1 Objetivos

- Aplicar el uso de técnicas de pre-procesamiento para lenguaje natural
- Utilizar y evaluar método clásicos de representación numérica
- Implementar un modelo de machine learning utilizando el algoritmo de NaiveBayes para la clasificación de correos como SPAM

2 Preámbulo

SPAM

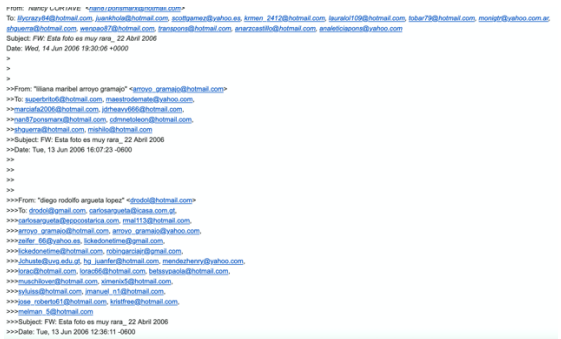
El correo no deseado se refiere a correos que no han sido solicitados, con remitente desconocido, falso o anónimo. Generalmente son mensajes de publicidad, pero algunos son peligrosos al combinarse con técnicas como el pishing y/o malware (troyanos, spyware, ransomware).

Aunque se ha reducido (de 69% en 2012 a 55% en 2018) aún ocupa un porcentaje significativo del tráfico de correo electrónico (en marzo de 2021 el SPAM ocupó el 45%).

¿Cómo se obtienen los correos electrónicos?

Se utilizan bots que recorren Internet en busca de direcciones en fuentes como:

- Foros, blogs
- Grupos de noticias
- Listas de correos
- Correos electrónicos con chistes, cadenas, etc. Si los usuarios reenvían estas cadenas sin ocultar direcciones, es fácil acumular una gran cantidad de correos.
- Brechas



Una cadena con un mensaje clic-bait "Esta foto es muy rara"

¿Cómo se envían los mensajes?

- Software que recorren las listas de correos y envían el mismo mensaje
- Botnets con computadoras infectadas utilizadas para enviar correos
- Servidores de correo mal configurados, algunas organizaciones legítimas han caído en lista negra luego de que sus servidores de correo han sido utilizados para enviar SPAM.

3 Desarrollo

El laboratorio será desarrollado en parejas. Se debe entregar un enlace a un repositorio de Github con el código fuente del pre-procesamiento y los modelos de representación de texto e implementación de los modelos de clasificación, así como la explicación de las métricas de evaluación.

Parte 1 – Ingeniería de características

Exploración de datos

Para el laboratorio se proporcionan dos datasets distintos. Revise la data y realice las operaciones necesarias para **unificar** los datasets y que el dataset final contenga el mensaje del correo y la etiqueta que indique si es SPAM o no.

Muestre ejemplos de los datasets individuales y del dataset final.

Preprocesamiento

Aplique las técnicas de pre – procesamiento de lenguaje natural que considere necesarias (conversión de minúsculas, mayúsculas, eliminación de acentos, expansión de contracciones, eliminación de stop words, etc.)

Representación de texto

Utilice los modelos de BoG (para n = 1,2) y TF-IDF. Muestre algunos ejemplos de los mensajes en su representación numérica.

Parte 2 – Implementación del modelo

Separación de datos

- Datos de entrenamiento: 70%
- Datos de prueba: 30%

Implementación

Utilice el algoritmo multinomial de NaiveBayes para entrenar el modelo con cada uno de los modelos de representación numérico. Muestre los valores obtenidos para las siguientes métricas:

- Matriz de confusión
- Precision
- Recall
- F1 Score

Conclusión

Compare los valores para cada modelo de representación numérico. ¿Qué modelo produjo el mejor resultado, BoG o TF-IDF? ¿A que se debe la mejora? Explique los valores obtenidos en las métricas para el mejor modelo.