Data Mining en Social Media

Javier Cortés Broch
jacorbro@masters.upv.es

July 15, 2017

Resumen

Queremos predecir el género y la variadad de unos tweets dados en un dataset. En la práctica vamos a intentar predecir con la mayor precisión posible, para ello haremos una limpieza de datos, es decir, eliminaremos palabras del vocabulario que consideremos no influyentes para la predicción del modelo. La segunda parte del trabajo será la elección del modelo, en relación tiempo y accuraccy obtenido. Probaremos varios modelos y haremos varios métodos de selección del vocabulario para obtener el mejor resultado.

1 Introducción

Tenemos un dataset con tweets etiquetados por género y variedad .

Tenemos un total de 2800 tweets para el training y 1400 tweets para el test. El objetivo trata de superar la precisión sobre género de 66'43% y variedad de 77'21%.

2 Dataset

El dataset está compuesto con ficheros XML separados en training y test.

Ejecutando el código proporcionado, transformamos estos ficheros XML para poder utilizarlos en R.

3 Propuesta del alumno

Para mejorar la precisión de las predicciones junto con el código proporcionado en R para la práctica, realizamos las siguientes transformaciones:

- Transformación de los textos a minúsculas
- Eliminar signos de puntuación

- Eliminar stopwords
- Eliminar palabras que terminen con s, para evitar plurales
- Eliminar aquellas palabras que tengan longitud 1.
- Eliminar las urls de que contienen imágenes.

Una vez finalizada la limpieza, procedemos a la elección del algoritmo.

Entrenamos el módelo con support vector machine, pero no supera la precisión inicial que queremos obtener.

En segundo lugar probamos un módelo Random Forest y observamos que al incrementar el número de arboles mejora nuestro módelo, en relación tiempo de espera y accuracy decidimos quedarnos con 100 arboles.

Decidimos hacer un módelo de Naive Bayes pero no conseguimos mejorar el accuracy utilizado por Random Forest.

Por ultimo probamos el algoritmo de Redes Neuronales pero el tiempo es excesivamente largo y decidimos no terminar con la ejecución.

4 Resultados experimentales

Depués de la elección del algoritmo Random Forest con 100 árboles. Obtenemos los resultados de 70.86% y 85% para genero y variedad respectivamente.

5 Conclusiones y trabajo futuro

Hemos conseguido mejorar la predicción y el accuracy obtenido es mejor que el objetivo planteado. Hemos observado como se comportan los algoritmos para tratar este tipo de problemas.

Como propuesta de mejora:

- Agrupar las palabras por su raíz.
- Calcular solo con las palabras mas utilizadas.
- Observar como afecta la longitud de las palabras.