Asignatura Text Mining en Social Media. Master Big Data

David Montull Crespo
damoncre@inf.upv.es

Abstract

La tarea que se debía realizar consistía en resolver un problema de Author Profiling. En concreto, se pretendía realizar dos tipos distintos de profiling, uno por género de la persona y otro por la variedad de idioma del español que se estaba utilizando. Para poder realizar dicho análisis se partía de una serie de ficheros XML, tanto para training como para test, que contenían una serie de tweets de distintos usuarios, tanto hombres como mujeres, que escriben en distintos idiomas de habla hispana.

En esta tarea, se pretenden utilizar distintos algoritmos de Machine Learning para predecir, como se ha dicho anteriormente, si un tweet ha sido escrito por un hombre o una mujer o bien si un tweet ha sido escrito por alguien de España, Argentina u otros pases de latinoámerica. Para ello, previamente se han de transformar los datos en un formato que pueda ser aplicado a modelos de Machine Learning, tales como Random Forest o Bayesian Generalized Linear Model .

1 Introducción

El problema de Author Profiling pretende identificar ciertos perfiles de un autor u autores partiendo de un dataset. Para el problema presentado en clase se pretende analizar una serie de tweets de la red social Twitter. Estos son, mensajes de un máximo 140 caracteres. Todos estos tweets, están escritos en distintos países de habla hispana por hombres y mujeres. Estos tweets, que se proporcionan en ficheros XML , forman juntos el dataset que utilizaremos para tratar de identificar para un texto concreto, el género de la persona que lo escribe as como la variedad de español que está utilizando.

2 Dataset

El dataset se denomina PAN-AP17 y un subset del mismo ha sido extraído del siguiente enlace . Este dataset consta de tweets de miles de autores de ambos géneros de distintos países de habla hispana. En cada fichero XML aparecen los tweets de un único autor. Estos ficheros XML han sido extrados en base a una serie de ficheros en formato JSON. Para el estudio de este problema, utilizaremos el subset proporcionado que se ha citado anteriormente. En concreto utilizaremos 300 autores (200 para training y 100 para test) y 100 tweets por autor.

3 Propuesta del alumno

El problema que se ha indicado anteriormente se divide en dos partes, el estudio de variedad y del género. Debido a la falta de tiempo para realizar este trabajo por grupos, el autor de este artículo se ha basado más en el estudio de la variedad, aunque los mtodos expuestos pueden ser utilizados también para el análisis de género.

Para intentar conseguir una mejor accuracy en el estudio de la variedad hemos utilizado el algoritmo svmLinear con y sin eliminar stopWords. Adicionalmente, hemos realizado lo mismo pero utilizando cross validation.

Así pues, estas sin las pruebas que hemos realizado:

1- Resultados sin eliminar stopwords (SVMLinear)

2- Resultados eliminando stopwords (SVMLinear)

3- Cross validation (sin eliminar stopwords, SVM)

4- Cross validation (eliminando stopwords, SVM)

Las stopwords que hemos utilizado y que pensabamos que no eran significativas para el análisis de la variedad son:

(”si”,”q”,”gracias”,”hoy”,”ser”,”da”,”mejor”,”jaja”,”xd”)
.

4 Resultados experimentales

Los resultados experimentales que hemos obtenido para las pruebas citadas en el apartado anterior son:

1- Accuracy: 0.7721
2- Accuracy: 0.7693
3- Accuracy: 0.7721
4- Accuracy: 0.7729

Podemos observar que utilizando las stopwords y utilizando el método cross validation con SVM hemos obtenido el mejor resultado.

5 Conclusiones y trabajo futuro

Tras la realizacin de esta tarea hemos podido observar que resolver un problema de Author Profiling no es una tarea fácil, ya que requiere de bastante trabajo previo para pasar los datos al formato que nos interesa para poder aplicar los algoritmos de Machine Learning y que requiere de la instalación correcta de muchas librerías distintas para poder ejecutar cada uno de los modelos distintos que deseemos probar. Adicionalmente, cabe mencionar que el procesamiento de lenguaje natural (NLP) es una tarea compleja, ya que no es nada sencillo detectar en un texto cuando es ambiguo, utiliza ironías o emociones.

Como trabajo futuro se podría utilizar una lista de stopwords ms grande para obtener una mejor accuracy, así como intentar otros algoritmos de Machine Learning para ver como se comporta con estos y si alguno de ellos consigue una mejor accuracy que el que hemos obtenido utilizando la opción 4.

References

[Aho and Ullman1972] Alfred V. Aho and Jeffrey D. Ullman. 1972. *The Theory of Parsing, Translation and Compiling*, volume 1. Prentice-Hall, Englewood Cliffs, NJ.