# Clustering

## Clasificación no supervisada.

Parece lógico que nos pregunten por qué razón no utilizamos clasificación no supervisada si es muy difícil etiquetar instancias.

Antes de responder, veamos cómo podría implementarse esto sobre un conjunto de entrenamiento con una relación de clases 1:1000.

Partimos de D documentos que forman el corpus

## K-medias.

Podríamos plantearnos el algoritmo de k-medias. Para ello cada documento de la muestra de tamaño *D* se identifica por un vector de atributos que, debería estar relacionado con los términos del vocabulario cuyas componentes podrían ser los tf*-idf* de cada término del documento.

Este vocabulario se ha obtenido tras la depuración, eliminación de palabras vacías y troncalización (*stemming*).

Tras esto, cada documento estaría representado por un vector (*d*=1,…, D)

En el problema que estamos tratando que incluye desequilibrios de clase, resulta inimaginable que el algoritmo de *clustering* permita clasificar por sí solo tuits de odio.

Si recordamos el funcionamiento de este algoritmo:

Para aplicar K-medias hemos de

1. Elegir una medida de distancia.
2. Elegir k centroides iniciales coincidentes con otros tantos elementos aleatoriamente seleccionados el conjunto
3. Asignar el resto de las instancias al centroide (*cluster*) más próximo.
4. Recalcular los centroides para cada cluster y reasignar las instancias.
5. Iterar esto hasta que se consiga una solución en la que ninguna instancia se reclasifique.

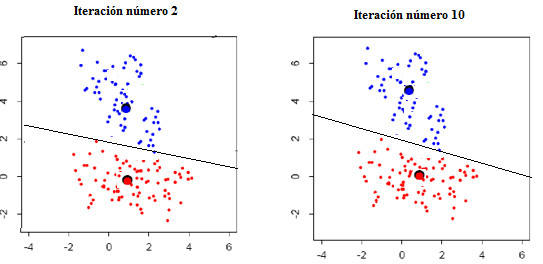
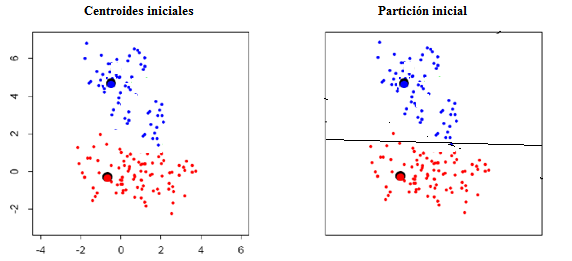


Fig. . Algoritmo 2-medias

Respecto a la distancia los más comunes son:

1. Distancia euclídea:
2. Distancia coseno:
3. Coeficiente de correlación de Pearson

Siendo los dos últimos los que mejor resultado dan.

### Problemas con el algoritmo.

Recordemos que nos enfrentamos a un problema de clasificación *binario* con clases *desequilibradas*. Por tanto, k debería ser 2.

Resulta imposible que, utilizando el procedimiento algorítmico descrito con selección aleatoria de centroides iniciales se clasifiquen las instancias en las clases de odio y neutra. Cabría esta posibilidad si asignásemos previamente los centroides iniciales con unos valores aproximados a los verdaderos de las dos clases, pero esto requeriría un etiquetado previo de las instancias que, en definitiva, es lo que hacemos directamente.

Sí cabría pensar en utilizar el *clustering* tras el filtrado inicial para tratar de descubrir nuevas instancias que, de nuevo, serían etiquetadas a mano para aplicar finalmente un clasificador supervisado, proceso en cierto modo similar al de ampliar vocabulario a partir de los tuits de odio que usamos.

## La ley de Zipf

Asimismo, teniendo en cuenta la ley de Zipf que nos dice que si ordenamos las palabras de un idioma con N términos por su frecuencia de aparición y llamamos k a la posición que ocupa en dicha ordenación

*La frecuencia de aparición de la palabra que ocupa el orden k sigue una*[*distribución*](https://es.wikipedia.org/wiki/Distribuci%C3%B3n_de_probabilidad)*que puede aproximarse por*

Y, por tanto, la frecuencia relativa entre dos términos consecutivos es:

Y respecto al más frecuente:

Y siendo α≈1 tendremos que, aproximadamente:



Motivo por el cual suele limitaremos el tamaño de vocabulario a las primeras 1.000 o 2.000 palabras clasificadas por su *tf-idf*.