# Método de los coreanos para etiquetar tuits de entrenamiento [1]

Antes de proceder a su estudio, es preciso recordar algunos conceptos sobre *búsqueda y* *recuperación de información*.

## El Modelo de Espacio Vectorial en Búsqueda y Recuperación de Información.

Se conoce como *modelo de espacio vectorial* [2] a un modelo algebraico utilizado para filtrado, recuperación, indexado y cálculo de relevancia de información. Representa documentos en lenguaje natural de una manera formal mediante el uso de *vectores* (de identificadores, por ejemplo términos de búsqueda) en un espacio lineal multidimensional.

Muchas de las tareas de recuperación de información como la búsqueda, agrupamiento o categorización de textos tienen como primer objetivo procesar documentos en lenguaje natural. El problema que surge es que los algoritmos que pretenden resolver estas tareas necesitan representaciones internas explícitas de los documentos. En el área de recuperación de información normalmente se usa una expresión vectorial, donde las componentes del vector representan términos, frases o conceptos que aparecen en el documento.

En este aspecto la representación más adoptada es la conocida como *bolsa de palabras*: una colección de documentos compuesta por n documentos indexados y m términos representados por una matriz documento-término de *D* x *T*. Donde los *D* vectores renglón representan los *D* documentos; y el valor asignado a cada componente refleja el *peso* o frecuencia ponderada que produce el *término*, frase o concepto *ti* (*i=1,…,T[[1]](#footnote-1)*) en la representación semántica del documento *j*.

Una consulta (*query*) puede representarse también por un vector:

En esta representación vectorial de documentos el éxito o fracaso se basa en la ponderación o peso de los términos.

Este procedimiento permite establecer una puntuación entre consultas y cada documento

En el modelo propuesto por  [Salton](https://en.wikipedia.org/wiki/Gerard_Salton), Wong and Yang[[2]](#footnote-2)  [1] cada elemento del *vector de pesos* del documento se define como:

Donde TF es la *frecuencia del término* *t* en el documento *d* (parámetro local)

IDF es la *frecuencia inversa de documento[[3]](#footnote-3)*

Donde el denominador - *dft*.- es el número de documentos en que aparece el término t.

La lógica tras la ponderación *tf-id* es clara: a mayor TF, más peso tendrá el término en cuestión para clasificar el documento, pero si el término se repite en la mayoría de los documentos (IDF-1 elevada), su relevancia discriminatoria se ve reducida.

En otras palabras, *tf-idftd* asigna al término *t* en el documento *d* un peso

1. que es mayor cuando *t*  aparece *muchas veces* en *pocos documentos;*
2. es menor cuando aparece *pocas veces en un documento* o aparece en muchos documentos y
3. Mucho menor cuando el *término* aparece en *casi todos los documentos*.

El cómputo de la similitud entre un documento y una consulta se realiza a través del coseno del ángulo de los vectores y :

Puede verse que, siendo los pesos ≥0, y la similitud entre documento y búsqueda está comprendida entre 0 y 1.

## Esquema del sistema.

El esquema de los coreanos es el siguiente:



Figura 1 Secuencia del sistema

En el que se pueden ver cuatro etapas principales que nosotros respetaremos:

* Preprocesado
* Creación de palabras claves para seleccionar conjunto de entrenamiento.
* Selección de atributos.
* Clasificación.
  1. **Preprocesado.**

En esta fase, se eliminan metadatos de los mensajes.

Cada *tuit* se segmenta en *oraciones* (sentencias).

Se extraen los nombres contenidos de cada sentencia usando un *POS Tagger[[4]](#footnote-4)*.

* 1. **Crear conjuntos de palabras claves de entrenamiento.**

Dado que no existen documentos de entrenamiento, definimos *manualmente* términos clave para cada categoría (más de una)[[5]](#footnote-5) como se ve en la Tabla 1 siguiente.

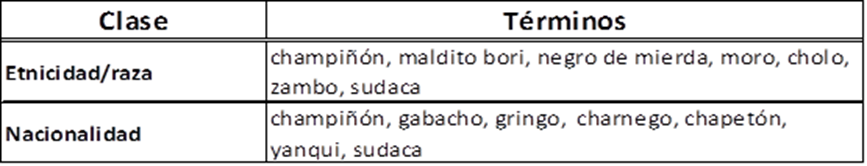


Tabla 1 Ejemplos de palabras clave por clase

Las *sentencias* que contienen *términos clave* predefinidos se eligen como representativas. El resto de sentencias se consideran no relevantes para la asignación de clase al texto.

Las *sentencias representativas* son las que contienen términos clave de la categoría. Sin embargo, existen sentencias erróneas dentro de las representativas que, a pesar de contener términos clave, no pueden clasificarse como pertenecientes a ella. Para eliminarlas, priorizaremos las sentencias representativas mediante un indicador del peso de cada sentencia. Para ello utilizan la combinación de los indicadores:

* Frecuencia del término (*Term Frequency TF*) y
* Frecuencia de documento inversa (*Inverse Document Frequency IDF*)

Donde

La combinación de ambos indicadores[[6]](#footnote-6) () nos da el peso *wij del* término *ti* en el documento *jº*

que, nos sirve para calcular el *peso de la sentencia* (oración) *k* de *N* términos de la clase *j* como:

y priorizar las sentencias. Se sigue un método empírico eligiendo las de más peso.

En esta fase, hemos seleccionado el conjunto de *sentencias de entrenamiento*, es decir aquellas que determinan si un *tuit* es de clase1 o clase2.

* 1. **Selección de atributos y clasificador de textos.**

El sistema no utilizará, en principio, más características que las palabras[[7]](#footnote-7) aun cuando cabe utilizar atributos adicionales del texto como uso de mayúsculas, signos de exclamación, etc.

Una de las mayores dificultades en la clasificación de textos es la dimensionalidad del espacio de atributos que, en principio, se compone de todas las palabras o términos que aparecen en las *sentencias de entrenamiento*.

Para ello podemos proceder de múltiples maneras [4]:

1. Estableciendo *umbrales de frecuencia documental*. Calculamos la frecuencia documental de cada término en el corpus de entrenamiento y eliminamos aquellos que aparezcan un número de veces inferior a un valor umbral (*raros*). Sin embargo, términos que aparecen en pocos documentos pueden ser eficaces para discriminar entre clases como hemos visto antes.
2. Por la *Ganancia de Información* que se consigue por añadir un término al vocabulario. Se eliminarían aquellos que proporcionen una menor ganancia. Este procedimiento tiene un elevado coste computacional.
3. Estadístico χ2. El estadístico χ2 se usa frecuentemente en problemas relacionados con *tablas de contingencia* bidimensionales para la prueba de independencia, consistente en comprobar si dos características cualitativas están relacionadas entre sí (por ejemplo: ¿la frecuencia de un término *t* está relacionada con la clase *c* a que está el *tui*t asignado?).

Si consideramos esta *tabla de contingencia* de un término *t* y una clase *c* mostrada en la Tabla 2 se observa que:

* α es el número de veces que aparece t en c.
* β es el número de veces que t ocurre sin c.
* γ es el número de veces que ocurre c sin t
* δ el número de veces que no aparecen ni t ni c
* D es el número de documentos.

En este caso una medida de la *bondad del término* será:

Este estadístico tiene un *valor nulo si t y c son independientes*.



Tabla 2 Tabla de contingencia

Calcularemos para cada clase el estadístico χ2 entre cada término del vocabulario del conjunto de sentencias de entrenamiento y la clase y combinaremos los valores de cada una de ellas en dos valores:

Este último valor mide la bondad de una palabra en una selección global.

Elegiremos este último método *en el caso de que el vocabulario inicial sea muy extenso* y su tamaño se elige priorizando los términos mediante el estadístico χ2 con respecto a cada clase.

* 1. **Medidas de rendimiento.**

Si los documentos se asignan solo a una clase[[8]](#footnote-8), se usan las conocidas magnitudes

* Precisión (p): porcentaje de los documentos recuperados (clasificados como relevantes) que realmente lo son (aciertos).
* Exhaustividad (r): porcentaje de los documentos relevantes existentes en la población que han sido recuperados correctamente.

Un valor p=1 nos dice que todos los elementos recuperados como relevantes, lo son, pero no nos dice nada acerca de si hemos recuperado todos los documentos relevantes (r).

La Figura 2 muestra gráficamente estos valores y la Figura 3 representa cómo habitualmente están negativamente relacionados.

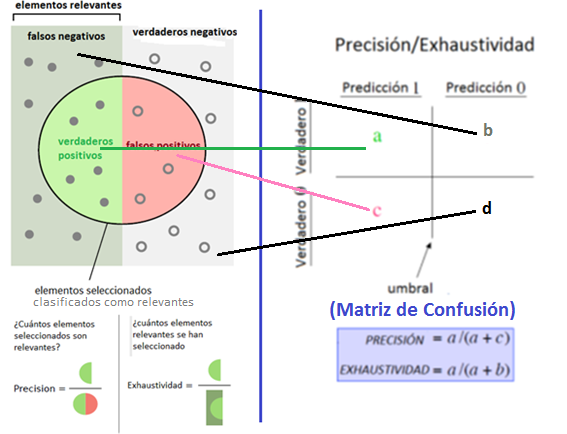


Figura 2 Medidas de rendimiento

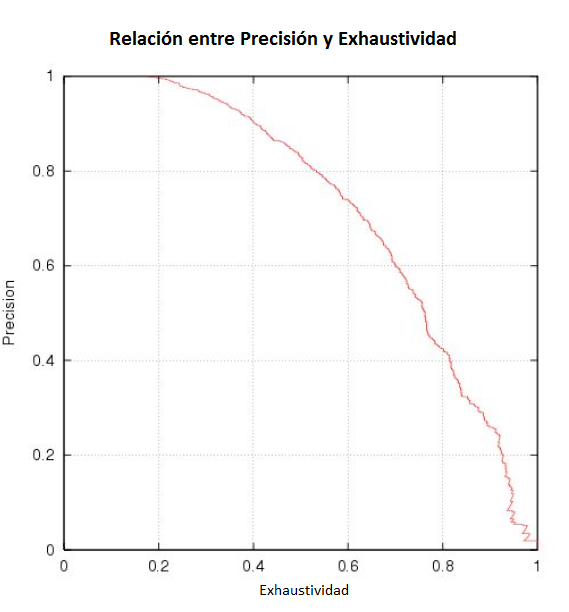
******

Figura 3 Relación Precisión-Exhaustividad

Ambos indicadores se combinan equilibradamente mediante su *media armónica* en F:

Otro indicador que se está utilizando cada vez más es el *ROC[[9]](#footnote-9)* habitual en Medicina y Biología para hablar de la detección de falsos positivos y negativos.

Ahora a la exhaustividad - - se la denomina *sensibilidad.* Como se ve*,* es la

y se introduce la *especificidad:*

De manera que

Y es

Por el teorema de la probabilidad total sabemos que:

Si dibujamos el gráfico que relaciona ambas magnitudes, obtenemos la Figura 4 en que el *ROC* es el *área bajo la curva* que puede tomar valores entre 0 (no acierta nunca) y 1 (la predicción acierta siempre).

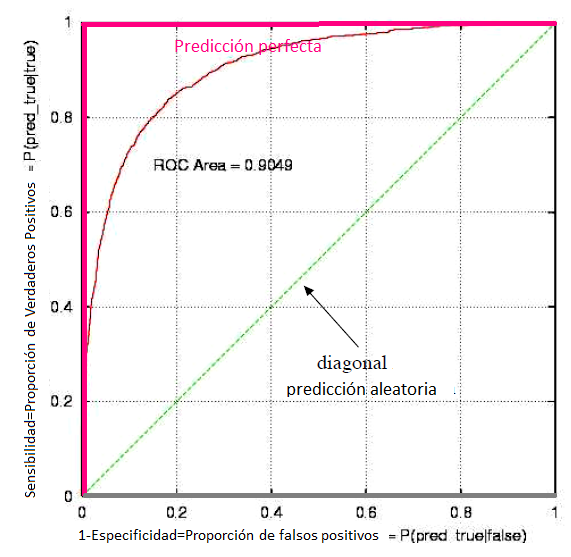


Figura 4 Indicador ROC

La clasificación para clases no mutuamente excluyentes se llama *clasificación multietiqueta*, *multivariable* o *cualquiera de* (*any-of*). Ahora un documento puede pertenecer a más de una clase simultáneamente, a una sola clase o a ninguna de las clases. La decisión que se tome respecto a una clase no afecta a las otras. A veces se dice que las clases son independientes, pero esto es engañoso porque las clases son raramente independientes en sentido estadístico (nacionalidad y raza no son independientes.

Resolver estos problemas de clasificación requiere:

* construir un clasificador para cada clase, son un conjunto de entrenamiento para cada una de ellas en que los documentos están etiquetados como pertenecientes a la clase o no pertenecientes a la misma;
* aplicar a cada documento de prueba cada clasificador por separado. La decisión de un clasificador no debe influir en la de los restantes.

# Referencias

|  |  |
| --- | --- |
| [1] | J. S. Youngjoong Ko, «Automatic text categorization by unsupervised learning,» de *Proceedings of the 18th conference on Computational linguistics*, 2000. |
| [2] | de *An Introduction to Information Retrieval*, Cambridge, U.K., Cambridge University Press, 2008, pp. 117-118. |
| [3] | A. W. y. C. S. Y. G. Salton, «A Vector Space Model for Automatic Indexing,» *Communications of the ACM,* vol. 18, nº 11, p. pages 613–620, 1975. |
| [4] | J. O. P. Yiming Yang, «A Comparative Study on Feature Selection in Text Categorization,» de *ICML '97 Proceedings of the Fourteenth International Conference on Machine Learning*, 1997. |

## Ilustraciones.

[Figura 1 Secuencia del sistema 2](#_Toc473119365)

[Figura 2 Medidas de rendimiento 6](#_Toc473119366)

[Figura 3 Relación Precisión-Exhaustividad 6](#_Toc473119367)

## Tablas.

[Tabla 1 Ejemplos de palabras clave por clase 3](#_Toc473120058)

[Tabla 2 Tabla de contingencia 5](#_Toc473120059)

1. T es el *tamaño del vocabulario* de la colección de documentos. [↑](#footnote-ref-1)
2. Ponderación tf-idf. [↑](#footnote-ref-2)
3. La base de logaritmos es irrelevante. [↑](#footnote-ref-3)
4. <http://www.nltk.org/book/ch05.html> [↑](#footnote-ref-4)
5. En <https://www.hatebase.org/> hay un vocabulario en español de 77 términos (palabras y n-gramas) denigratorios utilizados en España y América. [↑](#footnote-ref-5)
6. Ver [1] [↑](#footnote-ref-6)
7. Usan *unigramas* o *bag-of-words.* [↑](#footnote-ref-7)
8. Como es el caso en que las clases no son excluyentes. [↑](#footnote-ref-8)
9. Receiver Operator Characteristic. [↑](#footnote-ref-9)