Data Mining en Social Media. Máster Big Data Analytics.

David Selma Herrero
daselher@masters.upv.es

Abstract

El reto que se nos plantea es predecir a que genero (hombre, mujer) o variedad (país) pertenecen los tweets. Para ello se nos facilitan dos carpetas, una para training y otra para test. En estas dos carpetas tenemos un fichero donde podemos identificar el autor (nombre del fichero XML), genero (si es hombre o mujer), y su variedad etiquetadas. Este fichero se llama "truth.txt". Y por ultimo el resto de ficheros XML, un fichero XML por autor.

Para ello se realizan una serie de puntos en el siguiente orden:

- Entender los datos que manejamos y el problema al que nos enfrentamos.
- Estudio de la limpieza de datos, ya que dependiendo de que se limpie podemos mejorar o empeorar los resultados.
- Estudio y elaboración del algoritmo.
- Estudio de los modelos a aplicar a nuestro caso en particular.
- Uno de los parámetros a fijarnos principalmente: Accuracy y kappa.

Por ultimo, se han realizado diferentes pruebas con los diferentes modelos para obtener el mejor resultado en un tiempo razonable.

1 Introducción

El objetivo de este problema es superar la predicción base de genero y variedad.

- Genero: 66'43 %
- Variedad: 77'21 %

Se nos proporcionara un dataset para realizar una mejor predicción. Y para ello habrá que superar dichos parámetros.

2 Dataset

Se nos proporcionan un dataset ya separado en training y test, en dos carpetas diferentes. Cada uno de ellos contiene ficheros XML, uno por cada autor. Y un fichero llamado "truth.txt" donde tendrá nombre del autor, genero y variedad.

La carpeta de training contiene un total de 2800 autores y la carpeta de test tiene un total de 1400 autores. Es decir, un 66'66 % de training y un 33'34 % de test.

Cada autor tiene 100 tweets.

Por una parte la carpeta de training contiene 1400 autores de genero masculino y 1400 autores de genero femenino. Por otra parte la carpeta de test contiene 700 autores de genero masculino y 700 autores de genero femenino.

Se puede observar que este dataset de training esta compuesto por 400 autores de cada uno de los países que tiene llamados Colombia, Argentina, Spain, Venezuela, Peru, Chile y Mexico. También el dataset de test tiene 200 autores de cada uno de los países citados anteriormente.

3 Propuesta del alumno

Para tener una predicción mas exacta y precisa se propone primero una limpieza de datos y adaptación de la información para este problema en los siguientes puntos:

- Transformar las palabras a minúsculas (tolower).
- Eliminación de signos de puntuación (removePunctuation).
- Eliminación de números (removeNumbers).
- Eliminación de palabras stopwords (removeWords de swlist = "es").
- Eliminación de preposiciones (removeWords).

- Eliminación de espacios en blanco extra. Varios caracteres de espacio en blanco se contraen a un solo espacio en blanco (stripWhitespace).

Se realiza una bolsa de palabras adecuada para variedad y genero.

Una vez se finalice la limpieza adecuada y elaborado el algoritmo que se considera que obtendrá una mayor precisión en la predicción. Se procede a aplicar los siguiente modelos que hemos considerado apropiados para este caso:

- Redes Neuronales
- Naive Bayes
- Random Forest

En principio se entreno el modelo con una máquina de soporte de vectores con diferentes parámetros pero no hubo la mejora que se esperaba para la resolucin de este problema.

De modo alternativo, cuando se probo con los modelos Redes Neuronales, Navive Bayes y Random Forest. Se observa que Random Forest daba mejores predicciones al incrementar el número de árboles a 100.

Para este problema se ha tenido en cuenta que no se este haciendo sobreajuste.

Por otra parte nos fijamos en Accuracy y en el Kappa. El Kappa es importante que este lo mas cercano a 0, ya que si no es asi significa que el resultado es altamente probable debido al azar.

4 Resultados experimentales

Los resultados experimentales obtenidos tras la realización de este problema en las predicciones de gnero y variedad son:

- Genero: 72'21 %
- Variedad: 88'71 %

Estos resultados se obtienen al aplicar Random Forest con una configuración de 100 arboles. El cálculo se ha realizado dentro de un tiempo razonable y esperado.

Por falta de tiempo se procedió a ejecutar Random Forest con diferentes arboles y esto llevo todo el día dando peores resultados a los anteriormente.

5 Conclusiones y trabajo futuro

- Conclusiones: Los resultados obtenidos han mejorado con respecto a la predicción base dada inicialmente. La precisin de gnero ha mejorado en un 9'40 % y la de variedad un 15'20 %.
- Trabajo futuro:
  - Realización de la bolsa de vocabulario con las palabras mas utilizadas en cada uno de los países, es decir, palabras propias de cada país o genero. Pudiendo así identificar a que variedad pertenece. O en el caso del genero las palabras mas utilizadas por el genero femenino y también las palabras mas utilizadas por el genero masculino.
  - Longitud de las palabras ya que se piensa que un determinado genero o variedad escribe con mayor longitud los tweets.
  - Sacar la raíz o lexema de las palabras para predecir a que variedad pertenecen los tweets.
  - Quitar la raíz y quedarnos con las ultimas letras de las palabras para predecir a que genero pertenecen los tweets.
  - Leer tweets en grupos de dos o mas palabras detectar patrones que se puedan obtener en determinado genero o variedad.
  - Dar peso a determinadas palabras. Esta idea viene porque se piensa que aquellas palabras mas frecuentes o propias de un determinado genero o variedad tiene que tener mas importancia y mayor impacto en la predicción con respecto a aquellas palabras que son comunes en los dos géneros o que estén en diferentes variedades. Las palabras comunes en ambos géneros y diferentes variedades tendrían que tener menor peso por su poca importancia en la predicción.

6 Referencias

http://topepo.github.io/caret/index.html

https://topepo.github.io/caret/available-models.html