Para realizar un muestreo de todos los archivos de la fuente de datos que podemos revisar e este punto del proyecto, primero definimos las rutas de los distintos archivos a leer:

files\_address\_base = "/content/drive/MyDrive/Colab Notebooks/data/raw\_texts/"

source\_files\_addresses = [

  files\_address\_base + "\_C3\_81frica.txt",

  files\_address\_base + "Aeropuerto.txt",

  files\_address\_base + "Agricultura.txt",

  files\_address\_base + "Alan\_Turing.txt",

  files\_address\_base + "Antiguo\_Egipto.txt",

  files\_address\_base + "Arte.txt",

  files\_address\_base + "Asia.txt",

  files\_address\_base + "Babilonia\_(ciudad).txt",

  files\_address\_base + "Biblia.txt",

  files\_address\_base + "Biodiversidad.txt",

  files\_address\_base + "Capitalismo.txt",

  files\_address\_base + "Ciencia.txt",

  files\_address\_base + "Colombia.txt",

  files\_address\_base + "Consumo.txt",

  files\_address\_base + "Cristianismo.txt",

  files\_address\_base + "Cristo.txt",

  files\_address\_base + "Cultura.txt",

  files\_address\_base + "Democracia.txt",

  files\_address\_base + "Diccionario.txt",

  files\_address\_base + "Divina\_comedia.txt",

  files\_address\_base + "Ecologismo.txt",

  files\_address\_base + "Energ\_C3\_ADa\_nuclear.txt",

  files\_address\_base + "Enrique\_VIII\_de\_Inglaterra.txt",

  files\_address\_base + "Escritura.txt",

  files\_address\_base + "Est\_C3\_A9tica.txt",

  files\_address\_base + "Estado.txt",

  files\_address\_base + "Estados\_Unidos.txt",

  files\_address\_base + "Europa.txt",

  files\_address\_base + "Filosof\_C3\_ADa.txt",

  files\_address\_base + "Finanzas.txt",

  files\_address\_base + "Fruta.txt",

  files\_address\_base + "Gabriel\_Garc\_C3\_ADa\_M\_C3\_A1rquez.txt",

  files\_address\_base + "Gastronom\_C3\_ADa\_de\_M\_C3\_A9xico.txt",

  files\_address\_base + "George\_Patton.txt",

  files\_address\_base + "Grecia.txt",

  files\_address\_base + "Guerra\_Fr\_C3\_ADa.txt",

  files\_address\_base + "Henry\_Ford.txt",

  files\_address\_base + "Holocausto.txt",

  files\_address\_base + "Homo\_sapiens.txt",

  files\_address\_base + "Inteligencia\_artificial.txt",

  files\_address\_base + "Italia.txt",

  files\_address\_base + "Jorge\_Isaacs.txt",

  files\_address\_base + "Juda\_C3\_ADsmo.txt",

  files\_address\_base + "Juglar.txt",

  files\_address\_base + "Julio\_Garavito.txt",

  files\_address\_base + "Liberalismo.txt",

  files\_address\_base + "Libro.txt",

  files\_address\_base + "Ludwig\_van\_Beethoven.txt",

  files\_address\_base + "Nevado\_del\_Ruiz.txt",

  files\_address\_base + "Nikola\_Tesla.txt",

  files\_address\_base + "Odisea.txt",

  files\_address\_base + "Organizaci\_C3\_B3n\_de\_las\_Naciones\_Unidas.txt",

  files\_address\_base + "Per\_C3\_ADodo\_helen\_C3\_ADstico.txt",

  files\_address\_base + "Pizza.txt",

  files\_address\_base + "Pueblos\_germ\_C3\_A1nicos.txt",

  files\_address\_base + "Real\_Academia\_Espa\_C3\_B1ola.txt",

  files\_address\_base + "Reino\_Unido.txt",

  files\_address\_base + "Revoluci\_C3\_B3n\_Industrial.txt",

  files\_address\_base + "Rueda.txt",

  files\_address\_base + "Salsa\_(g\_C3\_A9nero\_musical).txt",

  files\_address\_base + "Saxof\_C3\_B3n.txt",

  files\_address\_base + "Segunda\_Guerra\_Mundial.txt",

  files\_address\_base + "Sublime.txt",

  files\_address\_base + "Thomas\_Alva\_Edison.txt",

  files\_address\_base + "Valle.txt"

]

Una vez realizado esto, procedemos a realizar el muestreo de cada archivo mediante el siguiente ciclo:

for address in source\_files\_addresses:

  print("Sample from " + address[len(files\_address\_base):] + ":\n")

  sample\_txt\_file(address)

  print("-----------------------------------------------------------------")

A continuación, presentamos una lista con el muestreo de los primeros 200 caracteres de cada fuente de datos:

Sample from \_C3\_81frica.txt:

Reunión (Francia)

Mayotte (Francia)

Santa Elena, Ascensión y Tristán de Acuña

Madeira

Pantelaria

Islas Canarias

Ceuta

Melilla

África es el tercer continente más extenso, tras Asia y América. Est

-----------------------------------------------------------------

Sample from Aeropuerto.txt:

Un aeropuerto es un aeródromo con terminal para el viaje de transporte aéreo en aeronave. Las funciones del aeropuerto son variadas, entre ellas el aterrizaje y despegue de aeronaves, embarque y desem

-----------------------------------------------------------------

Sample from Agricultura.txt:

La agricultura (del latín agri ‘campo’ y cultūra ‘cultivo’, ‘crianza’)[1]​[2]​ es el conjunto de actividades económicas y técnicas relacionadas con el tratamiento del suelo y el cultivo de la tierra p

-----------------------------------------------------------------

Sample from Alan\_Turing.txt:

Alan Mathison Turing (Paddington, Londres; 23 de junio de 1912-Wilmslow, Cheshire; 7 de junio de 1954) fue un matemático, lógico, informático teórico, criptógrafo, filósofo y biólogo teórico británico

-----------------------------------------------------------------

Sample from Antiguo\_Egipto.txt:

El Antiguo Egipto o Egipto Antiguo fue una civilización de la Antigüedad, que se originó a lo largo del cauce medio y bajo del río Nilo, cuya historia abarca más de tres milenios. Se la considera una

-----------------------------------------------------------------

Sample from Arte.txt:

El arte (del latín ars, artis, y este calco del griego τέχνη téchnē)[1]​ es entendido generalmente como cualquier actividad o producto realizado con una finalidad estética y también comunicativa, me

-----------------------------------------------------------------

Sample from Asia.txt:

Países transcontinentales: Armenia Azerbaiyán Chipre Egipto Georgia Indonesia Kazajistán Rusia Turquía

Asia es el continente más grande y poblado de la Tierra. Con cerca de 45 millones de km², supone

-----------------------------------------------------------------

Sample from Babilonia\_(ciudad).txt:

Babilonia fue una antigua ciudad de la Baja Mesopotamia situada cerca de la actual ciudad de Hilla (Irak). Fue la capital del antiguo Reino babilónico, y por varios siglos fue considerada como un impo

-----------------------------------------------------------------

Sample from Biblia.txt:

La Biblia (del latín biblĭa, y este del griego βιβλία biblía, ‘libros’)[1]​ es un conjunto de libros canónicos que en el cristianismo y en otras religiones, se consideran producto de inspiración divin

-----------------------------------------------------------------

Sample from Biodiversidad.txt:

La biodiversidad o diversidad biológica es, según el Convenio Internacional sobre la Diversidad Biológica, el término por el que se hace referencia a la amplia variedad de seres vivos sobre la Tierra

-----------------------------------------------------------------

Sample from Capitalismo.txt:

El capitalismo es un orden o sistema social y económico que se encuentra en constante movimiento, derivado del usufructo de la propiedad privada sobre el capital como herramienta de producción, que se

-----------------------------------------------------------------

Sample from Ciencia.txt:

La ciencia (del latín scientĭa, 'conocimiento') es un sistema que organiza y construye el conocimiento a través de preguntas comprobables y un método estructurado que estudia e interpreta los fenómeno

-----------------------------------------------------------------

Sample from Colombia.txt:

Colombia, oficialmente República de Colombia, es un país soberano situado en la región noroccidental de América del Sur, actualmente es miembro de la Comunidad Andina y se constituye en un estado

-----------------------------------------------------------------

Sample from Consumo.txt:

Consumo es la acción y efecto de consumir o gastar, sean productos, bienes o servicios, como por ejemplo la energía, entendiendo por consumir, como el hecho de utilizar estos productos y servicios pa

-----------------------------------------------------------------

Sample from Cristianismo.txt:

El cristianismo (del latín christianismus, y este del griego χριστιανισμός)[2]​ es una religión abrahámica monoteísta basada en la vida y enseñanzas de Jesús de Nazaret. Es la religión más extensa de

-----------------------------------------------------------------

Sample from Cristo.txt:

Cristo (del latín Christus, y este del griego antiguo Χριστός, Christós)[1]​ es una traducción del término hebreo «Mesías» (מָשִׁיחַ, Māšîaḥ), que significa «ungido»,[2]​ y que se emplea como título o

-----------------------------------------------------------------

Sample from Cultura.txt:

El término cultura (del latín cultūra)[1]​[2]​ tiene muchos significados interrelacionados, es decir, es un término polisémico. Por ejemplo, en 1952, Alfred Kroeber y Clyde Kluckhohn recopilaron una l

-----------------------------------------------------------------

Sample from Democracia.txt:

La democracia (del latín tardío democratĭa, y este del griego δημοκρατία dēmokratía)[1]​ es una manera de organización social que atribuye la titularidad del poder al conjunto de la ciudadanía. En sen

-----------------------------------------------------------------

Sample from Diccionario.txt:

El diccionario es una obra donde se puede consultar palabras o términos y se proporciona su significado, definición, etimología, ortografía, fija su pronunciación, separación silábica y forma gramatic

-----------------------------------------------------------------

Sample from Divina\_comedia.txt:

La Divina comedia (en italiano moderno: Divina Commedia, en toscano: Divina Comedìa), también conocida simplemente como Comedia, es un poema escrito por Dante Alighieri. Se desconoce la fecha exacta e

-----------------------------------------------------------------

Sample from Ecologismo.txt:

El ecologismo (en ocasiones llamado el movimiento verde o ambientalista) es un variado movimiento político, social y global, que defiende la protección del medio ambiente.[1]​

Habitualmente, el ecolo

-----------------------------------------------------------------

Sample from Energ\_C3\_ADa\_nuclear.txt:

La energía nuclear o atómica es la que se libera espontánea o artificialmente en las reacciones nucleares. Sin embargo, este término engloba otro significado que es el aprovechamiento de dicha energía

-----------------------------------------------------------------

Sample from Enrique\_VIII\_de\_Inglaterra.txt:

Enrique VIII (28 de junio de 1491-28 de enero de 1547) fue rey de Inglaterra y señor de Irlanda desde el 22 de abril de 1509 hasta su muerte. Fue el segundo monarca de la casa Tudor, heredero de su pa

-----------------------------------------------------------------

Sample from Escritura.txt:

La escritura es un sistema de representación gráfica de un idioma, por medio de signos trazados o grabados sobre un soporte. En tal sentido, la escritura es un modo gráfico específicamente humano de

-----------------------------------------------------------------

Sample from Est\_C3\_A9tica.txt:

La estética (del griego αἰσθητική [aesthetic], ‘sensación’, ‘percepción’, y este de[aísthesis], ‘sensación’, ‘sensibilidad’, e -ικά [-icá], ‘relativo a’) es la rama de la filosofía que estudia la esen

-----------------------------------------------------------------

Sample from Estado.txt:

Un Estado es una organización política constituida por un conjunto de instituciones burocráticas estables, a través de las cuales ejerce el monopolio del uso de la fuerza (soberanía) aplicada a una po

-----------------------------------------------------------------

Sample from Estados\_Unidos.txt:

Estados Unidos[nota 2]​ (EE. UU.; en inglés, United States o US),[nota 3]​ cuyo nombre oficial es Estados Unidos de América (EUA; en inglés, United States of America o USA),[12]​ es un país sobera

-----------------------------------------------------------------

Sample from Europa.txt:

Países transcontinentales:

Europa es un continente ubicado enteramente en el hemisferio norte y mayoritariamente en el hemisferio oriental. Las fronteras de Europa están situadas en la mitad occident

-----------------------------------------------------------------

Sample from Filosof\_C3\_ADa.txt:

La filosofía (del griego antiguo φιλοσοφία 'amor a la sabiduría' derivado de φιλεῖν [fileîn] 'amar' y σοφία [sofía] 'sabiduría';[1]​ trans. en latín como philosophĭa)[2]​ es una disciplina académica y

-----------------------------------------------------------------

Sample from Finanzas.txt:

Las finanzas son una rama de la administración y de la economía que estudia el intercambio de capital entre individuos, empresas, o Estados y con la incertidumbre y el riesgo que estas actividades con

-----------------------------------------------------------------

Sample from Fruta.txt:

Se denomina fruta a aquellos frutos comestibles obtenidos de plantas cultivadas o silvestres que, por su sabor generalmente dulce-acidulado, su aroma intenso y agradable y sus propiedades nutritivas,

-----------------------------------------------------------------

Sample from Gabriel\_Garc\_C3\_ADa\_M\_C3\_A1rquez.txt:

Gabriel José García Márquez (Aracataca, 6 de marzo de 1927-Ciudad de México, 17 de abril de 2014)[nota 1]​[2]​ ( escuchar) fue un escritor y periodista colombiano. Reconocido principalmente por sus

-----------------------------------------------------------------

Sample from Gastronom\_C3\_ADa\_de\_M\_C3\_A9xico.txt:

La gastronomía mexicana es el conjunto de platillos y técnicas culinarias de México que forman parte de las tradiciones y vida común de sus habitantes, enriquecida por las aportaciones de las distint

-----------------------------------------------------------------

Sample from George\_Patton.txt:

Tercer Ejército de Estados UnidosDecimoquinto Ejército de Estados Unidos

Expedición Punitiva

Primera Guerra Mundial

Segunda Guerra Mundial

George Smith Patton, Jr. (San Gabriel, California; 11 de

-----------------------------------------------------------------

Sample from Grecia.txt:

Grecia (en griego moderno, Ελλάδα, Elláda, AFI: [eˈlaða] ( escuchar); en griego antiguo, Ἑλλάς, Hellás), oficialmente República Helénica (en griego, Ελληνική Δημοκρατία, Ellinikí Dimokratía, AFI:

-----------------------------------------------------------------

Sample from Guerra\_Fr\_C3\_ADa.txt:

La Guerra Fría fue un enfrentamiento político, económico, social, ideológico, militar e informativo iniciado tras finalizar la Segunda Guerra Mundial entre el bloque Occidental (occidental-capitalista

-----------------------------------------------------------------

Sample from Henry\_Ford.txt:

Henry Ford (Dearborn, Míchigan; 30 de julio de 1863-Ib., 7 de abril de 1947) fue un empresario y emprendedor estadounidense, fundador de la compañía Ford Motor Company y padre de las cadenas de produc

-----------------------------------------------------------------

Sample from Holocausto.txt:

El Holocausto[1]​ —también conocido en hebreo como השואה, Shoá, traducido como «La Catástrofe»—, conocido en la terminología nazi como «solución final» —en alemán, Endlösung— de la «cuestión judía»,[2

-----------------------------------------------------------------

Sample from Homo\_sapiens.txt:

El Homo sapiens (del latín homo, «hombre», y sapiens, «sabio»), hombre o ser humano es una especie del orden de los primates perteneciente a la familia de los homínidos. También son conocidos bajo la

-----------------------------------------------------------------

Sample from Inteligencia\_artificial.txt:

La inteligencia artificial (IA) es, en informática, la inteligencia expresada por máquinas, sus procesadores y sus softwares, que serían los análogos al cuerpo, el cerebro y la mente, respectivamente,

-----------------------------------------------------------------

Sample from Italia.txt:

Italia (en italiano, Italia, pronunciado /i'talja/ ( escuchar)), oficialmente la República Italiana (en italiano, Repubblica Italiana pronunciado /re'pub:lika ita'ljana/), es un país soberano tran

-----------------------------------------------------------------

Sample from Jorge\_Isaacs.txt:

Partido Conservador

Jorge Ricardo Isaacs Ferrer (Santiago de Cali, 1 de abril de 1837-Ibagué, 17 de abril de 1895) fue un novelista, escritor y poeta colombiano del género romántico. Jorge Isaacs vi

-----------------------------------------------------------------

Sample from Juda\_C3\_ADsmo.txt:

El término judaísmo se refiere a la religión, tradición y cultura del pueblo judío. Históricamente, es la más antigua de las tres religiones abrahámicas,[5]​ grupo que tiene como base e incluye el c

-----------------------------------------------------------------

Sample from Juglar.txt:

Un juglar[a]​ —antiguamente también joglar[1]​— era un artista ambulante en la Edad Media. A cambio de dinero o comida, ofrecía su espectáculo callejero en las plazas públicas, y en ocasiones era cont

-----------------------------------------------------------------

Sample from Julio\_Garavito.txt:

Julio Garavito Armero (Bogotá, 5 de enero de 1865-Bogotá, 11 de marzo de 1920) fue un astrónomo, matemático, economista, poeta e ingeniero colombiano.[1]​ Sus investigaciones contribuyeron al desarrol

-----------------------------------------------------------------

Sample from Liberalismo.txt:

El liberalismo es una filosofía y movimiento político que defiende la libertad individual, la igualdad ante la ley y la limitación de los poderes del Estado. Como doctrina económica propugna la inicia

-----------------------------------------------------------------

Sample from Libro.txt:

Un libro (del latín liber, libri) es una obra impresa, manuscrita o pintada en una serie de hojas de papel, pergamino, vitela u otro material, unidas por un lado (es decir, encuadernadas) y protegidas

-----------------------------------------------------------------

Sample from Ludwig\_van\_Beethoven.txt:

Ludwig van Beethoven[a]​ (Bonn, Arzobispado de Colonia; 16 de diciembre de 1770[b]​-Viena, 26 de marzo de 1827) fue un compositor, director de orquesta, pianista y profesor de piano alemán. Su legado

-----------------------------------------------------------------

Sample from Nevado\_del\_Ruiz.txt:

El Volcán del Ruiz , Nevado del ruiz , también conocido como Mesa de Herveo,[4]​ y en la época precolombina como Cumanday, Tabuchía y Tama,[1]​ se conoce como Volcán del Ruiz quizás en honor a Alfonso

-----------------------------------------------------------------

Sample from Nikola\_Tesla.txt:

Nikola Tesla (en serbio, Никола Тесла; Smiljan, Imperio austríaco, actual Croacia; 10 de julio de 1856-Nueva York, 7 de enero de 1943) fue un inventor, ingeniero eléctrico y mecánico serbio nacional

-----------------------------------------------------------------

Sample from Odisea.txt:

La Odisea (en griego: Ὀδύσσεια, Odýsseia) es un poema épico griego compuesto por 24 cantos, atribuido al poeta griego Homero. Se cree que fue compuesta en el siglo VIII a. C. en los asentamientos que

-----------------------------------------------------------------

Sample from Organizaci\_C3\_B3n\_de\_las\_Naciones\_Unidas.txt:

La Organización de las Naciones Unidas (ONU), también conocida simplemente como Naciones Unidas (NN. UU.), es la mayor organización internacional existente. Se creó para mantener la paz y seguridad in

-----------------------------------------------------------------

Sample from Per\_C3\_ADodo\_helen\_C3\_ADstico.txt:

Se denomina período helenístico, helenismo o periodo alejandrino (por Alejandro Magno) a una etapa histórica de la Antigüedad cuyos límites cronológicos vienen marcados por dos importantes acontecimie

-----------------------------------------------------------------

Sample from Pizza.txt:

La pizza es una preparación culinaria que consiste en un pan plano, habitualmente de forma circular, elaborado con harina de trigo, levadura, agua y sal (a veces aceite de oliva) que tradicionalmente

-----------------------------------------------------------------

Sample from Pueblos\_germ\_C3\_A1nicos.txt:

Los pueblos germánicos o germanos son un histórico grupo etnolingüístico de los pueblos originarios del norte de Europa que se identifican por el uso de las lenguas germánicas (un subgrupo de la fam

-----------------------------------------------------------------

Sample from Real\_Academia\_Espa\_C3\_B1ola.txt:

La Real Academia Española (RAE)[3]​ es una institución cultural española privada financiada con fondos públicos con sede en Madrid, España. Esta y otras veintitrés academias de la Lengua correspondien

-----------------------------------------------------------------

Sample from Reino\_Unido.txt:

El Reino Unido (United Kingdom en inglés),[nota 1]​ o de forma abreviada RU (UK en inglés), oficialmente Reino Unido de Gran Bretaña e Irlanda del Norte (United Kingdom of Great Britain and Northe

-----------------------------------------------------------------

Sample from Revoluci\_C3\_B3n\_Industrial.txt:

La Revolución Industrial o Primera Revolución Industrial es el proceso de transformación económica, social y tecnológica que se inició en la segunda mitad del siglo XVIII en el Reino de Gran Bretaña,

-----------------------------------------------------------------

Sample from Rueda.txt:

La rueda es un elemento circular y mecánico que gira alrededor de un eje.[1]​ Puede ser considerada una máquina simple, y forma parte del conjunto denominado elementos de máquinas.

Es uno de los inve

-----------------------------------------------------------------

Sample from Salsa\_(g\_C3\_A9nero\_musical).txt:

La salsa es un género musical bailable resultante de la síntesis del son cubano y otros géneros de música caribeña y estadounidenses como el jazz y el blues. La salsa se consolidó como un éxito comer

-----------------------------------------------------------------

Sample from Saxof\_C3\_B3n.txt:

El saxofón, también conocido como saxófono o simplemente saxo, es un instrumento musical cónico, de la familia de los instrumentos de viento-madera, generalmente hecho de latón, que consta de una boqu

-----------------------------------------------------------------

Sample from Segunda\_Guerra\_Mundial.txt:

Invasión alemana de Polonia

 Alemania nazi

Reino de Italia

 Imperio del Japón

Francia de Vichy

Reino de Hungría

Reino de Rumania

Albania

Eslovaquia

Reino de Bulgaria

Croacia

Mengjiang

Manc

-----------------------------------------------------------------

Sample from Sublime.txt:

Lo sublime es una categoría estética, derivada principalmente de la célebre obra Περὶ ὕψους (Sobre lo sublime) del crítico o retórico griego Longino (o Pseudo-Longino), y que consiste fundamentalmente

-----------------------------------------------------------------

Sample from Thomas\_Alva\_Edison.txt:

Thomas Alva Edison (Milan, Ohio; 11 de febrero de 1847-West Orange, Nueva Jersey; 18 de octubre de 1931) fue un inventor, científico y empresario estadounidense.[1]​[2]​[3]​ Desarrolló numerosos dispo

-----------------------------------------------------------------

Sample from Valle.txt:

Un valle (del latín vallis) es una llanura entre montañas o alturas, una depresión de la superficie terrestre entre dos vertientes, con forma inclinada y alargada, que conforma una cuenca hidrográfica

-----------------------------------------------------------------