# Data Science

## Pretratamiento de tweets

### Recogida de tweets

Se ha obtenido un conjunto de tweets mediante la API de Twitter en lenguaje R.

#### Criterios de filtrado previo

Se ha centrado la búsqueda de tweets en un área centrada en Vitoria-Gasteiz con un radio de 10km.

El tema al que hacen referencia los tweets se ha centrado en aquellas cuestiones que se han identificado como frecuentes en las quejas vertidas en el buzón del ciudadano del Ayuntamiento de Vitoria – Gasteiz. Esto se hace porque es más probable encontrar quejas sobre dichos temas que en el universo de tweets generados en Vitoria. Ello no significa, no obstante, que todos los tweets recogidos sean quejas, por lo que será necesaria una posterior clasificación. Las palabras clave utilizadas han sido las siguientes:

* Accesibilidad
* Accidente
* Alcantarilla
* Anillo verde
* Aparcabicis
* Asco
* Ataque
* Atasco
* Autobús
* Basura
* Basureros
* Bici
* Bus
* Calles
* Ciclista
* Contaminación
* Contenedor
* Green
* Jardines
* Limpieza
* Medio ambiente
* Olor
* Palomas
* Papelera
* Pelea
* Pintadas
* Problema
* Ratas
* Reciclaje
* Reciclar
* Renta
* Robos
* Rotonda
* Ruido
* Suciedad
* Sucio
* Tráfico
* Tranvía
* Violencia
* Vitoria
* Vitoria Gasteiz

#### Obtención de tweets mediante código en R

Se ha utilizado la API de Twitter para R, especificando, como se ha comentado anteriormente:

* la geolocalización mediante el parámetro *geocode*
* las palabras claves mediante el parámetro *searchString*
* el número máximo de tweets a devolver mediante el parámetro *n*

### Clasificación manual de tweets

Para que la el algoritmo de clasificación de tweets pueda predecir si un tweet es una queja o no, debe ser previamente entrenado y testeado con una colección de tweets que han sido clasificados manualmente.

#### Criterios de clasificación

La consideración de un tweet como queja presenta una serie de dificultades que deben ser manejadas de una manera sistemática. El criterio principal es considerar como demanda ciudadana cualquier tweet que plantee una queja, información o sugerencia que los gestores del ayuntamiento puedan utilizar para mejorar el desempeño municipal. En casos que presentan dudas, se han aplicado los siguientes criterios de clasificación:

* No se consideran quejas que no hagan referencia a Vitoria – Gasteiz
* Las quejas relativas al funcionamiento de los grupos políticos del ayuntamiento y a partidos políticos no pueden ser solucionados por los gestores técnicos y, por tanto, no pueden considerarse quejas en este proyecto.

Por otra parte, la asignación del tema de la queja es más compleja porque existen varias categorías de clasificación, a saber:

* Limpieza: aquellas demandas que hacen referencia exclusivamente a la limpieza viaria.
* Espacio público: todas las demandas referidas al estado del espacio público que no tengan que ver con la limpieza. Por ejemplo, el estado del mobiliario urbano, parques, arbolado, edificios, etc.
* Administración pública: quejas referidas a la actuación del ayuntamiento y empresas públicas en la gestión de trámites administrativos
* Movilidad: quejas o información sobre los servicios de transporte público, movilidad peatonal o ciclista e incidencias viarias.
* Delincuencia: quejas referidas a la percepción de la delincuencia. Por ejemplo, inseguridad, robos, agresiones, estafas, violencia de género, etc.

Idealmente, la clasificación de un mismo tweet debería realizarse por varios evaluadores, de manera que exista un consenso sobre la clasificación correspondiente. En este caso, dadas las limitaciones de tiempo, cada tweet ha sido evaluado por un solo evaluador, algo que se corregirá en fases posteriores del proyecto.

El conjunto de tweets identificado como queja es bajo. De los aproximadamente 1400 tweets clasificados manualmente, solamente 200 han sido identificados como quejas, dando lugar a un dataset desbalanceado que deberá ser tratado para corregirlo.

En fases posteriores del proyecto, sería conveniente abordar la clasificación de tweets mediante una campaña online de colaboración ciudadana. De esta forma, además de obtener un dataset con el que entrenar mejor al modelo de clasificación mucho más amplio y en el que cada tweet es evaluado por varios evaluadores, se fomenta la sensación de que el ayuntamiento tiene en cuenta la opinión de la ciudadanía en la gestión municipal.

### Limpieza de tweets

#### Limpieza de caracteres

Se han eliminado todos los caracteres extraños de los tweets, esto es:

* tags (por ejemplo @vitoriagasteiz)
* enlaces
* signos de puntuación
* números
* tabulaciones
* espacios al principio y al final
* caracteres no alfanuméricos
* emoticonos
* espacios en blanco

Además, todas las mayúsculas se han convertido a minúsculas

#### Lematización

Para la lematización, se ha recurrido a un diccionario español de palabras raíz desarrollado por el proyecto *Global Glosary*. Por tanto, la lematización se realiza comparando las palabras contenidas en cada tweet con el diccionario y devolviendo las palabras raíz correspondientes (verbos en infinitivo, adjetivos y sustantivos en singular, etc. Por ejemplo, los verbos “agotáramos” y “agotaría” se sustituirían por su infinitivo “agotar”)

#### Eliminación de palabras comunes

Se ha optado por eliminar las palabras comunes después de comprobar que palabras que intuitivamente relacionamos con quejas como “no” tienen una frecuencia relativa similar entre los tweets quejas y los tweets no queja. Por tanto, no tienen una influencia decisiva en la creación del algoritmo de clasificación.

## Entrenamiento y testeo de un modelo Naïve Bayes para clasificación como “queja” o “no queja”

### Tratamiento del dataset de entrenamiento y testeo

El clasificador Naïve Bayes obtiene una precisión baja cuando el dataset utilizado en el entrenamiento y testeo del modelo no son los adecuados. En consecuencia, el dataset debe ser corregido para obtener una buena precisión. Los principales problemas a este respecto que hemos encontrado durante esta fase del proyecto han sido los siguientes:

* Contenía muchos más tweets clasificados manualmente como no-queja que como queja, en una proporción aproximada de 7:1
* Cada palabra del corpus se repite pocas veces en los diferentes tweets y, por tanto, la matriz término-documento obtenida es demasiado dispersa.

#### Corrección del desbalanceo queja/no-queja

El dataset original contenía un total de 1393 tweets de los que 208 eran quejas (15%). Se han probado 3 proporciones queja:no-queja diferentes (la proporción original 1:7, 3:7 y 1:1). Se ha optado por reducir el número de no-quejas hasta alcanzar una proporción del 50%. Los tweets eliminados se han elegido de manera aleatoria.

#### Corrección de la dispersión de la matriz término-documento

El dataset original contenía 4049 palabras diferentes, cada una de las cuales constituye una columna de la matriz término-documento. El objetivo de esta corrección es reducir el número de elementos vacíos en dicha matriz.

Se han probado 3 métodos diferentes para reducir la matriz término documento:

* removeSparseTerms: elimina los términos dispersos de una matriz término-documento, descartando aquello que tienen al menos determinado porcentaje de elementos. Función de la librería *tm.*
* topfeatures: devuelve una lista de los términos más frecuentes de una matriz término-documento, especificando un número de términos a devolver. Función de la librería *quanteda.*
* findFreqTerms: devuelve los términos que se repiten con una frecuencia mayor o igual a la especificada. Función de la librería *tm.*

Para ello, se han seleccionado las palabras que se repiten 2 o más veces en el conjunto de tweets. De esta manera, se consigue reducir el número de palabras diferentes y, por tanto, de columnas hasta 479.

### Validación cruzada de K iteraciones

Se ha utilizado la validación cruzada de K iteraciones o *K-fold cross-validation* para mejorar la precisión del modelo predictivo. Uno de los subconjuntos se utiliza como datos de prueba y el resto (K-1) como datos de entrenamiento. El proceso de validación cruzada es repetido durante k iteraciones, con cada uno de los posibles subconjuntos de datos de prueba. Finalmente se realiza la media aritmética de los resultados de cada iteración para obtener un único resultado. En nuestro proyecto, se ha escogido un valor de K=5.

### Analisis de sensibilidad de los parámetros del algoritmo de Naïve Bayes

La fase de ajuste en Naive Bayes consiste en encontrar mediante iteración el mejor valor del parámetro para el suavizado de Laplace, de entre un conjunto de valores candidatos, probando los distintos rendimientos en el conjunto de validación. Nuestras observaciones concluyen que el mejor valor para Laplace en esta fase de nuestro proyecto es 1.

### Codigo del algoritmo

## Entrenamiento y testeo de un modelo Naïve Bayes para clasificación temática

### Tratamiento del dataset de entrenamiento y testeo

Al igual que con el clasificador queja/no-queja, hay que corregir el dataset para el clasificador temático. En este caso, el dataset no está muy desbalanceado entre los diferentes temas, que aparecen as siguientes veces:

* Limpieza: 43 veces
* Espacio público: 29 veces
* Administración pública: 25 veces
* Movilidad: 85 veces
* Delincuencia: 26 veces

Por tanto, solamente se ha corregido la dispersión de la matriz término-documento.

#### Corrección de la dispersión de la matriz término-documento

El dataset original contenía 911 palabras diferentes y, consecuentemente, el mismo número de columnas. Se han reducido con la función findFreqTerms de la librería *tm*, seleccionando las palabras que se repiten 2 o más veces en el conjunto de tweets hasta quedarnos únicamente con 289 palabras.

### Validación cruzada de K iteraciones

Se ha procedido de igual forma que con el clasificador queja/no-queja.

### Analisis de sensibilidad de los parámetros del algoritmo de Naïve Bayes

Tras probar con distintos valores del parámetro de Laplace, nuestras observaciones concluyen que la mejor precisión en esta fase de nuestro proyecto se consigue con un Laplace = 0.

### Codigo del algoritmo

## Bibliografía

<http://www.lrec-conf.org/proceedings/lrec2014/pdf/292_Paper.pdf>

<http://www.vanessafriasmartinez.org/uploads/socialcomFinal.pdf>