PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ
FACULTAD DE CIENCIAS E INGENIERÍA

INF265 – APLICACIONES DE CIENCIAS DE LA COMPUTACIÓN (2018-1)
TAREA ACADÉMICA – SEGUNDA PARTE

Objetivo general:

El objetivo de la tarea académica es poner en práctica los conceptos y técnicas involucrados en el Aprendizaje Automático (Aprendizaje Supervisado o Clasificación) y el Procesamiento de Lenguaje Natural.

Parte A: Clasificación de Textos

A partir de conjuntos de documentos textuales anotados para clasificación, se va a contrastar dos enfoques de representación vectorial de documentos: (1) Vectores en base a frecuencias de apariciones de términos en el conjunto de documentos (*Bag of Words*, *Term Frequency*, TF-IDF); y (2) Vectores de documentos compuestos por vectores de palabras con información semántica (*Word Embeddings*)

Primer paso: (**Laboratorio 8**)
Limpieza y procesamiento de textos para la posterior generación de vectores

Segundo paso: (**Laboratorio 9**)
Generación de vectores, clasificación y comparación de resultados

¿Qué tipo de conjuntos de datos (datasets) se puede usar?
Aquellos cuyo contenido principal sea una descripción textual (por ejemplo, una opinión de un comprador sobre un producto electrónico), y que además, se disponga una atributo objetivo o etiqueta anotada sobre la cual se pueda clasificar el comentario (siguiendo el ejemplo anterior: el comentario debe tener una etiqueta que permita clasificar al comentario como positivo, negativo o neutro, si es que se hace un análisis de polaridad).

Lista de *datasets* sugeridos (pero pueden optar por otros que cumplan lo anteriormente indicado)**:**

- Ver punto 1 sobre "Text Classification": https://machinelearningmastery.com/datasets-natural-language-processing/
- Ver enlace de Google Drive al final del post: https://medium.com/@surmenok/large-text-classification-datasets-765cc40fece7
- Conjuntos textuales del repositorio UCI (verificar si tienen la anotación adecuada para clasificación):

https://archive.ics.uci.edu/ml/datasets.html?area=&att=&format=&numAtt=&numIns=&sort=nameUp&task=&type=text&view=table

- Conjuntos para NLP en general (verificar que esté anotado para seleccionar uno): https://github.com/niderhoff/nlp-datasets

Parte B: Generación automática de textos

A partir de un conjunto de documentos textuales, se entrenará un modelo (red neuronal) para que aprenda a generar nuevos textos en base al dominio original de los documentos seleccionados. Por ejemplo:

- Si se usa un conjunto de novelas, se espera que el modelo pueda generar un pasaje "coherente" en cierta medida con el contexto de la novela (personajes, entornos, diálogos, etc.).
- Si se usa un conjunto de publicaciones de redes sociales de un usuario o un grupo de usuarios, se espera que el modelo pueda generar un nuevo mensaje/publicación emulando al usuario o grupo de usuarios sobre el cuál aprendió.

Sugerencias para el conjunto de datos a usar:

- Mientras más largo sea el contenido textual, mejor

Entregable del Laboratorio 10

Fragmentos preliminares de nuevos textos generados automáticamente, y un breve análisis descriptivo de qué elementos del dominio textual se pudieron capturar y cuáles no. Además de comentar qué tan coherente es el resultado del texto generado.

Estos fragmentos deben ser mejorados para el entregable final del reporte, ya que dispondrán de mayor tiempo para que el modelo entrene y pueda mejorar su "comprensión" del dominio, es decir, de la secuencia de caracteres y términos

Entregable Final de la TA – Parte 2

Aparte de las notas de los Laboratorio 8, 9 y 10, se debe entregar un breve reporte (no más de 4 páginas) describiendo los 2 experimentos y discutiendo sus resultados.

- Para la Parte A, el énfasis debe realizarse sobre el resultado de comparación de clasificación usando los 2 paradigmas de vectores diferentes.
- Para la Parte B, se deben concentrar en discutir hasta qué nivel el texto generado tiene coherencia con el dominio sobre el cual se ha generado. Será importante que anexen ejemplos de fragmentos generados automáticamente, con diferentes tiempos de entrenamiento.

Fecha de entrega de reporte: 1 semana antes del día del Examen 4