Proyecto: Clasificación de Textos en Lenguaje Natural

Objetivo: Construir un sistema para la detección de cybertrolls en twitter a partir de un corpus con tweets de cybertrolls y tweets de usuarios normales.

Aviso importante: El corpus a tratar contiene expresiones altamente ofensivas. Si algún alumno prefiere trabajar un problema alternativo puede hacerlo comunicándoselo al profesor.

Contenidos:

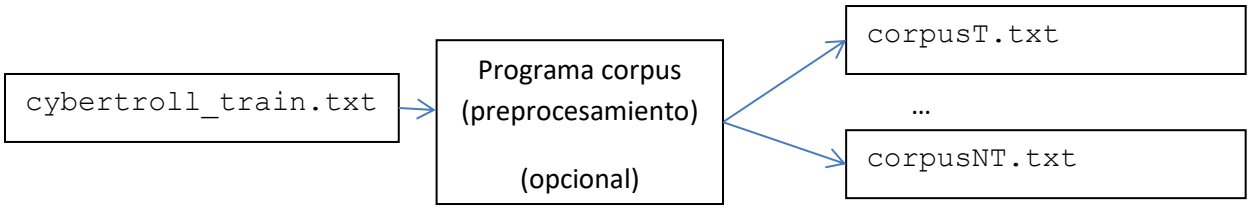
Parte 1 Estimación de probabilidades en el modelo del lenguaje

En esta parte se estimarán las probabilidades del modelo del lenguaje para las clases troll (T), y notroll (NT).

1.1 Creación de los corpus

Utiliza el fichero cybertroll_train.txt en el campus virtual. Tienes 16435 tweets clasificados en las categorías: troll y not_troll

Crea 2 corpus con nombre `corpus<T o NT>.txt` con los mensajes de cada categoría. Cada línea del fichero de salida en el corpus debe tener la siguiente estructura:
`<cadena con texto del fichero>`



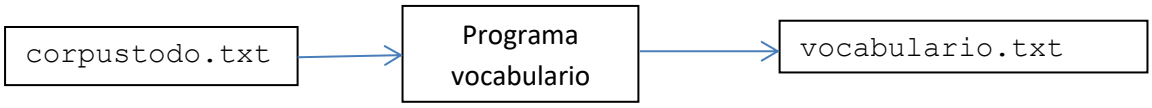
Crea también el fichero `corpustodo.txt` concatenando todos los corpus

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.

Entregable

En el Campus Virtual

- **Programas:**
  - Corpus (opcional), Vocabulario
- **Ficheros:**

`corpusT.txt,corpusnT.txt,corpustodo.txt, vocabulario.txt`

Nota

- Lenguaje de programación libre.

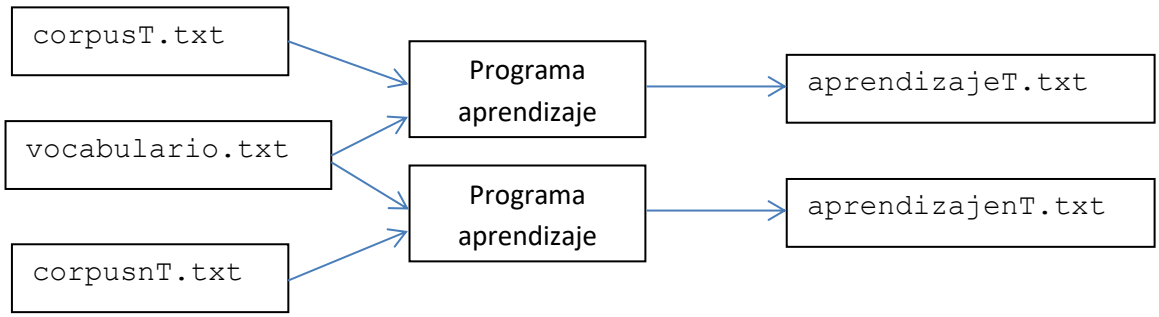
1.3 Estimación de probabilidades

La estimación de las probabilidades se escribirá en un fichero de texto llamado `aprendizaje<T o nT>.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>
```



Entregable

En el Campus Virtual

- **Programas:**
  - Aprendizaje(fuentes)
- **Ficheros:**
  - `vocabulario.txt, aprendizajeT.txt, aprendizajenT.txt`

-

- Lenguaje de programación libre.