Práctica: Clasificación de Textos en Lenguaje Natural

Objetivo: Construir un Sistema para la clasificación de los mensajes que emiten las empresas en las redes sociales. Los mensajes se dividen en informativos (afirmaciones objetivas sobre la empresa o sus actividades), diálogo (respuestas a los usuarios, etc.) o acciones (mensajes que piden votos a los usuarios, que piden clicks en enlaces, etc.)

Contenidos:

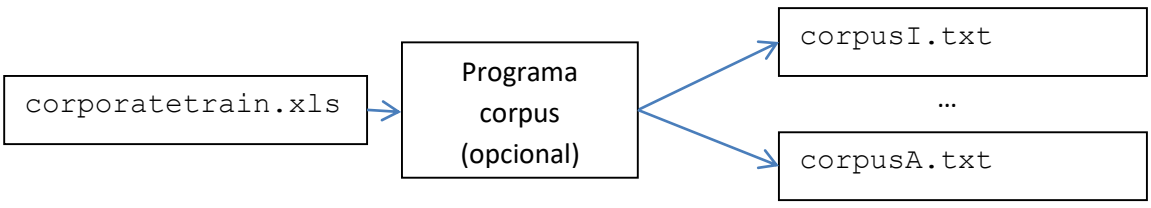
Parte 1 Estimación de probabilidades en el modelo del lenguaje

En esta parte se estimarán las probabilidades del modelo del lenguaje para las clases informativo (I), diálogo (D) y acción (A)

1.1 Creación de los corpus

Utiliza el fichero Excel `corporatetrain.xlsx` proporcionado en el campus virtual. Tienes 1716 mensajes de empresas clasificadas en las categorías: informativo, diálogo y acción

Crea 3 corpus con nombre `corpus<inicial_categoría>.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>`



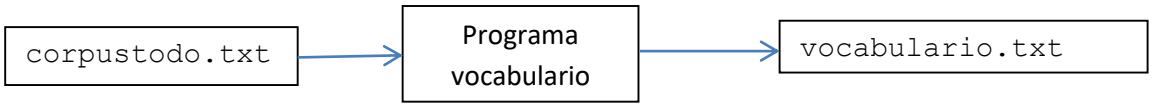
Crea también el fichero `corpustodo.txt` concatenando todos los corpus

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:**

`corpusI.txt,…,corpusA.txt,corpustodo.txt, vocabulario.txt`

Nota

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

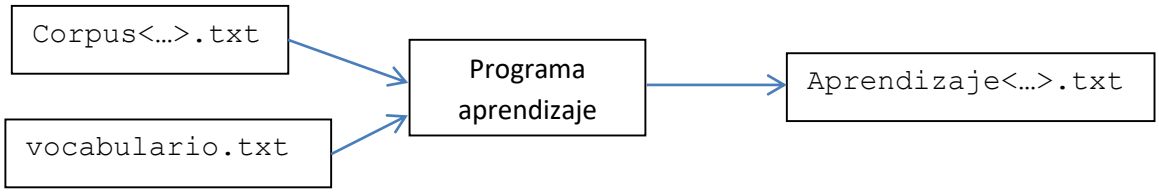
1.3 Estimación de probabilidades

La estimación de las probabilidades se escribirá en un fichero de texto llamado `aprendizaje<inicial_categoría>.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:**
  - Corpus, Aprendizaje(fuentes)
- **Ficheros:**
  - `vocabulario.txt, aprendizaje<inicial_categoría>.txt,`

Nota

- **Máximo grupos de 2 alumnos por práctica**. No puedes repetir con quien ya hayas trabajado.
- Lenguaje de programación libre.