EMGVD (EYP3707)
Profesor: Pedro Luiz Ramos
Ayudante: Diego Bernal

Ayudantía 8
Primer Semestre - 2025

Enunciado I

En el año 1950 Alan Touring publica el paper *"COMPUTING MACHINERY AND INTELLIGENCE"* en el cual postula la idea del juego de la imitación, lo que hoy en día conocemos como test de touring, el cual en simples palabras plantea la conjetura de si un humnao es capaz de dicernir entre un texto generado por una maquin y otro generado por una persona.

Tomando esta conjetura la ayudantía de hoy estara enfocada en trabajar con textos en especifico poemas algunos hechos por grandes poetas como *Pablo Neruda, Gabriela Mistral* y otros generados de forma artificial con IA.

Con ello plantearemos la siguiente variable binaria la cual fue realizada de forma supervisada.

$$Y_p: \begin{cases} 1 & \text{si el poema } p \text{ fue generado por una IA.} \\ 0 & \text{e.o.c} \end{cases}$$

Con esa conjetura trabajaremos con el vectorizador TFIDF para las palabras contenidas en un poema, este algoritmo realiza lo siguiente:

$$\text{TF}(t,d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}}$$

Donde:

1. $f_{t,d}$: es el número de veces que aparece el término t en el documento d.
2. $\sum_{t' \in d} f_{t',d}$ es el total de términos en el documento d.

La frecuencia inversa es:

$$\text{IDF}(t,D) = \log\left(\frac{N}{|\{d \in D : t \in d\}|}\right)$$

Donde:

1. N: es el número total de documentos en el corpus D.
2. $\{d \in D : t \in d\}$ es el número de documentos en los que aparece el término t.

$$\text{TFIDF}(t,d,D) = \text{TF}(t,d) \times \text{IDF}(t,D)$$

1. Primero cargue los datos y realice un preprocesamiento para generar la variabe objetivo.
2. Segundo reaalice una función que logre tokenizar cada poema quitando puntos, espacios y stop-words.
3. Tercero separe en train y test el set de datos.
4. Transforme los tokens en vectores con tfidf.
5. Aplique un modelo de SVM linear y vea como clasifica, plotee la matriz de confusión.
6. Utilice el algoritmo de gridsearchcv para encontrar el mejor modelo en base al accuracy.
7. Pruebe con nuevos textos como se clasifican, comente.