Práctica con PAN-AP 17.
Asignatura Text Mining en Social Media.
Master Big Data

Elena Tejadillos
elena.tejadillos@gmail.com

Abstract

En esta práctica, se parte de un conjunto de tuits que nos proporcionó el profesor de la asignatura. Tomando este conjunto de tuits como fuente inicial, sse debe predecir el sexo de la persona que escribió el tuit y la procedencia del mismo. Para ello, se tratan los textos eliminando acentos. Después, se estudia la lista de palabras más frecuentes. Entre ellas se descartan 9 como ”stopwordsz se eliminan de la lista al considerarse que no son relevantes para el cálculo de variedad. Finalmente, con el datasest para training y para test definido, se género dos 4 datasets, 2 para genero (uno de training y otro de test) y oros 2 para variedad. Se aplicaron 3 diferentes modelos (Suppert Vector Machine, Naive Bayes con validación cruzada y RandomForest). Finalmente nos quedamos con el modelo que mejor .ªccuracy”genera que es el RandomForest.

1. Introducción

.*Author profiling”* es el área de estudio que da la posibilidad de conocer rasgos de una persona a partir de los textos que escribe. En el caso conreto de esta práctica, se quiere averguar el sexo y el país de origen de los autores de un conjunto de tuits. Para realizar el estudio, se han proporcionado un conjunto de tuits. A partir de este conjunto de tuits, se dividirán en dos subconjuntos, uno utilizado para el *”training”* de los modelos y el otro para el *”test”*.

Además se extraerá una bolsa de X palabras más frecuentes en esos tuits que se utilizarán en el estudio. Finalmente se aplicarán diversos modelos para averiguar el sexo y el país de origen de los autores de los tuits.

2. Dataset

Se proporciona la información nesaria para el dataset PAN-AP’17 a través de un fichero zip. Este fichero contiene los tuits de 300 autores distintos: 200 autores para training y 100 autores para test. Al descomprimirlo, se generan 2 carpetas llamadas ”trainingz ”test”.

La carpeta ”training”se almacenarán 2800 ficheros xml que contiene los tuits a analizar y un fichero ”truth.txt”que contiene la lista de títulos de los ficheros xml y el género y páis a que corresponden.

La carpeta ”testçontiene la misma información pero con tan solo 1400 ficheros xml. Estos tuits son los se van a clasificar dependiendo del género (male/female) o en uno de los 7 paises a los que pueden pertenecer:

Argentina Chile
Colombia España
Mexico Perú
Venezuela

Al cargar toda la información en el programa, se tendrán 2 dataset, uno para training y otro para test. Estos datasets se han creado de manera que los datos estén balanceados, lo que significa que habrá el mismo número de tuits para hombres y para mujeres como para cada uno de los países incluidos en la muestra.

TUITS	TRAINING		TEST	
	MALE	FEMALE	MALE	FEMALE
ARGENTINA	200	200	100	100
CHILE	200	200	100	100
COLOMBIA	200	200	100	100
ESPAÑA	200	200	100	100
MEXICO	200	200	100	100
PERÚ	200	200	100	100
VENEZUELA	200	200	100	100
TOTAL	1400	1400	700	700
	2800		1400	

Figura 1: Distribución del Dataset original.

Una vez que se tiene ambos datasets cargados, se hará una limpieza de datos eliminando los signos de puntuación de los textos de los tuits.

3. Propuesta del alumno

El código de la práctica ya viene preparado para aplicar uno de los métodos más utilizados para estimar la importancia de un término, el conocido sistema Term Frecuency (TF). Está pensado para calcular la importancia de un término en función de su frecuencia de aparición en un documento.

Siguiendo este método, se genera una lista de 1000 palabras que son las que aparecen con mayor frequencia en los tuits.
Se decide esetudiar las 100 primeras y se imprime en una gráfica la lista de estas 100 palabras ordenada de mayor a menor frequencia.

Se decide excluir alguna de estas palabras a las que se denominarán *stopwords*. Las *stopwords* son palabras vacías , un listado de términos (preposiciones, determinantes, etc.) considerados de escaso valor semántico, que cuando se identifican en un documento se eliminan.

La supresión de todos estos términos evita los problemas de ruido y supone un considerable ahorro de recursos, ya que aunque se trata de un número relativamente reducido de elementos con una elevada tasa de frecuencia.

Tras un análisis inicial, se decide incluir como *stopwords* las siguientes palabras:

- La palabra más utiliza y por ello, considerada como con escaso valor semántico ***"mas"***
- Palabras de una sola letra ***"d"*** o ***"q"*** (preposión a o abreviación)
- Onomatopeyas universales (***"jajaja"***)
- Nombres propios o genéricos (***"Trump"*** o ***"video"***)

Indicar que la selección de palabras ha sido mas orientada a la variedad y no al género, donde se ha pensado si dichas palabras se dirían de igual manera en todos los paises, y no si se utiliza más los autores masculinos o femeninos.

A continuación, se aplican modelos de predicción para calcular el *Accuracy* que se puede obtener con ellos. Los modelos seleccionados son:

- Support Vector Machine (SVM)
- Naive Bayes con validación cruzada
- Random Forest

De todos ellos, Naive Bayes es el que peor resultado obtiene, y Random Forest el que obtenemos una mayor a *Accuracy*.

Aunque se mejora levemente el umbral dado por el profesor, se decide aplicar como propuesta de mejora aumentar la bolsa de palabras de 1000 a 2000, consiguiendo de esta forma el Accuracy más alta tanto para la predicción de género que de variedad en todos los modelos.

4. Resultados experimentales

Tras aplicar los 3 métodos elegidos con la bolsa de 2000 palabras, el mejor resultado obtenido tanto para género como para variedad es el generado por Random Forest con las siguientes:

- Género: 0.7313
- Variedad: 0.8679

Tal y como puede observarse en la siguiente gráfica, los resultados obtenidos por los tres métodos superan, aunque sea levemente, el umbral inicial proporcionado por el profesor, aunque es el Random Forest el que mayor *Accuracy* alcanza.

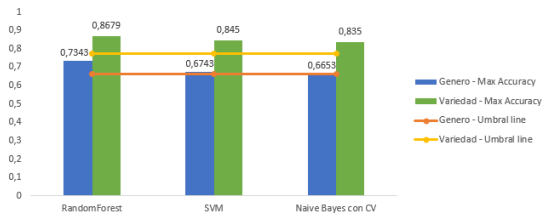

Figura 2: Comparativa de resultados obtenidos y umbrales.

5. Conclusiones y trabajo futuro

Las concluiones de la práctica ha sido que el mejor método para el análisis que hemos encontrado ha sido Random Forest, aunque eso

no ha sido la causa de mejora en el resultado de la predicción, sino el aumento de la bolsa de palabras de 1000 a 2000. Eso también ha conllevado a un aumento considerable en el cálculo de los modelos.

El estudio que se ha hecho de las palabras ha sido muy básico y muy orientado a la variedad, donde no se ha pensado en ningún momento en el género.

Se podría haber aplicado otro algoritmo como e l de S-stemmer. Este algoritmo es muy simple, y básicamente lo que hace es reducir las formas plurales al singular.

References

Mari Vallez y Rafael Pedraza-Jimenez (Universitat Pompeu Fabra) 2007. *El Procesamiento del Lenguaje Natural en la Recuperación de Información Textual y áreas afines. https://www.upf.edu/hipertextnet/numero-5/pln.html*