TALLER DE PROCESAMIENTO DE SEÑALES

1er Cuatrimestre 2026 - Trabajo Práctico Nº 9 - Multinomial Naive Bayes

---

Se desea decidir, de forma automática, a qué tópico pertenece un artículo científico a partir de su título y resumen.

(**a**) *Creación del dataset*: Se desean descargar 1000 artículos de arXiv de cada una de las siguientes categorías: `cs.CL`, `cs.CV`, `math.PR`, `physics.optics`, `q-bio.NC` y `econ.EM`. Para ellos se utilizará la API `"http://export.arxiv.org/api/query"`.

1. Utilizando `get` (request), solicitar el `xml` de un artículo. La API requiere una estructura:

```
params = {
    "search_query": f"cat:{category}",#Categoría a descargar
    "start": start, #Offset necesario para no descargar siempre el mismo artículo
    "max_results": max_results #Cantidad de artículos simultáneos a descargar
}
```

   ✍: La API posee un *rate limit*, es necesario espaciar las descargas.

2. Obtener de cada `xml` el título y resumen del artículo. Utilizar un parser de `xml`.
3. Generar una base de datos concatenando el título al resumen, y utilizando la categoría como etiqueta.
4. Definir los conjuntos de entrenamiento y testeo con las proporciones 80 % y 20 % de forma aleatoria.
5. Aplicar `CountVectorizer` (sklearn) al texto. Ajustar `max_df`, `min_df` y `stop_words` a criterio personal.

(**b**) *Multinomial Naive Bayes:*

1. Utilizando solamente numpy y scipy, implementar el clasificador MNB. El mismo debe contener los métodos `fit`, `predict` y `predict_proba`.
2. Reportar el *accuracy* tanto para entrenamiento como testeo.

(**c**) *Exploración de hiperparámetros:*

1. Repetir el punto (**b**) para diferentes valores de `max_df` y `min_df`.
2. Graficar el accuracy de testeo en función del `max_df` para un `min_df`= 0.0.
3. Graficar el accuracy de testeo en función del `min_df` para un `max_df`= 0.2.