Text Mining II. Máster Big Data Analytics

Miguel Ángel Sotomayor Fernández
masfworld@gmail.com

Abstract

El objetivo de este artículo es la generación de un modelo de predicción de la variedad del lenguaje y género de las personas en base a unos tweets. Para ello usamos un corpus donde la variedad del lenguaje y el género es conocido. La idea principal parte de una bolsa de palabras. Esta BOW se establece analizando todo el corpus. Una vez formada esta BOW, se ordena y se analiza las 1000 primeras con respecto a cada una de las entradas del corpus, calculando su frecuencia de aparición. Partiendo de este trabajo se ha analizado otras posibilidades, en concreto la generación N-gramas en vez de el BOW.

1. Introducción

El problema de *Author Profiling* se basa en la determinación del género, edad, lenguaje nativo, tipo de personalidad de uno o varios individuos. Este tipo de problema está en auge en muchas areas, tales como las redes sociales y análisis forenses, siendo en este último muy importante para determinan perfiles psicológicos de posibles sospechosos. Otro beneficio podría ser la identificación de los perfiles de los clientes para las empresas.

En este artículo no se pretende introducir una nueva metodología, sino por el contrario, usar diversas técnicas de manera eficiente para la identificación de ciertos parámetros que hagan posible la identificación de un sujeto en base a un texto escrito, en este caso tweets.

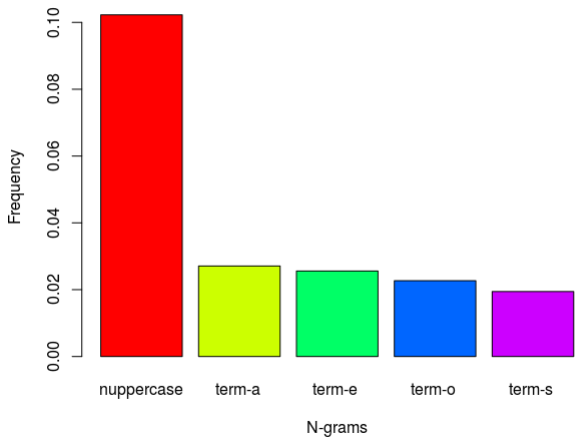
2. Dataset

El dataset origen es el conjunto de tweets (Hispatweets) proporcionado por el profesorado.

A partir de este dataset se ha generado una composición de N-gramas de 2 a 5 caracteres. En donde cada uno de esos N-gramas tiene una frecuencia de aparición dentro del corpus original, siendo este nuestro dataset de entrada al algoritmo de predicción. Por tanto, vamos a analizar el dataset correspondiente a los N-gramas y su frecuencia de aparición.

Para el análisis de este dataset se ha usado la herramienta R. El código fuente está dentro de la carpeta R de la solución.

Se va a analizar las medias de la aparición de cada uno de los N-gramas en el corpus. Se van a mostrar los 5 N-gramas más frecuentes:



Como podemos observar en la gráfica, los N-gramas de vocales comunes en la lengua castellana son muy frecuentes con respecto a otros. ”nuppercase”, es una medida adicional a los N-gramas que indica frecuencia de mayúsculas en los textos.

3. Propuesta del alumno

Se comenzó analizando la eficiencia de un BOW. Se vio que esta colección de palabras no representaba muy bien ciertos detalles, tales como terminaciones de palabras.

Por ejemplo, terminaciones como “ico”, “ito”, “ás”, “és” o “ís”, presente en palabras como “cochecito, perrico, hablás“ pueden ser utilizados por

individuos de una determinada zona geográfica. Por ejemplo, una palabra que podría identificar la procedencia es “vos”, ampliamente utilizado en Argentina.

Otro aspecto que los N-gramas pueden identificar más claramente que la bolsa de palabras son los géneros, ya que fácilmente identificará frecuencias de aparición en texto de terminaciones como ”la, lo”.

Hay otras características distintas a los N-gramas que se han tenido en cuentas. En concreto para este trabajo se ha tenido en cuenta las siguientes dimensiones:

- Frecuencia de comas.
- Frecuencia de puntos.
- Frecuencia de dos puntos.
- Frecuencia de exclamaciones.
- Frecuencia de interrogaciones.
- Frecuencia de mayúsculas.
- Frecuencia de palabras distintas: Para implementar esta característica de manera eficiente, se crea un diccionario donde las claves son todas las palabras de un texto. A continuación se cuenta el número de claves dentro del diccionario.
- Frecuencia de risas: Se identifica el número de palabras que pueden significar risa, por ejemplo este conjunto {”jeje”, ”jaja”, ”xdd”, ”hehe”, ”haha”, ”jiji”, risa”}.
- Terminaciones: Se identifica el número de terminaciones pertenecientes a este conjunto {”ás”, ”és”, ”ís”}.
- Terminaciones en “ico”: Se identifica el número de palabras con dicha terminación.
- Terminaciones en “ito”: Se identifica el número de palabras con dicha terminación.
- “vos”: Se identifica el número de apariciones de la palabra “vos”.

Para todas estas características se ha calculado su frecuencia con respecto al texto de entrada y se han añadido como dimensión adicional a la frecuencia de N-gramas.

Hay que tener presente los dos tipos de N-gramas. Por un lado tenemos los N-gramas de caracteres y por otro N-gramas de palabras. Por ejemplo, para el caso de las palabras, podemos considerar un trigrama el siguiente texto: ”Hola como estas”. Para este trabajo hemos considerado que los N-gramas de caracteres podían dar mejores resultados

4. Resultados experimentales

A continuación se muestra una tabla con los resultados en base a distintas pruebas que se han hecho con diferentes características:

Técnica	Alg. Aprendizaje	Country	Gender
N-gramas 3 carac.	Naive Bayes	39.975 %	49.290 %
N-gramas 3 carac.	Super Vector Machine	86.058 %	75.879 %
N-gramas 3 carac. + terminaciones “ás”, “és”, “ís” + terminaciones “ico”, “ito”+ “vos”	Naive Bayes	40.160 %	48.982 %
N-gramas 3 carac. + terminaciones “ás”, “és”, “ís” + terminaciones “ico”, “ito”+ “vos”	Super Vector Machine	85.746 %	75.694 %
N-gramas 3 carac. + terminaciones “ás”, “és”, “ís” + terminaciones “ico”, “ito”+ “vos”	Random Forest	75.694 %	71.931 %
N-gramas 2-5 carac.	Naive Bayes	33.682 %	47.008 %
N-gramas 2-5 carac.	Super Vector Machine	84.022 %	76.311 %
N-gramas 2-5 carac. + terminaciones “ás”, “és”, “ís”	Naive Bayes	39.975 %	47.008 %
N-gramas 2-5 carac. + terminaciones “ás”, “és”, “ís”	Super Vector Machine	83.529 %	76.558 %
N-gramas 2-5 carac. + terminaciones “ás”, “és”, “ís” + terminaciones “ico”, “ito”	Naive Bayes	39.913 %	48.365 %
N-gramas 2-5 carac. + terminaciones “ás”, “és”, “ís” + terminaciones “ico”, “ito”	Super Vector Machine	97.223 %	86.181 %
N-gramas 2-5 carac. + terminaciones “ás”, “és”, “ís” + terminaciones “ico”, “ito”+ “vos”	Naive Bayes	33.806 %	47.07 %
N-gramas 2-5 carac. + terminaciones “ás”, “és”, “ís” + terminaciones “ico”, “ito”+ “vos”	Super Vector Machine	84.207 %	76.620 %

Cuadro 1: Resultados (*Accuracy*)

Todas las pruebas llevan implícitas las siguientes características:

- Frecuencia de comas
- Frecuencia de puntos
- Frecuencia de dos puntos
- Frecuencia de exclamaciones e interrogaciones
- Frecuencia de risas

Se puede observar que la inclusión de los N-gramas supone un aumento “considerable”del *accuracy*. Casi todas las demás características no aportan mucho, algunas mejoran un poco, pero no es considerable el aumento. Quizás se puede observar una ligera mejoría en los N-gramas de 3 caracteres cuando examinamos la variación del lenguaje.

La verdadera característica que aporta valor es cuando ponemos la terminaciones “ico” e “ito”. Quizás al mezclarla con “vos” empeora porque esta última es sólo representativa de unas pocas zonas.

Como era de esperar si cambiamos un algoritmo probabilístico, como es el ”Naive Bayes”, por un Super Vector Machine mejoramos notablemente. Random Forest también es mejor que “Naive Bayes” pero no tanto como Super Vector Machine

5. Conclusiones y trabajo futuro

Se ha podido observar que los N-gramas son un buen punto de partida a la hora de realizar “Author profiling”. Sin embargo todavían quedan múltiples características a explorar y modelos que generar.

En el futuro sería recomendable probar a mezclar en un mismo dataset N-gramas de caracteres y de palabras, para más tarde aplicar un algoritmo PCA de reducción de dimensiones.

Por otro lado, también se podría incluir más variaciones lingüisticas de cada una de las zonas geográficas.

References

F. Rangel, P. Rosso, I. Chugur, M. Potthast, M. Trenkmann, B. Stein, B. Verhoeven, W, Daelemans. 2014. *Overview of the 2nd Author Profiling Task at PAN 2014*

Juan Soler Company 2013. *Author Profiling:Gender Identification*