

CLASIFICACIÓN DE TEXTOS: NAIVE BAYES VS NEURAL NETWORKS

David Pérez Gómez & Guzmán López Santamaría

EUITI - UPV/EHU

OBJETIVOS

Dado un conjunto de datos de autopsias verbales se quiere realizar un proceso de clasificación mediante el cual se prediga una clase para cada instancia. Este proceso se llevará a cabo mediante dos clasificadores diferentes y representando el conjunto de datos de dos formas distintas. El objetivo de este proyecto es analizar y comparar los resultados obtenidos para cada uno de los clasificadores y de las representaciones.

DATOS

El conjunto de datos es Verbal Autopsies, un conjunto de autopsias verbales realizadas en diversos países y que agrupa una gran variedad de enfermedades y causas de muerte. Serán estas las que nuestro clasificador utilice como clase a predecir. En cuanto a las entradas, utilizaremos el atributo correspondiente a la autopsia verbal en sí (gs_text34).

Flujo de cambios	Descripción
Raw	Conjunyo de textos original.
Limpieza textos vacíos	Se eliminan manualmente instancias duplicadas y aquellas que están vacías.
Minusculas	Se convierte todo el texto a letras minúsculas.
Stemmer	Se sustituye cada palabra por su raíz.
Tokenizer	?

REPRESENTACIÓN DE TEXTOS

TF-IDF

TF-IDF es una medida estadística que sirve para calcular la importancia de una palabra dentro de un texto. De forma simplificada, se considera que la importancia de una palabra aumenta cuantas más apariciones tenga en un documento dado, pero a su vez es inversamente proporcional a su ratio de aparición en otros documentos.

Document Embeddings

Mediante Document Embeddings se puede representar cada texto de un conjunto de estos como un vector con valores numéricos, es decir, se representa cada texto en un espacio vectorial de n dimensiones. Esto es especialmente útil en casos como el que nos atañe, la clasificación, pues los clasificadores no pueden trabajar directamente con textos y necesitan que estos sean representados de alguna otra forma, siendo un vector una de las más simples de hacerlo.

En nuestro caso utilizamos la librería Gensim para hacer esta transformación.

CLASIFICADORES

Naive Bayes

Como clasificador de referencia hemos decidido utilizar Naive Bayes. Este algoritmo calcula la probabilidad de pertenencia a cada clase para cada atributo en función de su valor. Con esto se está suponiendo que los atributos son independientes entre sí. Cuando se dispone a clasificar, realiza la predicción en función de las probabilidades calculadas para cada uno de los atributos de la instancia dada y selecciona la clase que maximice esta probabilidad.

Debido a que Naive Bayes asume una total independencia entre clases, consideramos que este clasificador es una buena base para comparar con otro clasificador más complejo.

Redes Neuronales

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

RESULTADOS

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

