TP4 - Machine Learning II

Modalidad del TP

Deberán elegir una tarea de las presentadas para alguno de los datasets y desarrollarla incluyendo:

- Un baseline
- Un dataset de validation/test para todos los experimentos
- Al Menos una implementación de alguna de las siguientes:
    - Redes neuronales
    - Reducción de dimensiones
    - Clustering

Por sobre eso se valora el uso de cualquier técnica de machine learning novedosa y potencialmente útil.

Formato de entrega

El formato de entrega es un video de **no más de 5 minutos (sin excepción)** explicando que se intentó, por qué y resultados. Junto con eso se deberá entregar el código y todo el desarrollo correspondiente.

Criterio de evaluación

La nota es a criterio del equipo docente, sobre una base de 25 en la que suman los puntos extra que les sobraran del TP3. Se aprueba con 60% (15 puntos). Cualquier nota por debajo de 60% va a reentrega a implementar algo sugerido por la cátedra para conseguir lo que le falte para aprobar. La única forma de reprobar de forma directa es no entregar un video explicando por lo menos el baseline utilizado.
Para el TP que tanto éxito tengan en la tarea es secundario, lo más importante es que puedan probar cosas y que ya sea el caso de que tengan éxito o no puedan explicarlas. Lo que es importante ya sea para las implementaciones que funcionen o no funcionen es que estén bien hechas, sean prolijas, correctas, que no fallen por errores de código.

Correctores y consultas

Este TP no tiene correctores individuales asignados, todas las consultas, sean o no de código corresponden al canal #consultas-tp4 y los motivamos a ayudarse entre ustedes.
Lo más normal es que no tengan idea de qué hacer en su tarea y es la idea que prueben cosas algo distintas. Después de investigar un poco que se puede usar, pueden preguntar en el canal de consultas para que les demos sugerencias sobre qué probar o que es una buena idea y cuál no.

Datasets para usar

Deep web ítems

Tenemos dos datasets sobre ítems de la deep web, uno para ítems entre 2013-2015 y otro del 2021.
Pueden encontrar docu y links de cada dataset armados por nosotros aca:

- Silkroad2
- Versus

Tarea 1: Clasificación del ítem en base al título y/o descripción

La tarea de NLP consiste en construir un clasificador que según el título y descripción obtenga la categoría del ítem. ¿Qué tan bien funciona en validación? ¿Qué tan bien funciona en el dataset de versus?

Tarea 2: Clasificación de imágenes de silkroad2

La tarea consiste en construir un clasificador que dada una imagen devuelva la categoría a la que corresponde. ¿Qué tan bien funciona la validación?

Tarea 3: Regresión sobre el precio del ítem

¿Se puede predecir el precio del ítem en base a su información? ¿Qué tan bien funciona en validación? ¿Qué tan bien funciona años después en versus?

Tarea 4: Predicción del score de la review

¿Qué tan difícil es predecir el score de la review? ¿Cuál es el modelo más sencillo que puede construirse con resultados aceptables? ¿Cuál es la menor cantidad de datos que puede usarse para tener datos aceptables?

Tarea 5: Predicción del score de la review entrenando sobre otros datos

¿Qué tan bien funcionan modelos entrenados en otros datos para esta tarea?

Tarea 6: Regresión sobre el precio de los shippings

¿Cuál es el modelo más simple que mejor funciona en validación? ¿Cuál es la calibración de los modelos y qué modelos consiguen la mejor? ¿Qué features utilizan?

Tarea 7: Extracción de tópicos no supervisada sobre las descripciones de los ítems

¿Qué tan fácil es asignarle tópicos a los ítems en base a sus descripciones? ¿Qué tan bien funciona? (y cómo medirlo?)

Tarea 8: Generar imágenes falsas en base a títulos/descripciones/categorías

¿Qué tan fácil es generar imágenes falsas que correspondan a un texto? ¿Cómo medir su performance?

Microsoft News Dataset

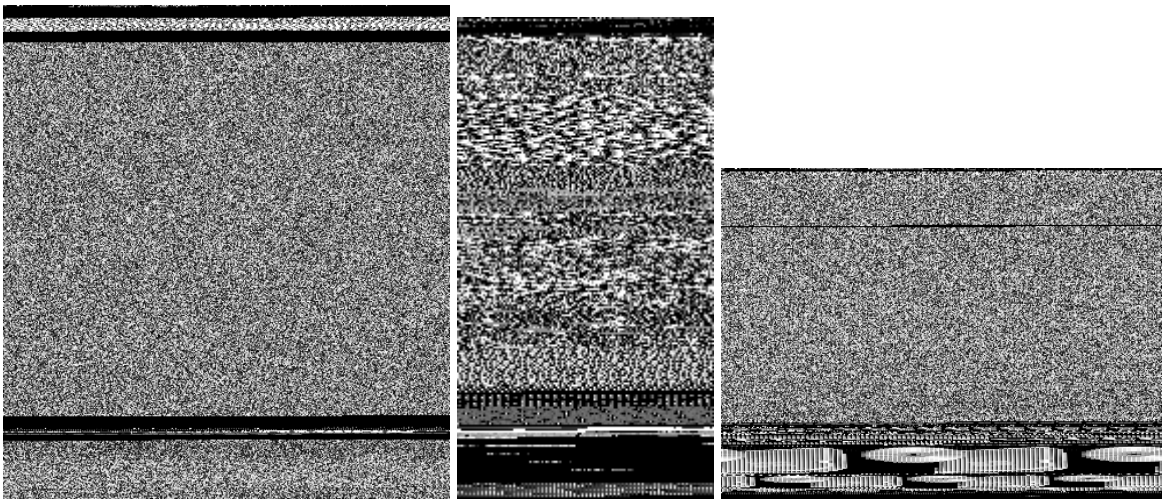
Es un dataset de microsoft sobre sistemas de recomendación que también incorpora información del contenido. Pueden ver la descripción aca.
El objetivo es predecir las impresiones que los usuarios van a hacer a las noticias, es un dataset raro en sistemas de recomendación ya que previo a este dataset había muy poca investigación sobre sistemas de recomendación en noticias.
El dataset puede bajarse acá.

Tarea: Sistema de recomendaciones

¿Qué tan bien funciona un baseline? ¿Qué tan fácil es lograr una buena métrica en validación y la competencia?

Malimg Malware Dataset

El dataset consiste de +9000 imágenes obtenidas de binarios de virus, clasificadas por tipo de virus.



Tarea: Clasificación NO supervisada

¿Qué tan bien funciona un clasificador no supervisado?

GTZAN Genre Collection

El dataset consta de 1000 audios de 30 segundos en formato .wav de 10 géneros musicales distintos.

Tarea: Construir un clasificador

¿Qué tan bien funciona en validación (10% random de cada género)? ¿Qué features usar?

SemEval 2022

La SemEval es una competencia anual de NLP orientada a la academia que intenta resolver problemas muy difíciles (que en general pero no siempre están alejados de la industria, pero cuyas soluciones ayudan como extensión a otros problemas similares de la industria). Las competencias siguen activas así que es una buena oportunidad para participar. De las 12 tasks disponibles elegimos las siguientes para que puedan encarar en el TP que son más sencillas. La tarea en todos los casos es *empezar* a participar de la competencia

Task 4: Patronizing and Condescending Language Detection

Detección de lenguaje condescendiente y paternalista, de forma binaria y por categorías.

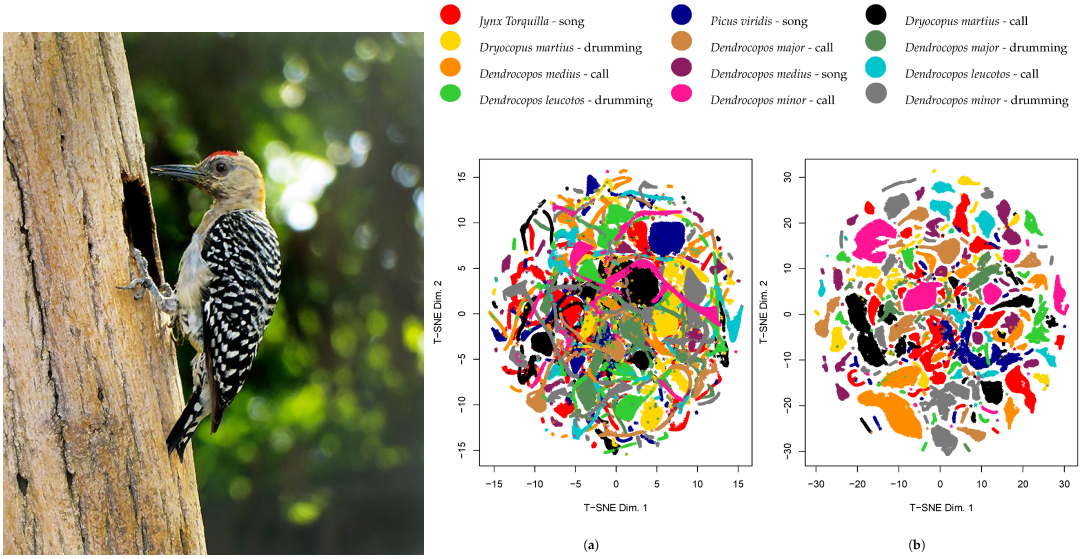
Task 6: iSarcasmEval - Intended Sarcasm Detection in English and Arabic

Detección de sarcasmo, queremos saber qué tan fácil es su detección. Recomendamos trabajarlo solo en inglés para el TP.

Task 7: Identifying Plausible Clarifications of Implicit and Underspecified Phrases in Instructional Texts

Esta es muy divertida, dado un manual de instrucciones para realizar una tarea sacado de wikihow, hay que recomendar palabras con algún score para ciertas palabras que faltan en los textos.

Canto de pájaros carpinteros



El dataset corresponde al siguiente paper: https://www.mdpi.com/2306-5729/2/2/18/htm
Y puede encontrarse para descargar aca: https://zenodo.org/record/574438

Tarea: Clasificación

Son 1669 audios, la tarea es clasificar lo mejor posible los audios o encontrar una reducción no supervisada que sea buena (como la de la imagen del paper, se puede intentar reproducir esa). Es un dataset poco conocido, así que estoy seguro que cualquier cosa que logren o duda que tengan podemos compartirla por diversión con sus autores, seguro se alegran de ver que se usa, siempre esas interacciones son divertidas :)