de https://fercontreras.com/conoce-que-es-un-yaml-e18e9d21ade4

Dewi, L. C., Meiliana, & Chandra, A. (2019). Social Media Web Scraping using Social Media Developers API and Regex. *Procedia Computer Science*, *157*, 444-449. https://doi.org/10.1016/j.procs.2019.08.237

Diab, A., & Barhoum, T. (2018). *Prevent XPath and CSS Based Scrapers by Using Markup Randomizer*. *5*(2), 11.

FEIST PUBLICATIONS, INC. V. TEL. RURAL SERVICIO CO. | FindLaw. (1991, marzo 27). https://caselaw.findlaw.com/us-supreme-court/499/340.html

Gheorghe, M., Mihai, F.-C., & Dârdal, M. (2018). Modern techniques of web scraping for data scientists. *Revista Romana de Interactiune Om-Calculator*, *11*(1), 63-75.

Hanretty, C. (2013). Scraping the Web for Arts and Humanities. *UNIVERSITY OF EAST ANGLIA*, 50.

Hernández, A. T., Vázquez, E. G., Rincón, C. A. B., & García, J. M. (2015). Metodologías para análisis político utilizando Web Scraping. *Research in Computing Science*, *95*, 113-121.

Hernández Herrero, C. (2014). *Aplicación de Técnicas de Web Scraping al Boletín Oficial de Castilla y León (Tesis de Grado)* [Grado en Ingeniería Informática de Servicios y Aplicaciones]. Universidad de Valladolid.

Hernández Sampieri, R., Baptista Lucio, P., & Fernández Collado, C. (2014). *Metodología de la investigación*. McGraw-Hill Interamericana.

Huaman Hilari, J. Z., & Quispe Ramos, M. A. (2019). *Modelo de búsqueda de productos alimenticios en supermercados online categoría abarrotes utilizando asistente virtual de tipo Chatbot y extracción de datos con Web Scraping*. Universidad Tecnlogica del Perú.

ISO/IEC 25010. (2013). Systems and software engineering—Systems and software Quality Requirements and Evaluation SQuaRE)—System and software quality models. *BSI Standards Publication*, 48.

Januzaj, Y., Luma, A., Aliu, A., Selimi, B., & Raufi, B. (2019). *WEB DATA SCRAPING TECHNIQUE AND PREPARATION FOR COMPARISON TECHNIQUES BETWEEN DIFFERENT DOCUMENTS*. *11*, 17.

Jarmul, K., & Lawson, R. (2017). *Python Web Scraping—Second Edition*. Packt Publishing. Recuperado de http://public.eblib.com/choice/publicfullrecord.aspx?p=4868542

Julian, L. R., & Natalia, F. (2015). The use of web scraping in computer parts and assembly price comparison. *2015 3rd International Conference on New Media (CONMEDIA)*, 1-6. https://doi.org/10.1109/CONMEDIA.2015.7449152

Khalil, S., & Fakir, M. (2017). RCrawler: An R package for parallel web crawling and scraping. *SoftwareX*, *6*, 98-106. https://doi.org/10.1016/j.softx.2017.04.004

Kiran, M., & Mownika, N. (2021). Machine learning integrated emotions detection on lockdowns in India using advanced web scraping. *Materials Today: Proceedings*. https://doi.org/10.1016/j.matpr.2021.01.460

Laaksonen, A. (2018). *Competitive Programmer's Handbook*. Recuperado de https://cses.fi/book/book.pdf

Landers, R. N., Brusso, R. C., Cavanaugh, K. J., & Collmus, A. B. (2016). A primer on theory-driven web scraping: Automatic extraction of big data from the Internet

for use in psychological research. *Psychological Methods*, *21*(4), 475-492. https://doi.org/10.1037/met0000081

Laurente Blanco, L. F., & Machaca Hancco, R. W. (2020). Modelamiento y proyección de la demanda de turismo internacional en Puno-Perú. *Revista Brasileira de Pesquisa em Turismo*, *14*(1), 34-55. https://doi.org/10.7784/rbtur.v14i1.1606

Marques, P., Dabbabi, Z., Mironescu, M.-M., Thonnard, O., Bessani, A., Buontempo, F., & Gashi, I. (2018). Detecting Malicious Web Scraping Activity: A Study with Diverse Detectors. *2018 IEEE 23rd Pacific Rim International Symposium on Dependable Computing (PRDC)*, 269-278. https://doi.org/10.1109/PRDC.2018.00049

Marres, N., & Weltevrede, E. (2013). SCRAPING THE SOCIAL?: Issues in live social research. *Journal of Cultural Economy*, *6*(3), 313-335. https://doi.org/10.1080/17530350.2013.772070

Martínez Rodríguez, J. L. (2012). *Método para la caracterización e indexación de contenidos en la Web a partir de roles en un dominio de interés especifico. (Tesis de Maestria)*. Centro de Investigación y de Estudios Avanzados del Instituto Politécnico Nacional.

Medrano, J. F. (2020). *Empleo de Minería de Texto para analizar la oferta de inmuebles a partir de avisos clasificados*. https://doi.org/10.13140/RG.2.2.14215.01444

Minguillón Alfonso, J. (2015). *Complejidad algorítmica Eficiencia de algoritmos y tipos abstractos de datos*. Universitat Oberta de Catalunya. Recuperado de https://www.studocu.com/en-us/document/universitat-oberta-de-catalunya/estructura-de-datos-y-algoritmos/lecture-notes/modulo-2-complejidad-algoritmica/2707725/view

Moskalenko, A. A., Laponina, O. R., & Sukhomlin, V. A. (2019). Developing a Web Scraping Application with Bypass Blocking. *Sovremennye Informacionnye Tehnologii i IT-obrazovanie*, *25*(2), 413-413-420. Directory of Open Access Journals. https://doi.org/10.25559/SITITO.15.201902.413-420

Muehlethaler, C., & Albert, R. (2021). Collecting data on textiles from the internet using web crawling and web scraping tools. *Forensic Science International*, 110753. https://doi.org/10.1016/j.forsciint.2021.110753

Muñoz Mandujano, M., Hernández Valerio, J. S., González Serrano, S. R., & Pérez Liévana, A. (2018). Web scraping para la recopilación de datos meteorológicos. *Revista NTHE*, *24*, 91-95.

Murillo, D., & Saavedra, D. (2017). Web Scraping de los Perfiles y Publicaciones de una Afiliación en Google Scholar utilizando Aplicaciones Web e implementando un Algoritmo en R Web Scraping of the Profiles and Publications of an Affiliation in Google Scholar using Web Applications and implementing an Algorithm in R. *4to Congreso Internacional AmITIC 2017*, 8.

PageSpeed Insights. (2021). Recuperado de https://developers.google.com/speed/pagespeed/insights/

Proyectos de ley emitidos por el Congreso de la República del Perú. (s. f.). Recuperado 12 de febrero de 2020, de http://proyectosdeley.pe/

Rizaldi, T., & Putranto, H. A. (2017). Perbandingan Metode Web Scraping Menggunakan CSS Selector dan Xpath Selector. *Teknika*, *6*(1), 43-46. https://doi.org/10.34148/teknika.v6i1.56

Sanchez, D. (2020). *PLATAFORMA DE RECOMENDACION DE HABILIDADES TECNOLOGICAS SEGUN PUESTO DE TRABAJO PARA PROFESIONALES DE TI, EN FUNCION DE LA DEMANDA EN LAS BOLSAS DE TRABAJO*

DIGITALES [Universidad de Lima]. Recuperado de https://repositorio.ulima.edu.pe/bitstream/handle/20.500.12724/12351/Sanchez_PLATAFORMA-RECOMENDACION-HABILIDADES.pdf?sequence=1&isAllowed=y

Sourav Sen, G. (2016). *Know Thy Complexities!* Recuperado de http://souravsengupta.com/research.html

Telstra Corporation Limited v Directorios telefónicos Company Pty Ltd [2010] FCA 44. (2010, febrero 8). Recuperado de http://www8.austlii.edu.au/cgi-bin/viewdoc/au/cases/cth/FCA/2010/44.html

TimeComplexity—Python Wiki. (s. f.). Recuperado 19 de junio de 2021, de https://wiki.python.org/moin/TimeComplexity

Toffler, A. (1974). *El «Shock» del futuro*. Plaza & Janés.

Torres, D., Villegas, R., & Vargas, E. (2017). Propuesta de Aplicación Web para el costeo Gastronómico. *Revista de Tecnologías Computacionales*, *1*(2), 46-52.

Ullah, H., Ullah, Z., Maqsood, S., & Hafeez, A. (2018). Web Scraper Revealing Trends of Target Products and New Insights in Online Shopping Websites. *International Journal of Advanced Computer Science and Applications*, *9*(6), 6.

Uriarte, J. I., Toro, G. R. R. M. de, & Larrosa, J. M. C. (2020). Web scraping based online consumer price index: The "IPC Online" case. *Journal of Economic and Social Measurement*, *44*(2-3), 141-159. https://doi.org/10.3233/JEM-190464

Valenciano López, J. (2015). *Auditoría mantenibilidad aplicaciones según la ISO/IEC 25000(Tesis de Grado)*. Universidad Complutense de Madrid.

Vàllez, M. (2017). Tesis doctoral – Síntesis. Exploración de procedimientos semiautomáticos para el proceso de indexación en el entorno web.

HIPERTEXT.NET. Anuario Académico sobre Documentación Digital y Comunicación Interactiva, *15*, 91-99. https://doi.org/10.2436/20.8050.01.50

Vernier, M., Cárcamo Ulloa, L., & Scheihing García, E. (2016). Diagnóstico de la estrategia editorial de medios informativos chilenos en Twitter mediante un clasificador de noticias automatizado. *Revista Austral de Ciencias Sociales*, *30*, 183-201. https://doi.org/10.4206/rev.austral.cienc.soc.2016.n30-09

Villarroel Colque, K. (2015). Infoxicación. *Revista de Investigación Scientia*, *4*(1), versión On-line.

Zhao, B. (2017). Web Scraping. En L. A. Schintler & C. L. McNeely (Eds.), *Encyclopedia of Big Data* (pp. 1-3). Springer International Publishing. https://doi.org/10.1007/978-3-319-32001-4_483-1

ANEXOS

Anexo N° 1: Archivo Robot.txt de las páginas a Scrapear

https://www.andenespuno.com/robots.txt

User-agent: *

Allow: /

Disallow: /cgi-bin/

Disallow: /zonasegura/

Disallow: /panel/

Disallow: /whois/

https://www.jumbotravelpuno.com/robots.txt

User-agent: *

allow: /

Sitemap: http://www.jumbotravelpuno.com/sitemap.xml

http://amarutours.com/robots.txt

User-agent: *

allow: /

Disallow: /wp-login

Disallow: /wp-admin

Disallow: //wp-includes/

Disallow: /*/feed/

Disallow: /*/trackback/

Disallow: /*/attachment/

Disallow: /author/

Disallow: /*/page/

Disallow: /*/feed/

Disallow: /tag/*/page/

Disallow: /tag/*/feed/

Disallow: /page/

Disallow: /comments/

Disallow: /xmlrpc.php

Disallow: /*?s=

Disallow: /*/*/*/feed.xml

Disallow: /?attachment_id*

Disallow: /amaru

Disallow: /mkt

titicacalake.com/robots.txt

User-agent: *

Disallow: /wp-admin/

Allow: /wp-admin/admin-ajax.php

https://punotours.com.pe/robots.txt

User-Agent: *

Disallow: /wp-admin/

Sitemap: http://punotours.com.pe/page-sitemap.xml

https://punoadventures.com.pe/robots.txt

User-agent: *

Crawl-delay: 5

https://www.kontikiperu.com/robots.txt

User-agent: *

Disallow: /wp-admin/

Allow: /wp-admin/admin-ajax.php

http://titicacaforyou.com/robots.txt

User-agent: *

Crawl-delay: 5

Anexo N° 2: Encuesta de satisfacción

ENCUESTA DE SATISFACCIÓN

Esta encuesta fue destinada a la evaluación del sitio web realizado a 10 colaboradores. la cual debe ser respondida marcando en cada casillero con una X según la escala de satisfacción con respecto a cada concepto.

Escala de satisfacción del sistema web:

5: Totalmente de Acuerdo. 4: De acuerdo. 3: Ocasionalmente. 2: En Desacuerdo y 1: Totalmente en Desacuerdo

N°	Concepto	Escala de Satisfacción				
		5	4	3	2	1
1	La información mostrada en el sitio web coincide con el de la página web oficial de la agencia de viajes					
2	El sitio web gopuno me resultó complejo.					
3	El sitio web gopuno me resultó fácil de usar.					
4	Necesitaría la ayuda de un experto para usar el sitio web.					
5	El tiempo de respuesta del sitio web fue el óptimo.					
6	Percibí que varias funciones del sitio web no son las correctas.					
7	Pienso que la mayoría de los usuarios pueden resolver sus dudas rápidamente.					
8	El sitio web me resultó pesado y complicado de usar.					
9	Me gustaría usar el sitio web para comparar la descripción e itinerario de los paquetes turísticos.					
10	¿La información mostrada en el sitio web Gopuno coincide con el de la página web oficial de la agencia de viajes?					

Fuente (karen sanchez)

a) ¿Cuánto tiempo emplea un usuario para encontrar un destino turístico de la región de puno en 3 páginas web diferentes de 3 agencias de viajes diferentes?

----------(minutos)

b) ¿Cuánto tiempo emplea haciendo uso del sitio web Gopuno un usuario para encontrar un destino turístico de la región de puno en 3 páginas web diferentes de 3 agencias de viajes diferentes?

----------(minutos)

Anexo N° 3: Ficha de validación

UNIVERSIDAD NACIONAL DEL ALTIPLANO
ESCUELA DE POSGRADO
MAESTRÍA EN INFORMÁTICA

FICHA DE VALIDACIÓN
INFORME DE OPINIÓN DEL JUICIO DE EXPERTO

DATOS GENERALES

1.1. Nombre de los instrumentos motivo de evaluación: **Encuesta de Satisfacción sobre Indexación de sitios web para optimizar la búsqueda de paquetes turísticos de la región de Puno basado en web scraping**.

1.2. Autor del Instrumento: Gerardino Juvenal Cauna Huanca

ASPECTOS DE VALIDACIÓN

Indicadores	Criterios	Deficiente				Regular				Bueno				Muy bueno				Excelente			
		0	6	11	16	21	26	31	36	41	46	51	56	61	66	71	76	81	86	91	96
		5	10	15	20	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
1. CLARIDAD	Está formado con lenguaje apropiado.															X					
2. OBJETIVIDAD	Está expresado en conductas observables.																X				
3. ACTUALIDAD	Adecuado al avance de la ciencia y la tecnología.																		X		
4. ORGANIZACIÓN	Existe una organización lógica.																X				
5. SUFICIENCIA	Comprende los aspectos en cantidad y calidad.															X					
6. INTENCIONALIDAD	Adecuado para valorar los instrumentos de investigación.																X				
7. CONSISTENCIA	Basado en aspectos teóricos científicos.															X					
8. COHERENCIA	Entre los índices e indicadores.															X					
9. METODOLOGÍA	La estrategia responde al propósito del diagnóstico.																X				
10. PERTINENCIA	Es útil y adecuado para la investigación.																X				

PROMEDIO DE VALORACIÓN: 80

OPINIÓN DE APLICABILIDAD: a) Deficiente b) Regular c) Bueno **d) Muy bueno** e) Excelente

Nombres y apellidos:	Reynaldo Sucari León	**DNI N°**	01341544
Dirección domiciliaria:	Jr. Jorge Basadre Nº 568 - Puno	**Teléfono celular:**	975126540
Grados Académicos:	Magister Scientiae en Informática y Doctor en Administración de la Educación		

Lugar y fecha: Puno, 08 de enero de 2021

UNA PUNO
Firmado digitalmente por SUCARI LEON Reynaldo FAU 20145496170 soft
Fecha: 2021.01.08 11:17:13 -05'00'

Dr. Reynaldo Sucari León
DOCENTE

UNIVERSIDAD NACIONAL DEL ALTIPLANO
ESCUELA DE POSGRADO
MAESTRÍA EN INFORMÁTICA

FICHA DE VALIDACIÓN
INFORME DE OPINIÓN DEL JUICIO DE EXPERTO

DATOS GENERALES

1.1. Nombre de los instrumentos motivo de evaluación: **Encuesta de Satisfacción sobre Indexación de sitios web para optimizar la búsqueda de paquetes turísticos de la región de Puno basado en web scraping**.

1.2. Autor del Instrumento: Gerardino Juvenal Cauna Huanca

ASPECTOS DE VALIDACIÓN

INDICADORES	CRITERIOS	Deficiente				Regular				Bueno				Muy bueno				Excelente			
		0	6	11	16	21	26	31	36	41	46	51	56	61	66	71	76	81	86	91	96
		5	10	15	20	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
1. CLARIDAD	Está formado con lenguaje apropiado.															✓					
2. OBJETIVIDAD	Está expresado en conductas observables.															✓					
3. ACTUALIDAD	Adecuado al avance de la ciencia y la tecnología.																		✓		
4. ORGANIZACIÓN	Existe una organización lógica.																✓				
5. SUFICIENCIA	Comprende los aspectos en cantidad y calidad.																✓				
6. INTENCIONALIDAD	Adecuado para valorar los instrumentos de investigación.															✓					
7. CONSISTENCIA	Basado en aspectos teóricos científicos.															✓					
8. COHERENCIA	Entre los índices e indicadores.																✓				
9. METODOLOGÍA	La estrategia responde al propósito del diagnóstico.																✓				
10. PERTINENCIA	Es útil y adecuado para la investigación.																✓				

PROMEDIO DE VALORACIÓN: 80

OPINIÓN DE APLICABILIDAD: a) Deficiente b) Regular c) Bueno d) **Muy bueno** e) Excelente

Nombres y apellidos:	**JUAN CARLOS JUAREZ VARGAS**	**DNI N°**	**40419555**
Dirección domiciliaria:	**JIRON 7 DE JUNIO 329 PUNO**	**Teléfono celular:**	**951007384**
Grados Académicos:	**Magister Scientiae en Informática y Doctor Scientiae en Administración**		

Lugar y fecha: Puno, 08 de enero de 2021

Universidad Nacional del Altiplano de Puno
Firmado digitalmente por JUAREZ VARGAS Juan Carlos FAU 20145496170 soft
Motivo: Soy el autor del documento
Fecha: 25.05.2021 21:44:59 -05:00

FIRMADO DIGITALMENTE
DR. JUAN CARLOS JUAREZ VARGAS

UNIVERSIDAD NACIONAL DEL ALTIPLANO
ESCUELA DE POSGRADO
MAESTRÍA EN INFORMÁTICA

FICHA DE VALIDACIÓN
INFORME DE OPINIÓN DEL JUICIO DE EXPERTO

DATOS GENERALES

1.1. Nombre de los instrumentos motivo de evaluación: **Encuesta de Satisfacción sobre Indexación de sitios web para optimizar la búsqueda de paquetes turísticos de la región de Puno basado en web scraping**.

1.2. Autor del Instrumento: Gerardino Juvenal Cauna Huanca

ASPECTOS DE VALIDACIÓN

INDICADORES	CRITERIOS	Deficiente				Regular				Bueno				Muy bueno				Excelente			
		0	6	11	16	21	26	31	36	41	46	51	56	61	66	71	76	81	86	91	96
		5	10	15	20	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
1. CLARIDAD	Está formado con lenguaje apropiado.																X				
2. OBJETIVIDAD	Está expresado en conductas observables.																X				
3. ACTUALIDAD	Adecuado al avance de la ciencia y la tecnología.																		X		
4. ORGANIZACIÓN	Existe una organización lógica.																X				
5. SUFICIENCIA	Comprende los aspectos en cantidad y calidad.															X					
6. INTENCIONALIDAD	Adecuado para valorar los instrumentos de investigación.																X				
7. CONSISTENCIA	Basado en aspectos teóricos científicos.																X				
8. COHERENCIA	Entre los índices e indicadores.															X					
9. METODOLOGÍA	La estrategia responde al propósito del diagnóstico.																X				
10. PERTINENCIA	Es útil y adecuado para la investigación.																X				

PROMEDIO DE VALORACIÓN: 80

OPINIÓN DE APLICABILIDAD: a) Deficiente b) Regular c) Bueno (d) Muy bueno e) Excelente

Nombres y apellidos:	JORGE CONDORI CHAVEZ	**DNI N°**	01221750
Dirección domiciliaria:	Jr. PUNO Nro 826 - PUNO	**Teléfono celular:**	951059555
Grados Académicos:	MSc. EN INFORMOTICA		

Lugar y fecha: Puno, 08 de enero de 2021

MSc. Jorge Condori Chavez
DOCENTE

Anexo N° 4: Directorio de agencias IPERI 2019

IPERÚ Puno
Esquina Jr. Deustua con Jr. Lima (Plaza de Armas)
E-mail: iperupuno@promperu.gob.pe
Telf. (051) 365088
Horario de atención: Lunes a Sabado 9:00 - 18:00 /
Domingos 09:00 - 13:00

AGENCIAS DE VIAJES 2019

CLASIFICACIÓN	RAZÓN SOCIAL	NOMBRE COMERCIAL	DIRECCIÓN	DISTRITO	PROVINCIA	REGIÓN	TELÉFONO FIJO	EMAIL	WEBSITE
Tour Operadores	Servicios Receptivos Titikaka E.I.R.L.	Servicios Receptivos Titikaka	Ps. Lima Nº 419 - Ofic. 207	Puno	Puno	Puno	(051) 36-9955	info@titikakatouroperador.com	www.titikakatouroperador.com
Tour Operadores	Serv. Turísticos All Ways Travel Titicaca Peru S.A.C.	All Ways Travel	Jr. Deustua N° 576 / Jr. Puno N°823 y 825 (Ref. Colegio Santa Rosa).	Puno	Puno	Puno	(51) 35-3979 / (051) 35-5552	allwaystravel@titicacaperu.com; sales@titicacaperu.com	www.titicacaperu.com/es/
Tour Operadores	Arcobaleno S.R.L.	Arcobaleno	Jr. Tarapacá N°355 Interior A	Puno	Puno	Puno	(051) 354402 / 951621659 (Eduardo Pineda) / 951621633	arcobaleno@titicacalake.com; gerencia@titicacalake.com; reservas@titicacalake.com	www.titicacalake.com
Tour Operadores	Viajes y Turismo Puno Travel E.I.R.L.	Puno Travel	Jr. Melgar Nº 173	Puno	Puno	Puno	(051) 35-2632	flaura@punotravel.com	www.punotravelavt.com
Tour Operadores	Agencia de Viajes y Turismo Edgar Adventures S.R.L.	Edgar Adventures	Jr. Lima N° 328	Puno	Puno	Puno	(051) 35-3444 / 36-9927	manager@edgaradventures.com	www.edgaradventures.com
Tour Operadores	León Tours E.I.R.Ltda.	Leon Tours	Jr. Ayacucho N°152	Puno	Puno	Puno	(051) 35-2771	leonpuno@hotmail.com	www.peru-titicaca.com
Tour Operadores	Peruvian Dream Tour Operator E.I.R.L.	Peruvian Dream	Jr. Horcapata Nro. 138 - Barrio Victoria	Puno	Puno	Puno	(051) 35-1657 / 951793366	peruviandream@gmail.com	www.perudreams.com
Tour Operadores	A.V.T Suri Explorer E.Ir.Ltda.	Suri Explorer	Jr. Teodoro Valcarcel Nº 158	Puno	Puno	Puno	(051) 36-8188/951634240	suriexplorer@hotmail.com; reservas@suriexplorer.com	www.suriexplorer.com
Tour Operadores	Guerra Nina, Enrique	Misterios del Titikaka	Jr. Teodoro Valcarcel Nº 135	Puno	Puno	Puno	(051) 35-2141	misteryagency@hotmail.com	www.travelagencypuno.com
Tour Operadores	Pirámide Tours S.A.C.	Pirámide Tours	Jr. Rosendo Huirse Nº 128	Puno	Puno	Puno	(051) 36-4125	piramide@titikaka.com	www.titikaka.com/espanol/index02.htm
09/05/2019	Amaru Tours E.I.R.L.	Amaru Tours	Jr. Tarapacá Nº 260 - Of. 103	Puno	Puno	Puno	(051) 35-3112 / 951751084	reservas@amarutours.com	www.amarutours.com
Tour Operadores	A.V.T. Cusi Expeditions E.I.R.Ltda.	Agencia de Viajes y Turismo Cusi Expeditions	Jr. Tarapacá N°300 (Ref. Frente al Hotel Sillustani).	Puno	Puno	Puno	(051) 36-9072	operaciones_cusi@hotmail.com	x
Tour Operadores	Sacred Lake Servicios Turísticos S.C.R.L.	Sacred Lake	Urbanización Chanu Chanu - Mz. H Lote 03	Puno	Puno	Puno	(051) 35-6085 / 951755337	info@sacredlaketitikaka.com	www.sacredlaketitikaka.com
Tour Operadores	Viajes y Turismo Gesam E.I.R.L.	Viajes y Turismo Gesam	Jr. Revolución Nro. 106 - Ag. Br. Alto Orkapata (A Espaldas Del Club De Madres De Alto Ork)	Puno	Puno	Puno	951589951 / 951753077	vytgesam@hotmail.com	x
Tour Operadores	Latin Reps E.I.R.L.	Latin Reps	Jr. Arequipa 736 - Interior A	Puno	Puno	Puno	(051) 36-4887	latinrepsperu@latinrepsperu.com	www.latinrepsperu.com
Tour Operadores	A.V.T. Titikaka Adventures E.I.R.L.	Titikaka Adventures E.I.R.L.	Jr. Santiago Giraldo Nº 222 - Cercado	Puno	Puno	Puno	951402502	reservas@titikakaadventures.com	www.titikakaadventures.com
Tour Operadores	Southern Cross E.I.R.Ltda.	Southern Cross	Jr. Tarapacá N°238	Puno	Puno	Puno	914345732 / 951628989	carbajalia@gmail.com	x
Tour Operadores	Comunidad Campesina de la Isla Taquile	Munay Taquile	Av. Titicaca Nº 508	Puno	Puno	Puno	(051) 35-1448	munay_taquile@hotmail.com	www.taquile.net
Tour Operadores	Expediciones Las Balsas S.C.R.L.	Expediciones Las Balsas	Jr. Lima Nº419 of. 213 - 2do Piso	Puno	Puno	Puno	(051) 36-4362 / 951622891	administracion@balsastours.com; expediciones@balsastours.com	www.balsastours.com
Tour Operadores	Jumbo Travel E.I.R.L.	Jumbo Travel	Jr. Independencia N°349 (Ref. Frente al Arco Deustua)	Puno	Puno	Puno	(051) 36-4928	reservas@jumbotravelpuno.com / richardariaslopez@hotmail.com	www.jumbotravelpuno.com
Tour Operadores	Agencia de Viajes y Transporte Turístico Kollasuyo Travel E.I.R.L	Kollasuyo Travel	Jr. Santiago Giraldo N° 164	Puno	Puno	Puno	(051) 36-8642/951524686	kollasuyotravel@hotmail.com	www.kollasuyotravel.com
Tour Operadores	Empresa de Servicios Turísticos American S.C.R.L.	American Tours	Jr. Lambayeque Nº 144	Puno	Puno	Puno	(051) 36-6122	reservas@hotelbuho.com	x
Tour Operadores	Kolla Tour Representaciones Turísticas E.I.R.L	Kolla Tour	Jr. Moquegua Nº 679 - Barrio Victoria	Puno	Puno	Puno	(051) 36-9863	gerencia@titikakakolla.com; operaciones@titikakakolla.com	www.titikakakolla.com
Agencia Minorista	A.V.T. Nayra Travel S.C.R.L.	Nayra Travel	Jr. Lima Nº 419 Of. 105	Puno	Puno	Puno	(051) 36-4774	reservas@nayratravel.com; info@nayratravel.com	www.nayratravel.com
Agencias Minoristas	Representaciones Turísticas Andenes S.R.L.	Andenes Reps	Jr. Cajamarca Nº 678	Puno	Puno	Puno	(051) 36-5704 /978470082	andenes_reps@hotmail.com; reservas@andenespuno.com	www.andenespuno.com
Agencias Minoristas	Uros Travel E.I.R.L	Uros Travel	Av. Titicaca Nº 579	Puno	Puno	Puno	951608011	urostravelpuno@hotmail.com/ jose.uros@hotmail.com	x
Tour Operadores (Sucursal)	Lima Tours S.A.C.	Lima Tours	Jr. Tacna N°147 (4to piso)	Puno	Puno	Puno	(051) 35-2001	litopuno@limatours.com inbound@limatours.com.pe	www.limatours.com.pe

Tour Operadores	Agencia de Viajes y Turismo Kontiki Tours E.I.R.L.	Kontiki Tours	Jr. Melgar N°188	Puno	Puno	Puno	(051) 35-3473	administracion@kontikiperu.com	www.kontikiperu.com
Tour Operadores	Kafer Viajes y Turismo E.I.R.L.	Kafer Travel	Jr. Juan José Calle N°172, Barrio Porteño	Puno	Puno	Puno	(051) 35-2701	kafer@speedy.com.pe	www.kafer-titicaca.com
Tour Operadores	Mundo Inka Sertur S.C.R.L	Mundo Inka Sertur	Av. Sesquicentenario N°576	Puno	Puno	Puno	(051) 36-6350	transmundo@terra.com.pe	x
Tour Operadores	Qallari E.I.R.L.	Qallari	Jr. Puno N°633 / 2do Piso	Puno	Puno	Puno	(051) 36-6809	puno@gaston_sacaze.com	x
Tour Operadores	Quimbaya Tours E.I.R.L	Quimbaya Tours	Jr. Lima N°419 Of. 305	Puno	Puno	Puno	(051) 36-3417	pe-alicia-hermoza@quimbaya-tours.com	x
Agencias Mayoristas	Solmartour S.A.	Solmartour	Jr. Independencia N°151 Departamento B 201	Puno	Puno	Puno	(051) 35-2901	administracion.puno@solmar.com.pe	www.solmar.com.pe
Agencias Mayoristas	Agencia de Viajes y Tuismo Giant Trip E.I.R.L.	Giant Trip	Jr. Lima N°440	Puno	Puno	Puno	(051) 35-3214	gianttrip1@gmail.com	www.gianttrip.com
Tour Operador	Agencia de Viajes y Turismo Titicaca For You E.I.R.L	Titicaca For You	Jr. llave N° 356	Puno	Puno	Puno	(051) 35-1105 / 998595252	titicacaforyou.puno@gmail.com	www.titicacaforyou.com
Tour Operador	Servicios Turísticos Titicaca Experiences S.A.C.	Servicios Turísticos Titicaca Experiences	Jr. Ayacucho N° 774 - Barrio San Antonio	Puno	Puno	Puno	946656462 / 950766721	derly_mira@hotmail.com / reservas@titicacaexperiences.com / titicacaexperienses@gmail.com	www.titicacaexperiences.com
Tour Operador	Inca´s Paradise Travel Agency E.I.R.L.	IP Travel	Jr. Lima Nº 419 Of. 105	Puno	Puno	puno	(051) 36-4774 / 953556680	info@incasparadise.com	www.incasparadise.com
Tour Operador	Agencia de Viajes Amaya Travel Peru	Amaya Travel Peru	Jr. Puno N°501	Puno	Puno	Puno	051 63-3535 / 973586325	laketiticacatravel@gmail.com	www.laketiticaca.travel
Tour Operador	Lago del Cielo S.A.C.	Inka Lake	Jr. Cajamarca N°619	Puno	Puno	Puno	(051) 62-4475	lagodelcieloperu@gmail.com / reservas@incalake.com	www.incalake.com

Fecha de actualización: 31/01/2019

Anexo N° 5: Matriz de consistencia

PROBLEMA	OBJETIVOS	HIPÓTESIS	VARIABLES	INDICADOR
Problema General ¿El uso del software para indexar sitios web optimizara la búsqueda de paquetes turísticos de la región de Puno utilizando la tecnología web Scraping? **Problemas específicos** - ¿ En qué medida el análisis de la estructura DOM de las páginas web de las agencias de viaje contribuye al desarrollo del algoritmo para la extracción de la información relevante? - ¿En qué medida el rendimiento de un sitio web basado en la experiencia de usuario facilita la búsqueda de paquetes turísticos? - ¿La norma ISO/IEC 25000 permite establecer el grado de satisfacción de un producto de software?	**Objetivo general** Desarrollar un software para la indexación de sitios web y optimizar la búsqueda de paquetes turísticos de la región de Puno basado en web Scraping **Objetivos específicos** - Desarrollar un algoritmo para extraer la información relevante de las páginas web de las agencias de viaje que operan en la región de Puno. - Implementar y determinar el rendimiento un sitio web basado en la experiencia de usuario que facilite la búsqueda de los paquetes turísticos - Establecer el grado en que el producto satisface a un	**Hipótesis general** La implementación del software para la indexación de sitios web basado en web scraping optimiza la búsqueda de paquetes turísticos de la región de Puno. **Hipótesis específicas** - El análisis de la estructura DOM facilita la implementación del algoritmo para la extracción de la información relevante de las páginas web de las agencias de viaje de la región Puno. - El rendimiento de un sitio web basado en la experiencia de usuario facilita la búsqueda de paquetes turísticos - La norma ISO/IEC 25000 permite establecer el grado de satisfacción de un producto de software.	**Variable independiente** Indexación de sitios web **Variable independiente** Búsqueda de paquetes turísticos de la región de Puno	Calidad de uso ISO/IEC 25010 Dimensiones: • Efectividad • Eficiencia • Satisfacción Tiempo de búsqueda de los paquetes turísticos de la región de Puno. Dimensiones: • Tiempo.

	usuario final basado en la norma ISO/IEC 25000			

Anexo N° 6: Archivo main.py

```
import argparse
import logging
logging.basicConfig(level=logging.INFO)
import news_page_objects as news

from common import config

import re
import datetime
import csv
from requests.exceptions import HTTPError
from urllib3.exceptions import MaxRetryError


logger = logging.getLogger(__name__)
is_well_formed_link = re.compile(r'^https?://.+/.+$')
is_well_formed_link2 = re.compile(r'^https?://.+/.+/.+$')
is_root_path = re.compile(r'^/.+$')

def _news_scraper(news_site_uid):
        host = config()['news_sites'][news_site_uid]['url']
        logging.info('Iniciando el scraper para {}'.format(host))
        homepage = news.HomePage(news_site_uid. host)
        articles = []
        for link in homepage.article_links:
                article = _fetch_article(news_site_uid.host.link)
                if article:
                        logger.info('Articulo encontrado!!!')
                        articles.append(article)

        _save_articles(news_site_uid. articles)
```

```
def _save_articles(news_site_uid. articles):
        now = datetime.datetime.now().strftime('%Y_%m_%d')
        out_file_name = '{news_site_uid}_{datetime}_articles.csv'.format(
                news_site_uid = news_site_uid.
                datetime = now)
        csv_headers = list(filter(lambda property: not property.startswith('_'). dir(articles[0])))
        with open(out_file_name. mode='w+') as f:
                writer = csv.writer(f)
                writer.writerow(csv_headers)

                for article in articles:
                        row = [str(getattr(article. prop)) for prop in csv_headers]#cambios en el article page
                        writer.writerow(row)


def _fetch_article(news_site_uid.host.link):
        logger.info('Comenzando a buscar articulos en {}'.format(link))

        article = None
        try:
                article = news.ArticlePage(news_site_uid. _build_link(host. link))
        except (HTTPError. MaxRetryError) as e:
                logger.warning('Error al buscar el articulo'. exc_info=False)

        if article and not article.body:
                logger.warning('Articulo invalido. No hay Body')
                return None
        return article
```

```
def _build_link(host. link):
        if is_well_formed_link.match(link):
                return link
        elif is_well_formed_link2.match(link):
                return link
        elif is_root_path.match(link):
                return '{}{}'.format(host. link)
        else:
                return '{host}/{uri}'.format(host=host.uri=link)

if __name__ == '__main__':
        parser = argparse.ArgumentParser()

        news_site_choices = list(config()['news_sites'].keys())

        parser.add_argument('news_site'.
                                        help='El nuevo Site que necesitas scrapear'.
                                        type=str.
                                        choices=news_site_choices)
        args = parser.parse_args()

        _news_scraper(args.news_site)
```

Anexo N° 7: Archivo page.py

```
import bs4
import requests
from common import config

class NewsPage:
    def __init__(self. news_site_uid.url):
        self._config = config()['news_sites'][news_site_uid]
        self._queries = self._config['queries']
        self._html = None
        self._visit(url)
        self._url=url


    def _select(self. query_string):
        return self._html.select(query_string)


    def _visit(self. url):
        response = requests.get(url)
        response.raise_for_status()

        self._html = bs4.BeautifulSoup(response.text. 'html.parser')


class HomePage(NewsPage):
    def __init__(self. news_site_uid.url):
        super().__init__(news_site_uid. url)


    @property
    def article_links(self):
        link_list = []
```

```
            for link in self._select(self._queries['homepage_article_links']):
                if link and link.has_attr('href'):
                    link_list.append(link)
            return set(link['href'] for link in link_list)


class ArticlePage(NewsPage):
    def __init__(self. news_site_uid. url):
        super().__init__(news_site_uid. url)

    @property
    def body(self):
        result = self._select(self._queries['article_body'])
        return result[0].text if len(result) else ''

    @property
    def title(self):
        result = self._select(self._queries['article_title'])
        return result[0].text if len(result) else ''

    @property
    def url(self):
        result= self._url
        return result if len(result) else ""
```