Perros vs Gatos

Dos amigos están discutiendo sobre si en *Internet* se habla más sobre **gatos** o sobre **perros**. No se ponen de acuerdo y deciden hacer un programa que resuelva ese problema.

Entrada

El programa recibe como entrada un fichero **domains.txt** que contiene un dominio por cada fila. Ejemplo:

domain.txt

sitio1.com
sitio2.net
sitio3.org
sitio4.com

Salida

El programa devuelve por la salida solamente una palabra "**perro**" o "**gato**" después de hacer un análisis de cuál de los dos términos es el más usado en los dominios datos.

Objetivo

El objetivo de este ejercicio no es programar eso. Sino reflexionar sobre todos los posibles detalles y problemas con los que te tendrías que enfrentar si tuvieras que hacerlo.

O sea, lo que tienes que enviarme como respuesta es:

a) un documento que describa paso por paso (sin tirar código) lo que hay que ir haciendo (se acepta algún que otro pedacito de pseudo código).

b) además (y esto es lo más importante) intentar predecir todas las cosas con las que vas a "chocar" si intentaras implementar eso.

La estructura de la respuesta tiene que ser algo como:

1. primero tenemos que hacer A
2. luego tenemos que hacer B
3. aquí entonces tenemos una decisión que tomar porque puede suceder C o D
   a. en el caso que ocurra C hacemos E
   b. en el caso que ocurra D hacemos F
4. luego tenemos que hacer G que tiene varias fases
   a. la primera parte es G.1 que consiste en ….
   b. luego viene G.2 que hay que tener en cuenta M y N
   c. y por último G.3
5. una vez llegados a este punto tenemos varias forma de atacar el problema
   a. variante 1: tiene tales ventajas y tales desventajas
   b. variante 2: tiene estas cosas buenas y estas otras malas
   c. yo escojo la variante 1 modificada por estas razones que explico a continuación ….
6. y por último hay que aplicar el algoritmo X y eso nos dará la respuesta esperada.

RESPUESTA:

Precondiciones:

1- Se desea contar, en el contenido de páginas web (e.d.; dominios o direcciones de internet), la cantidad de veces que aparece la palabra gato o perro.
   Es importante tener en cuenta que las palabras asociadas a la palabra **gato** pueden ser diminutivos, aumentativos, o variaciones en género y número (p.ej.; gatito, gatote, gatos, gatas). De igual forma ocurre con las palabra asociadas a la palabra **perro**.
2- Los dominios que aparecen en el documento **domain.txt** son direcciones a páginas web (URL) y para obtener estas páginas es necesario usar una biblioteca o software de tercero que permita actuar como araña de búsqueda de internet (p.ej., curl o libcurl: tiene varias versiones en distintos lenguajes de programación).
3- Un dominio puede tener varios subdominios. Por ejemplo dado el dominio *sitio3.net* puede contener los subdominios *sitio3.net/ventas, sitio3.net/ventas/mascotas.*
   Asumimos en la respuesta, al no especificarse, que se debe buscar en los subdominios asociados al dominio inicial (que aparece en cada línea del documento **domain.txt**).
4- Una página web asociada al dominio o subdominio tiene en su contenido direcciones (url) a dominios o subdominios como hipervínculos que no están asociados al dominio o subdominios.
   El procesamiento de estas direcciones o **urls** pudiera provocar un gasto excesivo en tiempo y memoria porque cada página a procesar contiene sus propios enlaces a otros subdominios diferentes a los que aparecen en el documento **domain.txt**.

   Por esta razón en este ejercicio asumimos que en las páginas web de hipervínculos a dominios o subdominios que no pertenecen a los iniciales (presentes en el documento **domain.txt**), sólo se tendrá en cuenta su contenido y no los hipervínculos que pueda contener.

Como entrada se debe tener como datos o variables iniciales:

cantidad_palabras_gato = 0 (cantidad total de palabras relacionados con gato en internet)

cantidad_palabras_perro = 0 (cantidad total de palabras relacionados con perro en internet)

lista_dominios_analizados = vacía (lista de dominio que han sido analizados)

1- Primero tenemos que leer el archivo domain.txt línea por línea y guardar la información presente en estas líneas (la dirección de los dominios) en una lista o arreglo.
2- Luego para cada elemento (dirección de dominio) en la lista.
3- **Si** la dirección de dominio no está en ***lista_dominios_analizados*** entonces añadir la dirección de dominio en la lista de ***lista_dominios_analizados*** para evitar que no sea analizado nuevamente.
   Sino continuar con el próximo elemento de la lista
4- Obtener la página web asociada a la dirección del dominio usando una biblioteca o programa como curl.

5- Extraer de la página web todos los enlaces o hipervínculos presentes en ella.

6- Dividir los enlaces en dos listas: La lista D estará asociada a los enlaces que son subdominios de la dirección de dominio que se está analizando, la lista E estará asociada a los enlaces que nos son subdominios del dominio que se está analizando.
   i. Solo se incluirán en la lista D y E si estos enlaces a subdominios u otros dominios externos no han sido analizado antes (e.d., no estén insertados en ***lista_dominios_analizados***)
7- Obtener el texto que no contiene etiquetas html de la página asociada al dominio que se está analizando usando alguna biblioteca que realice **scrapy** (e.d., extracción de elementos html y su contenido).

8- Para el contenido que se encuentra en la página buscar la lista de palabras que están asociada a **gato**, empleando una biblioteca o marco de trabajo que realice el Procesamiento del Lenguaje Natural (p.ej.; spaCy o NLTK), para extraer todas las palabras (diminutivos, aumentativos, similares en género y número).

9- Aumentar el contador global de palabras en internet sobre **gato** asociado a la variable ***cantidad_palabras_gato*** con la cantidad de palabras encontradas (tamaño de la lista).

10- Para el contenido que se encuentra en la página buscar la lista de palabras que están asociada a **perro**, empleando una biblioteca o marco de trabajo que realice el Procesamiento del Lenguaje Natural, para extraer todas las palabras (diminutivos, aumentativos, similares en género y número).

11- Aumentar el contador global de palabras en internet sobre **perro** asociado a la variable ***cantidad_palabras_gato*** con la cantidad de palabras encontrada (tamaño de la lista).

12- Para la lista D de subdominios presente en la página web asociado al dominio analizado ir al paso 3.

Nota: Este paso no está relacionado con otros pasos y puede ser paralelizabe al compartida con varios hilos de ejecución las variables **cantidad_palabras_gato**, **cantidad_palabras_perro**, **lista_dominios_analizados** y con el uso de una cola de trabajos o

dominios a procesar y una lista o pool de hilos de ejecución procesando de forma concurrente (y mientras el pool de hilos no esté lleno) las direcciones que aparecen en lista D y E; actualizando los datos en las variables antes mencionadas.

13- Para la lista E de enlaces externos presentes en la página web asociado al dominio ir al paso 3 y saltar al paso 7.

14- Mostrar la cantidad de palabras relacionadas con **gato** y con **perro** en internet (valores almacenados en las variables **cantidad_palabras_gato**, **cantidad_palabras_perro**) y determinar cuál es el mayor.