Práctica: Clasificación de Textos en Lenguaje Natural

Objetivo: Construir un Sistema para la detección de mensajes relevantes sobre desastres (incendios, bombas, terremotos, etc.) en Twitter (en oposición por ejemplo a un comentario casual, una observación sobre una película o algo no desastroso).

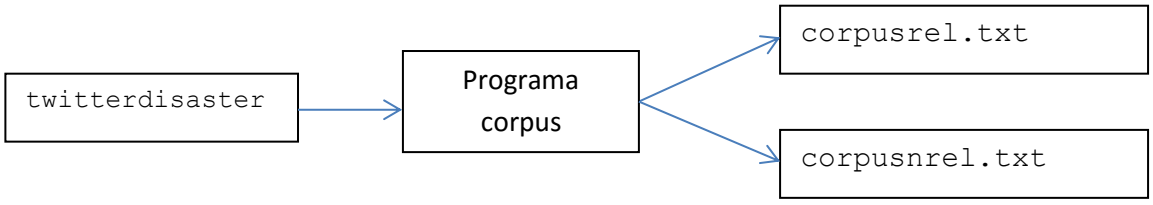
Contenidos:

Parte 1 Estimación de probabilidades en el modelo del lenguaje

En esta parte se estimarán las probabilidades del modelo del lenguaje para las clases relevante y no relevante

1.1 Creación de los corpus

Utiliza el fichero Excel `twitterdisaster.xlsx` proporcionado en el campus virtual. Tienes 10806 comentarios de los cuales 4654 son relevantes y 6152 son no relevantes. Crea dos corpus con nombre `corpus<rel o nrel>.txt`. El primero con los comentarios relevantes y el segundo con los comentarios relevantes. Cada línea del fichero de salida en el corpus debe tener la siguiente estructura:
`Texto:<cadena con texto del fichero>`



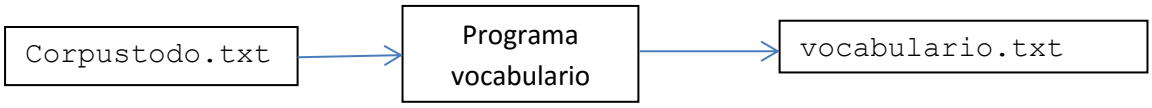
Crea también el fichero `corpustodo.txt` concatenando `corpusrel.txt` y `corpusnrel.txt`

1.2 Creación del vocabulario

Halla el vocabulario del problema. Para ello examina el fichero `corpustodo.txt` y obtén las palabras del vocabulario a partir del texto (tokenization).

Debes generar un fichero de salida `vocabulario.txt` con cabecera

```
Numero de palabras:<Número entero>
Palabra:<cadena>
```



Las palabras de `vocabulario.txt` estarán ordenadas alfabéticamente.

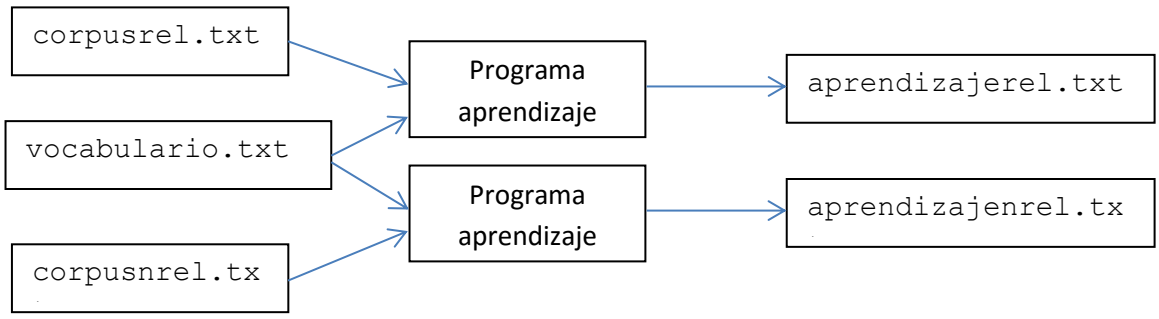
1.3 Estimación de probabilidades

La estimación de las probabilidades se escribirá en un fichero de texto llamado `aprendizaje<rel o nrel>.txt`. En el fichero de texto debe aparecer:
Cabecera:

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

Por cada palabra de `vocabulario.txt`, su frecuencia 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>
```

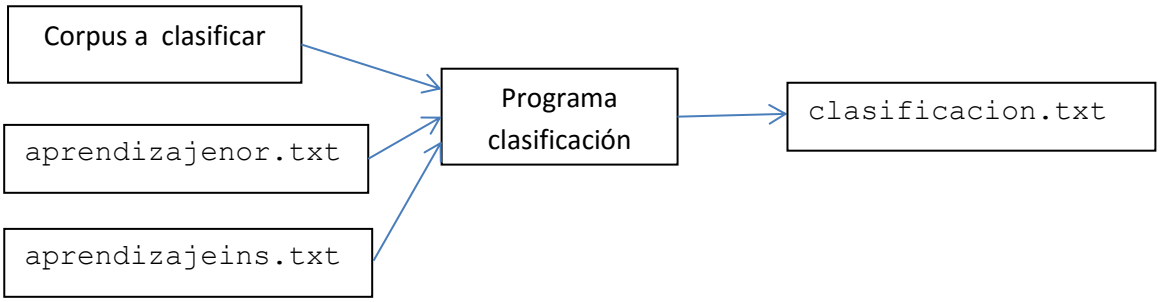


Parte 2 Clasificación

En esta parte se clasificarán los documentos presentes en un corpus.

Escribe un programa que tome como entrada las estimaciones de probabilidad de cada palabra y un corpus con documentos a clasificar (con el formato de corpus de 1.1) y devuelva los documentos clasificados en un fichero `clasificación.txt` donde cada línea del fichero de salida con el corpus tenga la siguiente estructura:

`Clase:<rel o nrel> Texto:<cadena con texto del fichero>`



Deberás clasificar `corpustodo.txt` generando el fichero `clasificación.txt`

Entregable

En el Campus Virtual

- **Programas:**
  - Corpus, Vocabulario, Aprendizaje,Clasificación
- **Ficheros:**

`corpusrel.txt,corpusnrel.txt,corpustodo.txt,aprendizajerel.txt aprendizajenrel.txt, clasificacion.txt`

Nota

- **Obligatorio: 2 alumnos por práctica**. **No** puedes repetir con quien ya hayas trabajado en grupo
- Lenguaje de programación libre.