
Universidad
Tecmilenio®

Universidad Tecmilenio

Procesamiento de lenguaje natural y visión computacional

Anotación y evaluación

Introducción

Existen distintas medidas de desempeño para evaluar algoritmos en procesamiento de lenguaje natural y aprendizaje automático, de las más populares en clasificadores de textos son: *Log-Loss, Accuracy, AUC (Area Under Curve), precision, recall*, entre otros. Dependiendo del métrico seleccionado será la interpretación del desempeño del clasificador.

En este tema conocerás:
•Métricos de evaluación de clasificadores de textos.
•En qué consiste un Corpus anotado.

Explicación

Los métricos pueden clasificarse en tres tipos: de umbral, de probabilidad y de clasificación. Por ejemplo, para medir la capacidad de generalización, es decir, evaluar la calidad del clasificador cuando se prueba con datos que no fueron parte del conjunto de entrenamiento, puede utilizarse la exactitud y el índice de error (*accuracy, error rate*, respectivamente en inglés). La elección de los métricos determina la forma en la que se mide y compara el desempeño de los clasificadores (Sanya, Bosch y Paquette, 2020).

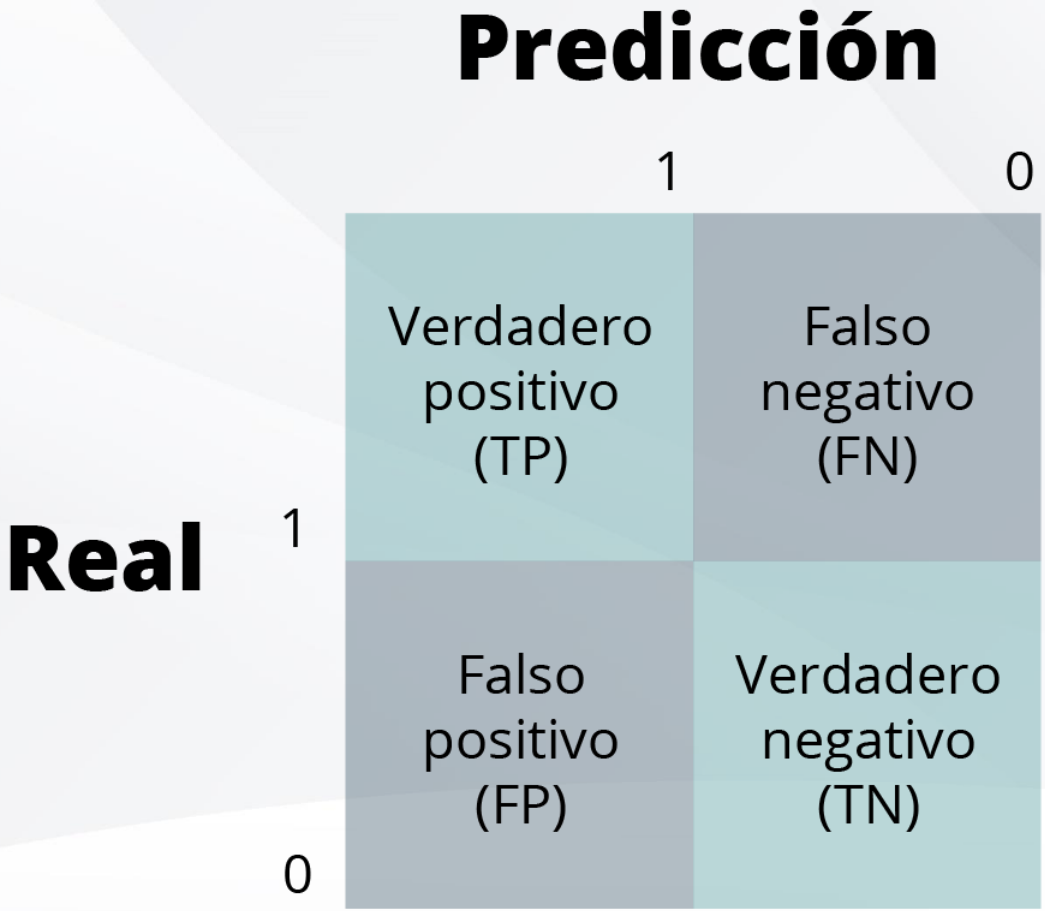
Explicación

Existen cuatro términos:

Verdaderos positivos (TP). El modelo predice positivo y la clase real del dato es positivo. Por ejemplo, el caso de una persona que tiene gripa y el modelo clasificó su caso como gripa.
Verdaderos negativos (TN). El modelo predice negativo y la clase real del dato es negativo. Por ejemplo, el caso de una persona que NO tiene gripa y el modelo clasificó su caso como NO gripa.
Falsos positivos (FP). El modelo predice positivo y la clase real del dato es negativo. Se clasificó falsamente como positivo. Por ejemplo, el caso de una persona que NO tiene gripa y el modelo clasificó su caso como gripa.
Falsos negativos (FN). El modelo predice negativo y la clase real del dato es positivo. Se clasificó falsamente como negativo. Por ejemplo, el caso de una persona que tiene gripa y el modelo clasifico su caso como NO gripa.

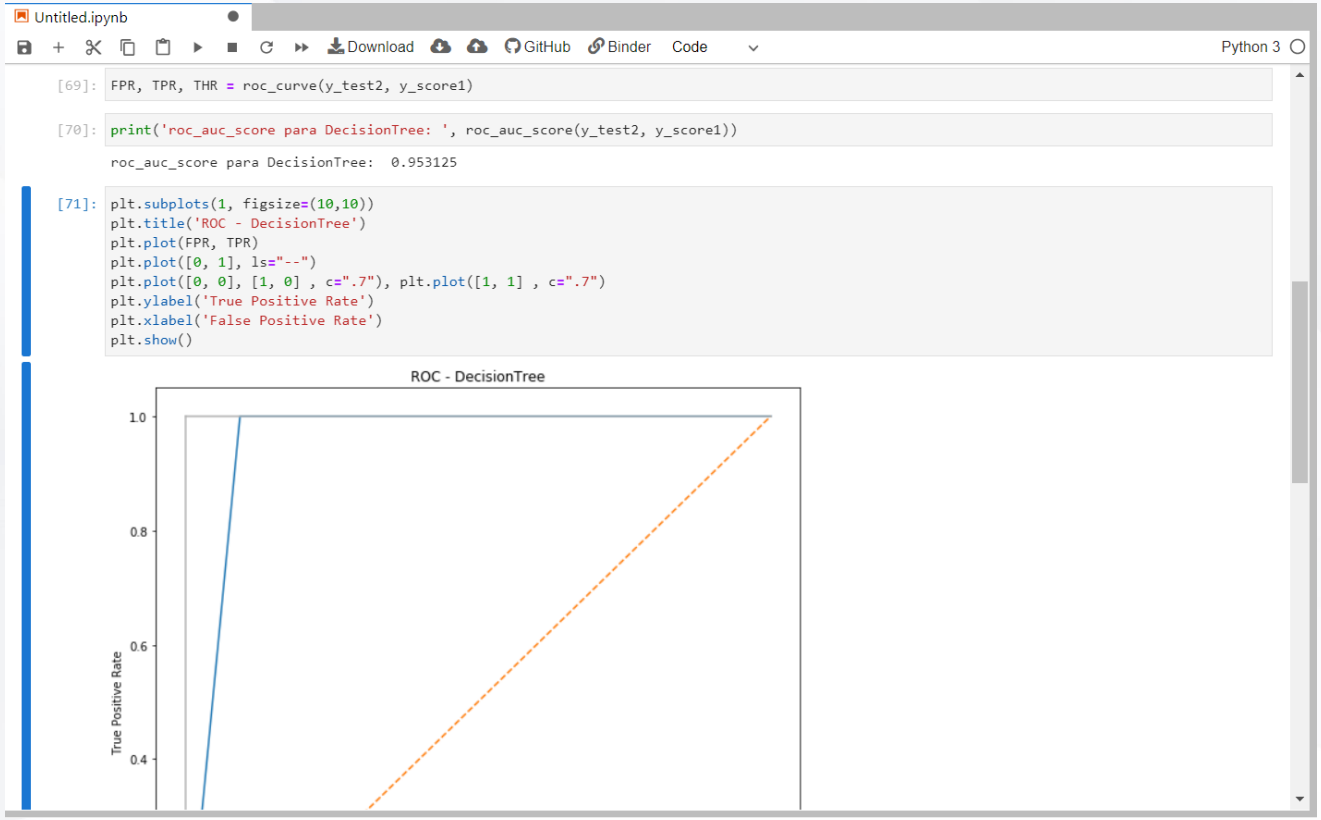
Explicación

La **matriz de confusión** es una de las formas más sencillas e intuitivas de determinar la exactitud y precisión de un modelo, además de ser una valiosa herramienta de visualización del desempeño. En problemas de clasificación binaria, la matriz de confusión se forma con una matriz rectangular de 2x2, donde los renglones representan los datos de salida del modelo y las columnas los valores que el modelo predice (algunos autores pueden invertir los valores de las columnas y renglones, por lo que es importante leer la documentación al interpretarla).



Explicación

La curva **ROC** (***Receiver Operating Characteristics***) grafica la tasa de verdaderos positivos (sensibilidad) en función de la tasa de falsos positivos (1 - *especificidad*) a distintos umbrales de cierto parámetro. Se construye a partir de la función de distribución acumulada de los verdaderos positivos (TP) en el eje Y versus la función de distribución acumulada de los falsos positivos (FP) en el eje X. El área bajo la curva ROC (**ROC AUC**, por sus siglas en inglés) es un métrico utilizado para evaluar el desempeño de clasificadores, a mayor valor, mejor el desempeño de un clasificador al distinguir entre clases. Este es un métrico recomendado cuando existen clases desbalanceadas en el conjunto de datos.



Actividad

1. Crear un modelo, a partir de los datos de entrenamiento, utilizando algún corpus de scikit-learn tales como “wine dataset”, “diabetes”, etc.
2. Con el modelo construido, realiza predicciones utilizando el conjunto de datos de prueba y almacena los resultados en una variable.
3. Importa el módulo metrics de scikit-learn y evalúa el modelo construido utilizando la matriz de confusión, la exactitud, la precisión, la sensibilidad, y el Valor F1.
4. Obtener la gráfica ROC, así como su AUC.
5. Realiza modificaciones a los parámetros del modelo utilizando el apartado de supervised learning en el sitio oficial de scikit-learn en Internet.
6. Vuelve a calcular las métricas con estas modificaciones y analiza las diferencias. ¿Puede identificar por qué son diferentes? ¿Cuál es mejor y por qué? ¿Qué métrica te sirve más para identificar claramente las diferencias?

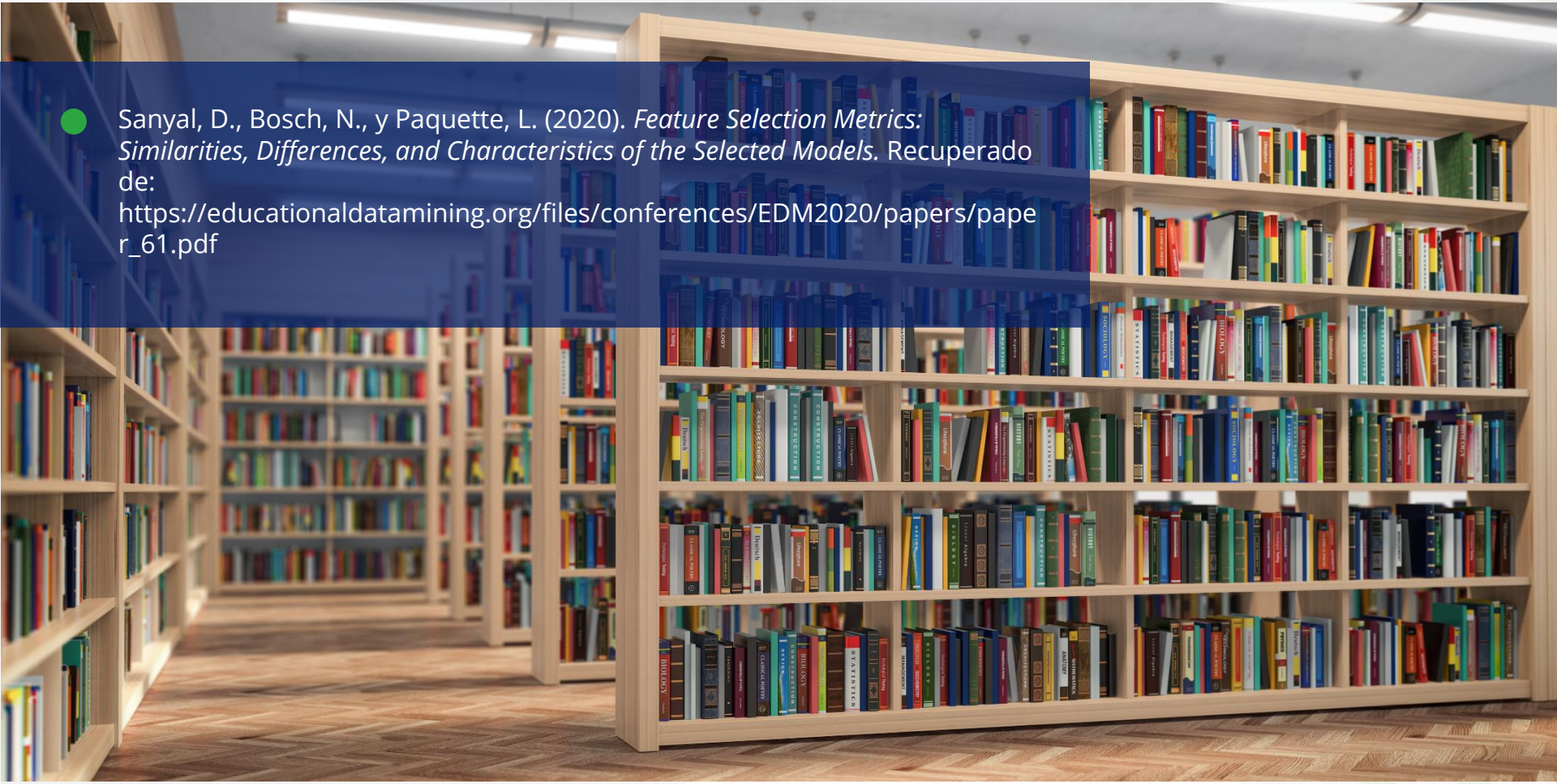
Cierre

Recuerda que el métrico utilizado para evaluar el desempeño del clasificador depende de la naturaleza del problema.

Los métricos *recall* y *precision* son los más utilizados para evaluar el desempeño de clasificadores, sin embargo, es importante estar familiarizado con distintas técnicas de evaluación para asegurar el éxito de un proyecto de clasificación.

Bibliografía

- Sanyal, D., Bosch, N., y Paquette, L. (2020). *Feature Selection Metrics: Similarities, Differences, and Characteristics of the Selected Models.* Recuperado de: https://educationaldatamining.org/files/conferences/EDM2020/papers/paper_61.pdf



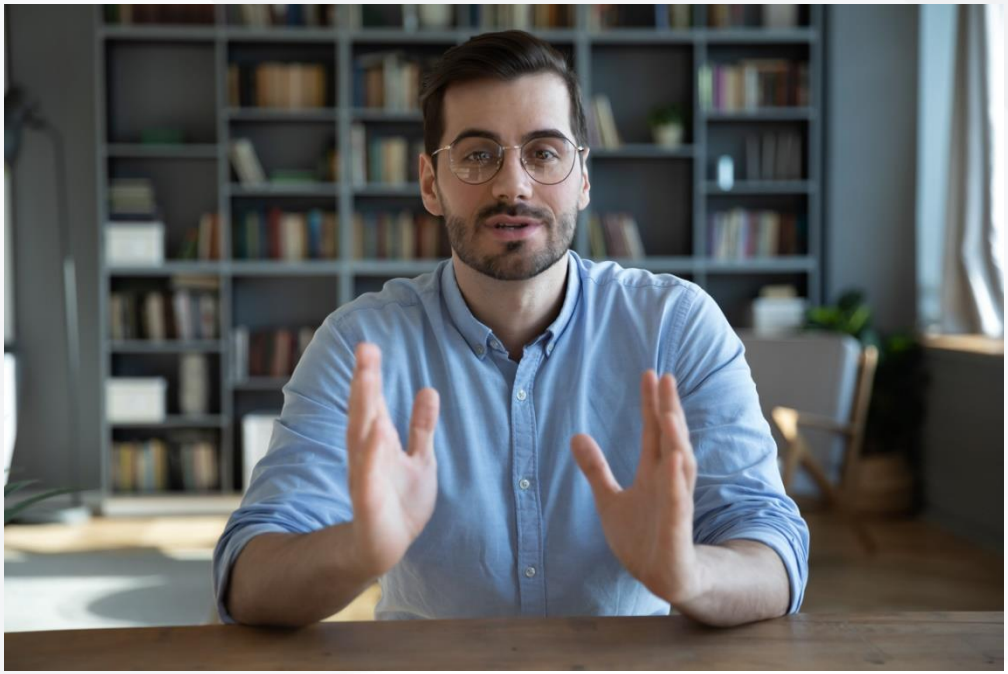
Universidad Tecmilenio

Procesamiento de lenguaje natural y visión computacional

Procesamiento sintáctico

Introducción

Algunas de las soluciones para las tareas de procesamiento de lenguaje natural se enfocan solo en las palabras.
El estudio de la gramática data de aproximadamente dos mil años, r ejemplo, la gramática de Panini proporcionó un análisis completo y la teoría gramatical occidental ha estado influenciada por esta gramática.



Explicación

La gramática libre de contexto tiene cuatro componentes y se define formalmente como:
G = (N,T,P,S)
Donde:

- N = Conjunto de símbolos no terminales.
- T = Conjunto de símbolos terminales.
- P = Conjunto de reglas de producción de la forma $A \rightarrow \beta$, donde $A \in N$ y $B \in (T \cup N)$.
- S = Símbolo distinguido o axioma inicial que es un símbolo de inicio.

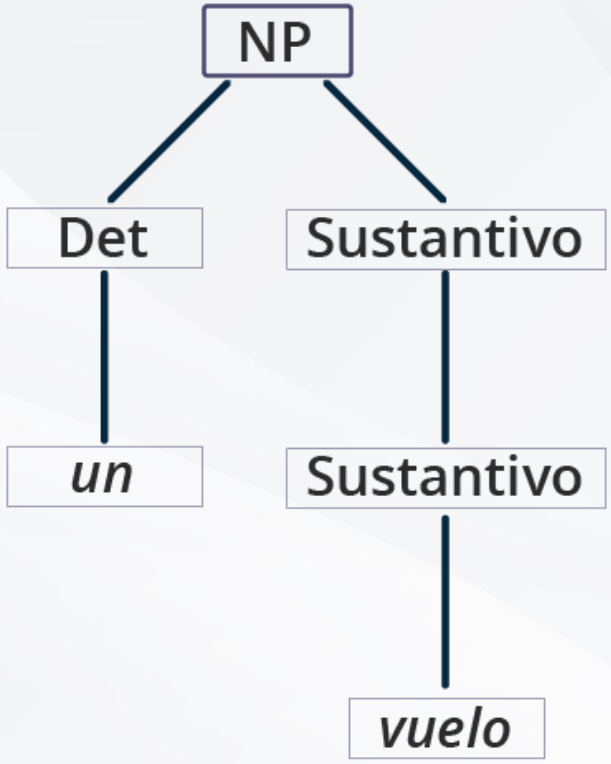
Explicación

Por convención, el lado izquierdo de la primera producción o regla es el símbolo de inicio de la gramática; por lo regular se utiliza la letra S y esta misma letra será el nodo raíz del árbol de derivación. La gramática del ejemplo incluye producciones que involucran distintas categorías sintácticas y se muestran en la figura .

Símbolo	Significado	Ejemplo
S	Oración	La chica comió
NP	Frase nominal	Un gato
VP	Verbol frasal	Darse prisa
PP	Frase preposicional	Con un teléfono
Det	Determinante	El
N	Sustantivo	Gato

Explicación

El **árbol de derivación** permite representar la derivación de cualquier cadena de palabras. Es un grafo formado por nodos y arcos, donde dos nodos distintos están conectados por arcos. Para la cadena de palabras “un vuelo” el siguiente sería un árbol de derivación:



Explicación

Una forma de encontrar al sujeto de la oración se muestra en la figura 4, donde para encontrar el "*NP* que es un hijo de *S*" basta con encontrar el subárbol cuya etiqueta es "*S*" y de ahí el nodo contenido en él con la etiqueta "*NP*".

```
[145]: print(arboles[0])

       (S (NP Lily) (VP (V saw) (NP (Det a) (N telescope))))

[146]: [s for s in arboles[0].subtrees()]

[146]: [Tree('S', [Tree('NP', ['Lily']), Tree('VP', [Tree('V', ['saw']), Tree('NP', [Tree('Det', ['a']), Tree('N', ['telescope'])])])]),
        Tree('NP', ['Lily']),
        Tree('VP', [Tree('V', ['saw']), Tree('NP', [Tree('Det', ['a']), Tree('N', ['telescope'])])]),
        Tree('V', ['saw']),
        Tree('NP', [Tree('Det', ['a']), Tree('N', ['telescope'])]),
        Tree('Det', ['a']),
        Tree('N', ['telescope'])]
```

Actividad

Genera una serie de oraciones o utiliza algunas de algún libro y conviértelas a su forma de árbol de derivación.
Practica con diferentes analizadores sintácticos en NLTK como el *shift-reduce, regex* y *left-corner*.

Cierre

Las oraciones tienen una estructura que puede representarse en forma de árbol. Una gramática es la caracterización compacta de un conjunto infinito de oraciones potenciales.
El análisis sintáctico es un proceso que permite encontrar los árboles de derivación de una oración bien formada.

Bibliografía