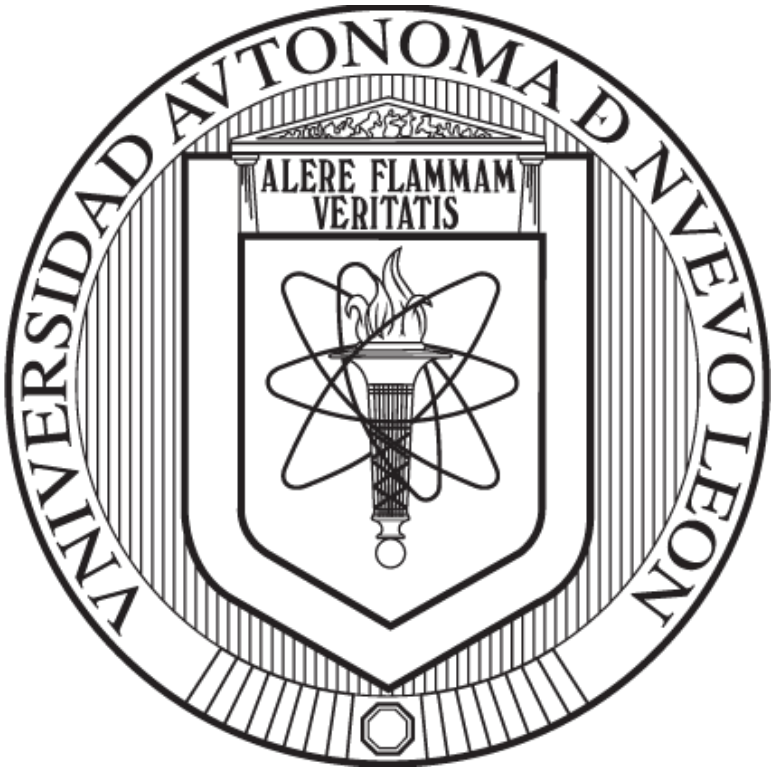
UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN



Reporte de tarea

Preprocesamiento de datos

Autor: **Karla Cureño Vega** Matrícula: **2085376**

Materia: **Procesamiento y Clasificación de Datos** Profesor:**Mayra Cristina Berrones**

Actividad: **Tarea 1** Fecha: **19 de mayo de 2022**

Índice

1. Introducción

El correo electrónico es una de las herramientas más antiguas y útiles que nacieron con internet y a pesar de que han surgido nuevo métodos de comunicación, hoy en día, el correo electrónico es ampliamente utilizado debido a la versatilidad en la transmisión del mensaje, sin embargo, también ha traído consigo una serie de problemas importantes, siendo dos de ellos, el aumento de tráfico de la red y el robo de identidad debido a correos electrónicos de carácter malicioso llamados correos spam.

2. Descripción del conjunto de datos

El conjunto de datos del cuál se extraerá un subset al de la base de datos "Enron Email Dataset". Contiene datos de alrededor de 150 usuarios con un total de alrededor de 500,000 mensajes de correo electrónico. Originalmente el dataset se hizo público por la Comisión Federal Regulatoria de Energía de Estados Unidos durante la investigación alrededor del colapso de la empresa Enron.

El subconjunto a utilizar consta de 5,854 correos, de los cuales 1,496 están catalogados como spam y el resto corresponden a correos electrónicos legítimos.

3. Preprocesamiento de datos

Se utilizan los siguientes pasos y métodos para preprocesar los emails:

- **Conversión a minísculas**: El cuerpo completo del email se convierte a minísculas, con la finalidad de ignorar la capitalización de las palabras. Para realizar este paso se hace uso de la función `lower()` de Python.
- **Remoción de caracteres no alfábeticos**: Se remueven las no-palabras y signos de puntación. Todos los espacios en blanco (tabs, líneas nuevas, espacios) se recorta a un cáracter de espacio individual. En este paso se utiliza una expresión regular para reemplazar estos caracteres con las funciones `compile` y `sub` de la biblioteca `re`.
- **Remoción de "stop-words"**: Se quitan preposiciones, pronombres y palabras conocidas como "stop-words". Para este paso se utiliza la lista de "stop-words"de la biblioteca `nltk` en el idioma inglés.
- **Lematización**: Consiste en dada una forma flexionada (es decir, en plural, en femenino, conjugada, etc), hallar el lema correspondiente. El lema es la forma que por convenio se acepta como representante de todas las formas flexionadas de una misma palabra. Para esto se utiliza `WordNetLemmatizer` de la biblioteca `nltk` de Python.

4. Resultados

En la figura 1 se observa un email que se usará como ejemplo para comparar los resultados del preprocesamiento. En esta figura, se aprecia el email original. En este email se puede observar la presencia de muchos signos de puntuación, caracteres especiales y números, todo esto no nos aporta información relevante sobre el mensaje, por lo que se realiza el preprocesamiento para limpiar el email y extraer la información más relevante del mismo.

```
Subject: rice program in financial engineering\ndear vince ,\ntony and i are looking forward
to meeting with you tomorrow\n( thurs . ) at 10 a . m . regarding the rice iniative in
financial\nengineering . attached to this message is our draft plan for\nthe proposed
undergraduate program ; i will also bring a copy\nto the meeting in the morning .\nsee you
tomorrow !\nbest regards ,\nkathy\nkatherine bennett ensor\nprofessor and
chairperson\ndepartment of statistics , ms 138\nrice university\nhouston , tx 77251 -
1892\nensor @ rice . edu\nphone : ( 713 ) 527 4687\nfax : ( 713 ) 285 5476\n- draft - plano 2
. doc
```

Figura 1: Correo electrónico original.

Después de haber aplicado el preprocesamiento descrito en la sección anterior el correo electrónico resultante se puede observar en la figura 2. Este mensaje ya está listo para ser analizado debido a que se ha removido la información no relevante.

```
rice program financial engineering dear vince tony looking forward meeting tomorrow thurs
regarding rice iniative financial engineering attached message draft plan proposed
undergraduate program also bring copy meeting morning see tomorrow best regard kathy katherine
bennett ensor professor chairperson department statistic rice university houston tx ensor rice
edu phone fax draft plano doc
```

Figura 2: Correo electrónico después de haberle aplicado el preprocesamiento de texto.

Debido a que el conjunto de datos estudiado contiene datos etiquetados dentro de dos diferentes categorías se decidió analizar cada categoría por separado para obtener más información resultante del preprocesamiento de los textos.

En primer lugar se observa en la figura 3 la nube de palabras correspondiente a los correos electrónicos etiquetados como spam. En la misma se puede observar el uso de palabras frecuentes tales como “free”, “money”, “new”, “account”. Esto también se observa en la figura 4 donde se muestra un gráfico de barras con la frecuencia de las 15 palabras más utilizadas en nuestro subconjunto de datos de correos electrónicos catalogados como spam. Se puede inferir que el uso frecuente de estas palabras en estos correos de tipo fraudulentos es debido a que usualmente en estos correos se nos contacta con publicidad falsa sobre algún nuevo producto, o anunciando algún premio o dinero gratis que se puede reclamar haciendo click en algún sitio web o creando una cuenta.

Figura 3: Nube de palabras de los correos spam.

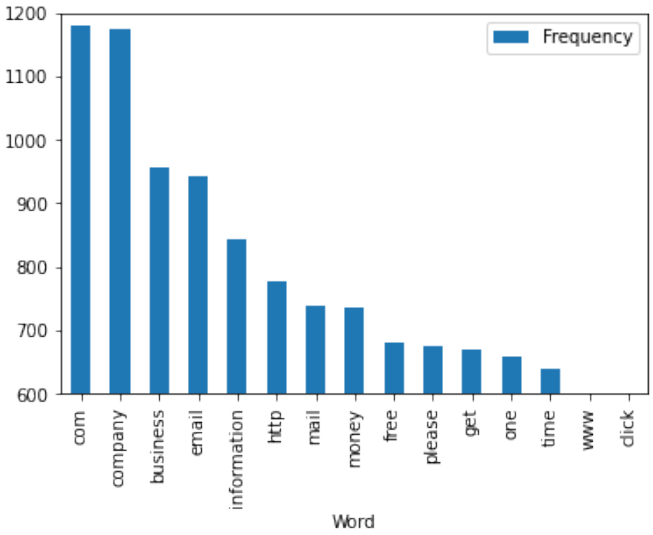


Figura 4: Gráfico de barras de las 15 palabras más utilizadas en correos spam.

Por otro lado, en la figura 5 se observa la nube de palabras correspondiente al subconjunto de datos de correos electrónicos etiquetados como legítimos. En esta se observa que las palabras destacadas son muy diferentes a las de los correos spam. Se complementa esta información con la gráfica de frecuencias de las palabras más utilizadas en estos correos que se observa en la figura 6. De estas dos figuras se puede concluir que las palabras más relevantes en este tipo de correos son tales como 'enron", "please", "research", "thanks", "risk", "market". A diferencia de las palabras destacadas en los correos spam, estas palabras denotan una intención de negocio clara dentro de la empresa Enron. Se puede asumir que estas palabras se utilizan en contexto coorporativo para pedir información sobre el mercado o los riesgos asociados a alguna decisión dentro de la compañía.

Figura 5: Nube de palabras de los correos legítimos (non-spam).

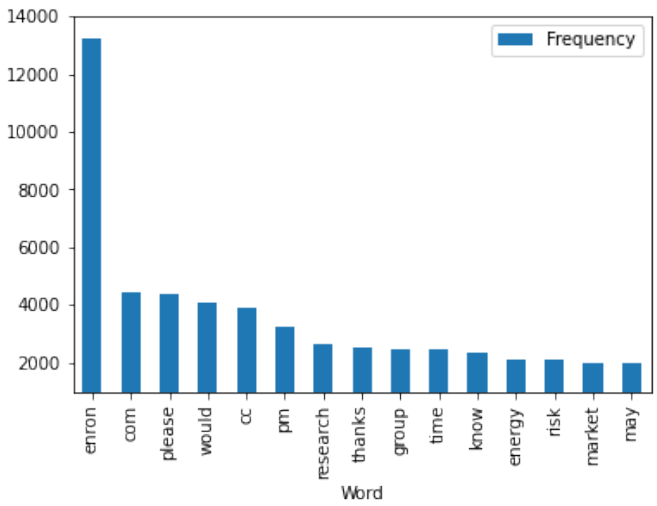


Figura 6: Gráfico de barras de las 15 palabras más utilizadas en correos legítimos (non-spam).

5. Conclusiones

El preprocesamiento de textos es un paso fundamental para poder comenzar un análisis. Con el mismo, es asegurado que se extrae únicamente la información relevante al análisis para cada texto preprocesado. Esto evita que en etapas posteriores se utilice capacidad computacional no necesaria analizando datos irrelevantes para la finalidad del análisis.

Los resultados obtenidos con el preprocesamiento realizado en esta actividad podrían ser utilizados para desarrollar un modelo capaz de etiquetar los correos electrónicas en spam o no spam de manera automática.

Referencias

[1] Cohen, W. (2015). Enron Email Dataset. [Online]. Disponible en: `https://www.cs.cmu.edu/~./enron/`

[2] Cureno, K. (2021). Preprocesamiento de Datos. [Online]. Disponible en: `https://github.com/karlacuv/MCD_Procesamiento/blob/main/Tarea1_Preprocesamiento.ipynb`

[3] Berrones, M. (2021). Preprocesamiento de Texto. [Online]. Disponible en: `https://github.com/mayraberrones94/FCFM/blob/master/Semana_1_Pre_procesamiento_de_datos.ipynb`

[4] NLTK Project. (2022). NLTK: Natural Language Toolkit. [Online]. Disponible en: `https://www.nltk.org/`

[5] NLTK Project. (2022). NLTK: Natural Language Toolkit. [Online]. Disponible en: `https://www.nltk.org/`

[6] PyPI. (2020). wordcloud PyPI. [Online]. Disponible en: `https://pypi.org/project/wordcloud/`