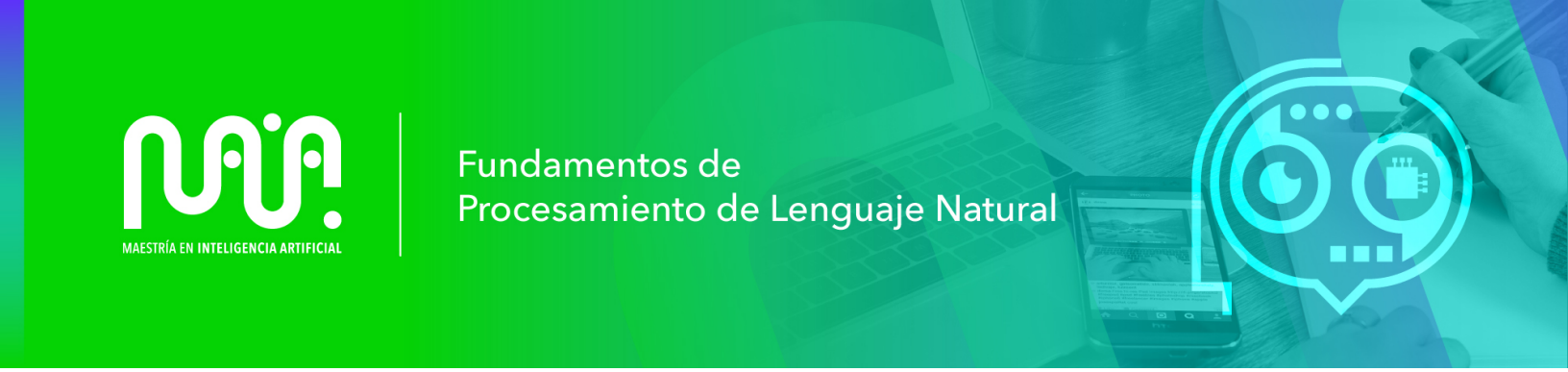

Instrucciones de Competencia Semana 8

Rubén Manrique

Universidad de Los Andes, Colombia

1 Introducción

Esta guía le será útil para desarrollar de manera satisfactoria el proyecto final del curso. Por favor léala atentamente. La guía está estructurada de la siguiente forma:

Contents

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

En la semana 8, usted junto con su grupo deben participar en una competencia de Kaggle que fue específicamente diseñada para este curso. La competencia consiste en una tarea de clasificación en donde dispone de secuencias de texto, con el objetivo de crear un modelo capaz de asignar el autor correspondiente a cada una de las sequencias.

3.1 Construcción del dataset

Para construir el dataset, se descargaron del proyecto Gutenberg una serie de libros de 5 autores. Para conformar las secuencias se realizaron los siguientes pasos de procesamiento:

1. Se tokenizó cada uno de los libros por oraciones.

2. Se ignoró el texto que no hace parte del contenido del libro (Metadata y licencias).
3. Se ignoraron las oraciones con menos de 4 palabras.
4. Se crearon secuencias de texto de mínimo 15 palabras, por lo tanto aquellas oraciones más pequeñas se concatenaron con otras hasta cumplir esta condición.

3.2 Detalles del dataset

El dataset se compone de secuencias de texto extraídas de los siguientes libros, ordenados según su autor:

- **Charles Dickens:**
  - David Copperfield
  - Oliver Twist
  - The Pickwick Papers
- **George Eliot:**
  - Middlemarch
  - Silas Marner
  - The Mill on the Floss
- **Jane Austen:**
  - Mansfield Park
  - Northanger Abbey
  - Pride and Prejudice
- **Lewis Carroll:**
  - Alice's Adventures in Wonderland
  - Phantasmagoria and Other Poems
  - Sylvie and Bruno
  - The Hunting of the Snark: An Agony in Eight Fits
  - Through the Looking-Glass
- **William Shakespeare:**
  - Hamlet, Prince of Denmark
  - Macbeth
  - Romeo and Juliet
  - A Midsummer Night's Dream
  - The Tempest

El **dataset de entrenamiento** contiene 60,484 secuencias de texto, distribuidas de la siguiente manera:

Autor	**Número de secuencias**
Charles Dickens	26,096
George Eliot	16,263
Jane Austen	9,585
Lewis Carroll	4,448
William Shakespeare	4,092

Table 1: Distribución de secuencias en el dataset de entrenamiento

El **dataset de evaluación** contiene 5,115 secuencias de texto, distribuidas de la siguiente manera:

Autor	**Número de secuencias**
Charles Dickens	1,023
George Eliot	1,023
Jane Austen	1,023
Lewis Carroll	1,023
William Shakespeare	1,023

Table 2: Distribución de secuencias en el dataset de evaluación

4 Rúbrica de la nota del proyecto de la semana 7

La nota del proyecto de la semana 7 se compone de la forma descrita en la siguiente tabla:

Item	**Porcentaje de la nota**
F1 score mínimo de 0.5	60%
Percentil obtenido en la competencia	40%

Table 3: Composición de la nota del proyecto de la semana 7

4.1 F1 score mínimo de 0.5

Para obtener el 60% de la nota usted y su grupo deberán crear un modelo de clasificación que alcance un desempeño mínimo de 0.5 en el *f1-score macro* de la competencia de Kaggle. Podrá ver el desempeño de su modelo en la tabla de líderes de Kaggle. Sin embargo, se debe tener en cuenta que el desempeño observado podrá variar frente al obtenido al cerrar la competencia, debido a que se reserva un subset del set de evaluación para la evaluación final.

4.2 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 4: Distribución del porcentaje del rubro según desempeño obtenido

4.3 Cumplimiento de requisitos de la actividad

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:

- Para el entrenamiento del modelo solo se deben usar los datos proporcionados, no podrá usar datos de otras fuentes. Pero, puede hacer uso de embeddings (incrustaciones) preentrenados.
- Solo podrá usar modelos desarrollados con la librería **scikit-learn**.

- 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.

5 Instrucciones para la creación del clasificador

Las siguientes instrucciones le servirán de guía para el desarrollo satisfactorio de esta actividad:

1. Cargue los datos de entrenamiento recibidos en formato csv. Revise los nombres de las columnas y entienda la estructura general de estos datos.
2. Reserve una determinada cantidad de los datos como un set de validación que le permita verificar el desempeño del modelo.
3. Genere una representación de texto de las secuencias que le permita entrenar un modelo de clasificación.
4. Realice el entrenamiento del modelo de clasificación.
5. Evalúe el desempeño del modelo desarrollado.
6. Guarde el modelo. Use *pickle* o *joblib* para hacerlo.

   ```
   import joblib
   joblib.dump(best_model, "model.pkl")
   ```

7. Genere un archivo csv con las etiquetas para las secuencias del set de evaluación según las indicaciones de la competencia de Kaggle.
8. Realice la entrega de sus resultados en la competencia de Kaggle, y revise el desempeño obtenido y su posición en el tablero de líderes.
9. Pruebe otros modelos, técnicas de procesamiento, y representaciones de texto para intentar mejorar el desempeño obtenido y ascender en el tablero de líderes.

Nota importante: El modelo de clasificación creado debe ser alguno de los vistos en el curso. En caso de usar otros, estos deben ser modelos de machine learning clásico, es decir que no podrá hacer uso de arquitecturas profundas como las LSTMs, o los Transformers. Adicionalmente, usted y su grupo deben hacer uso exclusivo de las siguientes librerías para la creación del modelo:

- NLTK
- TextBlob
- Stanza
- spaCy
- Scikit-learn (sklearn)

6 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. El archivo donde se guardó el modelo de *scikit-learn* entrenado. Recuerde que el modelo debe guardarse haciendo uso de las librerías *pickle* o *joblib*.