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.