Text Mining en Social Media. Master Big Data. Pos tagger

Javier Vidal Tellols
javitel@inf.upv.es

13/07/2017

Abstract

En este artculo se va a proponer la implementacin de un programa capaz de realizar el anlisis de texto enriquecido proviniente de una muestra de tweets dada para realizar a posteriori una clasificacin del gnero y del origen del autor de dicho mensaje haciendo uso de una combinacin de anlisis morfosintctico empleando la libreria pos tagger de Stanford y una resolucin de nombres de dominios de las fuentes empleadas extendiendo las urls contenidas en los propios mensajes.

1 Introduccin

Durante el desarrollo del proyecto llegamos a la conclusin de que la inclusin de un anlisis del texto escrito del autor podra aportar riqueza para discernir los perfiles deseados, por ello decidimos incluir adems del resto de lneas de desarrollo, la implementacin de un programa capaz de extraer dicha informacin de los textos dados. Como ya se posea un conocimiento previo de la librera de Stanford se decidi hacer uso de ella. Por otro lado, por sugerencia de uno de los profesores se nos indico que en los enlaces a webs externas podra aportar informacin muy valiosa, como los dominios dnde se encuentran las fuentes que emplean o discernir contenidos webs que ciertos perfiles visiten asiduamente.

2 Dataset

Al plantearse el problema a resolver se decidi observar con una pequea muestra los datos que podan extraerse y la conclusin es que el tamao del problema influye enormemente en esta parte del proyecto, pues la funcion de clasificar las palabras del lenguaje natural no se resuelven de una manera sencilla, requeriendo de mucho tiempo de ejecucin para tener una muestra considerable aadiendo adems un overhead considerable en cada muestra a analizar. Por otro lado, el anlisis de las URLs puede ofrecer dos tipos de informacin, el dominio de origen de dicha web y

el propio dominio que puede aportar informacin sobre los intereses del usuario. El primer caso aporta infomacin sobre el origen de ste, el segundo en cambio, ofrece tamben la capacidad de discernir entre los gustos, probablemente muy significativo a la hora de clasificar por gnero. Por contra, la segunda opcin puede generar una matriz de coincidencia que tiene a infinito, que es la cantidad de dominios existentes en la web.

3 Propuesta del alumno

Los resultados vienen dados en colaboracin con la colaboracin del modelado realizada por Orscar Garibo Orts, decidimos focalizarnos cada uno de los integrantes del grupo en especializarnos en una parte del proyecto para permitirnos llegar a un nivel de profundizacin mayor.

Se implement un cdigo en Python en base al usado en clase para proceder a analizar los tweets dados para realizar el anlisis de los textos, la conclusin es que el tamao de la muestra es excesivo para ser procesado de esta manera, por lo que se decidi implementar en java aplicando los conocimientos previos del lenguaje para realizar dicha tarea implementando tcnicas de paralelizacin escogiendo como unidad mnima de tarea por hilo cada tweet de manera independiente, con lo que conseguimos realizar dicha operacin en un tiempo entre las dos y las 3 horas, la conclusin es que la librera aporta informacin muy detallada sobre las palabras, aadiendo demasiada variabilidad, por lo que finalmente se simplificaron las clasificaciones, discernir entre los tipos de palabras elementales.

A posteriori se decidi incluir en el mismo programa desarrollado debido a que las directivas de paralelizacin aportaban un mejor rendimiento debido a la sobrecarga que supone el tiempo de respuesta en las llamadas HTTP, considerando aquellas secuencias que indicaban una llamada HTTP, se extraan y se resolvan de manera iterada hasta obtener el destino del enlace. La clasificacin en un inicio se decidi realizar haciendo uso del dominio completo de los enlaces, pero se experimentaron problemas de rendimiento debido a la necesidad de mantener en memoria una extructura dinmica de tamao variable que fuese capaz de mantener toda la informacin necesaria. Posteriormente se simplific y solo se almacenaban el dominio de origen (.com, .es, .mx...) y aunque el rendimiento era notablemente mejor, la inestabilidad en las respuestas produjo errores a lo largo de la ejecucin.

4 Resultados experimentales

Los resultados van a ser desglosados en dos casos de uso, ya que se pretende clasificar el autor de un tweet segn su origen y su gnero, dos dimensiones en las que el impacto de la implementacin puede variar. Los resultados se muestran empleando Multi-Layer Perceptron (MLP), ya que es el modelo que mejores resultados ha presentado.

En la variedad no se ve demasiado afectada la inclusin de los datos de pos tagger aadiendo ruido al modelo planteado. Tal y como se puede observar a continuacin.

MLP	Accuracy
TfIdf 0.90 Unigrams no accents stem	91.64
TfIdf 0.90 Unigrams no accents TAGS	**91.00**

Table 1: MLP + Stem and TAGS.

En el caso del gnero, en cambio, la inclusin de los datos del pos tagger han producido un incremento considerable.

MLP	Accuracy
TfIdf 0.90 + TAGS + WC	**77.43**
TfIdf 0.90 + WC	75.14

Table 2: Common:MLP + keep punctuation.

El anlisis de URLs no ha sido concluyente, ya que el tamao de la muestra ha hecho imposible realizar el proceso completo sin que presentase problemas de estabilidad en las peticiones, generando errores en los resultados.

5 Conclusiones y trabajo futuro

La conclusin obtenida de este estudio nos indica que en efecto, el anlisis de lenguaje natural aporta un valor notable en la clasificacin en cunto a gnero, no obstante el coste computacional de dicho anlisis hace que el uso de estas tcnicas sean de manera controlada y en casos en los que o se requiera de un aumento de precisin a cualquier coste o que el tiempo de respuesta no sea un requisito indispensable.

Como trabajo futuro se observan dos grandes caminos, por un lado la simplificacin de los resultados del pos tagger, proponiendo una adaptacion posterior para no pdoer parte de la informacin del anlisis y mejorar la parte del sofware dedicada a la resolucin de las direcciones a pginas web.

References

[1] 2017-06-09. *https://nlp.stanford.edu/software/tagger.shtml*. Version 3.8.0 .