Instituto Tecnológico de Costa Rica, ITCR
Programa de Ciencia de Datos
Aprendizaje Automático
Quiz 03

Profesora: María Auxiliadora Mora

Tema: Procesamiento de lenguaje natural (NLP) con diversos algoritmos.

Entrega: Un archivo .zip que contenga un documento en formato Jupyter notebook bien documentado que incluya los ejercicios. A través del TEC-digital.

Modo de trabajo: Individual.

Introducción:

En esta tarea se aplicarán conceptos básicos de procesamiento de lenguaje natural (NLP) con aprendizaje automático para resolver problemas que involucran clasificación de textos.

El **objetivo del trabajo** es poner en práctica las **habilidades de investigación y el conocimiento adquirido durante el curso** sobre aprendizaje automático por medio de ejercicios prácticos que permitan a las y los estudiantes experimentar con el flujo de trabajo de un proyecto de NLP.

Objetivos de aprendizaje:

1. Fortalecer en las y los estudiantes las habilidades de investigación y documentación de resultados asociados a proyectos de ciencia de datos.
2. Experimentar con el flujo completo de trabajo requerido en proyectos de aprendizaje automático para realizar clasificación de textos.
3. Fortalecer capacidades en los estudiantes en el uso de bibliotecas de aprendizaje automático.

Ejercicios

Análisis de sentimientos utilizando aprendizaje automático.

En este ejercicio, para realizar el análisis de sentimientos se utilizará un conjunto de datos seleccionado por las personas estudiantes.

Realice las siguientes actividades:

1. Seleccione un conjunto de datos para realizar la clasificación de textos (puntos extra si el conjunto de datos está escrito en español).
2. Describa el problema, el objetivo del ejercicio y los datos a utilizar.
3. Preprocese el conjunto de datos, es decir:
    1. Verifique si existen registros con valores faltantes y de ser así elimínelos.

    2. Utilice expresiones regulares para eliminar los caracteres especiales.
    3. Elimine las "stop words".
    4. Convierta el texto del campo a clasificar a minúsculas.
4. Explore y visualice algunas estadísticas con gráficos de barras o pastel. Por ejemplo, cuente cuántos registros hay en cada clase y haga un histograma con el largo de los textos.
5. Utilizando **redes LSTM con la biblioteca de PyTorch** procese el conjunto de datos para clasificar los textos.
6. Defina los hiper-parámetros del proceso de entrenamiento, por ejemplo, la función de perdida, el optimizador, entre otros.
7. Entrene el modelo.
8. Grafique la curva de error, explique los resultados obtenidos y ajuste el modelo o el proceso de entrenamiento apropiadamente.
9. Evalúe el modelo resultante utilizando una matriz de confusión y métricas extraídas a partir de esta (ie. exactitud, precisión, exhaustividad y F1). Despliegue de forma gráfica la matriz de confusión para el cálculo de las métricas y explique los resultados obtenidos.
10. Analice los resultados, proponga mejoras y explique los cambios realizados al flujo de trabajo del proyecto para mejorar el rendimiento del modelo (aplique al menos dos cambios que efectivamente mejoren el rendimiento).
11. Genere y documente sus conclusiones (incluya al menos cuatro conclusiones importantes).
12. Todas las secciones del ejercicio deben estar bien documentadas (con encabezado en las funciones que describen qué hace cada una y descripción de los parámetros, además, porciones internas del código deben estar documentadas también).
13. Incluya una sección de referencias en formato APA al final del documento que incluya una referencia al conjunto de datos.

Rúbrica

Rubro	**Puntos**
Clasificación de textos	
Se seleccionó un conjunto de datos a clasificar en español	5% extra
Se cargaron y prepararon los datos para ser introducidos al modelo	5
Explore y visualice algunas estadísticas con gráficos de barras o pastel: cuente cuántos registros hay en cada clase y haga un histograma con el largo de los textos.	2
Aprendizaje profundo (DL): Se definieron los hiper-parámetros de entrenamiento, por ejemplo, función de perdida, el optimizador.	1
DL- se entrenó el modelo	2
DL- Se graficó la función de error con datos entrenamiento y prueba de todas las épocas (como vimos en clase).	3
Se evaluó el modelo resultante utilizando una matriz de confusión y métricas extraídas a partir de esta (ie. exactitud, precisión, exhaustividad y F1) y se despliegó de forma gráfica la matriz de confusión para el cálculo de las métricas.	3

Se aplicaron mejoras al flujo de trabajo del proyecto que tuvieron impacto positivo en el rendimiento del modelo.	3
Se generó y documentó todas las conclusiones (al menos 4 conclusiones interesantes)	2
Documentación de ambos ejercicios	
Se describe el problema y el objetivo del ejercicio.	2
Se describen los datos utilizados en el ejercicio.	2
Todas las secciones del código están debidamente documentadas (con encabezado en las funciones que describen qué hace cada una y descripción de los parámetros, además, porciones internas del código están documentadas también).	4
Se incluyen referencias en formato APA.	1