Título del proyecto

CHRISTIAN ALEJANDRO RENGIFO MEGIA1, DIEGO FERNANDO GONZALEZ ORTIZ2,
22971 - Inteligencia Artificial I - Grupo B1
Escuela de Ingeniería de Sistemas e Informática

Resumen

En la actualidad los correos basura o spam han tomado una importancia en el medio con el objetivo de estafar y dar publicidad innecesaria, o incluso para perjudicar el sistema.

Se han logrado implementar algoritmos que detectan y clasifican, pero esto no siempre es 100% fiable.

Introducción

Con el paso del tiempo se ha normalizado el hecho de tener que leer y eliminar correos no deseados con información no relevante

En un intento de evitar esto y hacerlo de forma automática se creó una herramienta que reconoce y oculta dichos correos sin la necesidad de un esfuerzo manual.

Pero para poder usar correctamente la herramienta es requerida la importancia de entrenarla primero, que sepa hacer lo que buscamos, pero el inconveniente principal de esto es nuestro lenguaje.

Proceso y método

1. Inicialmente se eliminan columnas innecesarias del dataframe y se agregó un identificador numérico para diferenciar los spam de los ham.

2. Se eliminan las palabras vacías (stopwords), signos de puntuación y espacios, estos se eliminan de la data ya que no aportan información.

3. Se separan las columnas spam y ham, y se procede a contarlas para poder graficar la lista de palabras más usadas de cada label.

4. Usando el modelo de bag of words y TF-IDF, se trata la data y luego se separa en 80% para training y 20% para test.

5. Se procede a usar los clasificadores (Multinomial, Naive Bayes, Gaussian Naive Bayes, Support Vector Classifier, y Stochsatic Gradient Decent)

6. Se usa pipeline y cross validation (GridsearchCV) que permite hacer fit y predict a los clasificadores

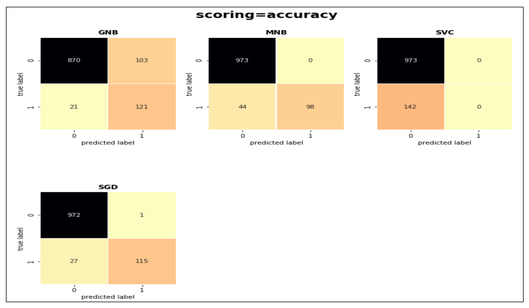
Resultados

Se logró clasificar, ordenar y acomodar de la mejor manera los datos para que posteriormente pudieran ser trabajados


```
0    [go, jurong, point, crazy, available, bugis, n...
1                        [ok, lar, joking, wif, u, oni]
2    [free, entry, 2, wkly, comp, win, fa, cup, fin...
3        [u, dun, say, early, hor, u, c, already, say]
4    [nah, dont, think, goes, usf, lives, around, t...
Name: text, dtype: object
```

Mediante la graficación de los datos, se logró estudiar de mejor manera los mismos, y se pudieron dar las primeras conclusiones de los mismos

En las matrices de confusión se logra apreciar cuál fue la efectividad del procedimiento y se logra igualmente identificar cual es la efectividad de los clasificadores y e esta manera poder concluir de la mejor manera

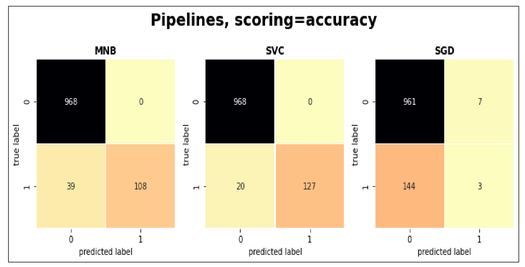


Conclusiones

1. Es indispensable que las redes sociales, los correos electrónicos dispongas de algoritmos adecuados, capaces de hacer frente al Spam

2. Hay gran variedad de métodos para entrenar los datos y de esta manera poder entender mejor su comportamiento

3. Mediante las matrices de confusión se puede visualizar de una mejor manera los resultados obtenidos del procedimiento



Trabajo Futuro

Este proyecto sirve como base para que en investigaciones futuras se pueda mejorar la eficiencia y eficacia del mismo

Información de contacto

Christian Rengifo Mejia , Email: alejoreme12@gmail.com:
Diego Fernando Gonzalez , Email: diego18_98@hotmail.com

Docente: Gustavo Garzón, gustavo.garzon@saber.uis.edu.co

Referencias Bibliográficas (en formato APA)

1. Mayo, M., 2018. *Preprocesamiento De Datos De Texto: Un Tutorial En Python*. [online] Medium. Available at: <https://medium.com/datos-y-ciencia/preprocesamiento-de-datos-de-texto-un-tutorial-en-python-5db5620f1767> .
2. sitiobigdata.com. 2019. *Machine Learning Procesamiento De Texto - Sitiobigdata.Com*. [online] Available at: <https://sitiobigdata.com/2019/12/23/machine-learning-procesamiento-de-texto/#>
3. Kaggle.com. 2019. *SMS: Spam Or Ham (Beginner)*. [online] Available at: <https://www.kaggle.com/dejavu23/sms-spam-or-ham-beginner.