Preprocesamiento de datos, integración y gestión de datos mediante una DB NOSQL

Raúl Carlomagno, Franco Catania

**Resumen.** El presente trabajo consiste en aplicar diferentes técnicas de minería de datos a un dataset proveniente de Tweeter. Realizando desde un análisis exploratorio de datos hasta minería de texto, pasando por consultar una base de datos nosql, preprocesar datos y la construcción de un modelo predictivo. Realizando esta pila de tareas, se extraerá cierto conocimiento de los datos que a simple vista no son identificables.

1 Introducción

La aplicación de técnicas de datamining es realizada sobre datasets de diferentes temas y origenes. A su vez los datos pueden ser estructurados, por ejemplo cuando son extraidos de una base de datos relacional, o no estructurados, por ejemplo una imagen o discurso de una persona.

Para el trabajo en cuestión se partió de 2 datasets provenientes de tweeter. El primero de ellos consiste en un stream de 600 tweets capturados con R con el término “gripe” y guardados en formato [Newline Delimited JSON](http://ndjson.org/). El otro dataset consiste en 600 elementos tambien en formato NDJSON, que hacen referencia a los usuarios que originaron los tweets.

Como primera aproximación se realizó un análisis exploratorio de los datos para comprender y adentrarnos en la información capturada. Para realizar esa tarea se parsearon los archivo en formato JSON y se trabajó con dataframe en memoria. A su vez, hubo algunos problemas de encoding con los archivos, pero fueron solucionados. Ya con los datasets en memoria se pudieron describir los datos, conocer las distribuciones de las variables continuas y sus respectivas correlaciones. Los archivos JSON también fueron cargados en MongoDB para realizar consultas sobre los mismos.

Una vez conocidos los datos se realizó una tarea de preprocesamiento para eliminar variables que no afectaban a la variabilidad de los datos, debido a correlacion por ejemplo. Tambien se crearon algunas variables nuevas que podrían ayudar a la construcción de un modelo de predicción. Como variable target binaria (“popular”) se utilizó el criterio positivo si es que el tweet tenia minimo algún retweet y si fue por los menos una vez favorito.

Ya con los datos preprocesados se creó un modelo de clasificación que ayude a predecir si el tweet es popular o no.

Por último y no menos importante se realizaron diferentes técnicas de text mining para concluir en una [Document-term matrix](https://en.wikipedia.org/wiki/Document-term_matrix) sobre el texto del tweet y conocer que términos fueron los más utilizados.

2 Materiales y Métodos

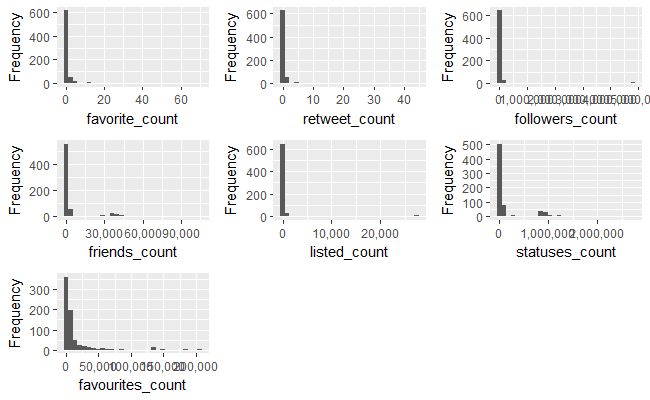
Estructura del dataset inicial de tweets:

|  |  |  |
| --- | --- | --- |
| Nombre | Tipo de variable | Descripcion |
| status\_id | Numerica discreta | Identificador del tweet |
| created\_at | Fecha con hora | Fecha de creacion del tweet |
| text | Texto | Texto del tweet |
| source | Categorica nominal (27) | Fuente que originó el tweet |
| is\_quote | Booleana | Flag si el tweet es una cita |
| is\_retweet | Booleana | Flag si el tweet es retweet |
| favorite\_count | Numerica discreta | Contador de veces que el tweet fue favorito |
| retweet\_count | Numerica discreta | Contador de veces que el tweet fue retweeteado |
| hashtags | Lista de texto | Hashtags que contiene el tweet |
| urls\_url | Lista de texto | Urls cortas en el tweet |
| urls\_t\_co | Lista de texto | Urls t.co (short url) en el tweet |
| urls\_expanded\_url | Lista de texto | Urls full en el tweet |
| media\_url | Texto | Url de la media del tweet |
| media\_t\_co | Texto | Url t.co (short) de la media del tweet |
| media\_type | Texto (“photo” o NA) | Tipo de media |
| ext\_media\_url | Texto | Url de la media del tweet |
| ext\_media\_t\_co | Texto | Url t.co (short) de la media del tweet |
| ext\_media\_expanded\_url | Texto | Url full de la media del tweet |
| mentions\_user\_id | Lista numérica discreta | Identificadores de los usuarios mencionados en el tweet |
| mentions\_screen\_name | Lista de texto | Nicknames de los usuarios mencionados |
| lang | Texto | Lenguaje del tweet |
| reply\_to\_status\_id | Numerica discreta | Identificador del tweet al que responde |
| reply\_to\_user\_id | Numerica discreta | Identificador del usuario al que le responde |
| reply\_to\_screen\_name | Texto | Nicknames de los usuarios a los que le responde |
| place\_url | Texto | Url del lugar donde se generó el tweet |
| place\_name | Texto | Nombre del lugar donde se genero el tweet |
| place\_full\_name | Texto | Nombre full del lugar donde se genero el tweet |
| place\_type | Categorica nominal (3) | Tipo de lugar donde se generó el tweet |
| country | Categorica nominal (8) | País donde se generó el tweet |
| country\_code | Categorica nominal (8) | Codigo ISO del país donde se generó el tweet |
| lng | Numerica continua | Longitud de la posición donde fue originado el tweet |
| lat | Numerica continua | Latitud de la posición donde fue originado el tweet |
| quoted\_text | Texto | Texto citado |

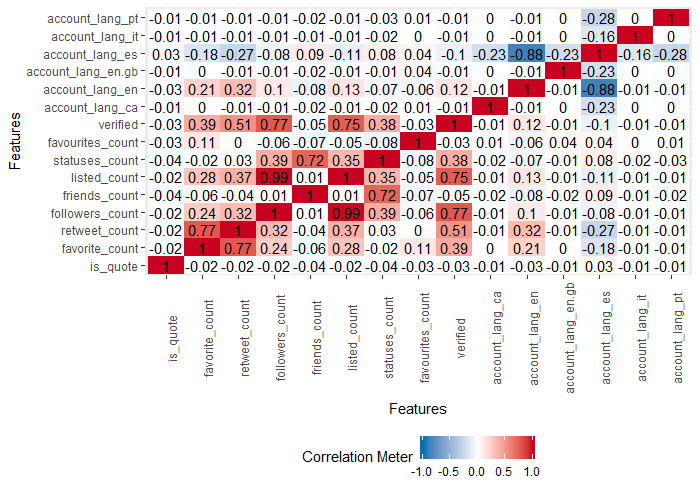
Estructura del dataset inicial de usuarios:

|  |  |  |
| --- | --- | --- |
| Nombre | Tipo de variable | Descripcion |
| user\_id | Numerica discreta | Identificador del usuario |
| screen\_name | Texto | Nickname del usuario |
| name | Texto | Nombre del usuario |
| location | Texto | Ubicación del usuario |
| description | Texto | Descripción del usuario |
| url | Texto | Url del usuario relacionada con su perfil |
| followers\_count | Numerica discreta | Cantidad de seguidores |
| friends\_count | Numerica discreta | Cantidad de amigos |
| listed\_count | Numerica discreta | Cantidad de listas publicas que el usuario es miembro |
| statuses\_count | Numerica discreta | Cantidad de tweets (incluyendo retweets) que hizo el usuario |
| account\_created\_at | Fecha con hora | Fecha de creación de la cuenta |
| verified | Booleana | Flag si la cuenta esta verificada |
| profile\_url | Texto | Url corta del perfil del usuario |
| profile\_expanded\_url | Texto | Url full del perfil del usuario |
| account\_lang | Categorica nominal (6) | Codigos ISO del lenguaje de la cuenta |
| profile\_banner\_url | Texto | Url del banner del usuario |
| profile\_background\_url | Texto | Url del fondo de perfil del usuario |
| profile\_image\_url | Texto | Url de la imagen perfil del usuario |

Distribución de variables:



Análisis de correlación:



Como parte del análisis e introducción a los datos se consultó la base de datos nosql MongoDB donde previamente se habían importado los 2 datasets en formato JSON. Consultando la base se obtuvo que:

* El 28,16% de los tweets tienen por lo menos una marca RT o fav
* Los usuarios Roxieli (946429732664602624), Fla17\_marielSF (451419035), Vacunar (479701640), Globovisión (17485551), Reportes24H (720571556909490177) son los usuarios con más tweets, poseen 3 tweets cada uno.
* 99 tweets fueron enviados desde la aplicación web de tweeter
* Hay 27 usuarios con cuenta en ingles, 565 en español, 3 en portugues, 1 en italiano, 2 en catalán y 2 en ingles británico.
* El usuario con mas followers es Globovisión de Venezuela y sus 3 tweets no están clasificados como “popular”, por lo tanto no hay una relación directa entre followers y popularidad de un tweet.

Habiendo realizado el análisis de los datos se procedió a trabajar sobre los datos y preprocesarlos para obtener un dataset mas “limpio” y optimo para la creación de un modelo de clasificación.  
Para ello se trabajó sobre los 2 datasets individualmente y luego se los unió respecto al campo user\_id, para llegar al dataset final procesado.

Del dataset de tweets se eliminaron varias variables y se crearon unas nuevas.  
A grandes rasgos, las variables que eran texto puro (salvo el texto del tweet) se eliminaron, como por ejemplo el nombre del usuario del tweet, todo lo relacionado al lugar de origen del tweet, las fechas (se pudo haber trabajado con las fechas y calcular por ejemplo una variable nueva que sea la diferencia de días entre la fecha de creación del tweet y la fecha de creación del perfil del usuario), los identificadores, etc. La variable source fue tenida en cuenta y preservada en el dataset.  
Las nuevas variables consisten en tomar las variables que eran listas originalmente, por ejemplo la lista de hashtags, de usuarios mencionados, de urls, etc. Y generar nuevas variables con la cantidad de ítems de esa lista. Tambien se crearon 2 variables, reply\_to\_user y reply\_to\_tweet, binarias, originadas por el id del tweet o id del usuario al que respondia el tweet.

Del dataset de usuarios se eliminaron también las texto puro, como urls, nombre, descripción, fecha de creación, etc. Se dejaron solo las variables de cantidades, la de cuenta verificada y el lenguaje de la cuenta.

Una vez unidos ambos datasets se trabajó sobre la posibilidad de cambiar el criterio de la variable target. Originalmente dado el enunciado se tuvo en cuenta la variable target como positiva si el tweet tuvo algún retweet y si fue marcado como favorito por lo menos una vez.

Se aplicó [análisis de componentes principales](https://es.wikipedia.org/wiki/Análisis_de_componentes_principales) sobre el dataset transformado para evaluar la posibilidad de tener otro criterio de calificación para un tweet popular.

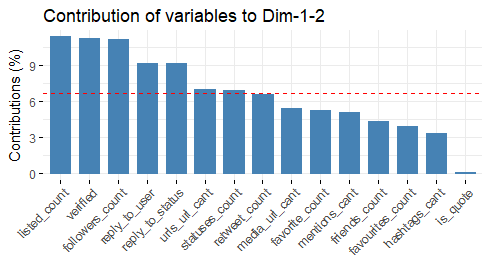
Un modelo de clasificación basado en el algoritmo [CART](https://es.wikipedia.org/wiki/Aprendizaje_basado_en_árboles_de_decisión) fue creado con el menor número de variables independientes posibles, comenzando a construirlo con las variables mas importantes una por una, este método es conocido [como Fast Forward Feature Selection](https://en.wikipedia.org/wiki/Feature_selection). El orden de importancia de las variables fue generado gracias a el algoritmo [Learning vector quantization](https://en.wikipedia.org/wiki/Learning_vector_quantization). Para evaluar el modelo se utilizó la métrica [AUC](https://es.wikipedia.org/wiki/Curva_ROC) (área bajo la curva).

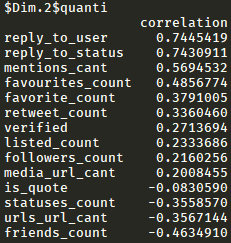
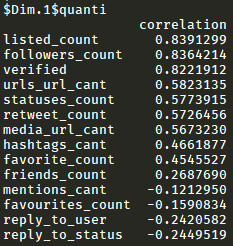
Se trabajó con herramientas de text mining para obtener una nube de texto representativa en frecuencia por cada termino del documento, las 50 mas importantes. Para llegar a la matriz de términos frecuentes hubo que realizar diferentes transformaciones sobre el corpus, filtrar palabras, remover espacios, puntos, realizar stemming, etc. Para poder obtener un corpus lo mas puro posible y realizar el análisis sobre el mismo.

3 Resultados obtenidos

Teniendo en cuenta el enunciado inicial de la variable “popular”, la distribución de sus valores es: NO (93,20%) y SI (6,80%). Por lo tanto si intentamos otra alternativa para la variable target deberá por un lado, respetar esa proporción y por otro lado, poseer casi los mismos valores para las muestras.

Realizar un análisis de componentes principales sobre el dataset nos brinda la siguiente información respecto a las contribuciones de las variables para los componentes 1 y 2:



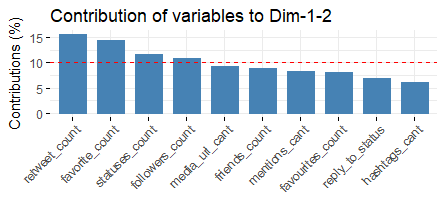


Con esta información brindada por el análisis de componentes principales y contrastada por el análisis de correlación, se procede a eliminar ciertas variables ya se por correlacion o por falta de contribucion:

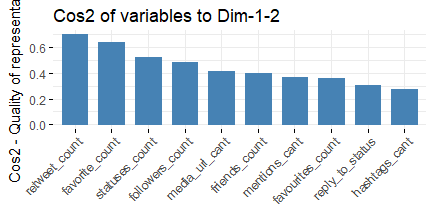
reply\_to\_user, listed\_count, verified, urls\_url\_cant y is\_quote

Una vez eliminadas las variables se procede a realizar un nuevo analisis de componentes para ver ahora que podemos tomar de las variables existentes para revaluar el concepto de “popular”.

La nueva contribución para los componentes 1 y 2 es:



y la calidad de representación de las variables para los componentes 1 y 2 (mensurada por la metrica cos2) es:



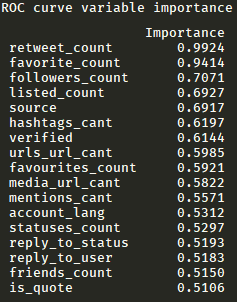
Lo que observamos es que originalmente era válida y acertada la propuesta de establecer las variables retweet\_count y favorite\_count representativas del concepto “popular”. Pero en 2do lugar, disponemos de las variables statuses\_count y followers\_count para establecer el nuevo criterio, estas variables pertenecían al dataset de usuarios, por lo tanto el nuevo criterio queda dependiente de variables del usuario.

La elección del nuevo criterio trae aparejado la cuestión del umbral que le debemos dar a estas variables para establecer el concepto de “popular”.

Para establecer el umbral se optó por el promedio de cada uno de estas variables, por lo tanto el concepto de “popular” queda definido tweet, cuyo usuario creador, posee cantidad de tweets mayor al promedio de tweets y cantidad de followers mayor al promedio de followers. De esta manera, la distribución de la variable “popular” es: NO (94,07%) y SI (5,93%). Luego se comparó esta nueva variable target “popular” con la original, el área bajo la curva de esta nueva variable comparada con la anterior fue de 0,89. Se podría afirmar que esta nueva variable respeta aproximadamente el comportamiento de la variable target original.

La siguiente parte del trabajo de creación de un modelo fue realizado con el dataset que poseía la variable target original “popular”, ya que al hacerlo con la nueva variable target propuesta, quedaba muy sesgado por el nuevo criterio de “popularidad” y la efectividad era del 100% con una sola variable independiente, se hicieron ambas pruebas y se decidió desarrollar la del dataset con la variable target original. Al dataset procesado se lo fraccionó en 2 partes (80% [train] y 20% [test]) y se entrenó un modelo de clasificación con el algoritmo Learning vector quantizitation, pero este modelo no se utilizó para clasificar, sino para obtener la importancia de las variables del dataset y construir un nuevo modelo de árboles de decisión CART realizando un proceso de Fast Forward Feature Selection, donde se va agregando de a una variable al modelo hasta que el modelo no necesito el agregado de más variables, ya que se va evaluando su performance con la variable agregada y si la performance no mejora con el agregado, nos quedamos con el modelo anterior sin la última variable agregada.

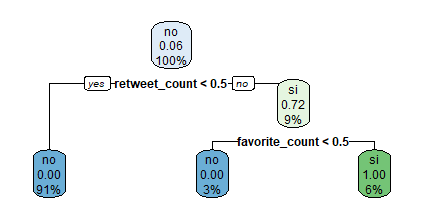
La importancia de las variables entregadas por LVQ fue:



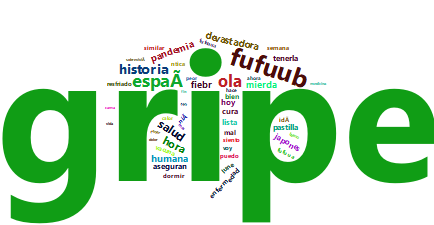
Próximo paso fue empezar a entrenar iterativamente modelos CART con cross validation = 10, agregando variables una por una hasta llegar al corte dependiendo de la performance, medida en este caso en AUC.

El primer modelo entrenado solo con la variable retweet\_count tuvo una AUC de 0,96, y ya el próximo modelo entrenado con retweet\_count + favorites\_count tuvo una AUC de 1.

En el siguiente grafico podemos ver el árbol elaborado por CART con las 2 variables independientes:



El resultado del proceso de text mining fue la siguiente nube de palabras, donde la frecuncia de la palabra está relacionada con el tamaño de la misma en la nube.



Referencias

|  |
| --- |
|  |
| Kohonen, T. (1995). Learning vector quantization. In *Self-Organizing Maps* (pp. 175-189). Springer, Berlin, Heidelberg. |
| Chicago |  |