

Instituto Politécnico Nacional

ESCOM

PLN

Islas Osorio Enrique

7CM3

Practice IV - Text classification

**Introducción.**

En esta práctica para clasificar texto primero se carga un corpus de texto en este caso fetch\_20newsgroups, luego se divide en conjuntos de entrenamiento y prueba. El texto se normaliza para eliminar el ruido, y luego se crea una representación del texto. Finalmente, se utiliza un método de aprendizaje automático para entrenar un modelo y predecir las instancias de prueba.

En la practica vamos a normalizar los textos con:

Tokenizacion: Es el proceso de dividir un texto en unidades significativas, llamadas tokens. Los tokens pueden ser palabras, números, símbolos o cualquier otra unidad de texto que tenga un significado específico para el contexto en el que se utiliza.

Limpieza de texto: Es el proceso de eliminar del texto elementos no deseados, como errores tipográficos, caracteres especiales, símbolos, stop words, etc. La limpieza de texto es importante para mejorar la calidad de los datos y la precisión de las tareas de procesamiento del lenguaje natural (NLP).

Stop words: Son palabras que aparecen con mucha frecuencia en un texto, pero que no tienen mucho significado para el contexto en el que se utilizan. Ejemplos de stop words en español son: "el", "la", "los", "las", "de", "a", "para", "en", "y", "que", etc. Las stop words suelen eliminarse del texto antes de realizar tareas de NLP, ya que pueden interferir con el rendimiento de los modelos.

Lematización: Es el proceso de reducir una palabra a su forma de base, o lexema. El lexema es la forma más básica de una palabra, sin sus afijos derivativos o flexivos. La lematización es importante para mejorar la precisión de las tareas de NLP, ya que permite agrupar las palabras que tienen el mismo significado, aunque estén flexionadas de forma diferente.

En la practica se hicieron pruebas sin normalizar, tokenizacion + stopwords, tokenizacion + lematización, tokenizacion + stopwords + lematización y , tokenizacion + stopwords + lematización + cleantext

También haremos uso de diferentes representaciones de texto.

Binarización: Representación de texto en la que cada palabra se representa como un número binario que indica si la palabra aparece en el texto o no.

Frecuencia: Representación de texto en la que cada palabra se representa como un número entero que indica la frecuencia con la que la palabra aparece en el texto.

TF-IDF: Representación de texto que combina la frecuencia y la importancia de las palabras.

Como vimos en la practica anterior, TF-IDF es el que mejor resultado dio

Finalmente, varios algoritmos de Machine Learning.

Logistic regression es un algoritmo de aprendizaje supervisado que se utiliza para problemas de clasificación binaria. Es un modelo lineal que predice la probabilidad de que una muestra pertenezca a una clase determinada.

Multinomial Naive Bayes es un algoritmo de aprendizaje supervisado que se utiliza para problemas de clasificación multiclase. Es un modelo probabilístico que asume que las variables predictoras son independientes entre sí.

K-nearest neighbors (KNN) es un algoritmo de aprendizaje no supervisado que se utiliza para problemas de clasificación, regresión y agrupamiento. KNN funciona comparando una nueva muestra con las muestras más cercanas en el conjunto de entrenamiento.

Random forest es un algoritmo de aprendizaje supervisado que combina múltiples árboles de decisión para mejorar la precisión. Random forest funciona entrenando un conjunto de árboles de decisión en subconjuntos aleatorios del conjunto de entrenamiento.

Multilayer perceptron (MLP) es un algoritmo de aprendizaje supervisado que se utiliza para problemas de clasificación y regresión. MLP es un tipo de red neuronal artificial que consta de una o más capas ocultas de neuronas.

**RESULTADOS**

Este es mi intento de árbol como el que vimos en clase, primero probe sin normalizar y tokenizacion + stopwords con los 3 tipos de representaciones, como el binario era el peor, en los siguientes no lo tome en cuenta; lo siguiente que probe fue tokenizacion + lematizar y tokenizacion + stopwords + lematizar como KNN es el algoritmo de ML que peor resultados daba no lo tome en cuenta para probar con una normalización completa tokenizacion + stopwords + lematizar + clean text, no me dio buen resultado por lo que me quede con tokenizacion + stopwords + lematizar que fue lo que mejor resultado me dio.

Con tokenizacion + stopwords + lematizar probe los algoritmos de ML que más se tardaba que eran MLP y SVC estos los probe con diferentes parámetros y el que mejor resultado dio fue con MLP.

Finalmente use SVD con tokenizacion + stopwords + lematizar, TFIDF y MLP,SVC con los mejores parámetros

Imagen en blanco y negro

Descripción generada automáticamente con confianza baja

Imagen que contiene Calendario

Descripción generada automáticamente

Imagen que contiene Calendario

Descripción generada automáticamente

Calendario

Descripción generada automáticamente con confianza baja

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente con confianza media

**Conclusión**

Al realizar varias combinaciones, la mejor combinación fue tokenización + stopwords + lematizar, representación TF-IDF y MLP con estos parámetros Texto

Descripción generada automáticamente

Nos dio un valor de 0.85, que fue el más alto. Como vimos en la práctica anterior, TF-IDF es una muy buena representación de texto, ya que le da valor a las palabras más importantes. Solo lematizar el texto no daba una gran mejora en comparación con solo quitar stopwords, pero al combinarlos dio un muy buen resultado, ya que se elimina el ruido. Cuando lo junté con cleantext, dio un peor resultado. Esto pudo ser porque tal vez quité cosas que eran importantes y no lo sabía. Finalmente, el algoritmo de MLP fue el que mejor resultado dio, porque es un buen algoritmo de aprendizaje automático.

Estas son las capturas de los resultados para que no se me perdieran

Imagen que contiene Texto

Descripción generada automáticamente

Este primero fue sin normalizar ni nada, se uso representación de texto TF-IDF y Logistic regression

Igual sin nada con TF-IDF

Imagen que contiene Calendario

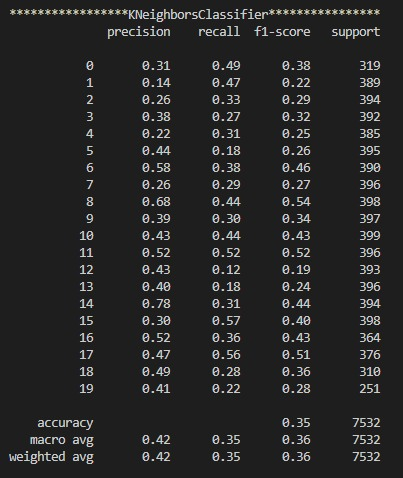
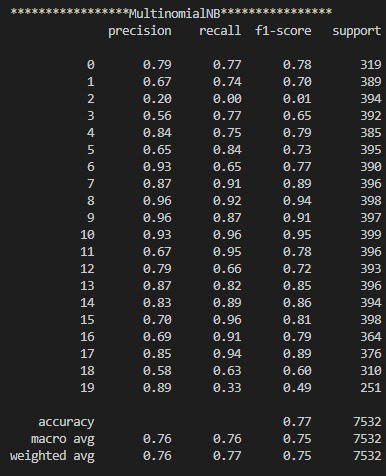
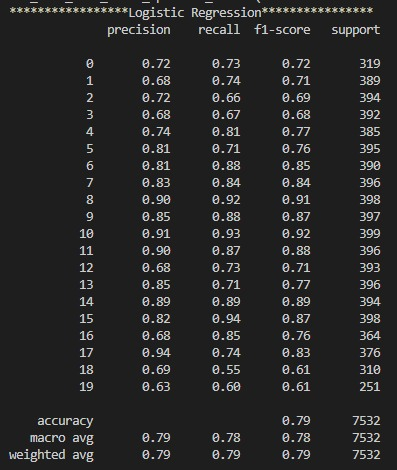
Descripción generada automáticamente

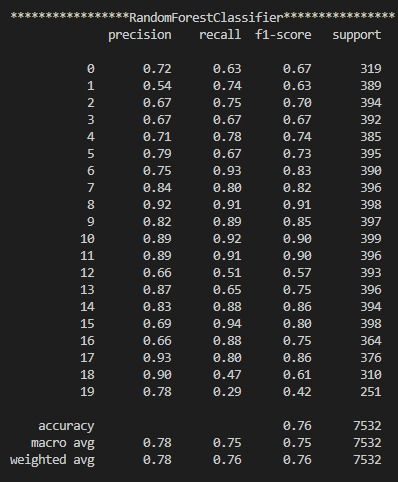
Imagen que contiene Texto

Descripción generada automáticamenteImagen que contiene Tabla

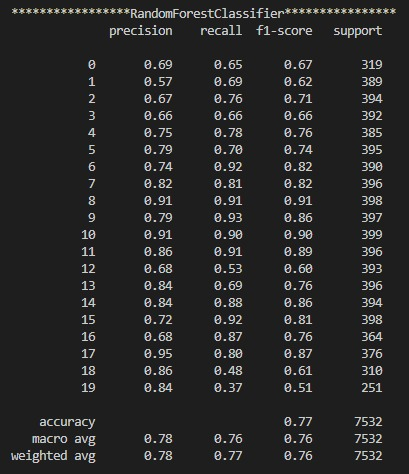
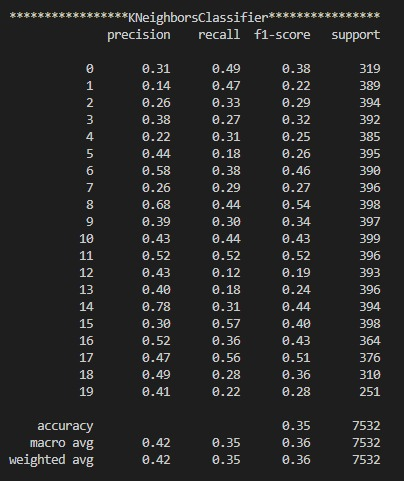
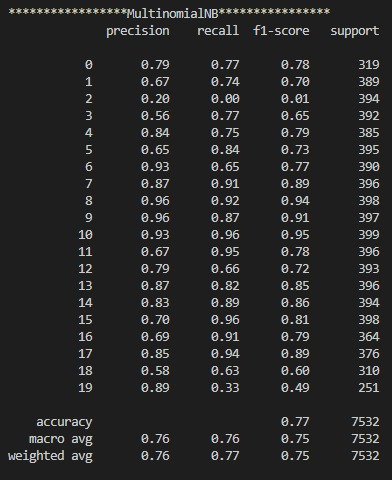
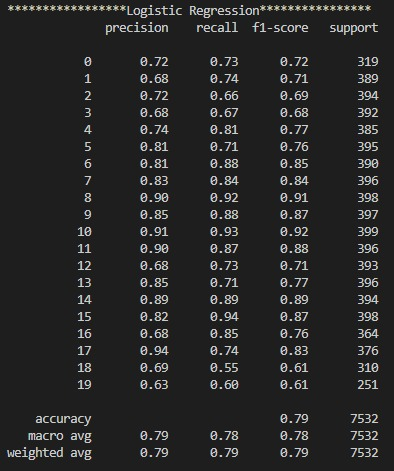
Descripción generada automáticamente

Frecuencia





BINARIO



Sin stopwords y TFIDF

Imagen que contiene Tabla

Descripción generada automáticamente

Imagen que contiene Tabla

Descripción generada automáticamenteImagen que contiene Tabla

Descripción generada automáticamenteImagen que contiene Tabla

Descripción generada automáticamente

Stopwrodsenglish and TFIDF

Tabla

Descripción generada automáticamente con confianza mediaImagen que contiene Tabla

Descripción generada automáticamenteTabla

Descripción generada automáticamente con confianza bajaImagen que contiene Tabla

Descripción generada automáticamente

STOPWORDS y BINARIO

Imagen que contiene Tabla

Descripción generada automáticamenteImagen que contiene Tabla

Descripción generada automáticamenteImagen que contiene Calendario

Descripción generada automáticamente

Imagen que contiene Texto

Descripción generada automáticamente

FRECUENCUA Y STOPWORS

Imagen que contiene Texto

Descripción generada automáticamenteImagen que contiene Tabla

Descripción generada automáticamenteImagen que contiene Tabla

Descripción generada automáticamenteImagen que contiene Calendario

Descripción generada automáticamente

LEMATIZADO FRECUENCIA

Imagen que contiene Tabla

Descripción generada automáticamenteImagen que contiene Tabla

Descripción generada automáticamenteImagen que contiene Tabla

Descripción generada automáticamenteImagen que contiene Calendario

Descripción generada automáticamente

LEMATIZAR TFIDF

Imagen que contiene Tabla

Descripción generada automáticamenteImagen que contiene Tabla

Descripción generada automáticamenteImagen que contiene Texto

Descripción generada automáticamenteImagen que contiene Tabla

Descripción generada automáticamente

LEM STOP TFDFIDF

Imagen que contiene Tabla

Descripción generada automáticamenteImagen que contiene Texto

Descripción generada automáticamenteImagen que contiene Calendario

Descripción generada automáticamenteImagen que contiene Tabla

Descripción generada automáticamente

LEM STOPWORDS FRECUENCIA

Imagen que contiene Tabla

Descripción generada automáticamenteImagen que contiene Tabla

Descripción generada automáticamenteImagen que contiene Tabla

Descripción generada automáticamenteImagen que contiene Calendario

Descripción generada automáticamente

LEM SW CT TFIDF

Imagen que contiene Tabla

Descripción generada automáticamenteImagen que contiene Tabla

Descripción generada automáticamenteImagen que contiene Tabla

Descripción generada automáticamenteImagen que contiene Tabla

Descripción generada automáticamente

LEM + STOP MLP

Imagen que contiene Tabla

Descripción generada automáticamenteImagen que contiene Tabla

Descripción generada automáticamente+

Imagen que contiene Calendario

Descripción generada automáticamente

Imagen que contiene Tabla

Descripción generada automáticamente

AHORA CON SVC

1.0

Imagen que contiene Tabla

Descripción generada automáticamente

10

Imagen que contiene Texto

Descripción generada automáticamente

100Imagen que contiene Tabla

Descripción generada automáticamente

MLP con SVD

Imagen que contiene Tabla

Descripción generada automáticamente