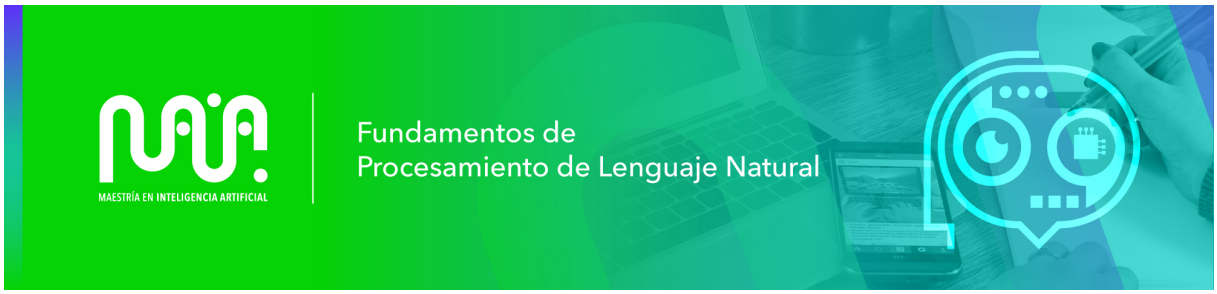

Instrucciones de Competencia Final

Curso Fundamentos de Procesamiento de Lenguaje Natural

1 Introducción

Bienvenido a la competencia final del curso **Fundamentos de Procesamiento de Lenguaje Natural**. En este desafío, los participantes deberán construir un sistema de **clasificación de textos por década de origen**. El objetivo es diseñar un modelo capaz de predecir la década en que fue escrito un párrafo, utilizando técnicas tradicionales de aprendizaje de máquina y procesamiento de lenguaje natural.

2 Resultados de Aprendizaje

Con esta actividad se busca que el estudiante pueda poner en práctica el desarrollo de una solución completa de *machine learning* para un problema de clasificación de texto. Tras realizar esta actividad se espera que el estudiante esté en capacidad de:

1. Proponer y seleccionar modelos de *machine learning* que permitan resolver un problema de clasificación de texto.
2. Identificar y aplicar diferentes técnicas de procesamiento de texto.
3. Representar el texto de manera que permita el uso de modelos de *machine learning*.
4. Hacer uso de los datos para entrenar y evaluar el desempeño del modelo.
5. Probar distintos modelos para encontrar la mejor solución para el problema particular.

3 Descripción de la Competencia

Usted y su grupo deberán participar en una competencia en Kaggle diseñada para este curso. La competencia consiste en clasificar textos en la década en que fueron escritos. El reto es construir un modelo que, dado un párrafo de entrada, prediga correctamente la década de origen. La competencia es exclusiva para estudiantes de la Maestría en Inteligencia Artificial de la Universidad de los Andes inscritos en el primer ciclo del 2025-20.

3.1 Requisitos técnicos

Recuerde que su modelo no tendrá validez si no cumple los requisitos que se mencionan a continuación. Por lo tanto, obtendrá una calificación de 0 en la actividad correspondiente al proyecto final. Los requisitos son los siguientes:

- No se permite el uso de técnicas de aprendizaje profundo. No está permitido el uso de modelos de lenguaje de tipo transformer.
- Solo podrá usar modelos de clasificación desarrollados con la librería `scikit-learn` y que no violen otros de los requisitos aquí establecidos.
- La competencia debe resolverse exclusivamente con modelos clásicos de aprendizaje de máquina y técnicas de PLN vistas en este curso. El uso de incrustaciones como Word2Vec o Glove estan permitidas.
- Para la revisión en Coursera debe subir un solo modelo, el cual debe corresponder al modelo que obtuvo el mejor resultado en la competencia de Kaggle.
- Se permite el uso de datos externos siempre que la licencia lo permita, puedan ser referenciados, descargados y distribuidos. NO se permiten datos sintéticos.
- Está prohibido utilizar los datos de prueba para entrenar modelos.
- Los resultados deben ser replicables a partir de los entregables en la plataforma de Coursera.

4 Envío de resultados por la plataforma Kaggle

Para enviar los resultados, genere un archivo CSV en el formato especificado y súbalo en la plataforma Kaggle. Una parte de los resultados se usará para el **score público** y otra parte para el **score privado**, que determinará el resultado final.

5 Entregables en Coursera

En la semana 8 del curso se habilitará una actividad en donde usted y su grupo deben subir los siguientes entregables correspondientes únicamente al modelo final con el que obtuvieron los mejores resultados en la tabla de líderes pública de la competencia de Kaggle:

1. Jupyter Notebook que se utilizó para realizar la implementación y el entrenamiento del clasificador de texto.
2. Si utilizo mas datos debe referenciarlos claramente, incluyendo la página donde los encontró y la licencia explicita.
3. El archivo donde se guardó el modelo de `scikit-learn` entrenado. El modelo debe guardarse haciendo uso de las librerías `pickle` o `joblib`.

6 Rúbrica de Evaluación

La nota del proyecto se compone de los siguientes criterios:

Table 1: Composición de la nota del proyecto final

Item	**Porcentaje de la nota**
Participación en la competencia	20%
Superar la línea base (ver en Kaggle)	40%
Percentil obtenido en la competencia	40%

6.1 Participación en la competencia

Para que sea considerada su participación en la competencia debe realizar al menos 5 envíos `diferentes` en la plataforma de Kaggle. La idea es evidenciar que el grupo trato de mejorar su enfoque inicial. Tiene dos semanas para cumplir con ese número mínimo de envíos. No hay limite máximo de envíos.

6.2 Superar la línea base

El equipo del curso preparo una linea base (i.e. modelo) que debe superar. Esta linea base la encuentra en el leaderboard de la competencia en Kaggle.

6.3 Percentil obtenido en la competencia

El 40% restante de la nota final se obtiene de acuerdo al desempeño del modelo desarrollado con su grupo en la competencia de Kaggle. Al cerrar la competencia se generan unos puntajes finales, y su nota en este rubro dependerá del desempeño obtenido. El porcentaje obtenido de este rubro se calcula de la siguiente manera:

Desempeño obtenido	**Porcentaje del rubro**
$<$ percentil 25	0%
$\geq$ percentil 25 y $<$ percentil 50	50%
$\geq$ percentil 50 y $<$ percentil 75	75%
$>$ percentil 75	100%

Table 2: Distribución del porcentaje del rubro según desempeño obtenido

6.4 Fechas Importantes

- Fecha de apertura de la competencia: **13/09/2025 00:00 AM**
- Fecha de finalización: **28/10/2025 11:59 PM**

7 Descripción de los Datos

Se proporcionan los siguientes archivos:

- **train.csv**: contiene alrededor de 31,400 ejemplos de entrenamiento con columnas:
  - `text` (str): texto a clasificar.
  - `decade` (int): década del texto expresada como los tres primeros dígitos del año (ejemplo: para el año de 1572 la década es 157).
- **eval.csv**: archivo de evaluación con columnas:
  - `id`: identificador único del ejemplo de prueba.
  - `text`: texto a clasificar.

7.1 Archivo de respuesta

El archivo de respuesta debe contener el siguiente formato:

```
ID,answer
2,150
5,162
6,173
...
```

Cada `id` en el conjunto de prueba debe tener un valor en la columna `answer`, indicando la década predicha por el modelo.

8 Enlaces de la Competencia

- **Enlace de Invitación**: Kaggle Invitation