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 1 Estimación de probabilidades en el modelo del lenguaje

En esta parte se estimarán las probabilidades del modelo del lenguaje para las clases Phishing y Safe_Email. Utiliza el fichero `Phi_train.csv` en el campus virtual. Tiene 15000 correos electrónicos formato:

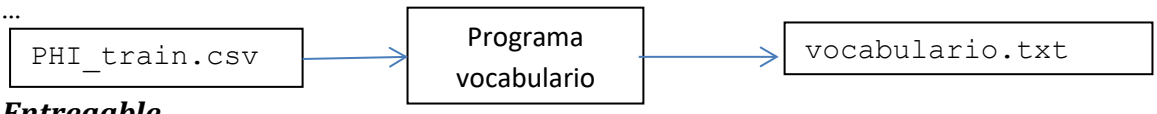
```
<numero de correo>;<texto>;<tipo> tipo=Phishing o Safe Email
```

1.1 Creación del vocabulario

Halla el vocabulario del problema. Para ello examina el fichero `PHI_train.csv`, obtén qué palabras están presentes en los correos (preprocesamiento y tokenización) y pon las palabras en el fichero `vocabulario.txt`. Si una palabra se repite ponla sólo una vez. Las palabras del fichero de vocabulario deben estar ordenadas alfabéticamente.

El fichero `vocabulario.txt` tendrá el formato:

```
Numero de palabras:<Número entero>
<palabra>
<palabra>
…
```



Entregable

En el Campus Virtual

- **Programas:**
  - Vocabulario
- **Ficheros:**
  `vocabulario.txt`

- ***Nota:*** Proyecto individual, lenguaje de programación libre, utilización de librerías libre. Se penalizará con un 50% no entregar los ficheros en el formato pedido.

Preprocesamiento
Tareas típicas:

- Pasar a minúsculas.
- Eliminación de signos de puntuación.
- Eliminación de palabras reservadas (stopwords).
- Eliminación de emojis y emoticonos o su conversión a palabras.
- Eliminación de URLs, etiquetas HTML, hashtags.
- Corrección ortográfica.
- Truncamiento: Reducir una palabra a su raíz (grito, grita, gritos, gritas ->grit).
- Lematización: Reducir una palabra a su forma canónica (dije,diré,dijéramos->decir).

Algunas stopwords en inglés:

a,able,about,across,after,all,almost,also,am,among,an,and,any,are,as,at,be,because,been,but,by,can,cannot,could,dear,did,do,does,either,else,ever,every,for,from,get,got,had,has,have,he,her,hers,him,his,how,however,i,if,in,into,is,it,its,just,least,let,like,likely,may,me,might,most,must,my,neither,no,nor,not,of,off,often,on,only,or,other,our,own,rather,said,say,says,she,should,since,so,some,than,that,the,their,them,then,there,these,they,this,tis,to,too,twas,us,wants,was,we,were,what,when,where,which,while,who,whom,why,will,with,would,yet,you,your.

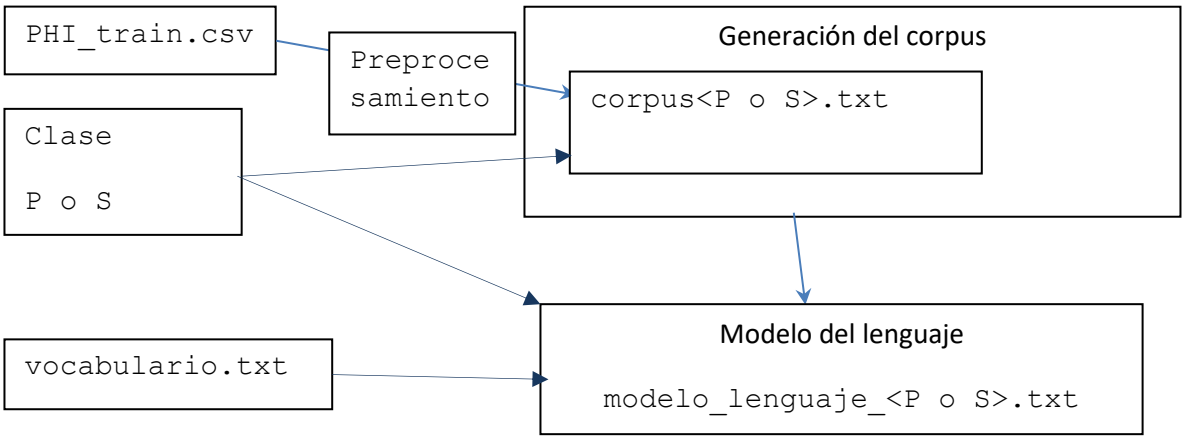
1.2 Estimación de probabilidades

La estimación de las probabilidades para los corpus correspondientes a las clases Phishing y Safe_Email. Se escribirá en un fichero de texto llamado `modelo_lenguaje_<P o S>.txt`. En el fichero de texto debe aparecer:
Cabecera:

```
Numero de documentos (noticias) del corpus :<número entero>
Número de palabras del corpus:<número entero>
```

Por cada palabra de `vocabulario.txt`, su frecuencia absoluta en el corpus y una estimación del logaritmo de su probabilidad mediante suavizado laplaciano con tratamiento de palabras desconocidas. Las palabras en los ficheros de aprendizaje estarán ordenadas alfabéticamente.

```
Palabra:<cadena> Frec:<número entero> LogProb:<número real>
```



Entregable

En el Campus Virtual

- **Programas:**
  - Aprendizaje
- **Ficheros:**

  `modelo_lenguaje_<P o N o T>.txt`.

- ***Nota:*** Proyecto individual, lenguaje de programación libre. Se penalizará con un 50% no entregar los ficheros en el formato pedido.