PROYECTO FINAL

Integrantes:

• Nicolas Quintero Sierra – 2220090.

• Jeferson Jair Acevedo Sarmiento – 2221790.

INTRODUCCIÓN

El avance tecnológico ha revolucionado la forma en la que nos comunicamos, brindando una amplia gama de herramientas, siendo los correos electrónicos una de las más utilizadas, ya sea a través de plataformas como Gmail, Hotmail u otras opciones disponibles. Sin embargo, esta comodidad en la comunicación también ha dado lugar a un problema que ha afectado a la gran mayoría de personas: los correos no deseados o spam. Estos mensajes, a menudo provenientes de remitentes desconocidos, inundan nuestras bandejas de entrada con ofertas y servicios, aunque en ocasiones llevan consigo intenciones fraudulentas o incentivos engañosos.

Este proyecto tiene como objetivo abordar este desafío mediante la implementación de un sistema de clasificación de correos electrónicos, asignándoles probabilidades de ser spam. La estrategia se basa en la identificación de palabras clave y frases dentro del contenido del correo que son típicamente asociadas con mensajes no deseados. Ejemplos de estas palabras clave podrían ser "Has ganado" o "Da clic en este link". Al analizar la presencia y la frecuencia de estas palabras clave, se clasificará en una alta o baja probabilidad de que el correo en cuestión sea spam.

Esta metodología busca proporcionar a los usuarios una herramienta efectiva para filtrar y distinguir entre correos legítimos y potencialmente perjudiciales, mejorando así la seguridad y la eficiencia en la gestión de la comunicación electrónica. La capacidad de discernir entre correos de confianza y aquellos con posibles intenciones maliciosas se convierte en una necesidad crucial en un entorno digital cada vez más interconectado.

ESTRATEGIA DE SOLUCIÓN

Para abordar el problema, se comenzó procesando el mensaje del correo electrónico con la ayuda de la clase “Array”. Se convierte todo el texto a minúsculas y se elimina cualquier signo de puntuación presente. Este paso asegura coherencia y elimina elementos irrelevantes. A continuación, se desglosa el mensaje en palabras para identificar posibles indicadores de spam. Esta descomposición facilita la identificación de palabras clave que podrían sugerir la naturaleza del mensaje.

En la construcción de este autómata no determinista (NFA), el alfabeto está compuesto únicamente por las letras del abecedario que conforman las palabras del diccionario de palabras

spam. Las transiciones entre los estados del autómata están determinadas por las letras que componen palabras específicas que son indicativos de spam, por ejemplo: "Gano", "Premio", "Dinero", "Obsequio", y otras similares. Cada una de estas palabras clave se convierte en un conjunto de transiciones desde un estado a otro.

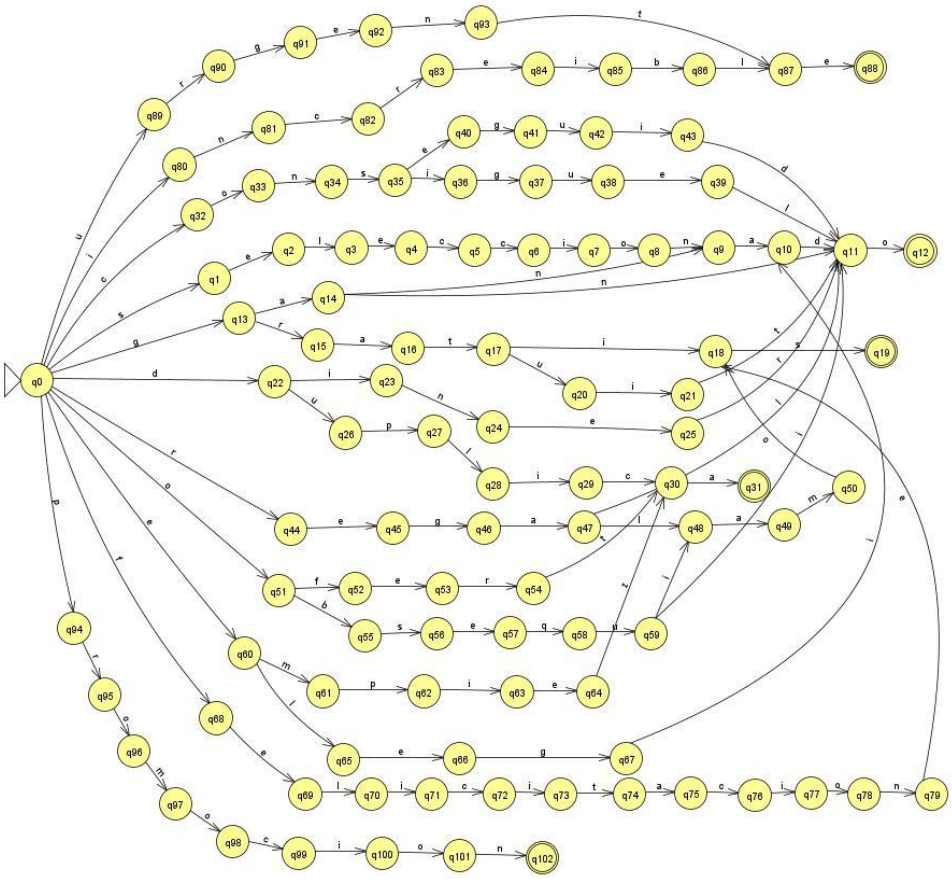
El proceso de funcionamiento del autómata implica procesar las letras del mensaje del correo una por una. Comienza en un estado inicial y avanza a través de las transiciones de acuerdo con las letras del mensaje. Si, en algún momento, el autómata alcanza un estado de aceptación después de procesar una o más letras, se concluye que el mensaje contiene al menos una de las palabras clave, y el correo electrónico se considera potencialmente spam. Por otro lado, si no se encuentra ninguna de las palabras clave durante el proceso, el autómata rechaza la cadena, indicando que el mensaje no es spam. La noción de que un mensaje tenga una alta o baja probabilidad de ser spam se introduce al contar la cantidad de palabras clave encontradas en el mensaje. Cada palabra clave aceptada incrementa esta probabilidad. Cuantas más palabras clave se encuentren, mayor será la probabilidad de que el mensaje sea spam y, por lo tanto, se considerará que es un mensaje no deseado.

MODELO E IMPLEMENTACION DE LA SOLUCION

Diccionario de palabras spam: Palabras que se encuentran frecuentemente en correos spam. El diccionario de palabras es el lenguaje del autómata.

1. Seleccionado
2. Gano
3. Ganado
4. Gratis
5. Gratuito
6. Dinero
7. Duplica
8. Consíguelo
9. Conseguido
10. Regalo
11. Regalamos
12. Oferta
13. Obsequio
14. Obsequiamos
15. Empieza
16. Elegido
17. Felicitaciones
18. Promoción
19. Increíble
20. Urgente

En la siguiente imagen se puede observar el modelo del autómata encargado de encontrar las palabras del diccionario de palabras spam.



Seguido a esto se encuentran los siguientes métodos:

1. Se encarga de procesar el mensaje y devolver un arreglo de palabras que no contenga ningún signo de puntuación o letras con tilde, además cada palabra la deja en letras minúsculas.

```
def processText(txt):
  dictionary = [
        ('á','a'),
        ('é','e'),
        ('í','i'),
        ('ó','o'),
        (',',''),
        ('.',''),
        ('¡',''),
        ('!',''),
        ('¿',''),
        ('?',''),
        ('(',''),
        (')',''),
        (';',''),
        (':','')]
  for c in dictionary:
    txt = txt.replace(c[0],c[1])
  txt = txt.lower().split(' ')
  return txt
```

2. Se encarga de contar cuantas palabras dentro del arreglo son aceptadas por el autómata (es decir, pertenecen al diccionario).

```
def countWords(txt):
  txt = processText(txt)
  aut = automata()
  num_words = 0
  for word in txt:
    if aut.accepts_input(word):
      num_words+=1
  return num_words
```

3. Se encarga de determinar, según la cantidad de palabras dentro del arreglo que pertenezcan al lenguaje del autómata, si la probabilidad de que el mensaje sea spam es baja o alta.

```
def findSpam(txt):
  num = countWords(txt)
  if num>=4:
    print('Hay una alta probabilidad de que el correo sea spam')
  else:
    print('Hay una baja probabilidad de que el correo sea spam')
```

VERIFICACION DEL FUNCIONAMIENTO

Para la verificación del funcionamiento del proyecto se utilizaron 4 correos (de ejemplo), 2 correos que no fueran spam, así como 2 correos que si lo fueran:

- Correos spam:
  1. Correo #1: ¡Felicitaciones, has sido elegido para disfrutar de nuestros beneficios exclusivos! Empieza a disfrutar de tus beneficios ahora. Hemos preparado una oferta especial solo para ti. Como obsequio de nuestra parte, te ofrecemos la oportunidad de obtener algo increíble. Esta promoción es limitada, así que es urgente que actúes cuanto antes. No querrás perderte esta increíble oportunidad.
  2. CORREO #2: ¡Urgente! Usted ha sido seleccionado para para ser parte de esta increíble oferta ¡Felicidades! Obsequiamos una gran numero de celulares, computadores, etc.… y queremos que tu participes. Si deseas participar, empieza enviándonos la siguiente información así podrás ser uno de los afortunados ganadores.
- CORREOS NO SPAM:
  3. Correo #3: Notas ejercicios clase tercer corte. Hola a todos(as), adjunto las 5 notas del tercer corte (N1 a N5) y la definitiva usando las tres mejores (DEF N). También se indican las notas de los dos primeros cortes (P1 y P2). El próximo viernes vamos a realizar el tercer previo en el salón de clase.

  4. Correo #4: Novedades Exclusivas para Nuestros Socios. Espero que este mensaje le encuentre bien. Queremos agradecerle por su continuo apoyo y confianza en nuestros productos y servicios. Estamos emocionados de compartirle algunas novedades exclusivas que hemos preparado para nuestros valiosos socios. Estas mejoras están diseñadas para proporcionarle una experiencia aún más satisfactoria con nuestros productos y servicios. Nos encantaría saber su opinión sobre estas novedades y cómo podemos seguir mejorando para satisfacer sus necesidades. ¿Estaría disponible para

una breve llamada o reunión la próxima semana? Agradecemos sinceramente su asociación y esperamos seguir sirviéndole de la mejor manera posible. Saludos cordiales.

Resultados:

Observaciones:

El proyecto funciona satisfactoriamente, pero está limitado a reconocer correos spam que contengan únicamente palabras que el autómata pueda aceptar, por lo que el proyecto se puede seguir mejorando para que el rango de correos que pueda determinar cómo spam sea más amplio de lo que es ahora.

CONCLUSIÓN

En conclusión, el proyecto presenta un enfoque innovador para abordar el persistente problema del spam en los correos electrónicos, proponiendo un sistema de clasificación basado en la identificación de palabras clave y la construcción de un autómata no determinista (NFA). El objetivo principal es proporcionar a los usuarios una herramienta efectiva que mejore la seguridad y la eficiencia en la gestión de la comunicación electrónica.

Sin embargo, es crucial reconocer que la efectividad de este sistema dependerá en gran medida de la precisión con la que se identifiquen las palabras clave asociadas al spam. La capacidad para adaptarse a nuevas formas de spam también es esencial, ya que las tácticas de los remitentes no deseados pueden evolucionar con el tiempo. La rigurosidad en la selección y actualización de las palabras clave será determinante para mantener la relevancia y eficacia del sistema a lo largo del tiempo.

BIBLIOGRAFIA

Team, B. (2022, August 19). *La lista definitiva de palabras de activación de spam en Emails*. Benchmark Email - Herramienta De Email Marketing. https://www.benchmarkemail.com/es/blog/la-lista-definitiva-de-palabras-de-activacion-de-spam-en-emails/