Instituto Tecnológico de Costa Rica
Escuela de Ingeniería en Computación, ITCR
Aprendizaje automático

Tarea programada 04 NLP y clustering

Entrega: A través del TEC-digital. Un archivo .zip que contenga los datos o referencia a estos, un documento en formato Jupyter notebook con todas las secciones solicitadas y código bien documentado.

Modo de trabajo: Grupos de 2 personas máximo.

Tecnología a utilizar: Python, PyTorch, Scikit-learn

Introducción:

En este trabajo se aplicarán conceptos básicos de análisis de agrupamientos o clustering para caracterizar un fenómeno de elección de las personas estudiantes utilizando el conjunto de datos de su preferencia y se reforzará la experiencia en el uso de algoritmos de aprendizaje automático para la clasificación de textos.

El **objetivo de los ejercicios** es poner en práctica las **habilidades de investigación y el conocimiento adquirido durante el curso** por medio de ejercicios prácticos que permitan a las y los estudiantes experimentar con algoritmos de aprendizaje supervisado y no supervisado.

Objetivos de aprendizaje:

1. Desarrollar habilidades de investigación y documentación de resultados.
2. Desarrollar un proyecto utilizando algoritmos de clustering aplicados a problemas de interés de las personas estudiantes.
3. Explorar el uso de algoritmos de clasificación de textos y evaluar los resultados.
4. Fortalecer capacidades en los estudiantes en el uso de bibliotecas de aprendizaje automático.

Las personas estudiantes deberán seleccionar uno de los siguientes ejercicios:

Ejercicio A) Análisis de agrupamientos (Clustering)
Aplique **al menos tres algoritmos de custering al conjunto de datos seleccionado.**

Objetivo: El objetivo del presente ejercicio es utilizar datos de Costa Rica para demostrar cuán efectivos y precisos pueden ser los algoritmos de *clustering* en la caracterización de un fenómeno particular.

1. Obtenga el conjunto de datos.
2. Preprocese, limpie y visualice los datos para conocerlos utilizando los métodos vistos en clase u otros.

3. Seleccione las variables a utilizar en el ejercicio (**al menos cinco variables**). Documente el motivo de la selección de acuerdo al problema en estudio.
4. **Utilice los tres algoritmos de clustering** para caracterizar los datos usando las variables seleccionadas.
5. Utilice métodos de ajuste de parámetros, por ejemplo, el **método del codo** para seleccionar el mejor K para el algoritmo K-Means o el Coeficiente de Silueta y vuelva a ejecutar el algoritmo usando los parámetros recomendados. Realice esto con todos los algoritmos.
6. Evalúe los modelos de clustering resultantes utilizando el Coeficiente de Silueta.
7. **Documente y compare los resultados de los algoritmos y genere conclusiones** (incluya al menos cuatro conclusiones importantes).
8. Incluya referencias bibliográficas en formato APA.

Ejercicio B) Análisis de sentimientos utilizando aprendizaje automático.

En este ejercicio, para realizar el análisis de sentimientos se utilizará el conjunto de datos seleccionado por las personas estudiantes para realizar el Q03. La tarea consistirá en aplicar diferentes algoritmos al mismo conjunto de datos y comparar los resultados.

Realice las siguientes actividades:

1. Describa el problema, el objetivo del ejercicio y los datos a utilizar.
2. El conjunto da datos ya fue pre-procesado para realizar la tarea anterior.
3. Explore y visualice algunas estadísticas con gráficos de barras o pastel. Por ejemplo, cuente cuántos registros hay en cada clase.
4. Extraiga características utilizando el método TF-IDF.
5. Utilizando **al menos tres algoritmos - uno de ellos de aprendizaje profundo con un perceptrón multicapa (MLP) -** procese el conjunto de datos para clasificar los textos.
6. Defina los hiper-parámetros del proceso de entrenamiento, por ejemplo, en caso del aprendizaje profundo, función de perdida, el optimizador, entre otros.
7. Entrene los modelos.
8. En caso del MLP, grafique la curva de error, explique los resultados obtenidos y ajuste el modelo o el proceso de entrenamiento apropiadamente.
9. Evalúe los modelos resultantes 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. Genere y documente conclusiones sobre los resultados obtenidos (incluya al menos cuatro conclusiones importantes).
11. 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).

12. Incluya una sección de referencias en formato APA al final del documento que incluya una referencia al conjunto de datos.

Rúbrica

Clustering	**Puntos**
Preprocese y visualice los datos para conocerlos utilizando los métodos vistos en clase u otros.	1
Seleccione las variables a utilizar en el ejercicio (al menos cinco variables). Documente el motivo de la selección de acuerdo al problema en estudio.	1
Utilice los tres algoritmos de clustering para caracterizar los datos usando las variables seleccionadas.	6
Utilice métodos de ajuste de parámetros con todos los algoritmos.	6
Evalúe los modelos de clustering resultantes utilizando el Coeficiente de Silueta.	3
Documente y compare los resultados de los algoritmos y genere conclusiones (incluya al menos cuatro conclusiones importantes).	2

Clasificación de textos	**Puntos**
Se extrajeron características utilizando el método TF-IDF y Word embedding en caso de DL.	2
Se entrenaron modelos con 2 algoritmos (sin incluir el MLP)	2
MLP- Se definieron los hiper-parámetros de entrenamiento, por ejemplo, función de perdida, el optimizador.	1
MLP- se entrenó el modelo	1
MLP- Se graficó la función de error con datos entrenamiento y prueba de todas las épocas (como vimos en clase).	3
Se evaluaron los modelos resultantes utilizando una matriz de confusión y métricas extraídas a partir de esta (ie. exactitud, precisión, exhaustividad y F1) y se desplegó 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 de los modelos.	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
Nota extra: Se aplica el modelo de transformers y se evalúan los resultados	10%