Clasificador Naive Bayes

Oliver Arturo Casas Pontanillo | A01645764 Erik Abraham Jajan Díaz | A01644648
José Luis Santos Montaño | A01781721

2025-09-14

Table of contents

https://github.com/olivercasas17/MA2014_NaiveBayes

1 Abstract

El siguiente trabajo toma un enfoque en la clasificación de textos en las historias de terror, el objetivo es clasificar las categorías a las que pertenecen las historias, basado en las frecuencias de las palabras mas utilizadas, realizando matrices sparse para capturar las mismas, con la eliminación de palabras innecesarias cómo stopwords o nombres propios, la dicotomización de los datos para mejorar el modelo, y trabajando con grandes volúmenes de relatos para desarrollar un modelo que alcanza una exactitud elevada, pero no demasiado eficiente, demostrando que la clasificación de estos cuentos puede ser difícil de trabajar.

2 Introducción

En este proyecto se realiza un análisis de historias paranormales recopiladas una página web de historias paranormales mediante web scraping, una técnica para la extracción automática de datos la cual permite acceder a volúmenes grandes de texto, para estudiar patrones y clasificar los relatos según su tipo. La información obtenida incluye el texto de la historia, la categoría asignada y el país de origen. El objetivo

principal es construir un clasificador automático capaz de predecir la categoría de una historia nueva a partir de su contenido textual.

Se utiliza el método de Naive Bayes, un clasificador de categoría de texto basado en el teorema de Bayes, recurriremos a la construcción de matrices sparse para el análisis, ideal para problemas de clasificación de texto debido a su simplicidad, eficiencia y buen desempeño en datasets grandes. Este enfoque permite calcular la probabilidad de que un relato pertenezca a cada categoría, considerando la frecuencia de las palabras en las historias.

3 Metodología

El presente estudio se basó en la construcción de un modelo de clasificación de texto a partir de datos obtenidos mediante scraping de una página web con historias de terror. Inicialmente, la información recolectada se organizó y almacenó en un data.frame, conservando las variables relevantes, tales como la categoría, el país de origen y el contenido textual. Esta etapa fue fundamental para garantizar la integridad de los datos y preparar la información para su posterior análisis.

Posteriormente, se realizó un preprocesamiento del texto con el objetivo de normalizarlo y transformarlo en una representación adecuada para los algoritmos de aprendizaje automático. Entre las técnicas aplicadas se incluyó la dicotomización, eliminación de palabras irrelevantes y construcción de una matriz de términos, que permitió convertir el texto en un formato cuantificable.

Con la matriz de términos lista, los datos se dividieron en conjuntos de entrenamiento y prueba, lo que posibilitó entrenar el modelo sobre una parte de la información y evaluar su desempeño sobre datos no vistos previamente. Para la clasificación se empleó un modelo de Naive Bayes, reconocido por su eficiencia en tareas de clasificación de texto y por su capacidad de manejar matrices dispersas de gran dimensión.

Finalmente, se generaron las predicciones sobre el conjunto de prueba y se evaluó el desempeño del modelo mediante métricas de clasificación, como la matriz de confusión, la exactitud, la precision y el recall. Esta metodología permitió establecer un flujo sistemático y reproducible para abordar la clasificación de textos en el dataset, asegurando que cada etapa estuviera claramente definida desde la recolección hasta la evaluación del modelo.

4 Aplicacion

Para realizar este análisis trabajamos en el archivo Naive_Bayes.qmd, despues de realizar la matriz de términos, pudimos hacer nuestro clasificador. Primero lo hicimos con la libreria e1071, y el resultado

muy bajo, en todas las metricas cercano al 0.1%. Despues lo realizamos con la libreria naivebayes, con la cual obtuvimos resultado practicamente identicos.

Por lo tanto, para mejorar nuestro modelo, decidimos dicotomizar nuestras categorías, ya que nuestra base de datos contaba con muchas y nuestro modelo naive bayes no estaba funcionando correctamente con esos parámetros, por lo que nos quedamos con Apparitions / Voices / Touches, que era la categoria con más muestras. Una vez dicotomizados las categorias, probamos de nuevo a realizar nuestros modelos, esta vez con la libreria e1071 subio la accuracy a 49%, y con la libreria naivebayes y modificando que la distribucion de las variables eran tipo poisson, subio el accuracy a 79%.

Por ultimo hicimos un cross-validation para encontrar el valor mas óptimo del laplace smoothing, sin embargo, se llego a la conclusion de que ese valor era 0, por lo que nuestro último modelo co 79% de accuracy se quedo siendo el mejor.

5 Conclusiones

Los resultados evidencian que la eficacia de Naive Bayes depende no solo del uso de librerías específicas, sino también de un adecuado preprocesamiento de los datos y de la correcta definición de las categorías de clasificación. Mientras que los modelos iniciales resultaron ineficaces, la estrategia de dicotomizar la variable objetivo y asumir una distribución Poisson para las variables permitió obtener un modelo con un rendimiento satisfactorio.

En conclusión, este ejercicio muestra cómo la combinación de técnicas de simplificación de categorías y ajustes en la distribución de las variables puede transformar un modelo con métricas insignificantes en un clasificador robusto, alcanzando un 79% de exactitud en la tarea planteada.

Este trabajo nos mostró que crear un clasificador de textos puede ser difícil debido a la enorme cantidad de variables a considerar para el análisis, y en caso de lograr realizarse, la exactitud del modelo es increiblemente baja por la complejidad del trabajo a reaizar haciendo que las probabilidades de que el modelo clasifique correctamente sean igual de probables a que el modelo se equivoque.

6 Referencias

https://www.yourghoststories.com/

https://www.youtube.com/watch?v=99Hkmfb2i80

https://arxiv.org/abs/1709.08314?