Proyecto: Clasificación de Textos en Lenguaje Natural

Objetivo: Construir un sistema para detectar automáticamente Phishing (obtener información confidencial de forma fraudulenta) en correos electrónicos.

Contenidos:

Parte 3 Clasificación

En esta parte se clasificarán los correos como seguros (`S`) o phishing (`P`).

Escribe un programa que tome como entrada las estimaciones de probabilidad de cada palabra en `modelo_lenguaje_<P o S>.txt` y pida un corpus con correos a clasificar: `PHI_test.csv` (con cada correo en una línea).

El programa debe clasificar todas los correos de `PHI_test.csv` y devolver los correos clasificados en dos ficheros:

- `clasificacion_alu<numero de alu>.csv` donde cada línea del fichero de salida tiene el formato:

`<primeros 10 caracteres del correo>,<lP en S>,<lP en P>,<S o P>`

`lP:` logaritmo neperiano de: la probabilidad del correo en la clase por la probabilidad de la clase, con 2 decimales.

`<S o P>` la clase en la que se clasifica el correo.

- `resumen_alu<numero de alu>.csv` donde cada línea del fichero de salida tiene el formato:

`<S o P>:` clase en la que se clasifica el correo.

Notas:

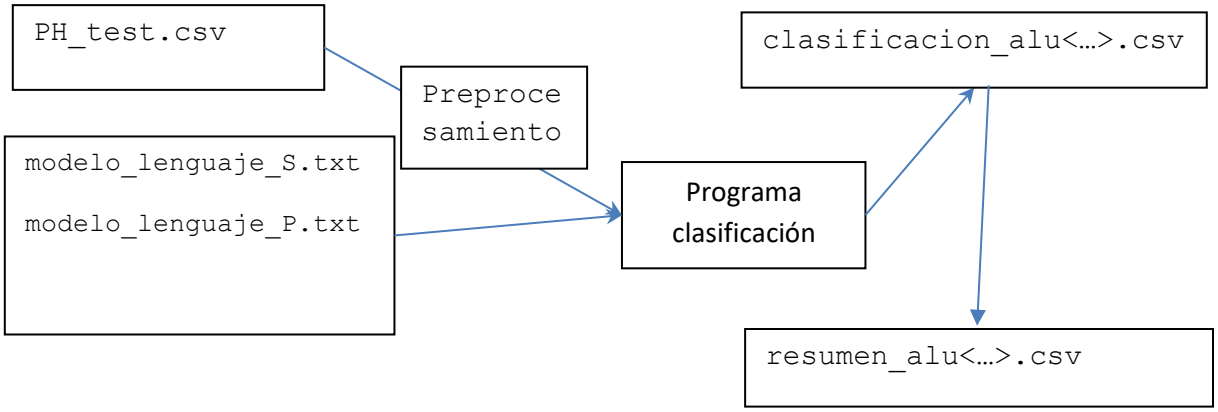
En los ficheros de salida no deben aparecer corchetes <,>

Los correos clasificados deben estar en el mismo orden de entrada.

Antes de subirlo:

1. Probar el programa con `PH_train.csv` quitando la clase y estimar el error de clasificación. Este error debe escribirse en el informe de la implementación.
2. Dividir `PH_train.csv` en dos ficheros: `PH_train_1.csv` con los primeros 10000 correos y `PH_train_2.csv` con las restantes. Entrenar el sistema con el primer fichero y estimar el porcentaje de error con el segundo.

<u>Se penalizará</u> con un 50% de la evaluación no ajustarse al nombre del fichero o al formato pedido.



Evaluación del Proyecto

- Entregables: Breve informe con la implementación: Preprocesamiento, librerías utilizadas, implementación de los programas. Estimación de errores 1 y 2. Programas y ficheros pedidos (3/10)
- Rendimiento del programa sobre el corpus que proporcionará el profesor (7/10):
  - 97-100% del porcentaje de acierto del mejor programa: 7 puntos
  - 95-97% del porcentaje de acierto del mejor programa: 6 puntos
  - 90-95% del porcentaje de acierto del mejor programa: 5 puntos
  - 87-90% del porcentaje de acierto del mejor programa: 4 puntos
  - 83-87% del porcentaje de acierto del mejor programa 3: puntos
  - 75-83% del porcentaje de acierto del mejor programa 2: puntos
  - Menos del 75% del porcentaje de acierto del mejor programa 0.5 puntos