Casa abierta al tiempo

UNIVERSIDAD AUTÓNOMA METROPOLITANA

Unidad Cuajimalpa

Plataforma WEB para la búsqueda y visualización de concordancias en documentos digitales.

Autor:

Emmanuel Ulisses González López

Profesores responsables:

M.C. Adriana Gabriela Ramírez de la Rosa

Dr. Esaú Villatoro Tello

TABLA DE CONTENIDOS

1 INTRODUCCIÓN.

Actualmente en esta era de la información es muy fácil para cualquier persona descargar y recolectar grandes cantidades de información de diferentes fuentes, como por ejemplo noticias de periódicos en línea, información producida en blogs, mensajes extraídos de redes sociales, y documentos digitales de diferentes formatos. Sin embargo, pocos son los usuarios que tienen la facilidad de hacer un análisis cuantitativo y/o cualitativo de todas estas fuentes de información, pues normalmente el poder hacerlo significa contar con ciertas habilidades de programación y/o tener acceso a herramientas especializadas. El principal problema para los interesados en investigar recursos masivos de información, es la búsqueda de palabras clave y el uso que se les da a estas palabras mediante un análisis de concordancias por medio de herramientas especializadas.

Dichas herramientas no están disponibles para la mayoría del público ya que tienen un alto costo de adquisición y eso muestra una gran limitante a la hora de querer realizar estos análisis. Agregado a esto, las pocas herramientas que existen tienen una gran cantidad de complejas características e interfaces poco amigables, lo que genera una dificultad de uso para usuarios poco especializados y en consecuencia, éstos no son capaces de sacar el máximo provecho de estas herramientas.

Para resolver esta limitante de accesibilidad y usabilidad se ha desarrollado una plataforma WEB de acceso gratuito. La cual cuenta con una interfaz gráfica amigable e intuitiva que permitirá a los usuarios subir archivos, crear corpus, calcular frecuencias, generar árbol de concordancias con el menor número de clics. La plataforma está dirigida para cualquier usuario con interés de realizar un análisis cuantitativo y cualitativo, con el fin de agilizar el proceso de investigación de algún tema en particular. Y así los usuarios podrán familiarizarse más con este tipo de herramientas y los alcances que tienen al analizar grandes cantidades de información.

Una de las características importantes de esta herramienta es su escalabilidad, ya que alguien más puede retomar el proyecto y añadir otras características de análisis de información y graficación de resultados con gran facilidad. Cada uno de sus componentes está diseñado de tal forma que modificarlos y/o añadir nuevas características a la plataforma WEB sea sencillo y no afecte al comportamiento entero de la aplicación.

Sin embargo lo que hace sobresalir a esta herramienta de las demás, es que hace énfasis en la visualización de estos resultados, al ser una herramienta WEB, hace uso de una biblioteca especializada para manipular datos y generar gráficas, estas gráficas facilitan la representación de la información de una manera más amigable para el usuario y esté pueda sacarle todo el provecho a esta característica.

1.1 OBJETIVO GENERAL.

Diseñar y desarrollar una herramienta WEB escalable que permita realizar un análisis cuantitativo y cualitativo por medio de la extracción de distribución de frecuencias y concordancias a grandes cantidades de información basadas en texto y generar la visualización de dichos análisis por medio de librerías de graficación.

1.2 OBJETIVOS PARTICULARES:

1. Diseñar e implementar una herramienta en línea que permita a diversos usuarios hacer compilaciones de corpus (conjunto de varios documentos digitales).
2. Implementar métodos que permitan hacer un análisis cuantitativo y cualitativo de un corpus, como el análisis de vocabulario a través de distribución de frecuencias.
3. Desarrollar un método que permita hacer la búsqueda y análisis de concordancias por medio de consultas simples, consultas de múltiples palabra y expresiones regulares.
4. Implementar esquemas de visualización de información que faciliten al usuario la visualización de los resultados de los análisis.

2 MARCO TEÓRICO.

En esta sección se hará mención de la definición de algunos conceptos clave necesarios para hacer más fácil la comprensión de las características de la plataforma WEB desarrollada en este trabajo y se mostraran ejemplos de las gráficas que la aplicación genera a partir de una colección de datos específica. Esta colección de datos formada a partir de 24 archivos de texto, que tienen en promedio 500 palabras cada archivo, representan documentos de noticias de desastres naturales y en particular de huracanes, los cuales fueron proporcionados por los asesores de este trabajo.

2.1 DISTRIBUCIÓN DE FRECUENCIAS.

Se llama distribución de frecuencias al número de observaciones por categoría de datos agrupados en clases mutuamente excluyentes. (Montgomery, Runger, & Medal, 1996). Una categoría es el resultado de una clasificación en pares del tipo (llave, valor). Esta plataforma WEB usa dicha categoría para representar una palabra como llave y el valor es el número de observaciones, es decir, la frecuencia con la que aparece en un conjunto de documentos.

Para calcular una distribución de frecuencias de palabras, la plataforma WEB lee cada palabra dentro de una colección de archivos y hace uso de la interface Map (java.io.Map) (Oracle, s.f.). en Java, que permite representar una estructura de datos para almacenar pares de valores "*clave/valor*", de tal manera que para cada *clave* tenemos un dato único, en este caso, una palabra, y el *valor* es el número de veces que se repite dicha palabra. A continuación se describe el algoritmo diseñado para calcular la distribución de frecuencias.

Algorithm 1 Frequency

```
procedure FREQUENCY(Archivos)                    ▷ Uno o varios archivos
frecuencias[](K, V)                              ▷ Lista de pares de valores
    for cada archivo en Archivos do
        palabras[] ← palabras en archivo
        for cada palabra en palabras[] do
            valor ← V en frecuencias(palabra, V)
            if valor :== NULL then
                frecuencias(K, ) ← palabra
                frecuencias( ,V) ← 1
            else
                frecuencias(palabra, V) ← V +1
            end if
        end for
    end for
    return frecuencias[](K, V)                   ▷ una lista de pares de valores
end procedure
```

Ilustración 1. Algoritmo diseñado para el cálculo de frecuencias.

Histogramas: En la ilustración 1 se muestra una gráfica de distribución de frecuencias en la que se puede observar que aparecen varias palabras ordenadas de mayor a menor, en donde las palabras con mayor frecuencia son las más utilizadas dentro de un documento o un conjunto de documentos. Notar que, por ejemplo, la palabra "huracán" tiene una frecuencia de 29 mientras que el resto de las palabras están por debajo de 20. Esta representación nos indica qué palabras y conceptos son más usados dentro de un conjunto de documentos.

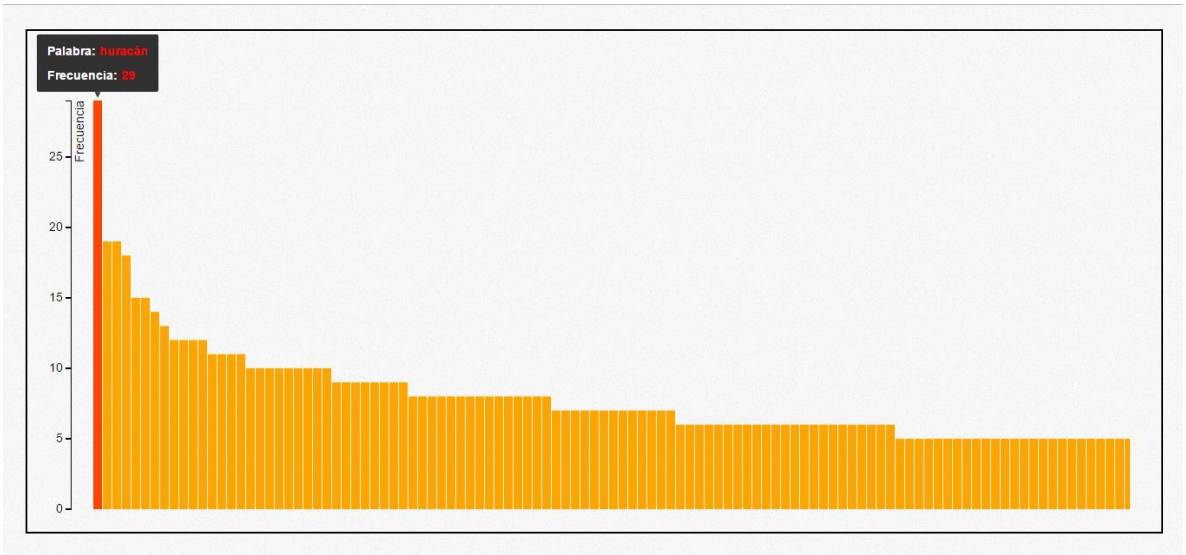


Ilustración 2. Ejemplo de distribución de frecuencias de palabras generado con la plataforma WEB desarrollada.

Nube de palabras: En la ilustración 2 se muestra una nube de palabras que son otra representación visual de una distribución de frecuencias donde el tamaño de las palabras es mayor cuando su frecuencia es sobresaliente. En el ejemplo de la ilustración 2, producida con el mismo conjunto de datos se puede identificar con mayor facilidad los conceptos más utilizados y se puede observar que palabras como “huracán”, “vientos”, “tormentas”, “inundaciones” son algunas de las más sobresalientes.

Ilustración 3. Nube de palabras generada con la plataforma WEB desarrollada

2.2 CONCORDANCIAS DE PALABRAS.

La concordancia es un recurso de la lingüística para marcar reglas gramaticales entre los diversos constituyentes mediante referencias cruzadas para poder identificar en qué contexto se está usando una palabra. (Bosque & Gutiérrez-Rexach, 2009).

En la ilustración 3 se muestra un ejemplo de las concordancias que existen alrededor del término “huracán” donde se puede observar claramente las palabras que rodean ese término y como se utiliza, y así determinar el contexto en el que se encuentran dichas palabras.

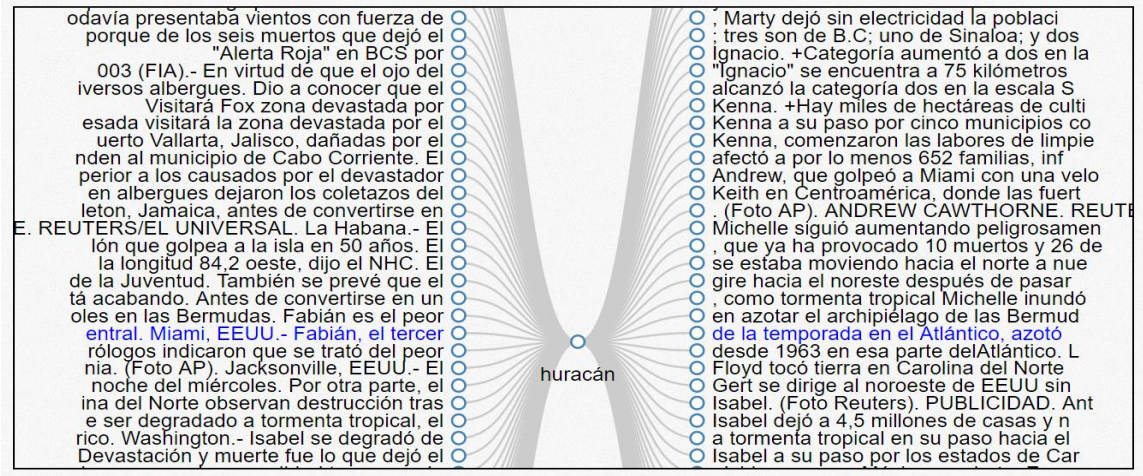


Ilustración 4. Ejemplo de un árbol de concordancias generado con la plataforma WEB

A continuación se describe el algoritmo diseñado para generar concordancias.

Algorithm 1 Concordances

Require: *Archivos*: Uno o varios archivos.
Require: *ContextoI*: Un entero que representa el número de caracteres a la izquierda de la consulta.
Require: *ContextoD*: Un entero que representa el número de caracteres a la derecha de la consulta.
Require: *Consulta*: Una plabra, multiples palabras o una expresión regular.

```
procedure CONCORDANCES(Archivos, ContextoI, ContextoD, Consulta)
    Concordancias[]
    for cada Archivo en Archivos do
        while Consulta se encuentra en Archivo do
            izquierdo ←NULL
            derecho ←NULL
            indiceI ← indice inicio donde encontro Consulta - ContextoI
            while indiceI < indice inicio donde encontro Consulta do
                izquierdo ← izquierdo + caracter en indiceI
                indiceI ← indiceI + 1
            end while
            indiceD ← indice final donde econtro Consulta
            limite ← indice final donde encontro Consulta + ContextoD
            while indiceD < limite do
                derecho ← derecho + caracte en indiceD
                indiceD ← indiceD + 1
            end while
            Concordancia ← nueva Concordancia(izquierdo, derecho)
            Conconrdancias agregar(Concordancia)
        end while
    end for
    return Concordancias[]                    ▷ una lista de concordancias
end procedure
```

Ilustración 5. Algoritmo diseñado para el cálculo de concordancias.

3 ESTADO DEL ARTE.

En esta sección se mostrará una tabla comparativa y una breve descripción de las diferentes aplicaciones que realizan análisis de textos digitales y sus principales características, junto con las características de este proyecto.

- TexStat: es un programa de escritorio simple para el análisis de textos digitales, que soporta diferentes formatos, cuenta con características básicas, como el cálculo de distribución de frecuencias, concordancias, y el uso de expresiones regulares, es de código libre, y el software es de uso gratuito, no cuenta con visualización gráfica muestra los resultados en texto plano. (Hüning, 2014).
- Nvivo: es un programa de escritorio con una gran cantidad de características, con el mismo propósito, acepta una variedad más amplia de documentos digitales, así como soporte para audio y vídeo, la adquisición del software es mediante una licencia de compra, y su código no es libre, cuenta con visualización gráfica. (International, 2017).
- Tropes Analysis: es un programa de escritorio diseñado especialmente para la clasificación semántica y análisis cualitativo de documentos digitales, de igual forma soporta un extenso grupo de formatos digitales, es de uso gratuito y su código no es libre, cuenta con visualización gráfica. (Pierre Molette, 2014).
- Sketch Engine: es un servicio WEB con características que se centran en determinar partes del lenguaje (sustantivos, verbos, adjetivos) y categorías gramaticales (singular, plural, presente, pasado, etc.) para buscar combinaciones de palabras e incluso estructuras gramaticales, no cuenta con visualización gráfica. (Adam Kilgarriff, 2003).

A continuación se mostrara una tabla comparativa, donde se puede observar las características más relevantes que contienen algunas de las herramientas existentes, y las presentes en la plataforma WEB desarrollada en este proyecto. Cabe mencionar que la herramienta se diseñó e implementó para ser un sistema más grande y que por el momento solo contará con características básicas de análisis, que son el cálculo de frecuencias de palabras y la generación de concorcondancias. De igual forma la plataforma incorpora técnicas de visualización de resultados de análisis, que facilitan el estudio cuantitativo y cualitativo, cosa que la mayoría de las herramientas disponibles no considera; ver Tabla 1.

Características	TextSTAT	Nvivo	Tropes Analysis	Sketch Engine	Plataforma WEB desarrollada
Código abierto	✓	×	×	×	✓
Soporta expresiones regulares	✓	✓	✓	✓	✓
Extrae concordancias	✓	✓	✓	✓	✓
Es de software gratuito	✓	×	✓	×	✓
Cálculo distribución de frecuencias	✓	✓	✓	✓	✓
Soporta audio y vídeo	×	✓	×	×	×
Análisis semántico	×	×	✓	✓	×
Genera gráfica de nube de palabras	×	✓	×	×	✓
Genera gráfica de barras	×	✓	×	×	✓
Genera gráfica de árbol de concordancias	×	✓	×	×	✓
Es una plataforma WEB.	×	×	×	✓	✓
Soporte para archivos					
PDF	×	✓	✓	✓	✓
DOC	✓	✓	✓	✓	✓
TXT	✓	✓	✓	✓	✓

Tabla 1. Tabla comparativa de las diferentes herramientas y la propuesta de desarrollada

Las características de otras herramientas como el análisis semántico de oraciones y el soporte para audio y video no se van a implementar en la propuesta ya que requiere de procesos más precisos y detallados para estudiar las propiedades del lenguaje y el significado de las palabras. Cabe mencionar que cada una de las herramientas presentes en la Tabla 1 fueron descargadas y usadas para delimitar que y como se podría realizar la propuesta aquí desarrollada.

4 DESARROLLO E IMPLEMENTACIÓN.

En esta sección se va a describir el trabajo realizado para la elaboración de la plataforma WEB y su implementación, análisis de requerimientos, que incluye el diseño de la base de datos, el diseño de los componentes de la plataforma y los diagramas de casos de uso.

4.1 DISEÑO DE LA BASE DE DATOS.

En la ilustración 4 se muestra el diagrama entidad relación de la base de datos. La base de datos se diseñó de tal manera que los usuarios, que van usar la plataforma, podrán subir archivos, eliminar archivos, crear y eliminar Corpus de manera sencilla, y que al mismo tiempo, ellos podrán ver qué archivos y qué Corpus tienen en todo momento. La base de datos cuenta con cinco entidades, las cuales son "ingreso", "archivo", "usuario", "corpus", "archivos_corpus". Cada una de estas entidades almacenara información relevante tanto para los usuarios, como para el sistema (para ver más detalles ver Anexo 1).

- "ingreso": almacena solo las credenciales de acceso a la plataforma para cada usuario que se registró.
- "archivo" almacena todos los archivos y las referencias de a qué usuario le pertenece cada archivo y los detalles de cada archivo.
- "usuario": almacena información personal, como nombres, apellidos y la dirección donde se almacenarán sus archivos (este atributo solo es visible para el sistema y el usuario no puede acceder o modificarlo).
- "corpus": almacena información de los corpus qué el usuario haya creado, como nombre, descripción y las referencias de a que usuario pertenece cada corpus.
- "archivos_corpus": esta entidad es usada por el sistema para hacer la unión de los archivos que pertenecen a un corpus y un usuario en particular.

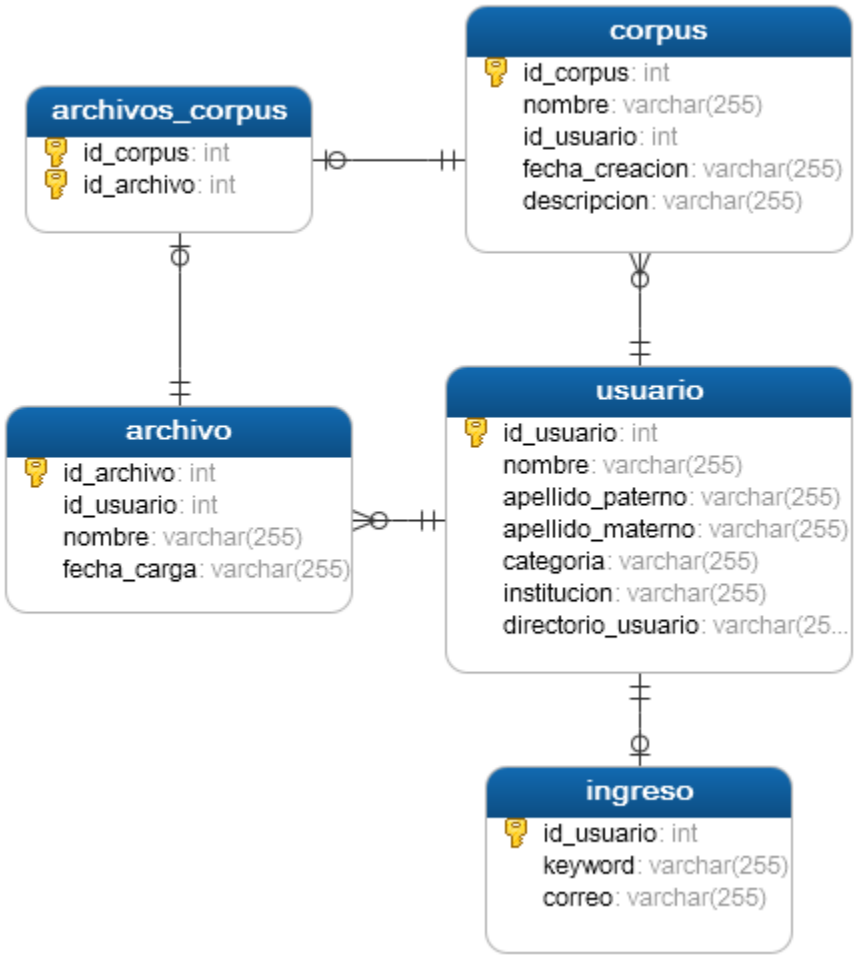


Ilustración 6. Diagrama entidad relación de la base de datos de la plataforma WEB.

4.2 ARQUITECTURA GENERAL DE LA PLATAFORMA.

En la ilustración 5 se muestra el diagrama de componentes de la plataforma WEB desarrollada, donde podemos observar que cuenta con cinco componentes, los cuales son "Gestor de usuarios" "Base de datos", "Gestor de archivos", "Conversor de formato", "Módulo de visualización", "Operaciones de análisis de textos". Cada uno de éstos realizan las acciones con las que la plataforma cuenta y sus dependencias con otros componentes (para más detalles ver Anexo 2).

- "Gestor de usuario": Este componente se encarga de los registros de usuarios nuevos en la base de datos y posteriormente la validación de usuarios registrados, y tiene una relación directa con el usuario.

- “Base de datos”: La base de datos se utiliza para almacenar la información de los usuarios y todo el contenido digital que deseen subir.
- “Gestor de archivos”: El gestor de archivos tiene la función de permitirle al usuario gestionar el contenido digital que desee subir o eliminar de su cuenta. También la aplicación tiene acceso a funciones como consultar archivos o corpus cuando el usuario desea hacer algún tipo de análisis efectuado por el componente de operaciones de análisis de textos.
- “Conversor de formato”: El conversor de formato tiene la funcionalidad de convertir cualquier formato en texto plano para facilitar el trabajo al módulo de operaciones de análisis.
- “Módulo de visualización”: Este componente tiene la funcionalidad de graficar las operaciones de análisis y tiene una dependencia directa con el componente de operaciones de análisis de textos.
- “Operaciones de análisis”: Este componente es el que realiza el cálculo de frecuencias y la extracción de concordancias a partir de uno o varios documentos digitales. Se comunica con el gestor de usuarios y el conversor de formato.

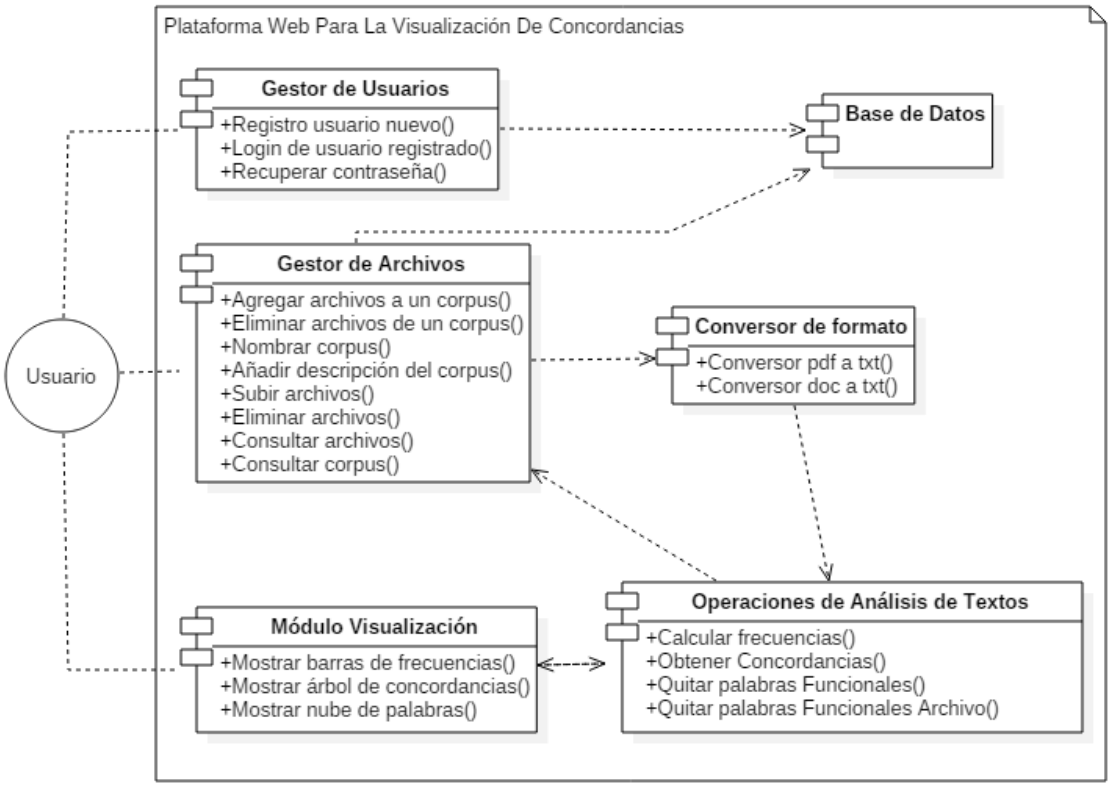


Ilustración 7. Diagrama de componentes de la plataforma WEB.

4.3 CASOS DE USO.

A continuación en las ilustraciones 8, 9 y 10 se mostrarán los casos de uso de las diferentes páginas con las que cuenta la plataforma WEB las cuales son “Página de inicio”, “Página de prueba”, “Página principal para usuarios registrados”. Ver anexo 3.

Casos de uso para la “Página de inicio”:

- Entrar: Esta operación es para usuarios ya registrados y su función es desplegar un formulario donde el usuario ingresara su cuenta y contraseña para poder acceder a la plataforma principal.
- Probar: Esta operación re direcciona al usuario a una página nueva, en donde un usuario no registrado podrá hacer uso de las funciones básicas de la plataforma.
- Registrarse: Esta operación desplegara al usuario un formulario el cual deberá llenar para ser un usuario registrado y hacer uso de todas las funcionalidades de la plataforma.

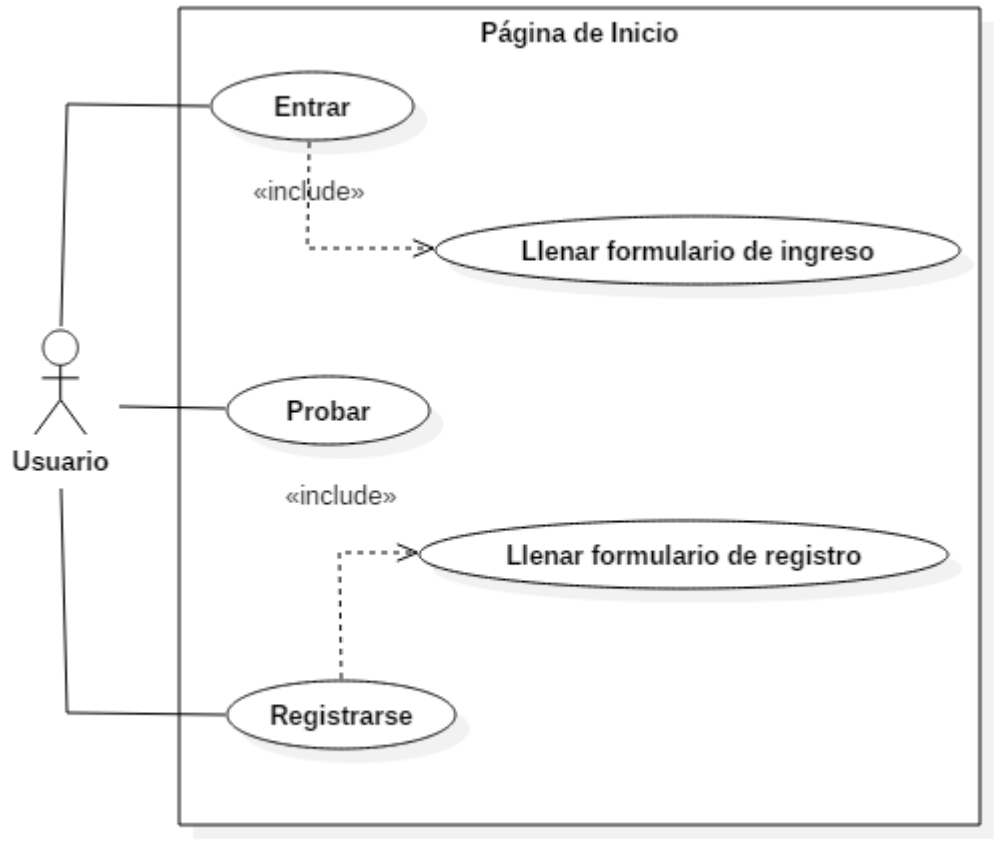


Ilustración 8. Diagrama de caso de uso para la página de inicio de la plataforma WEB.

Casos de uso para la “Página de prueba”:

- Escribir texto: Esta operación es para que el usuario escriba un texto limitado al que se le puede aplicar alguna de las operaciones de análisis.
- Pestaña de cálculo de concordancias: Esta operación mostrara el formulario para generar concordancias.
- Pestaña de cálculo de frecuencias: Esta operación mostrara dos botones que generan diferentes gráficos de una distribución de frecuencias.
- Limpiar área de texto: Esta operación borrará todo el contenido del área de texto para escribir o pegar contenido nuevo.
- Copiar texto de artículo: Esta operación copiará el texto de alguno de los artículos que la aplicación proporciona para facilitar el uso de la página de prueba.
- Pegar texto: Esta operación pegará el texto al que se le realizará alguna de las operaciones de análisis en el área de prueba.
- Generar nube de palabras: Esta operación requiere que se haya llenado el área de texto, si está lleno producirá una gráfica de nube de palabras con la información que haya sido escrita.

- Generar grafico de barras: Esta operación requiere que se haya llenado el área de texto, si está lleno producirá una gráfica de barras con la información que haya sido escrita.
- Generar árbol de concordancias: Esta operación requiere que se haya llenado el área de texto y el formulario de la pestaña de cálculo de frecuencias, si está lleno producirá una gráfica de árbol de las concordancias que haya encontrado con la información escrita en el área de texto.

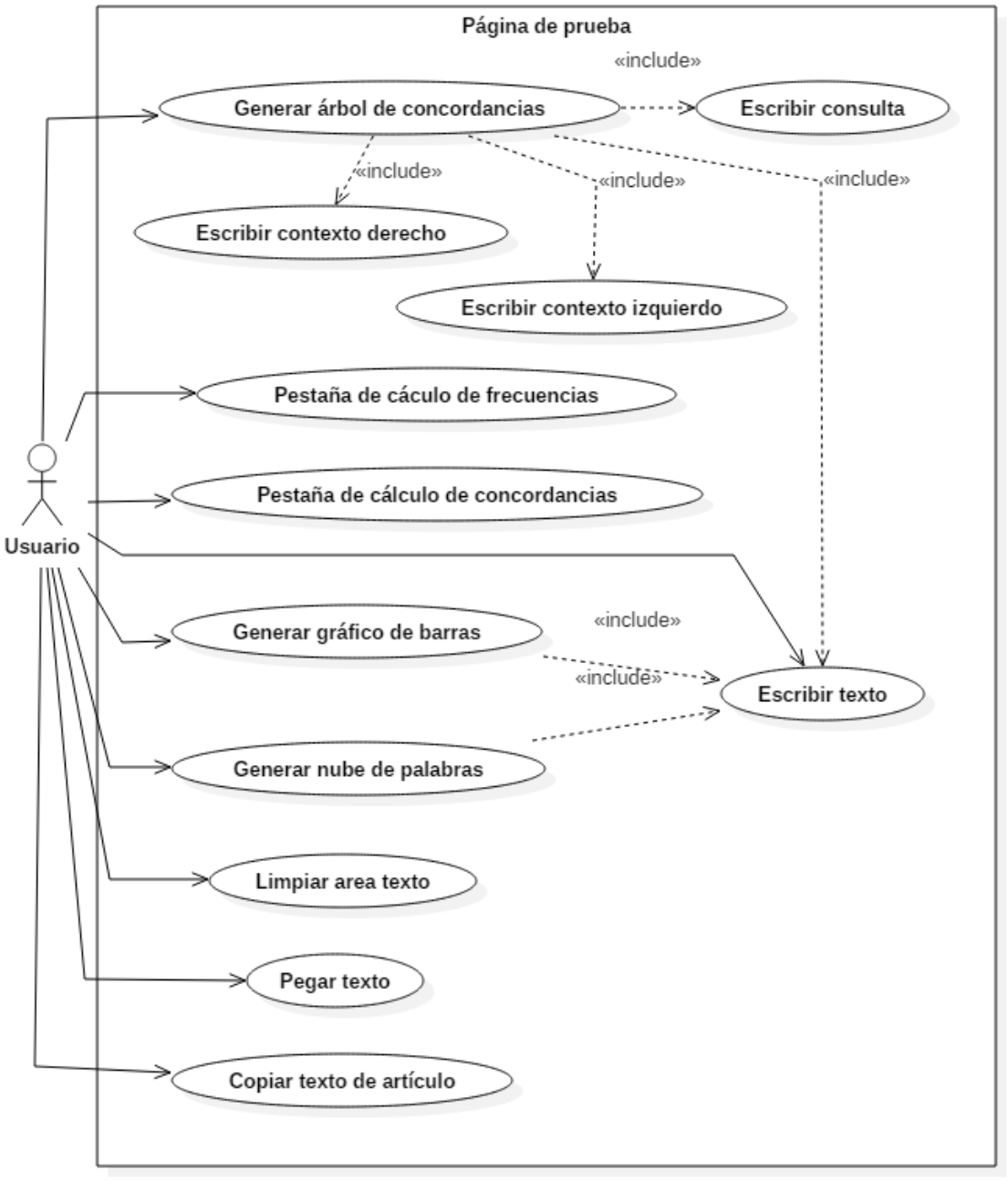


Ilustración 9. Diagrama de caso de uso para la página de prueba de la plataforma WEB.

Casos de uso para la “Página principal para usuarios registrados”:

- Inicio: Esta operación mostrará la página de inicio donde se le da la bienvenida al usuario registrado y se mostrará un pequeño manual de cómo usar la plataforma WEB.

- Frecuencias: Esta operación mostrara una pestaña de las opciones de análisis para el cálculo de frecuencias y se mostrara en todo momento los archivos y corpus que el usuario tenga en dos tablas diferentes.
- Concordancias: Esta operación mostrara una pestaña de las opciones y el formulario de la consulta para la generación de concordancias de igual forma se mostrara en todo momento los archivos y corpus que tenga el usuario para seleccionar aquellos a los que quiera realizar le algún tipo de análisis.
- Salir de la plataforma: Esta operación cerrara la sesión actual de un usuario y lo regresara a la página de inicio de la plataforma.
- Subir archivos: Esta operación subirá los archivos que se hayan seleccionado por medio de la operación “Seleccionar archivos a subir”.
- Seleccionar archivos a subir: Esta operación abrirá un explorador de archivos para que el usuario seleccione todos los archivos que quiera subir, esta operación requiere que el usuario este en la pestaña de frecuencias.
- Eliminar palabras vacías ES: Esta operación eliminara todas las palabras vacías del idioma español que encuentre al realizar un cálculo de frecuencias, esta operación requiere que el usuario este en la pestaña de frecuencias.
- Eliminar palabras vacías EN: Esta operación eliminara todas las palabras vacías del idioma ingles que encuentre al realizar un cálculo de frecuencias, esta operación requiere que el usuario este en la pestaña de frecuencias.
- Subir archivo de palabras vacías: Esta operación es para que el usuario pueda subir un archivo propio de palabras vacías al realizar un cálculo de frecuencias, esta operación requiere que el usuario este en la pestaña de frecuencias.
- Eliminar archivo de palabras vacías: Esta operación es para limpiar el campo del archivo de palabras vacías, esta operación requiere que el usuario este en la pestaña de frecuencias.
- Crear corpus: Esta operación creara un corpus a partir de uno o varios archivos seleccionados, esta operación requiere que el usuario haya seleccionado varios archivos de la tabla de archivos y estar en la pestaña de frecuencias o concorcondancias.
- Eliminar seleccionados: Esta operación eliminara los archivos o corpus seleccionados dependiendo en que tabla se haya activado, esta operación requiere que el usuario haya seleccionado uno o más archivos o uno o más corpus para proceder con su eliminación.
- Generar gráfico de barras: Esta operación generara un gráfico de barras a partir de las opciones de consulta que haya especificado el usuario, esta operación requiere que el usuario haya seleccionado uno o más archivos o uno o más corpus.
- Generar nube de palabras: Esta operación generara una nube de palabras a partir las opciones de consulta que haya especificado el usuario, esta operación requiere que el usuario haya seleccionado uno o más archivos o uno o más corpus.

- Generar árbol de concordancias: Esta operación generara un gráfico en forma de árbol, esta operación requiere que el usuario haya seleccionado uno o más archivos o uno o más corpus y que haya llenado correctamente el formulario de la consulta.

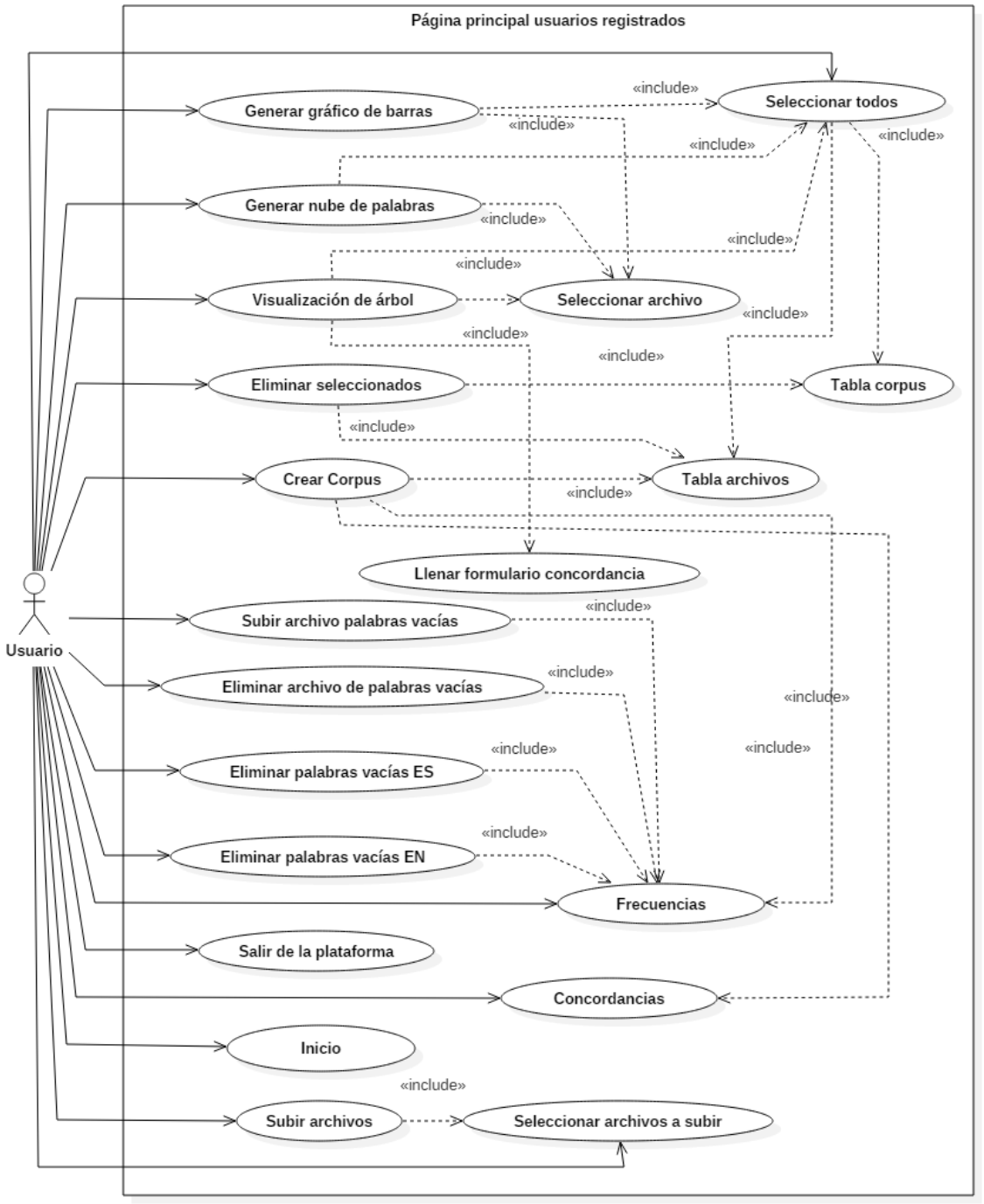


Ilustración 10. Diagrama de caso de uso para la página de usuarios registrados de la plataforma WEB.

5 DISEÑO DE LA INTERFAZ.

En esta sección se muestra el diseño de la interfaz de la plataforma WEB con capturas de pantalla de sus diferentes páginas, y una descripción de las partes que componen cada página. A lo largo de esta sección se habla de "palabras vacías" las cuales son palabras sin significado como, artículos, pronombres, preposiciones o aquellas palabras comúnmente más usadas en un lenguaje.

5.1 PÁGINA DE INICIO DE LA PLATAFORMA WEB.

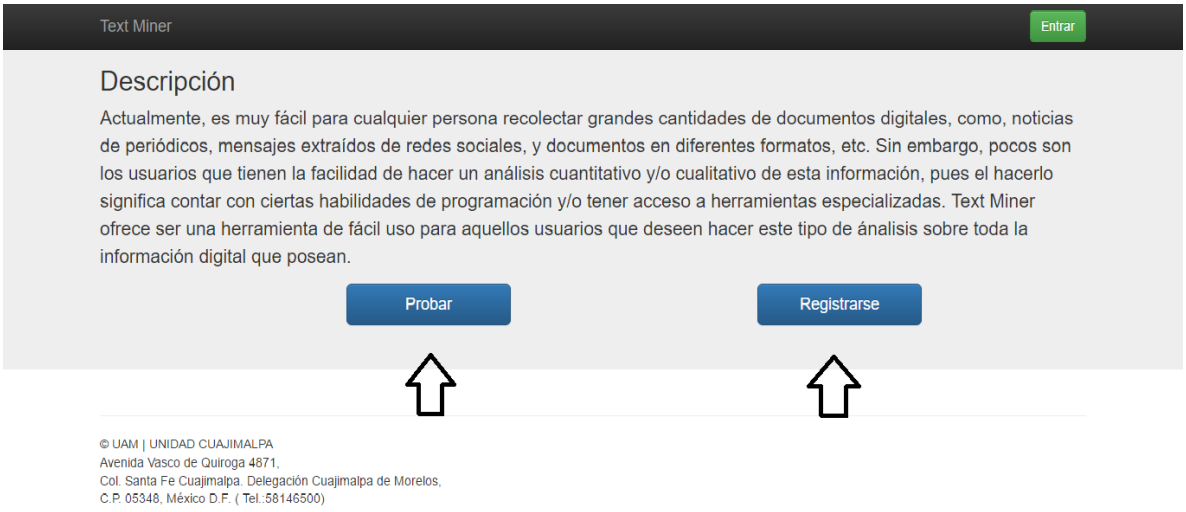


Ilustración 11. Página principal donde se describe el objetivo de la aplicación, con botones para su registro, ingreso o prueba.

La intención de la página mostrada en la ilustración 11 es describir a grandes rasgos qué es y para quién está dirigida la aplicación. Se puede observar que cuenta con 3 botones marcados con flechas, los cuales son "Probar", "Registrarse" y "Entrar".

- Probar: lleva a la página de prueba. Ver ilustraciones 14, 15 y 16.
- Registro: muestra una ventana emergente con un formulario que el usuario debe llenar para poder registrarse y hacer uso de la aplicación y la totalidad de sus funciones. Ver ilustración 12.
- Entrar: muestra una ventana con un formulario donde el usuario ingresa su correo y contraseña para poder ingresar a la página de usuarios registrados. Ver ilustración 13.

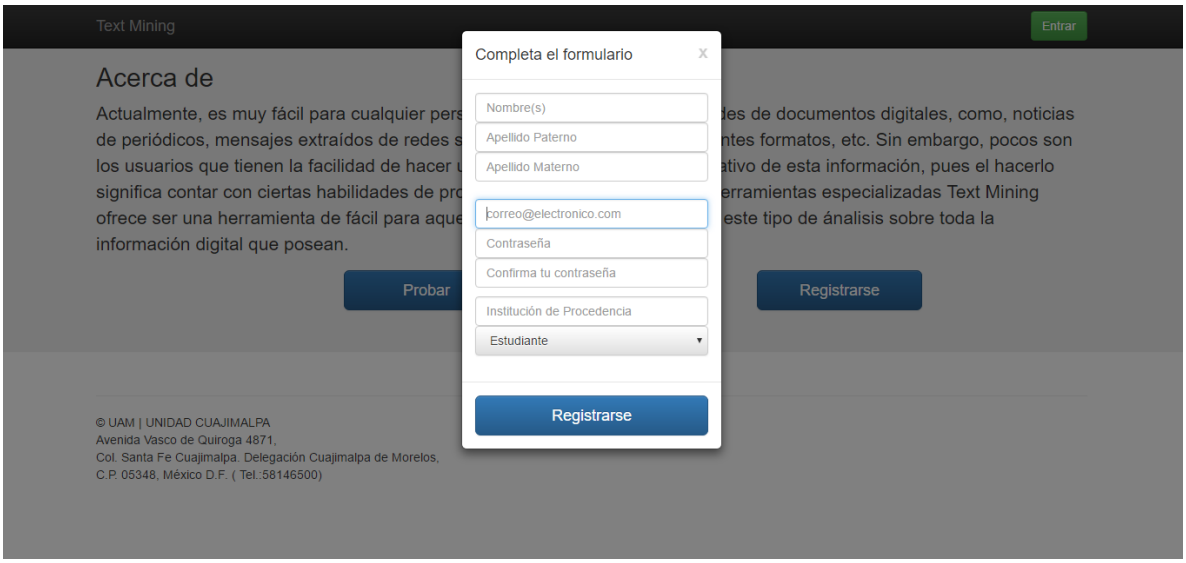


Ilustración 12. Formulario para el registro de un usuario nuevo.

Ilustración 12. En esta ilustración se muestra el formulario que aparece tras presionar el botón de registrarse ya sea en la página de inicio o en la página de prueba. Los campos que debe llenar son obligatorios y son: nombre(s), apellido materno, apellido paterno, correo electrónico, su contraseña, una confirmación de la contraseña, la institución a la que pertenece el usuario, y una lista de categorías (estudiante, profesor o externo) para saber la cantidad de usuarios de cada categoría que se registran.

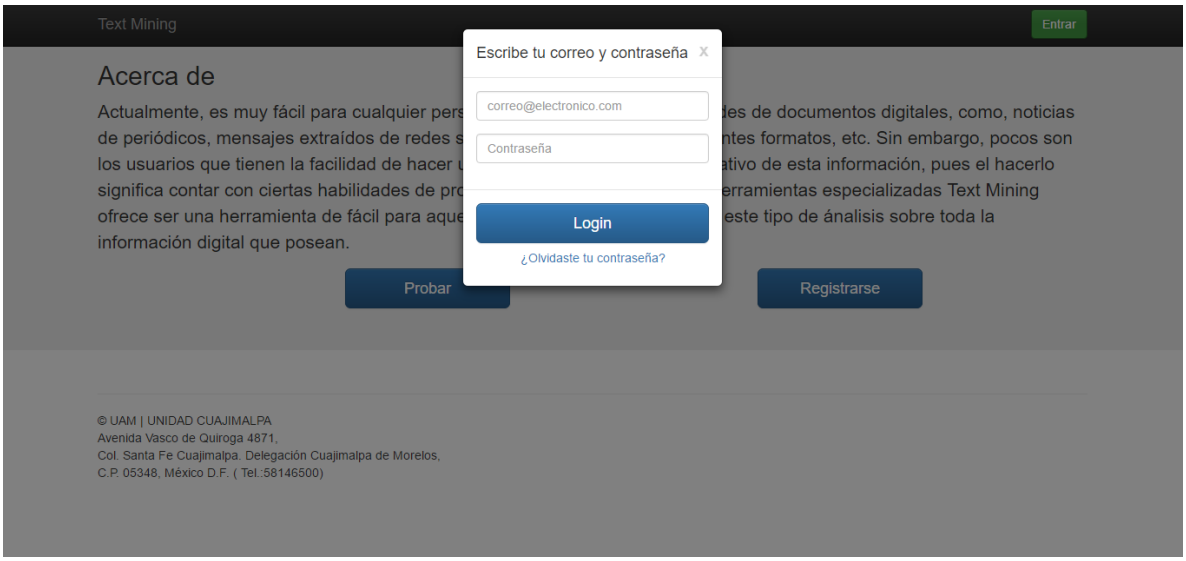


Ilustración 13. Formulario para el ingreso de un usuario registrado.

Ilustración 11. En esta ilustración se muestra el formulario que aparece tras presionar el botón de entrar, los campos del formulario solo son el correo electrónico y la contraseña de igual forma son obligatorios.

5.2 PÁGINA DE PRUEBA DE LA PLATAFORMA WEB.

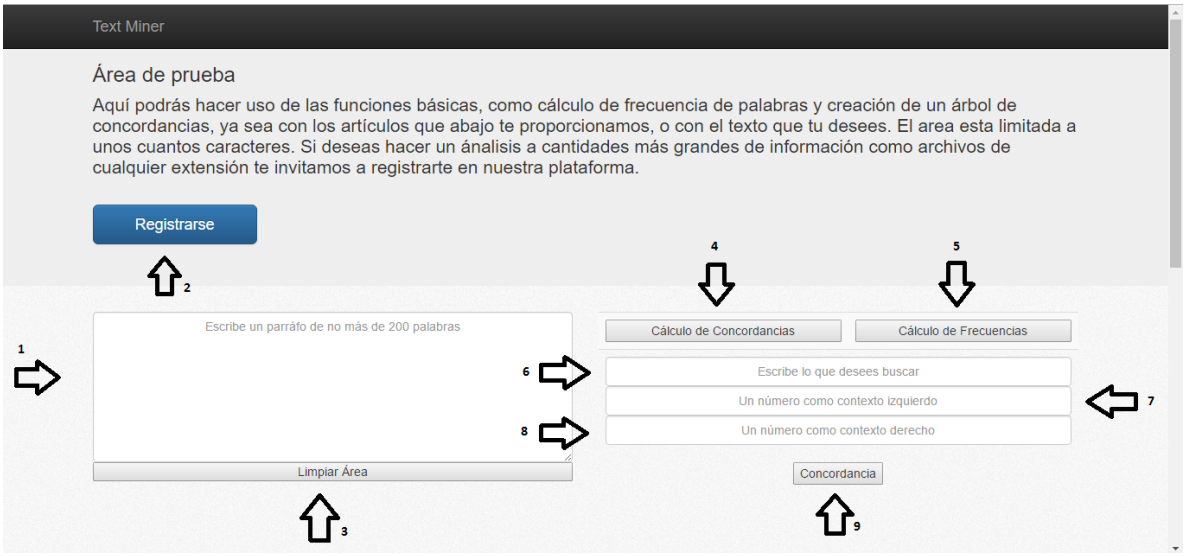


Ilustración 14. Captura de pantalla de la página de prueba con la pestaña de cálculo de concordancias activa.

La intención de la página mostrada en la Ilustración 14 es ofrecer al usuario un área en la cual pueda hacer uso de las diferentes características básicas de la plataforma WEB.

1. Área texto de prueba donde el usuario podrá ingresar una limitada cantidad de caracteres para hacer uso de las características básicas de la aplicación o pegar uno de los artículos que se le proporcionan en la misma página. Ver ilustración 16.
2. Botón para registrarse. Una vez que el usuario haya probado la aplicación y sea de su agrado podrá registrarse aquí mismo. Ver ilustración 12.
3. Botón para limpiar el área texto de prueba. Este botón borrara todo el contenido que haya sido escrito en esta área.
4. Botón para cambiar la pestaña de análisis. En esta ilustración la pestaña activa es la de cálculo de concordancias.
5. Botón para cambiar la pestaña de análisis, de concordancias a cálculo de frecuencias. Ver ilustración 15.
6. Campo de entrada para consultar una palabra, numero o expresión regular.
7. Campo para ingresar un número que representa la cantidad de caracteres que tendrá el contexto izquierdo a partir de la consulta que se realizó.
8. Campo para ingresar un número que representa la cantidad de caracteres que tendrá el contexto derecho a partir de la consulta que se realizó.
9. Botón que generará la visualización de árbol de una concordancia. Ver ilustración 17.

Ilustración 15. Captura de pantalla de la página de prueba con la pestaña de cálculo de frecuencias activa.

Ilustración 13. En esta ilustración se muestran los mismos componentes que la ilustración 12 con la diferencia de que en esta se tiene activa la pestaña de cálculo de frecuencias.

1. Este botón generara un gráfico de barras a partir del texto que se haya insertado en el área de texto y aparecerá inmediatamente debajo del área de prueba. Ver ilustración 18.
2. Este botón genera un gráfico de nube de palabras a partir del texto que se haya insertado en el área de texto y aparecerá inmediatamente debajo del área de prueba. Ver ilustración 19.

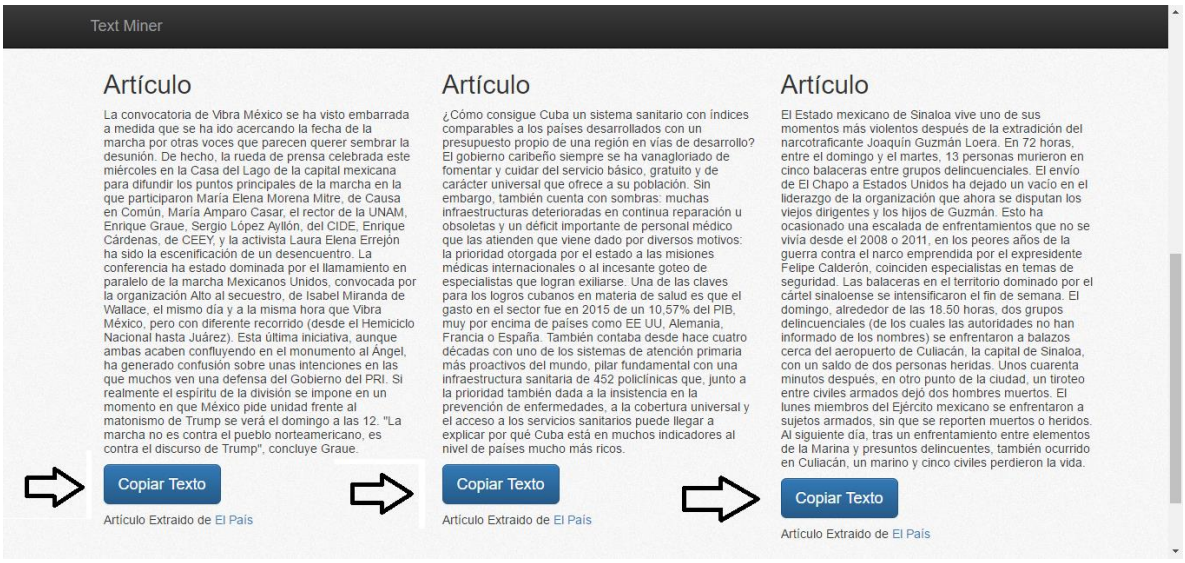


Ilustración 16. Artículos de prueba que la aplicación proporciona para hacer uso de las operaciones de análisis.

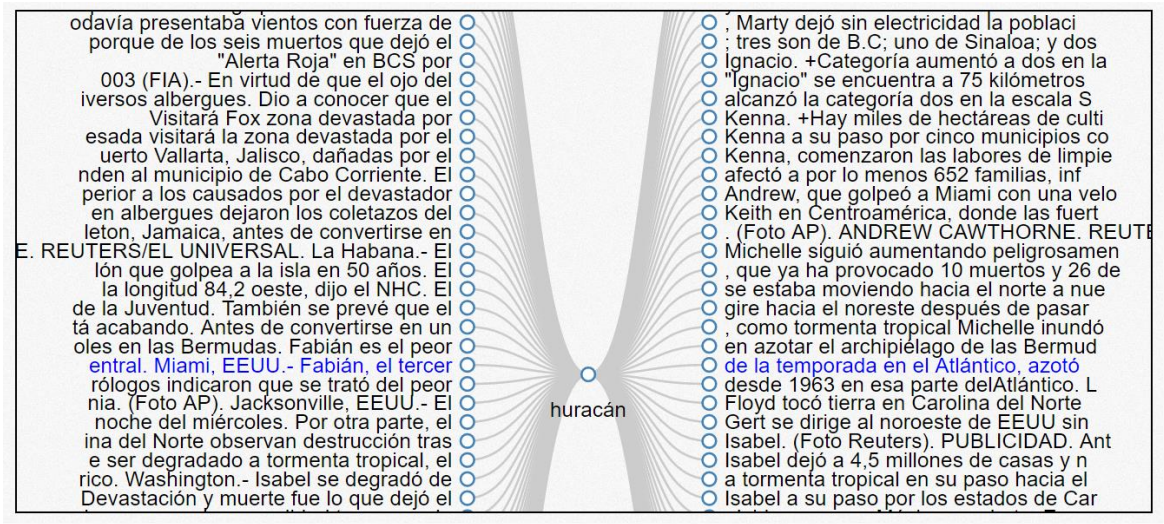


Ilustración 17. Graficación de una concordancia de palabras.

Ilustración 17. En esta ilustración se puede ver la generación de una concordancia de palabras, donde se puede observar de color azul el contexto izquierdo, el contexto derecho y al centro del árbol, la palabra que se consultó. De esta forma podemos apreciar el contexto en el que se encuentra dicha palabra y como se usa. Esta aparece inmediatamente debajo del área de prueba.

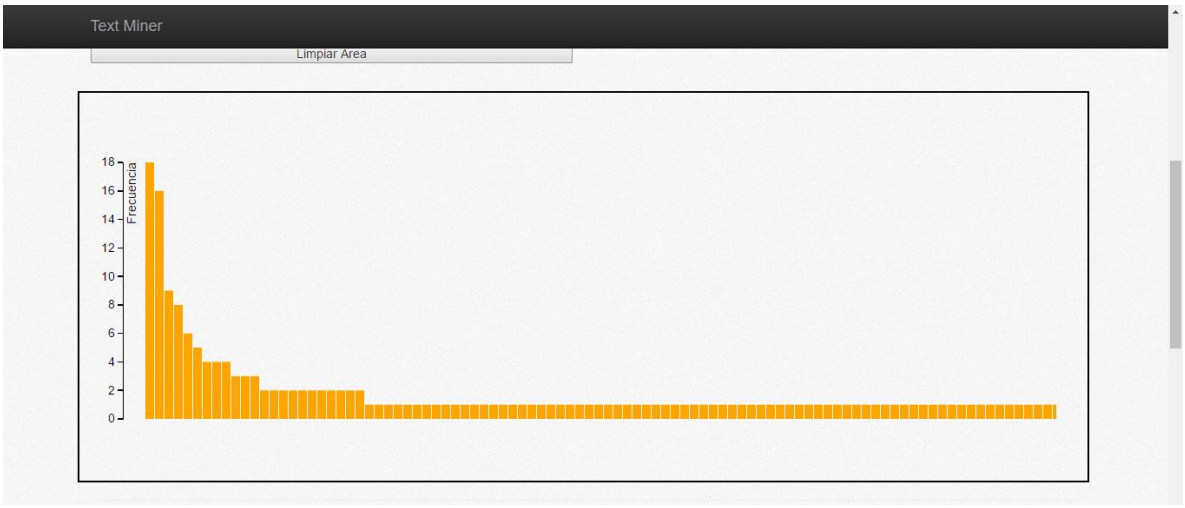


Ilustración 18. Generación de un gráfico de barras en la página de prueba.

Ilustración 18. Aquí se puede observar una forma de visualización de un cálculo de frecuencias, donde las palabras que más se repiten dentro de un artículo son las que tienen una mayor frecuencia y así darnos una idea de los conceptos que se manejan.

Ilustración 19. Graficación de una nube de palabras en la página de prueba.

Ilustración 19. Aquí podemos observar otra forma de visualización de un cálculo de frecuencias donde el tamaño de las palabras es proporcional a la frecuencia con la que aparece en el texto que se introduce en el área de prueba.

5.3 PÁGINA PRINCIPAL PARA USUARIOS REGISTRADOS DE LA PLATAFORMA WEB.

Ilustración 20. Página principal para usuarios registrados en la pestaña de inicio.

Ilustración 20. En esta ilustración se muestra la página principal para usuarios registrados donde se llega luego de presionar el botón de entrar en la página de inicio y haber ingresado sus datos de ingreso. La página de inicio le da una bienvenida al usuario que ingresó y se describe una breve guía de lo que se puede hacer y de cómo debe hacerse.

1. Pestaña de inicio activa en esta ilustración.
2. Pestaña de cálculo de frecuencias.
3. Pestaña de cálculo de concordancias.
4. Botón para seleccionar archivos que se desean subir.
5. Botón para efectuar la carga de los archivos anteriormente seleccionados.
6. Botón para cerrar la sesión del usuario actual.

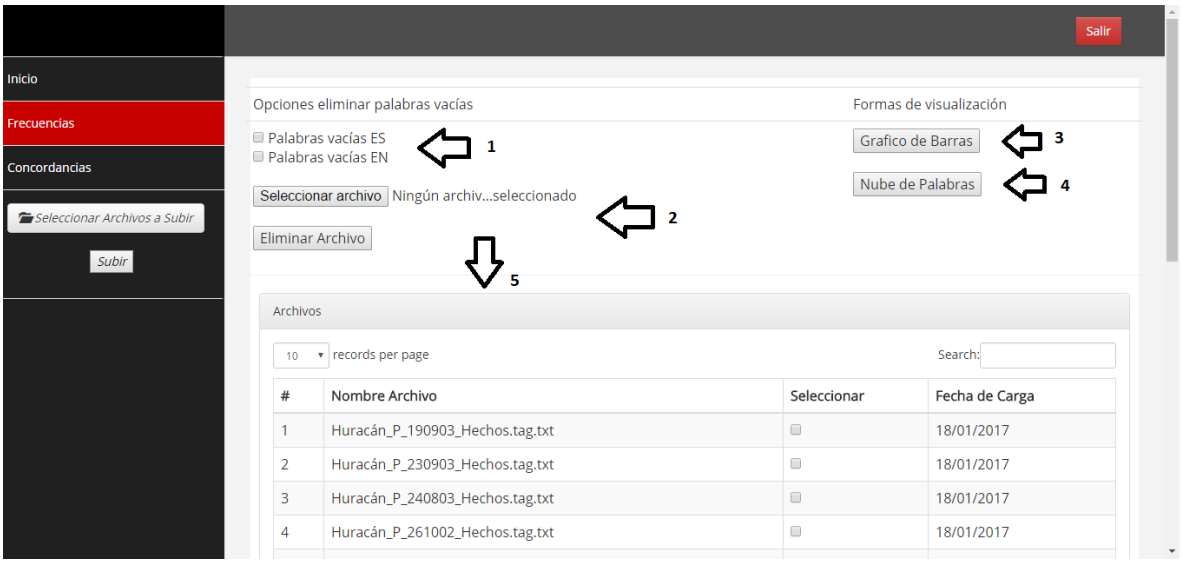


Ilustración 21. Pestaña de frecuencias para usuarios registrados.

Ilustración 21. En esta ilustración se puede observar que la pestaña activa es la de frecuencias. Aquí es donde el usuario puede seleccionar archivos o corpus que ya haya subido o creado con anterioridad para aplicarles un cálculo de frecuencias.

1. Opciones para eliminar palabras vacías en idioma inglés o español.
2. Opción para que el usuario pueda subir un archivo con sus palabras vacías propias y un botón para eliminarlo una vez usado.
3. Botón para generar el grafico de barras a partir de la consulta del usuario.
4. Botón para generar una nube de palabras a partir de la consulta del usuario.
5. Tabla que se mostrará en todo momento con los archivos y corpus del usuario, si las pestañas de frecuencias o concordancias están activas.

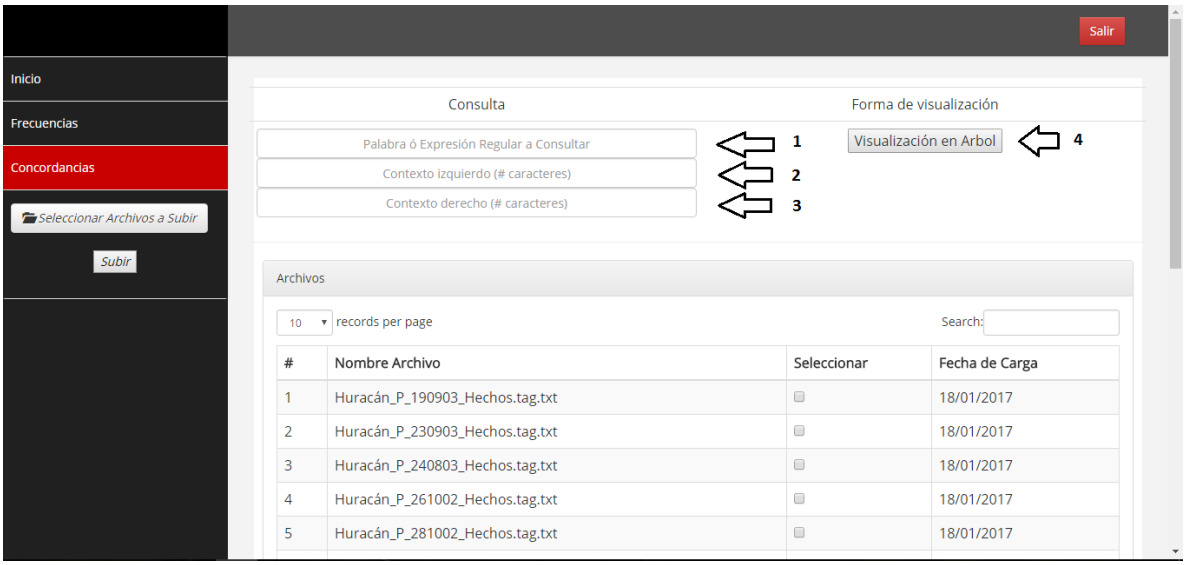


Ilustración 22. Pestaña de concordancias para usuarios registrados.

Ilustración 22. En esta ilustración se puede observar que la pestaña activa es la de concordancias. Aquí es donde el usuario puede seleccionar archivos o corpus que ya haya subido o creado con anterioridad para aplicarles un cálculo de frecuencias.

1. Entrada de texto para la consulta, puede ser un número, una palabra o una expresión regular.
2. Entrada de texto que debe ser un número, el cual representa el contexto izquierdo que se quiere mostrar a partir de la palabra que se está consultando.
3. Entrada de texto que debe ser un número, el cual representa el contexto derecho que se quiere mostrar a partir de la palabra que se está consultando.

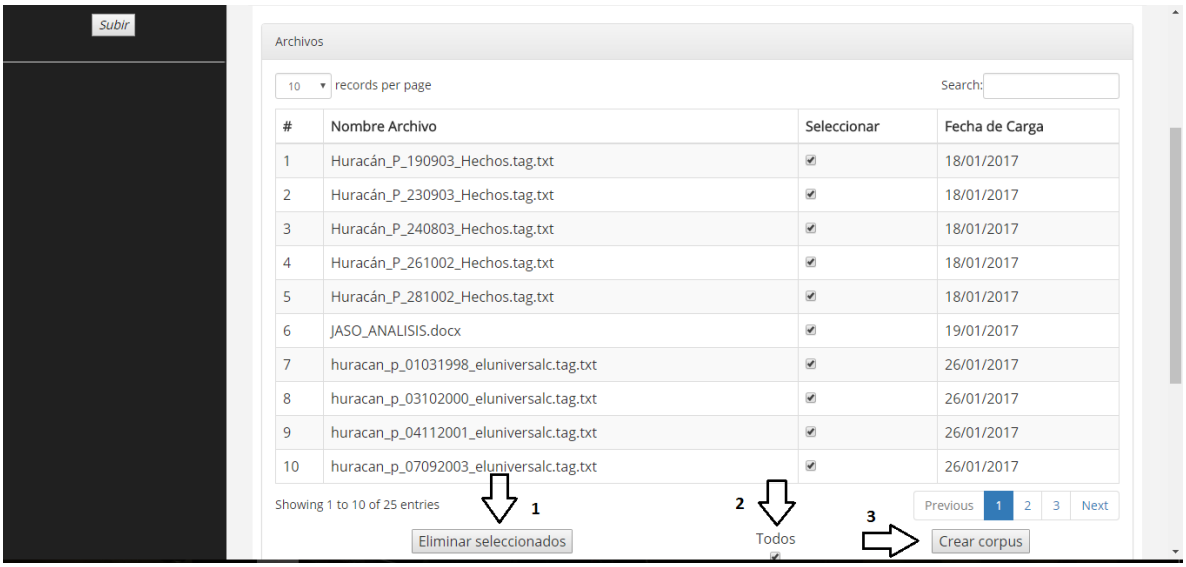


Ilustración 23. Tabla de archivos para un usuario registrado.

Ilustración 23. En esta ilustración se puede apreciar una tabla que contiene todos los archivos pertenecientes a un usuario, esta es una tabla dinámica que puede ordenar los archivos de diferentes formas y mostrar diferentes cantidades de archivos.

1. Botón para eliminar archivos, este botón le permite al usuario eliminar archivos una vez seleccionados.
2. Caja de selección que selecciona todo los archivos de la tabla para ahorrar tiempo al querer seleccionar varios archivos.
3. Botón para crear corpus, este botón creara un corpus a partir de los archivos seleccionados esto para facilitar el proceso de análisis a grandes cantidades de documentos.

Cabe mencionar que todas las gráficas generadas en la página de usuarios registrado abren una pestaña nueva en el explorador que esté usando el usuario.

6 TRABAJO A FUTURO Y CONCLUSIONES.

En esta sección se analizan los resultados obtenidos a lo largo del desarrollo de la plataforma WEB, realizando conclusiones finales y presentando las opciones de trabajo a futuro que permitirán darle continuidad a este proyecto.

6.1 CONCLUSIONES.

El uso de herramientas de minería de textos permite a los usuarios hacer un análisis cuantitativo y cualitativo de grandes cantidades de información. Sin embargo no todos los usuarios tienen los recursos para comprar una aplicación de este tipo y aquellas que sí están disponibles de manera gratuita normalmente son muy complejas para un usuario poco especializado.

A su vez la mayoría de las aplicaciones existentes no proporcionan esquemas de visualización de análisis de información lo cual resulta poco amigable para el usuario poder hacer uso de esos resultados para sacar conclusiones de la calidad del contenido de las colecciones de datos que se están analizando.

Se puede asegurar que este proyecto ha alcanzado sus objetivos, al desarrollar una plataforma WEB de uso gratuito que permite a los usuarios hacer colecciones de datos y realizar análisis cuantitativo y cualitativo a través de cálculo de frecuencias y generación de concordancias, integrando esquemas de visualización amigables para el usuario que le facilitara el proceso de investigación y análisis a grandes cantidades de documentos. La liga para la plataforma WEB desarrollada en este trabajo es la siguiente.

http://hao.cua.uam.mx:8080/Corpus/pages/index.html

6.2 TRABAJO A FUTURO.

Como se describió en secciones anteriores la plataforma WEB está diseñada de tal manera que sea escalable para que con el paso del tiempo se pueda crear un sistema mucho más grande y robusto con más características de minería de textos, análisis de información y la creación de sus respectivas visualizaciones gráficas.

Actualmente este proyecto está siendo retomado por un compañero de la licenciatura en tecnologías de la información quien está realizando otro tipo de análisis de información, y que en un futuro lo integrará a la plataforma WEB con su respectiva visualización gráfica.

7 BIBLIOGRAFÍA

Adam Kilgarriff, P. R. (2003). *Sketch Engine | language corpus managment and query system*. Obtenido de https://www.sketchengine.co.uk/

Bosque, I., & Gutiérrez-Rexach, J. (2009). *Fundamentos de gramática formal (1ª edición)*. Madrid: Akal.

Hüning, D. M. (2014). *TextSTAT - Simple Text Analysis Tool*. Obtenido de http://neon.niederlandistik.fu-berlin.de/en/textstat/

International, Q. (2017). *QSR International*. Obtenido de http://www.qsrinternational.com/nvivo-spanish

Montgomery, D. C., Runger, G. C., & Medal, E. G. (1996). *Probabilidad y estadística aplicadas a la ingeniería.* McGraw Hill.

Oracle. (s.f.). *Oracle*. Obtenido de https://docs.oracle.com/javase/7/docs/api/java/util/Map.html

Pierre Molette, A. L. (2014). *Semantic Search Engine, Text Analysis & Semantics*. Obtenido de http://www.semantic-knowledge.com/

8 ANEXOS.

8.1 ANEXO 1. DETALLES DE IMPLEMENTACIÓN DE LA BASE DE DATOS.

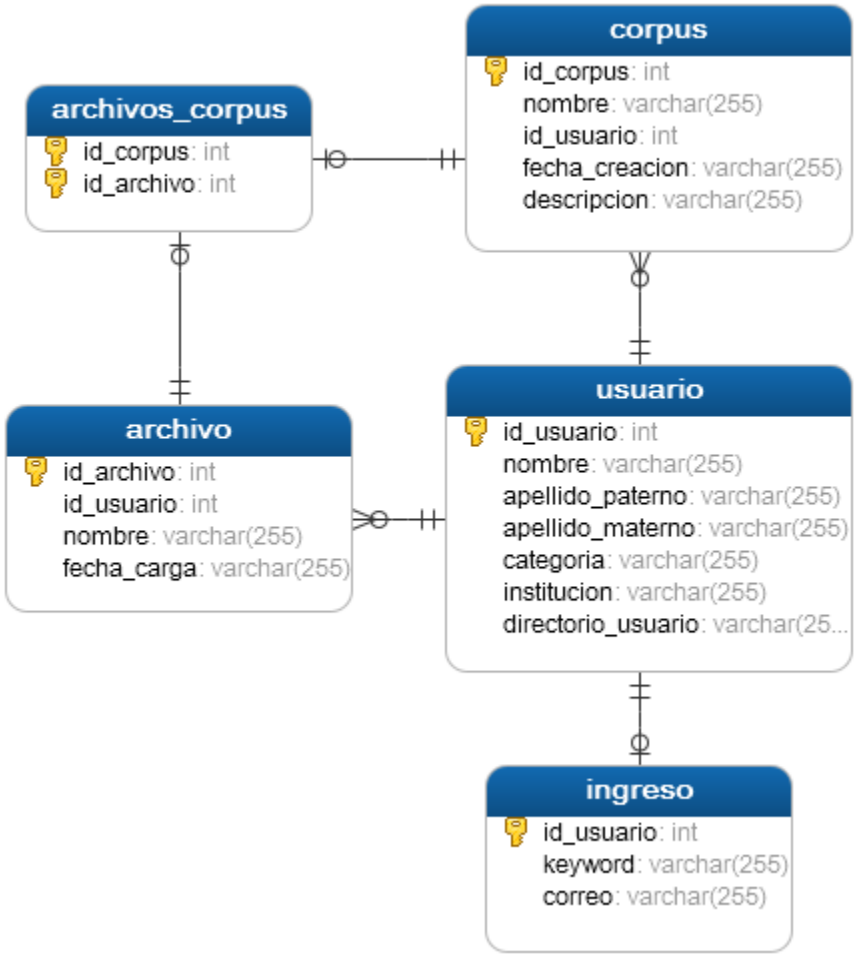


Ilustración 11. Diagrama entidad relación de la base de datos que usa la plataforma WEB.

La base de datos consta de cinco entidades con sus respectivos atributos para dar funcionalidad a la plataforma WEB, por medio de interacciones entre los diferentes componentes y acciones que el usuario realice a través de la aplicación. A continuación se describirá cada uno de las entidades, atributos y la funcionalidad que tienen.

Archivo: esta entidad va a estar encargada de almacenar todos los archivos que los usuarios suban a la plataforma.

- id_archivo: este atributo va a ser la llave primaria para cada archivo que se suba a la base de datos y es de tipo INT.
- nombre: este atributo se va a encargar de almacenar el nombre de los archivos y es de tipo VARCHAR.
- Id_usuario: este atributo está encargado de asociar este archivo a un usuario y es de tipo INT.
- fecha_carga: este atributo está encargado de almacenar la fecha en la que se subió el archivo.

Corpus: esta entidad va a contener los corpus de archivos que los usuarios creen, esta entidad es primordial para el análisis de los documentos.

- id_corpus: este atributo será la llave primara para cada corpus que el usuario vaya a crear y será de tipo INT.
- nombre: este atributo sirve para darle un nombre a un corpus y será de tipo VARCHAR.
- id_user: este atributo está encargado de asociar los corpus a un usuario en particular y es de tipo INT.
- fecha_creacion: este atributo está encargado de almacenar la fecha en la que fue creado dicho corpus.
- descripción: este atributo esta encarga de almacenar una descripción que el usuario podrá poner a la hora de crear un corpus para identificar en contenido de este.

Usuario: esta entidad es principalmente usada para almacenar información importante de los usuarios que emplean la plataforma.

- id_usuario: este atributo es la llave primaria para cada usuario que se registre en la base de datos y es te tipo INT.
- nombre: este atributo almacenará el nombre del usuario y es de tipo VARCHAR.
- apellido_paterno: este atributo almacenará el apellido paterno del usuario y es de tipo VARCHAR.
- apellido_materno: este atributo almacenará el apellido materno del usuario y es de tipo VARCHAR.
- categoría: este atributo almacenará el tipo de usuario que esté usando el sistema (estudiante, profesor, otro) y es de tipo VARCHAR.
- institución: este atributo guardara a que institución pertenece cada usuario y es de tipo VARCHAR.
- directorio_usuario: este atributo almacenar el directorio donde serán almacenados los archivos correspondientes a un usuario y es de tipo VARCHAR.

Ingreso: esta entidad es para almacenar los datos de inicio de sesión de los usuarios.

- id_user: este atributo será la llave primaria y la referencia de un usuario y sus datos de autentificación dentro de la página y es de tipo INT.
- keyword: este atributo almacenará la contraseña de acceso de un usuario en particular y es de tipo VARCHAR.
- correo: este atributo almacenar el correo electrónico de un usuario, en caso de ser necesario una recuperación de contraseña se enviará a este correo y es de tipo VARCHAR.

archivos_corpus: esta entidad es una tabla de unión para tener las referencias de que archivos pertenecen a un corpus en específico.

- id_archivo: es parte de la llave compuesta primaria y la referencia del archivo que pertenece al corpus asociado.
- id_corpus: es parte de la llave compuesta primaria y la referencia del corpus que se está asociando con los archivos.

8.2 ANEXO 2. DETALLES DE IMPLEMENTACIÓN DEL DIAGRAMA DE COMPONENTES.

En la ilustración 9 se muestra el diagrama de componentes de la plataforma WEB, donde podemos observar que cuenta con cinco componentes los cuales son “Gestor de usuarios” “Base de datos”, “Gestor de archivos”, “Conversor de formato”, “Módulo de visualización”, “Operaciones de análisis de textos” cada uno de estos componentes hace uso de diferentes clases para operar correctamente.

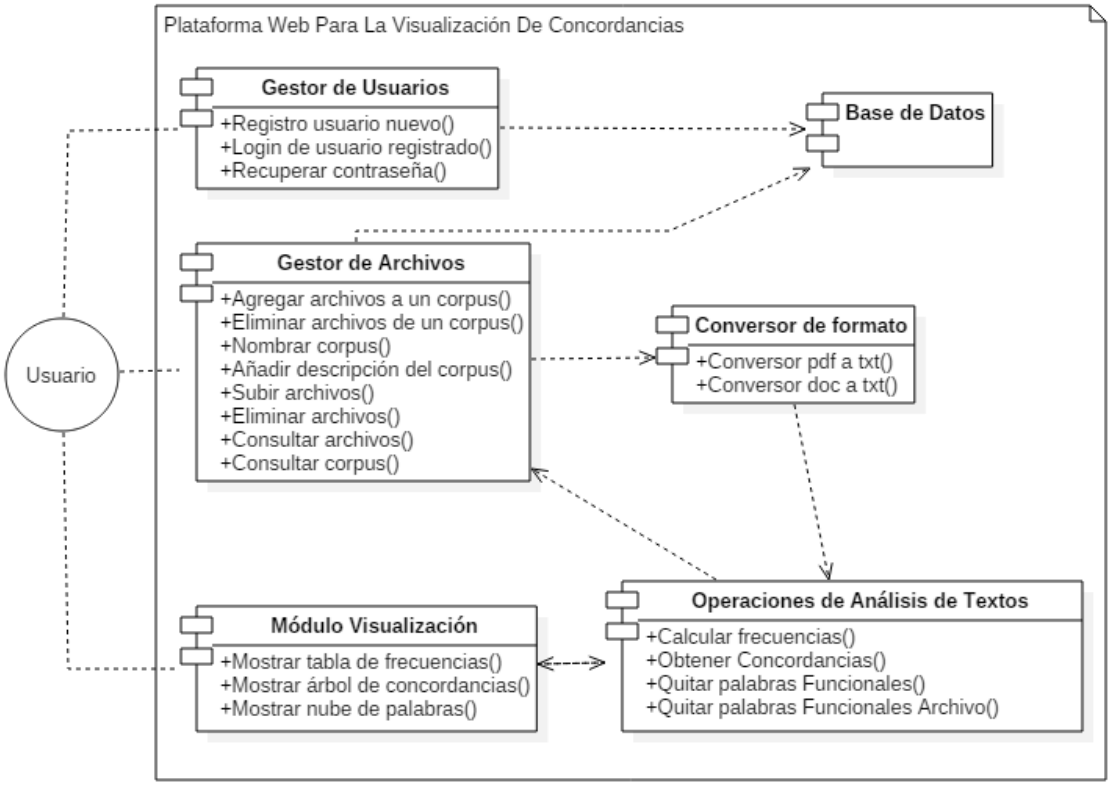


Ilustración 12. Diagrama de componentes de la herramienta WEB.

En la ilustración 10 se puede observar de manera general una abstracción de todos los objetos que la aplicación usa, los cuales son “Archivo”, “Corpus”, “Usuario”, “Session”, “Login”, “Concordancia”, cada uno de estos objetos se relaciona directamente con una entidad en la base de datos con excepción de la clase “Session” que se usa para manejar las sesiones del lado del cliente de la plataforma WEB. A continuación se va a describir el uso que tiene cada uno de estos objetos y que con la ayuda de la base de datos dan funcionalidad a la plataforma WEB.

- “Archivo”: El objeto archivo se usa para abstraer el nombre del archivo, la fecha en el que fue cargado, a que usuario pertenece cada archivo y un identificador único para cada archivo que es cargado en él servidor.
- “Corpus”: El objeto corpus se usa para abstraer el nombre del corpus, la fecha de creación, a que usuario pertenece cada corpus y un identificador único para cada corpus.
- “Usuario”: EL objeto Usuario se usa para abstraer los datos de un usuario como son, su nombre, apellido materno, apellido paterno, su categoría (alumno, profesor u otro), institución en la que labora o estudia, un directorio en el cual se van almacenar sus archivos y un identificador único para cada usuario que se registra.
- “Session”: El objeto Session se usa para abstraer los datos de sesión de un usuario cuando ingresa a la plataforma WEB, estos datos son, el correo del usuario, el nombre y su id. Esto para que cuando un usuario cree corpus o suba archivos se sepa a qué usuario pertenecen.
- “Concordancia”: El objeto Concordancia se usa para abstraer los datos de una concordancia, los cuales son el contexto izquierdo, contexto derecho y el centro. Se usa cuando un usuario extrae concordancias a partir de un conjunto de documentos la plataforma genera lista de concordancias para luego visualizarlas.
- “Login”: El objeto Login se usa para extraer los datos de ingreso de un usuario especifico, los cuales son el correo, la contraseña y el id del usuario. En el formulario para el ingreso solo se le solicita al usuario el correo y la contraseña, pero una vez que es validado un usuario se usa el id de ese usuario para crear una sesión dentro de la plataforma.

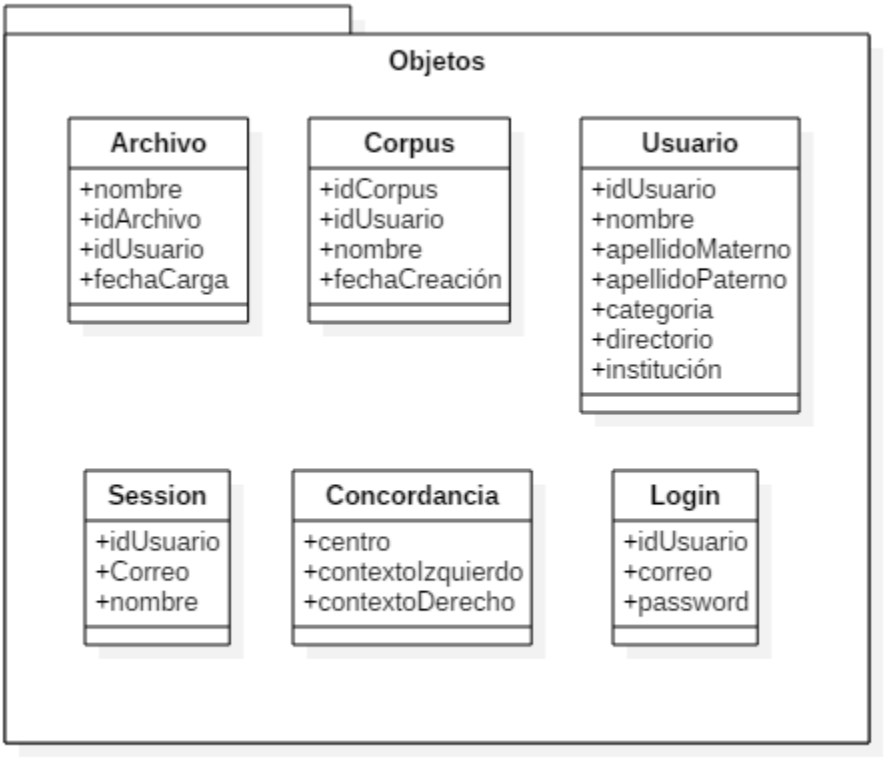


Ilustración 13. Diagrama de objetos que usa la aplicación.

En la ilustración 12 se puede muestra el diagrama de clases para el componente de operaciones de análisis de textos. Se puede observar la relación que tiene con otros objetos y clases dentro de la aplicación.

- La clase Frequency sobre escribe la interfaz "*compare"* para ordenar un conjunto de pares ordenados del tipo map <Key, Value>. Y hace uso de la clase MutableInteger para modificar el valor asociado a una llave (Key) dentro de la interface Map para incrementar en 1 la frecuencia con la que se encuentra una palabra.
- La clase Concordancias usa la clase Conversor que tiene la función de convertir cualquier formato de archivos, en texto plano para facilitar el procesamiento de su contenido, de igual forma tiene una lista de concordancias donde se almacenaran todas las concordancias generadas a partir de una consulta echa por el usuario.

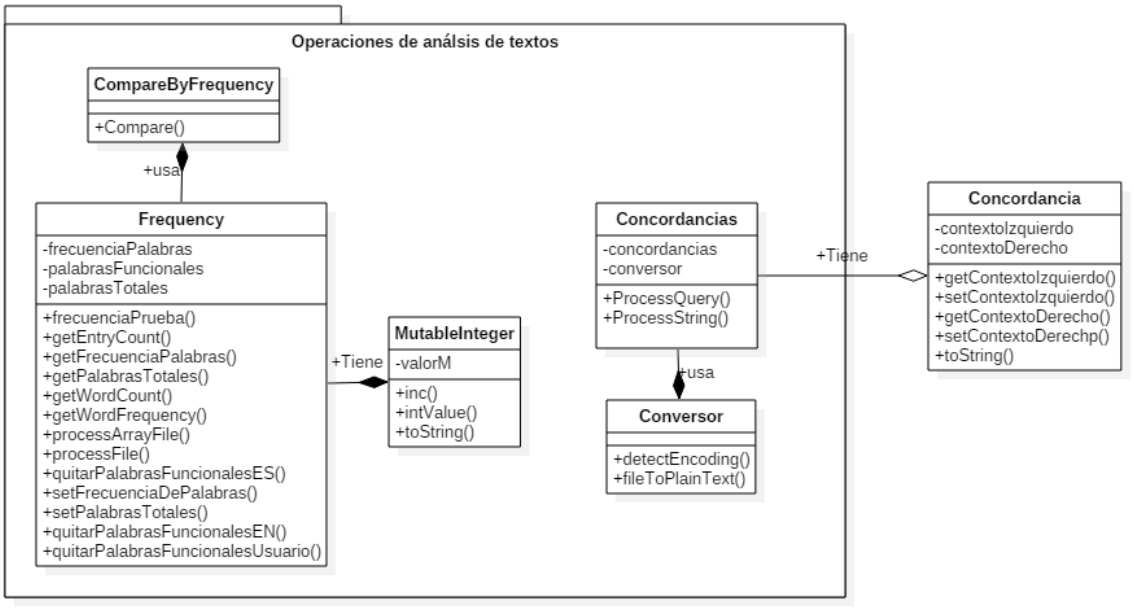


Ilustración 14. Diagrama de clases para el componente de operaciones de análisis de textos.

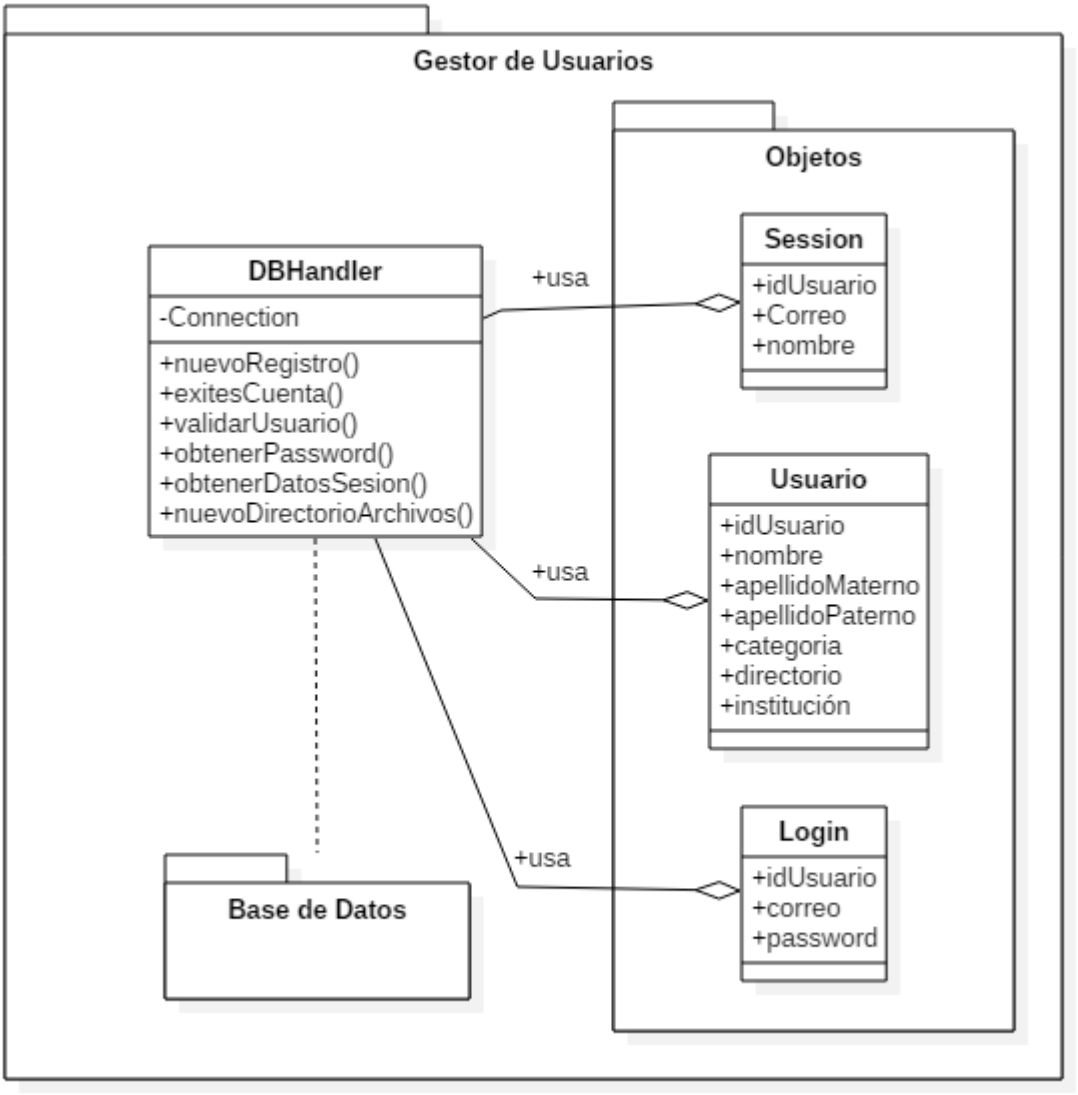


Ilustración 15. Diagrama de clases para el componente de gestor de usuarios.

En la ilustración 15 se puede apreciar el componente de gestor de usuarios, el cual hace uso de los objetos ‘Login’, ‘Usuario’ y ‘Session’, la clase ‘DBHandler’ que hace la conexión a la base de datos y las consultas para la validación, el acceso y el registro de usuarios a la plataforma WEB.

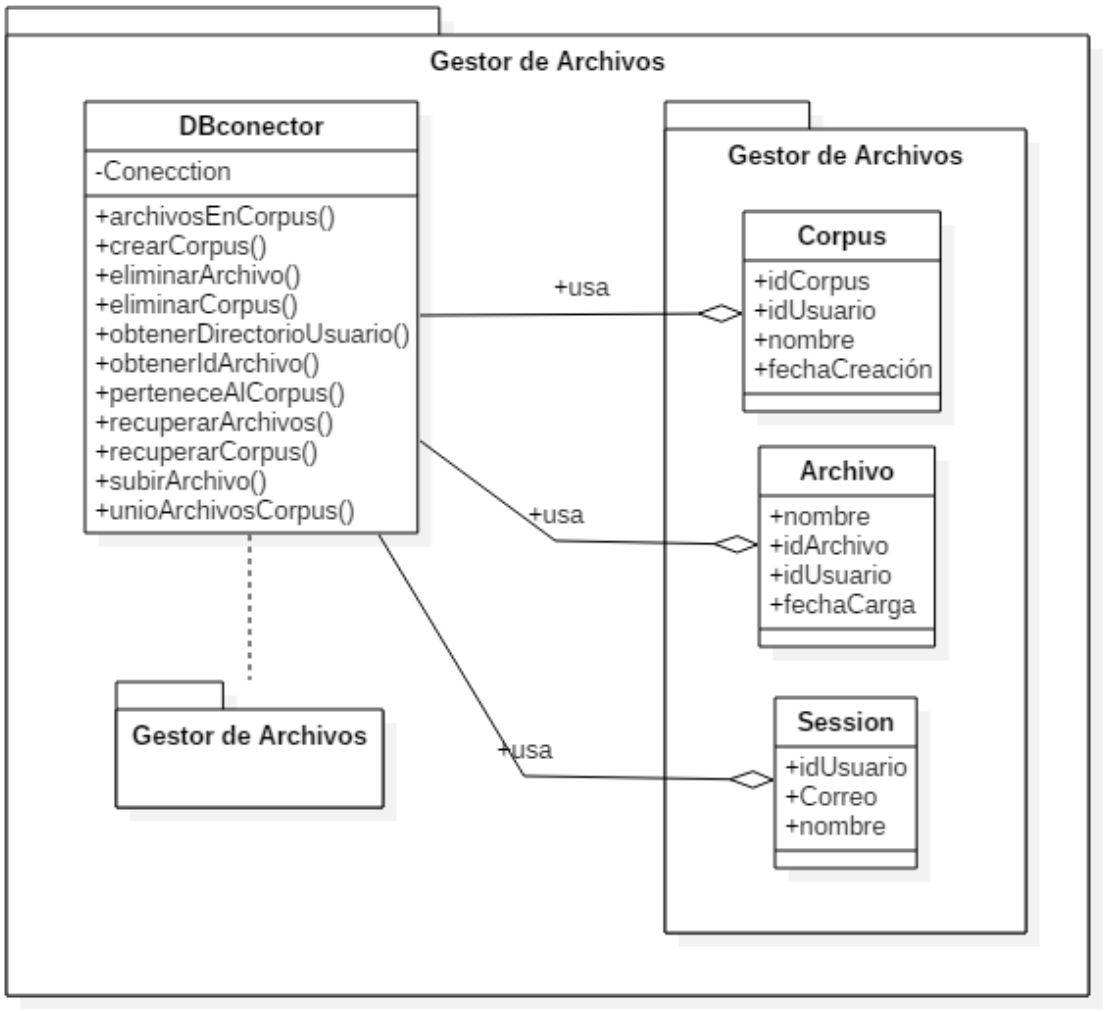


Ilustración 16. Diagrama de clases para el componente de gestor de archivos.

En la ilustración 16 se puede observar el componente de gestor de archivos, el cual hace uso de los objetos ‘Session’, ‘Archivo’, ‘Corpus’ y hace uso de la clase ‘DBHandler’ para hacer la conexión a la base de datos y hacer las acciones de ‘Subir archivos’, ‘Eliminar archivos’, ‘Crear Corpus’, ‘Recuperar Archivos’, ‘Eliminar Corpus’ todas estas asociadas al usuario cuando se realice algún evento desde la página WEB.

8.3 ANEXO 3. DETALLES DE LOS CASOS DE USO.

En esta sección se mostraran los casos de usos identificados al momento de hacer el análisis de requerimientos correspondiente para la plataforma WEB. En las siguientes tablas se definen los casos de usos, en los cuales se identifica al actor principal, precondiciones, post condiciones y los flujos básicos y alternos de cada caso de uso.

Caso de uso	Página de inicio - Entrar	
Actor Principal	Usuario	
Precondición	El usuario debe ingresar al sitio principal de la plataforma WEB, Visualizar que la página haya cargado y haber hecho clic en el botón “Entrar”.	
Post condición	El usuario visualizara un formulario en el cual debe ingresar su correo y contraseña, de ser correctos ingresará a la página de usuarios registrados.	
Descripción	La plataforma deberá comportarse como se describe en el siguiente caso de uso, cuando el usuario haya echo clic en el botón de “Entrar” de la página de inicio.	
Secuencia Normal	Paso	Acción
	1	Hacer clic en el botón “Entrar” para desplegar el formulario de ingreso.
	2	El usuario deberá ingresar su correo y contraseña.
	3	El sistema validara los datos de ingreso y le dará acceso a la página de usuarios registrados.
Secuencia alterna	Paso	Acción
	1	El sistema no le permitirá el acceso y tendrá que llenar de nuevo el formulario de ingreso si sus datos de ingreso son incorrectos.

Tabla 2. Tabla donde se describe el flujo del caso de uso "Entrar" en la página de inicio.

Caso de uso	Página de inicio - Probar	
Actor Principal	Usuario	
Precondición	El usuario debe ingresar al sitio principal de la plataforma WEB, Visualizar que la página haya cargado y haber hecho clic en el botón “Probar”.	
Post condición	El usuario visualizara una página totalmente nueva con las principales características de análisis de textos.	
Descripción	La plataforma deberá comportarse como se describe en el siguiente caso de uso, cuando el usuario haya echo clic en el botón de “Probar” de la página de inicio.	
Secuencia Normal	Paso	Acción
	1	Hacer clic en el botón “Probar” para re direccionar al usuario a la página de prueba.
	2	El usuario deberá visualizar una página totalmente nueva con varios campos para ingresar consultas para hacer análisis de un texto pequeño.

Tabla 3. Tabla donde se describe el flujo del caso de uso "Probar" en la página de inicio.

Caso de uso	Página de inicio – Registrarse	
Actor Principal	Usuario	
Precondición	El usuario debe ingresar al sitio principal de la plataforma WEB, Visualizar que la página haya cargado y haber hecho clic en el botón “Registrarse”.	
Post condición	El usuario visualizara un formulario que deberá llenar con sus datos personales para hacer uso de la página de usuarios registrados.	
Descripción	La plataforma deberá comportarse como se describe en el siguiente caso de uso, cuando el usuario haya echo clic en el botón de “Registrarse” de la página de inicio.	
Secuencia Normal	Paso	Acción
	1	Hacer clic en el botón “Registrarse” para desplegar el formulario de registro.
	2	El usuario deberá ingresar nombres, apellido materno, apellido materno, institución de origen, categoría de usuario, correo electrónico y contraseña.
	3	El sistema ingresara todos los datos del usuario nuevo y lo re direccionara a la página de usuarios registrados.
Secuencia alterna	Paso	Acción
	1	El sistema no le permitirá el registro y el acceso si no llena el formulario correctamente.

Tabla 4. Tabla donde se describe el flujo del caso de uso "Registrarse" en la página de inicio.

8.4 ANEXO 4. MANUAL TÉCNICO.

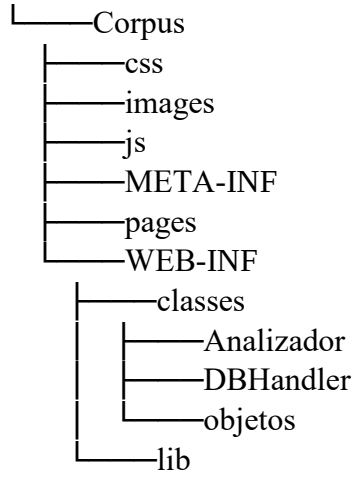
Esta sección se describe como instalar de manera local la plataforma, hacer modificaciones y/o agregar nuevas características a la plataforma WEB.

Requerimientos para su instalación:

- Servidor Apache Tomcat versión 7.0.72 o superior
- Mysql – MariaDB versión 10.1.3 o superior
- Java JRE versión 1.8.0 o superior
- Conexión a internet

Instalación del sistema de manera local en Windows:

- Descargar e instalar apache Tomcat (Servidor).
- Descargar e instalar el IDE de NetBeans.
- Descargar e instalar Mysql - MariaDB.
- Descargar e instalar la versión más reciente de Java JRE.
- Ejecutar NetBeas y abrir proyecto existente, Corpus (Archivos fuentes que se entregaran junto con este manual).
- Levantar el servicio del servidor, en una instalación normal se ubica en C:\tomcat\bin\starup.exe.
- Levantar el servicio de la base de datos, en una instalación normal se ubica en C:\mysql\bin\mysql.exe.
- Entramos a la línea de comando de Mysql y creamos la base de datos: CREATE DATABASE ‘Corpus’.
- Usamos la base de datos creada: USE ‘Corpus’.
- Ejecutamos nuestro archivo SQL de la base de datos: SOURCE directorio/del/archivo/archivo.sql (archivo proporcionado junto con los archivos fuentes).
- Si NetBeans se instala con el servidor tomcat no hay necesidad de levantar el servidor desde el directorio de su instalación. Desde netbeans se puede levantar el servidor en la sección de ‘Servers’ una vez levantado el servicio, clic derecho al proyecto ‘Corpus’ y dar clic en run automáticamente se abrirá un explorador web para visualizar la plataforma.
- Si se hace todo manualmente sin el uso de NetBeans se deben crear 2 directorios que siguen la siguiente estructura:



- La raíz es el nombre del proyecto que contiene los subdirectorios mostrados, el más importante es WEB-INF y sus subdirectorios, en el subdirectorio de clases se coloca la estructura de tus clases y los archivos .class previamente compilados, y en la carpeta lib, todas las librerías necesarias para el funcionamiento de la aplicación
- Una vez echo eso copiar y pegar toda la carpeta de corpus en el directorio de C:\tomcat\webapps\, levantamos el servicio de tomcat y accedemos mediante un browser a la página, normalmente es localhost:8080/Corpus/pagina.html.

Añadir nuevas características al código:

Para añadir nuevas característica a la plataforma se da por hecho que toda la información que se va a analizar ésta contenida en archivos de texto.

- Crear una clase que va a tomar como argumento principal argumente los archivos y el algoritmo que va a analizar la información para producir un resultado.
- Diseñar como va a representar esta información en formato JSON.
- Añadir la nueva característica a la plataforma WEB los botones de entradas si se requieren.
- Escribir el AJAX que capturara la información de entrada y que enviará al servidor para ser analizada y que recibirá la respuesta y almacenara el archivo JSON.
- Crear la visualización correspondiente a este nuevo análisis que debe recibir un archivo en formato JSON.
- Se pueden reutilizar los archivos JSP que hacen la consulta de los archivos de un usuario en específico y pasarlos como argumento a su clase creada, capturar los resultados, crear el JSON y enviarlo como response.
- Para agregar de manera definitiva las nuevas características se deben hacer previamente las modificaciones necesarias al código, compilar sus clases y subir los archivos modificados en el directorio del servidor el cual es ‘/var/lib/tomcat/webapps/Corpus/’.