PROYECTO - ENTREGA FINAL

Integrantes:

- Daniel Felipe Montenegro

Tema: Modelamiento del idioma español en bases relacionales

Introducción

El objetivo de este proyecto es modelar por medio de bases relacionales el idioma español, el programa principalmente será un analizador de texto el cual podrá recibir cualquier tipo de escrito; esto quiere decir que podrá recibir textos como un ensayo, un cuento, un párrafo, una frase o cualquier conjunto de palabras. La finalidad será analizar su orden, composición y estructura, para luego construir de acuerdo con estos parámetros un modelo el cual podrá tener futuras aplicaciones en áreas como **Procesamiento del Lenguaje Natural** y **Redes Neuronales**.

Descripción

El resultado de este proyecto será una librería de código abierto para el análisis de texto en **Python**, la librería podrá ejecutarse como un programa el cual permitirá crear scripts para agregar textos o frases a la base de datos y también visualizar su contenido. La biblioteca permitirá hacer uso de su información como de sus métodos para implementarlos en cualquier otro programa.

Herramientas utilizadas

Librerías Python

La librería para su correcto funcionamiento hace uso de otras cuatro librerías existentes en Python, estas librerías permiten la interconectividad de la herramienta, a continuación, se detalla el uso de cada librería:

- **Os:** Permite la lectura y escritura de archivos
- **Time:** Genera pequeños retardos para que la interfaz sea mas amigable
- **Requests:** Crea las solicitudes a las paginas web
- **Mysql.connector:** Es la librería oficial de Mysql para conectarse desde Python
- **BeautifulSoup:** Facilita el análisis de documentos html

La librería **Os** y la librería **Time** vienen por defecto distribuidas con **Python**, el resto se instalo por medio de **pip**.

Corpus de Referencia del Español Actual

El **Corpus de Referencia del Español Actual (CREA)** provee un listado de frecuencias de todas palabras existentes en la lengua

Web Scraping

Por defecto la base de datos se distribuirá con las primeras **121000 frecuencias del CREA**, sin embargo, para poner a prueba la herramienta procesamos con ella **3276 textos** (cuentos de diversos autores en el idioma español) que contenían **852302 frases** los cuales fueron recolectados de la pagina web www.ciudadseva.com, esto gracias a que en su documento **robots.txt** especificaba al momento del raspado que era posible extraer este contenido.

Esta biblioteca se formo por medio de una aplicación que esta integrada en la librería para hacer **Web Scraping** de cualquier pagina web.

MYSQL

Es el gestor de base de datos que utiliza el proyecto, los scripts de creación de la base de datos que se distribuyen con esta librería están implementados especialmente para su uso en MYSQL.

Resumen de la base de datos

La base de datos que se distribuye con la librería por defecto esta compuesta de la siguiente manera:

- **Entidades (5):** word, coding, phrase, mistake y text
- **Triggers (1):** frequent_mistake
- **Views (7):** crea, scripts, total_words, mistakes, codings, texts y phrases
- **Stored Procedures (7):** insert_crea, insert_word, insert_mistake, insert_coding, insert_text, insert_phrase y id_word
- **Users (2):** root y guest
- **Values (121000):** Las 121.000 primeras frecuencias del CREA

Modelo de la base de datos

La base de datos que se distribuye por defecto es modificable, sin embargo, el script de creación incluido solo plantea la necesidad de cinco entidades para su correcto funcionamiento.

Triggers

La base de datos tiene un trigger el cual se encarga de verificar cada vez que se inserta un error si la frecuencia del mismo es superior a **21**, si un error se repite mas de **21** veces el trigger se encarga de agregarlo como una palabra a la tabla **word** en la columna **script.**

Stored Procedures

La base de datos cuenta con siete procedimientos almacenados, seis de los cuales se utilizan para insertar datos y el ultimo se utiliza para encontrar el id de una palabra.

Usuarios

La base de datos cuenta con dos usuarios, el primero es el usuario por defecto “root” y el otro es el usuario “guest”. El script de creación para los usuarios invitados es el siguiente:

Values

Por defecto la librería tiene una carpeta **SYSTEM** la cual contiene los scripts de creación de la base de datos, uno de esos scripts es **modeler_initialization.sql** el cual posee las sentencias para ingresar las primeras **121.000 del CREA.**

Un ejemplo de esas sentencias es el mostrado a la izquierda, el cual es el llamado al procedimiento almacenado **insert_crea** que recibe como parámetro la palabra y las veces que se repite en el **CREA.**