**50% proyecto integrador**

**Desarrollo de un algoritmo para el análisis de sentimientos de comentarios de clientes sobre productos de computación de Linio Colombia.**

**Kevin Orlando Carreño Torres**

**Johan Eduardo Cala Torra**

**Rober Julián Quintero Oviedo**

**Estructura de Datos**

**Matemáticas Discretas II**

**Lenin Javier Serrano Gil**

**Julián Darío Miranda Calle**

**UNIVERSIDAD PONTIFICIA BOLIVARIANA**

**ESCUELA DE INGENIERÍA  
FACULTAD DE INGENIERÍA DE SISTEMAS E INFORMÁTICA**

**BUCARAMANGA**

**2020**

TABLA DE CONTENIDO

[1. Requerimientos de software 3](#_Toc51155735)

[1.2. Requerimientos Funcionales 3](#_Toc51155736)

[1.3. Requerimientos No Funcionales 6](#_Toc51155737)

[1.4. Requerimientos de Testing 7](#_Toc51155738)

[1.5. Restricciones 7](#_Toc51155739)

[1.5.1. Restricciones Hardware 7](#_Toc51155740)

[1.5.2. Restricciones Software 7](#_Toc51155741)

[2. Diseño de software 7](#_Toc51155742)

[2.2. Proceso de adquisición de datos 7](#_Toc51155743)

[2.3. Modelo Entidad – Relación 8](#_Toc51155744)

[2.4. Diccionario de datos 8](#_Toc51155745)

[2.5. Diagrama de clases 11](#_Toc51155746)

[3. Desarrollo de algoritmos 11](#_Toc51155747)

# **Requerimientos de software**

# **Requerimientos Funcionales**

|  |  |
| --- | --- |
| **Identificación del requerimiento:** | RF01 |
| **Nombre del Requerimiento:** | Adquisición de los datos |
| **Características:** | Se debe adquirir los datos mediante técnicas automatizadas para su almacenamiento en una base de datos local |
| **Descripción del requerimiento:** | Se deben adquirirse un mínimo de 50 productos, los cuales deben tener 3 comentarios donde se catalogue el producto con 4 o mas estrellas y 3 comentarios donde se catalogue el producto con 3 o menos estrellas |
| **Requerimiento NO funcional:** | * RNF01 |
| **Prioridad del requerimiento:**  Alta | |

|  |  |
| --- | --- |
| **Identificación del requerimiento:** | RF02 |
| **Nombre del Requerimiento:** | Procesamiento de Datos |
| **Características:** | Permitir el procesamiento de datos de texto para la estructuración, estandarización y limpieza de estos. |
| **Descripción del requerimiento:** | Cada uno de los comentarios de usuario adquiridos por medios automatizados debe permanecer inocuo y sin modificaciones tras su adquisición, de tal forma que todo carácter o cadena especial debe ser permitida, de tal forma que se presente al usuario un espacio sin sesgos frente a los comentarios analizados. |
| **Requerimiento NO funcional:** | * RNF01 * RNF03 |
| **Prioridad del requerimiento:**  Alta | |

|  |  |
| --- | --- |
| **Identificación del requerimiento:** | RF03 |
| **Nombre del Requerimiento:** | Extracción de atributos |
| **Características:** | Extraer los atributos de, disminución de la dimensionalidad y representación de los datos vectorialmente |
| **Descripción del requerimiento:** | de tal forma que pueda entrenarse un modelo de aprendizaje con la salida de estos procesos. |
| **Requerimiento NO funcional:** | * RNF01 * RNF03 |
| **Prioridad del requerimiento:**  Alta | |

|  |  |
| --- | --- |
| **Identificación del requerimiento:** | RF04 |
| **Nombre del Requerimiento:** | Ejecución de un proceso de entramiento y prueba del modelo |
| **Características:** | Debe pasar los datos limpiados y almacenados |
| **Descripción del requerimiento:** | Proceso de evaluación de su desempeño al ser expuesto a datos de prueba no conocidos y conocidos |
| **Requerimiento NO funcional:** | * RNF01 * RNF02 * RNF04 |
| **Prioridad del requerimiento:**  Alta | |

|  |  |
| --- | --- |
| **Identificación del requerimiento:** | RF05 |
| **Nombre del Requerimiento:** | Análisis de sentimientos |
| **Características:** | El algoritmo clasifica los sentimientos positivos y negativos a través de los comentarios |
| **Descripción del requerimiento:** | A través de los comentarios de usuarios correspondientes de la plataforma de linio Colombia, dependiendo del contexto identificado y analizado |
| **Requerimiento NO funcional:** | * RNF01 * RNF02 * RNF03 * RNF04 |
| **Prioridad del requerimiento:**  Alta | |

|  |  |
| --- | --- |
| **Identificación del requerimiento:** | RF06 |
| **Nombre del Requerimiento:** | Datos Estadísticos |
| **Características:** | El algoritmo debe permitir la visualización de estadístico básicos, |
| **Descripción del requerimiento:** | Permite la visualización de datos estadísticos básicos como la cantidad de comentarios clasificados como positivos y negativos y cantidad de comentarios que no fueron clasificadas correctamente |
| **Requerimiento NO funcional:** | * RNF01 * RNF02 * RNF03 * RNF04 |
| **Prioridad del requerimiento:**  Alta | |

# **Requerimientos No Funcionales**

|  |  |
| --- | --- |
| **Identificación del requerimiento:** | RNF01 |
| **Nombre del Requerimiento:** | Base de Datos |
| **Características:** | El sistema presentara una base de datos de manera local |
| **Descripción del requerimiento:** | El sistema debe tener una base de datos que permita almacenar los comentarios de entrenamiento y prueba del modelo |
| **Prioridad del requerimiento:**  Alta | |

|  |  |
| --- | --- |
| **Identificación del requerimiento:** | RNF02 |
| **Nombre del Requerimiento:** | Técnicas de Machine Learning |
| **Características:** | El algoritmo debe implementar una o varias técnicas de Machine Learning |
| **Descripción del requerimiento:** | el algoritmo debe ser desarrollado usando una o varias técnicas de Machine Learning, sea de aprendizaje supervisado o no supervisado, argumentando la escogencia sobre una base teórico-práctica. |
| **Prioridad del requerimiento:**  Alta | |

|  |  |
| --- | --- |
| **Identificación del requerimiento:** | RNF03 |
| **Nombre del Requerimiento:** | Data Wragnling |
| **Características:** | los datos almacenados en la base de datos local deben estar limpios |
| **Descripción del requerimiento:** | Limpiar los datos almacenados en la base de datos de la forma que no tengan emoticones, stop Words u otro tipo de caracteres o cadenas especiales |
| **Prioridad del requerimiento:**  Alta | |

|  |  |
| --- | --- |
| **Identificación del requerimiento:** | RNF04 |
| **Nombre del Requerimiento:** | Desempeño del modelo |
| **Características:** | Se debe mostrar el proceso de entrenamiento y prueba del modelo |
| **Descripción del requerimiento:** | de tal forma que se especifiquen los resultados de ambos procesos  Especificación de resultados de ambos procesos mediante métricas estandarizadas, como el Error Porcentual Absoluto Medio (MAPE), la precisión, la exactitud o el F1 Score. |
| **Prioridad del requerimiento:**  Alta | |

# **Requerimientos de Testing**

# **Hardware de Testing**

Computador con 2GB de RAM, 15 GB de espacio disponible de disco duro, procesador de 2Ghz.

# **Software de Testing**

Windows, versión de 7 a 10. Python, version 3.7. Microsoft Office Excel, aplicación de escritorio ejecutada por Jupyter Notbook 6.1.1

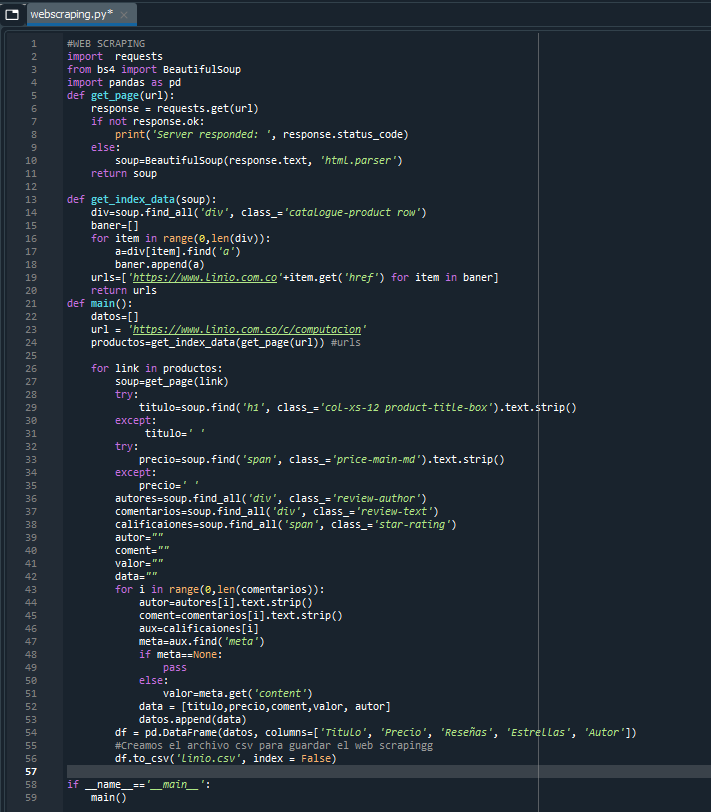
# **Restricciones**

# **Restricciones Hardware**

# **Restricciones Software**

# **Diseño de software**

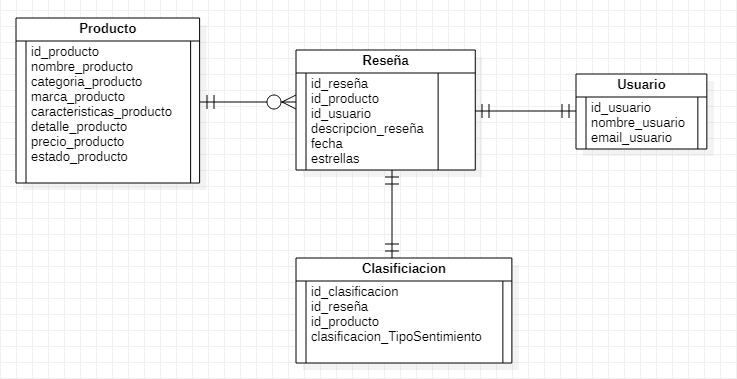
# **Proceso de adquisición de datos**



El programa fue desarrollado para realizar el Web Scraping en la plataforma Linio Colombia en la sección de computación para el posterior análisis de sentimientos de las reseñas de los productos, para cumplir esto desarrollamos los siguientes métodos:

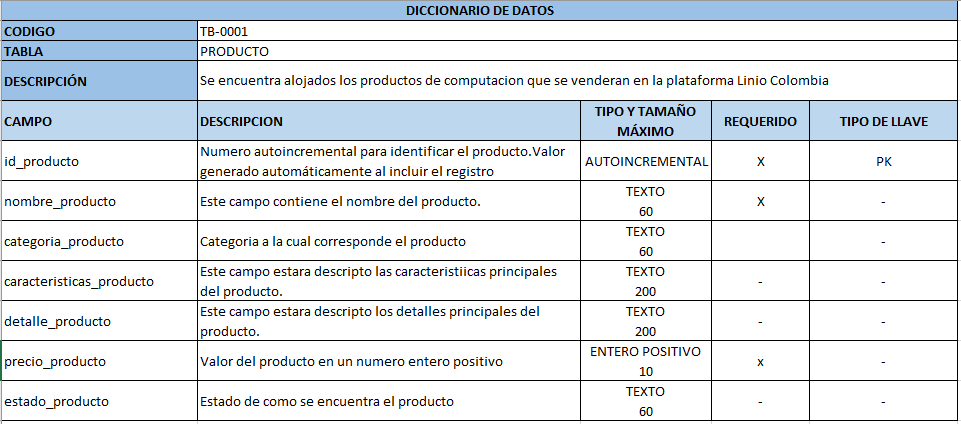
* **get\_page(url):** El cual tiene como parámetro de entrada la URL de la página con la que se realiza la conexión, donde se tiene como parámetro de retorno todo el código fuente de la página consultada para poder empezar a buscar las etiqueta que se desean(soup).
* **get\_index\_data(soup**): el cual tiene como parámetro soup que es el método de retorno del método anterior, este se encarga de encontrar todos los productos de la página y para así acceder a cada uno de los que se presenten en la página para posteriormente extraer los datos, el carácter de salida de este método es una lista que contiene las URLS del productos.
* **main():** Este es el método principal del programa donde su función es ejecutar los dos métodos anteriores, para así poder extraer por cada producto de dicha página, los atributos con el método “fin\_all”, para luego mediante un for teniendo en cuenta las reseñas de cada artículo, tomaremos por cada una su autor, reseña y calificación o estrellas y luego exportar los datos mediante un DataFrame donde sus columnas serán Titulo, Precio, Reseñas, Estrellas, Autor. Al completar este DataFrame lo exportamos a un archivo CSV para guardar nuestro datos para después procesarlos de manera correcta.

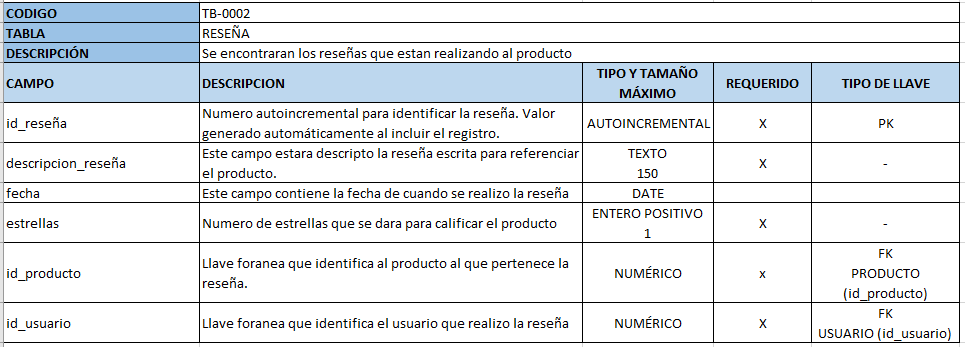
# **Modelo Entidad – Relación**

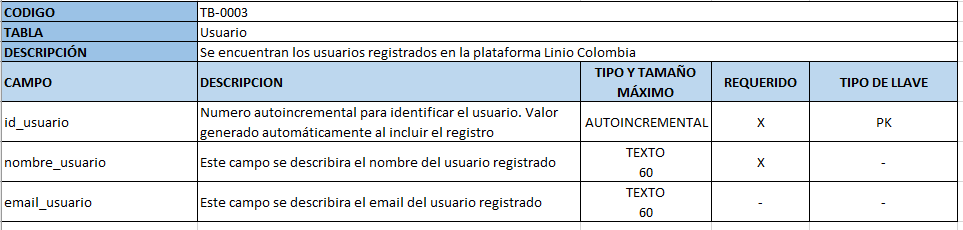


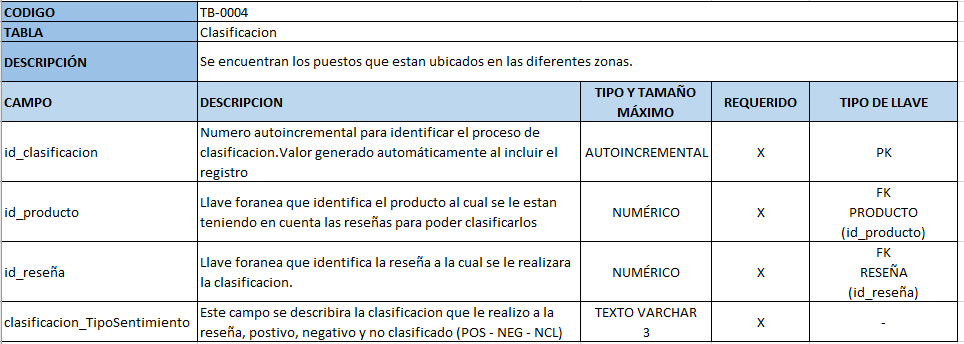
En este diagrama podremos observar el modelo que se tendrá en cuenta para el desarrollo del algoritmo, tomando de referencia la información relevante de la plataforma Linio Colombia en la categoría “computación”, primero nuestra tabla “Producto”, tendrá los atributos que serán las características principales, el cual tiene una relación de uno a muchos con la tabla “Reseñas”, la cual serán las reseñas escritas por los usuario dándole una calificación al producto, y estará relacionado uno a uno con la tabla “usuario”, puesto que una reseña solo la podrá realizar un usuario, además también estará relacionada con la tabla “clasificación”, con una relación uno a uno, pues tomara cada reseña, para luego poder analizar e identificar los sentimientos de ella, y así poder clasificarla según corresponda, ya sea positivo, negativo o no clasificado.

# **Diccionario de datos**

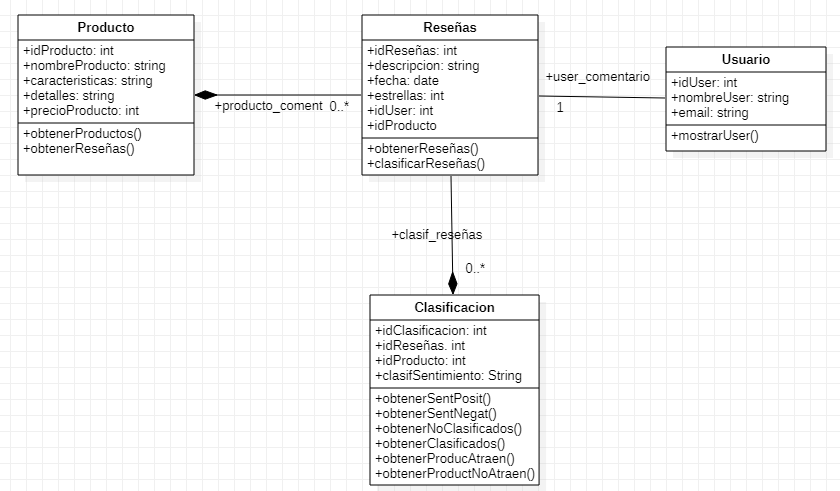








# **Diagrama de clases**



En este diagrama podemos observar las clases que se tendrán en cuenta para el desarrollo del algoritmo, “Productos” será la clase principal en donde estarán los datos necesarios respecto a los productos, además con los métodos se podrá obtener las reseñas sobre el producto.

En la clase “reseñas”, estarán todas las reseñas respectivas de cada producto diferenciadas por un id, en donde se realizará el método “clasificar reseñas”, para así llevar esa información a la clase “clasificación” en donde con ayuda de la implementación del algoritmos obtendremos las reseñas positivas, negativas, los productos que más atraen y menos atraen. En la clase usuario se tendrá para hacer referencia a la persona que está realizando la reseña.

# **Desarrollo de algoritmos**