Tarea 5

1. Entrenando un clasificador basado en texto

En la carpeta tienes dos archivos tipo csv. El primero, intervenciones.csv, tiene los discursos de todos los miembros de la convención constitucional realizados en el pleno de la convención, la información que está es el orador y el discurso. El segundo, leg_data_party.csv, tiene bastante información sobre los constituyentes. Interesan acá los campos nombres_cc, con el nombre del(a) convencional, y género, que puedes ser MASCULINO o FEMENINO.

Para esta tarea debes implementar una clase ClasText, con dos métodos:

1. Una método entrenar, que debe recibir una lista o un array de strings X, y una lista, array o series de números 0 o 1 y. Este método debe entrenar un clasificador que pueda predecir si la persona que escribió ese string fue un hombre 0 o una mujer 1.

2. Un método predecir, que reciba una lista o un array de strings, y entregue una lista, array o series de números 0 o 1.

Debes entrenar tu clase usando los textos de las intervenciones en la convención constitucional. Nota que para cada uno de esas frases, puedes usar el dataset leg_data_party.csv para averiguar si quien lo dijo es hombre o mujer. Con eso, puedes construir un set de entrenamiento con 1708 registros.

El método resultante debe entregar una accuracy media superior al 57 %, cuando se prueba con 5-fold cross-validation, de acuerdo a la funcionalidad cross_val_score de scikit learn, que debe llamarse como (para un dataset X con vector de clasificaciones 0/1 y.

```
import numpy as np
from sklearn.model_selection import cross_val_score
clf = <definir el clasificador>
np.mean(cross_val_score(clf, X, y, cv=5, scoring='accuracy'))
```

Importante: puedes verificar esto sin necesidad de llamar a tu clase, si no que hacer un módulo que lo verifique de forma aparte. Esto, por que clf debe ser un clasificador de sklearn, y no es parte de esta tarea extender esa clase (aunque se puede hacer sin problemas, y no es mala idea aprender a hacerlo).

2. Detalles administrativos

La entrega de esta tarea es el Viernes 3 de diciembre, a las 20:00 hrs, por cuestionario en SIDING. Debe entregarse un zip que contenga los archivos necesarios para correr un solo notebook, que ya debe venir con todos los pedazos de código ya ejecutados. El archivo debe llamarse **NumeroAlumno_Apellido_Nombre**.