Los archivos dispuestos por el cliente para la búsqueda de palabras pueden encontrarse en un repositorio público en Github, mediante el siguiente enlace:

[enlace de Github]

En total, fueron provistos 65 archivos de texto con textos para divulgación académica que son ilustrativos del conjunto de vocablos que forman la lengua española; por lo que se consideran suficientes para los posteriores procesos de análisis del proyecto. En total, estos archivos pesan 2,65 MB. Se supone que las palabras a encontrar en estos archivos son representativas de todo el idioma. Los archivos fueron obtenidos mediante webscrapping y una etapa inicial de procesamiento; la cual consistió en remover las etiquetas HTML y el contenido ajeno a la información divulgativa dispuesta para el usuario final del sitio web de donde fueron extraídos. Esta etapa inicial de procesamiento, sin embargo, no limpió completamente el resultado de caracteres especiales, y símbolos ajenos a las palabras del idioma español.

Los nombres de estos archivos se listan a continuación:

source\_files\_names = [

  "\_C3\_81frica.txt",

  "Aeropuerto.txt",

  "Agricultura.txt",

  "Alan\_Turing.txt",

  "Antiguo\_Egipto.txt",

  "Arte.txt",

  "Asia.txt",

  "Babilonia\_(ciudad).txt",

  "Biblia.txt",

  "Biodiversidad.txt",

  "Capitalismo.txt",

  "Ciencia.txt",

  "Colombia.txt",

  "Consumo.txt",

  "Cristianismo.txt",

  "Cristo.txt",

  "Cultura.txt",

  "Democracia.txt",

  "Diccionario.txt",

  "Divina\_comedia.txt",

  "Ecologismo.txt",

  "Energ\_C3\_ADa\_nuclear.txt",

  "Enrique\_VIII\_de\_Inglaterra.txt",

  "Escritura.txt",

  "Est\_C3\_A9tica.txt",

  "Estado.txt",

  "Estados\_Unidos.txt",

  "Europa.txt",

  "Filosof\_C3\_ADa.txt",

  "Finanzas.txt",

  "Fruta.txt",

  "Gabriel\_Garc\_C3\_ADa\_M\_C3\_A1rquez.txt",

  "Gastronom\_C3\_ADa\_de\_M\_C3\_A9xico.txt",

  "George\_Patton.txt",

  "Grecia.txt",

  "Guerra\_Fr\_C3\_ADa.txt",

  "Henry\_Ford.txt",

  "Holocausto.txt",

  "Homo\_sapiens.txt",

  "Inteligencia\_artificial.txt",

  "Italia.txt",

  "Jorge\_Isaacs.txt",

  "Juda\_C3\_ADsmo.txt",

  "Juglar.txt",

  "Julio\_Garavito.txt",

  "Liberalismo.txt",

  "Libro.txt",

  "Ludwig\_van\_Beethoven.txt",

  "Nevado\_del\_Ruiz.txt",

  "Nikola\_Tesla.txt",

  "Odisea.txt",

  "Organizaci\_C3\_B3n\_de\_las\_Naciones\_Unidas.txt",

  "Per\_C3\_ADodo\_helen\_C3\_ADstico.txt",

  "Pizza.txt",

  "Pueblos\_germ\_C3\_A1nicos.txt",

  "Real\_Academia\_Espa\_C3\_B1ola.txt",

  "Reino\_Unido.txt",

  "Revoluci\_C3\_B3n\_Industrial.txt",

  "Rueda.txt",

  "Salsa\_(g\_C3\_A9nero\_musical).txt",

  "Saxof\_C3\_B3n.txt",

  "Segunda\_Guerra\_Mundial.txt",

  "Sublime.txt",

  "Thomas\_Alva\_Edison.txt",

  "Valle.txt"

]

A continuación, para poder utilizar cada uno de estos archivos, basta con utilizar un programa que sea capaz de utilizar las rutas o nombres de cada uno para leerlos; esto, siempre que el acceso a los mismos no esté restringido. Por el presente proyecto, se realizará un script de Python que utilice las funcionalidades nativas de este lenguaje de programación (como la función open) para recorrer la lista de los nombres de los archivos y leerlos uno a uno.