MASTER BIG DATA ANALYTICS
Text Mining II
Author Profiling

Marco Antonio Melero Gallo
marmegal@hotmail.com

1 Introducción

Author Profiling es la tarea de identificar cualquier característica del autor de un texto a partir de su escritura. En este sentido puede identificarse la edad, el sexo, la lengua origen, nivel de estudios u otras muchas características. En el siguiente documento describiremos la aproximación llevada a cabo para determinar el sexo y la lengua origen de los autores de un conjunto de tweets, a partir del texto de los mismos.

2 Dataset

El conjunto de datos utilizado está compuesto de dos ficheros de texto, training.txt y test.txt, y de 7 directorios. Estos dos ficheros mantienen internamente una estructura idéntica y es que, cada línea que lo compone contiene, separados por los símbolos ':::', el identificador de un usuario de Twitter, el país del que procede (argentina, chile, colombia, espana, mexico, peru o venezuela) y el sexo del usuario (male, female o UNKNOWN, éste último en el caso de desconocer el sexo). Cada uno de estos ficheros será por tanto los que utilizaremos en nuestros modelos de Machine Learning, en la fase de training y de test. Los 7 directorios tienen los nombres de los posibles países de origen de los usuarios y se corresponden con los que aparecen en los ficheros antes comentados. Dentro de cada uno de ellos se encuentran 650 ficheros *.json. Éstos tienen por nombre el identificador de un usuario de Twitter, que se corresponderá con los que aparecen en los ficheros de training.txt y de test.txt. Por lo tanto, cada uno de estos ficheros contiene, en formato json, un conjunto de tweets del usuario que viene indicado en el nombre del fichero. La información de cada uno de estos tweets, es la que nos facilita Twitter (texto, identificador, geolocalización, fecha de creación, etc.), sin embargo, nosotros sólo utilizaremos el campo texto para llevar a cabo nuestra aproximación.

3 Propuesta del alumno

La aproximación llevada a cabo para resolver nuestro problema ha sido la basada en una bolsa de palabras con las 1000 más frecuentes del corpus de training. Se han creado 2 bolsas de palabras distintas, una para el sexo y otra para el país de origen. Para la construcción de la bolsa de palabras del sexo se ha utilizado un tokenizador de la librería NLTK (1), donde se mantenían las letras mayúsculas, no se reducía el tamaño de las palabras y se eliminaba cualquier mención a usuarios. Para la construcción de la bolsa de la palabras del país, se ha utilizado el mismo tokenizador que en el caso anterior, pero además se han eliminado las stopwords en español definidas en la librería NLTK (1). Hemos optado por separar las bolsas de palabras porque consideramos que en el caso del sexo, influyen muchas más variables que nos permiten diferenciar entre la escritura de hombres y mujeres y, por lo tanto, no hay que eliminar ninguna de ellas. En cambio, en el caso del país de origen nos basaremos principalmente en el valor semántico de las palabras, y no tanto en el conjunto de múltiples variables de todo el texto. Por ese motivo hemos eliminado las stopwords en español.
La creación de estas dos bolsas la hemos realizado en una fase inicial, almacenando dichas bolsas en los ficheros BOWSex.txt y BOWCountry.txt para su posterior utilización en fases posteriores. El motivo por el cual se almacenan estas bolsas es porque su construcción es algo lenta, y de esta manera se agiliza el posterior proceso de modelado y testeo.

Una vez creadas estas dos bolsas de palabras, procederemos a crear la matriz de características de los tweets de training. En dicha matriz, cada columna se corresponde con una de las 1000

palabras más comunes de los tweets, es decir, las que están almacenadas en la bolsa de palabras. Y cada fila, se corresponde con un autor. Por lo tanto, en cada línea se almacenará la frecuencia absoluta de aparición de las 1000 palabras más comunes en los tweets de un determinado autor. El orden en el que se colocan cada una de esta palabras siempre será el mismo.
Otras opciones alternativas al rellenado de la matriz con frecuencias absolutas han sido rellenar dicha matriz con frecuencias relativas (frecuencia absoluta de una palabra dividido entre el número total de palabras de un usuario) y frecuencias binarias (aparece un 1 si existe la palabra para ese usuario y un 0 en caso contrario).

Con todos estos datos, con los vectores donde se almacena el sexo y la nacionalidad de cada uno de los usuarios, y generando la matriz de características con los tweets de test, procedemos a entrenar nuestro modelo mediante los siguientes algoritmos de la librería scikit-learn (2):

* K-Nearest Neighbors (KNN)
* Gaussian Naive Bayes (GNB)
* Support Vector Machines (SVM)

El entrenamiento de nuestros modelos se llevará a cabo con los 3 tipos de matrices obtenidas (frecuencias absolutas, relativas y binarias) y con las dos bolsas de palabras (la del sexo y del país de origen).

4 Resultados experimentales

A continuación veremos unas tablas resumen que contienen los resultados obtenidos de las distintas pruebas realizadas. En ellas se desglosan los resultados con los distintos algoritmos de machine learning y las distintas frecuencias de aparición de las palabras, donde en cada uno de ellos se muestra el número de muestras mal clasificadas (columna KO), el número total de muestras y el accuracy obtenido.

Como se observa en la Tabla 1, en ninguno de los casos se supera el 52% de accuracy. Los resultados no son muy alentadores y, por lo tanto, habría que replantear algunas mejoras de nuestro modelo para conseguir mejorarlos. Es posible que estos resultados estén motivados por la generalidad planteada inicialmente en la creación de la bolsa de palabras para el sexo.

	Frecuencia	KO	Total	Accuracy
KNN	Absoluta	1059	1820	41.8%
	Relativa	1064	1820	41.5%
	Binaria	970	1820	46.7%
GNB	Absoluta	1101	1820	39.5%
	Relativa	1053	1820	42.1%
	Binaria	1291	1820	29.1%
SVM	Absoluta	1066	1820	41.4%
	Relativa	1066	1820	41.4%
	Binaria	862	1820	52.6%

Table 1: Resultados para la predicción del sexo

	Frecuencia	KO	Total	Accuracy
KNN	Absoluta	1193	1820	34.5%
	Relativa	1186	1820	34.8%
	Binaria	592	1820	67.5%
GNB	Absoluta	531	1820	70.8%
	Relativa	483	1820	73.5%
	Binaria	453	1820	75.1%
SVM	Absoluta	1558	1820	14.4%
	Relativa	1455	1820	20.1%
	Binaria	152	1820	91.6%

Table 2: Resultados para la predicción del país

En la Tabla 2 los resultados son muy variados, y van desde el 14.4% de accuracy hasta el 91.6%. En comparación con el tratamiento realizado para averiguar el sexo, en este caso se ha obtenido un mejor resultado, debido seguramente al pequeño tratamiento adicional que se ha hecho (la eliminación de las stopwords). Así, se produce un refinamiento a la hora de clasificar las muestras.

5 Conclusiones y trabajo futuro

A la vista de los resultados obtenidos no se puede concluir que, con los modelos utilizados y los algoritmos aplicados, se puedan clasificar claramente estas muestras, ya que consideramos que el número de pruebas para la validación de los resultados ha sido insuficiente (por falta de tiempo). Aún así, parece claro que el hecho de eliminar palabras carentes de sentido en determinados idiomas es un buen punto de partida para investigar. Por ese motivo, consideramos que una buena propuesta futura sera aplicar TF/IDF para clasificar el país de origen de los autores de los tweets. Otras posibles mejoras, tanto para averiguar el sexo como el país, pueden ser el filtrado previo de palabras, antes de tokenizar con la librería NLTK

(1). Para ello, podrían utilizarse expresiones regulares y eliminar url's, emoticonos y signos de puntuación.

References

[1] http://www.nltk.org

[2] http://scikit-learn.org

[3] https://docs.python.org/3/library/collections.html

[4] http://es.diveintopython.net/index.html