Proyecto Fin de Master

Clasificación automática de textos utilizando un algoritmo SVM

Luis Fernando Moreno

Introducción

La clasificación automática de textos es un tema que se ha estado estudiando con gran empeño en los últimos años, aunque hasta ahora ya hay mucho camino recorrido, debido a la amplitud, flexibilidad y gran cantidad de diversas variantes del lenguaje, no se ha conseguido un sistema que pueda cubrir satisfactoriamente las necesidades en esta rama.

En este trabajo se utiliza un algoritmo de Máquinas de vectores de soporte (o SVM por sus siglas en inglés). Para realizar la clasificación de una cantidad de cadenas de texto relativamente cortas, relacionadas con unas temáticas específicas y obtenidas de Twitter (o tweets), a los que se les aplicó un tratamiento previo y se prepararon para poder entrenar y aplicar el algoritmo sin inconvenientes

La elección de este algoritmo específico se realizó después de llevar a cabo un proceso de investigación sobre diversos algoritmos de clasificación de texto, en el cual se llegó a la conclusión que las Máquinas de vectores de soporte son el algoritmo que más se ajusta al problema que se plantea.

Objetivos

El objetivo inicial de este proyecto consistía en la idea de estudiar cuándo se hablaba en twitter sobre tráfico o contaminación en Madrid, y contrastar estos datos con datos públicos obtenidos de fuentes oficiales: el ayuntamiento de Madrid y la DGT. Debido al tamaño y comlejidad de este objetivo, se ha visto reducido y limitado a la implementación de un algoritmo SVM para la clasificación automática de Tweets que hablen bien de contaminación, tráfico o de ambos.

Aunque el objetivo se ha visto reducido en gran medida con respecto al original, las labores de limpieza de datos de calidad del aire y de tráfico se llevaron a cabo, y se pueden consultar en este repositorio.

Metodología

Adquisición de datos

Se adquirieron datos públicos desde Twitter en formato json, a través de la Universidad Complutense de Madrid, con un total de 2.323.778 registros con 54 campos cada uno, alcanzando este conjunto de datos un tamaño de 2,8Gb. Una gran parte de estos registros se encontraban en mal estado o incompletos, por lo que se descartaron.

Por cuestiones de compromiso y confidencialidad, el set de datos original no está disponible, sólo lo están unas pequeñas muestras y los sets de entrenamiento y test utilizados para entrenar y probar el modelo de SVM.

Limpieza y procesamiento previo de datos

Esta fase consistió en extraer los registros en buen estado, eliminar los campos innecesarios y dar un formato conveniente a los campos que iban a ser usados más adelante. Este proceso se llevó a cabo en el archivo Tweets_para_entrenar.ipynb

Cambio de formato:

Algunos de los campos venían en un formato inconveniente o poco manejable, se realizó la transformación de estos formatos a uno más conveniente.

Extracción de registros en buen estado:

Gran cantidad de los registros tenían carencia de información, estaban incompletos o eran ilegibles, estos registros fueron descartados, conservando sólo aquellos que podían ser utilizados para el proceso.

Eliminación de campos innecesarios:

Los registros iniciales tenían una gran cantidad de campos, muchos de los cuales no iban a ser de utilidad para las tareas realizadas posteriormente, de manera que se descartaron, conservando sólo los campos que iban a ser utilizados.

Procesamiento previo de texto:

Una vez realizados los pasos anteriores, se procedió a eliminar o modificar caracteres extraños de los tweets que se iban a analizar tales como tildes o virgulillas, que podían presentar problemas más adelante al implementar el algoritmo SVM. De igual modo se transformaron todas las letras a minúsculas.

Se realizó un filtrado en función de una lista de palabras clave, relacionadas con los temas de tráfico y contaminación, aunque no siempre se encuentran en comentarios relacionados con estas.

Una vez realizadas las tareas anteriores, se contaba con un set de registros relacionados con las temáticas de interés, se le añadieron algunos registros más para que al entrenar el modelo, pudiera aprender también a identificar registros que no encajaran.

Implementación del algoritmo SVM

Esta parte del proceso se lleva a cabo en el archivo Implementación_de_SVM.ipynb. En este se realizan varias tareas:

Tratamiento de texto:

Aunque en las fases previas ya se había realizado un tratamiento superficial de texto, ahora se realiza un tratamiento más intensivo que consiste principalmente en tokenización, lematización, eliminación de stop-words y caracteres no alfanuméricos.

Ranking de palabras relevantes:

Este paso consiste en, a partir de los valores IDFs para cada palabra, crear vectores de características a partir de los cuales se calcula la ganancia de cada palabra; básicamente consiste en otorgar valores a las palabras que cuantifiquen el peso que tienen estas al momento de realizar la clasificación de tweets.

Una vez realizado eso, se seleccionan las palabras más relevantes para simplificar el proceso de clasificación.

Implementación de SVM:

Se ajustan dos modelos de SVM: con kernel establecido por defecto en el paquete de sklearn y con kernel lineal. En la investigación previa se había llegado a la conclusión de que el kernel lineal es el que mejor funciona en estos casos, premisa que se comprobó al comparar los porcentajes de accuracy, obteniendo el kernel lineal un accuracy de 76,58%.

Conclusiones

Aunque el método de clasificación aún se aleja bastante de lo que puede considerarse satisfactorio, el algoritmo todavía clasifica una gran cantidad de registros correctamente en muy poco tiempo (más de 850000 registros en 10 minutos) lo que puede considerarse positivo para no ser un documento riguroso. Este trabajo puede ser utilizado como apoyo a una implementación más profunda, o como comparación con otros métodos de clasificación de textos, bien sean con otros algoritmos, o igualmente con SVMs iplementados de otras maneras.