Modelamiento de Espacio de Palabras en la Clasificación de Documentos

Word space model in the document clasification

Dr. Víctor Manuel Cornejo Aparicio

RESUMEN

El presente artículo pretende mostrar los fundamentos de la clasificación de documentos aplicando el método del modelo de espacio de palabras, donde se demostrara por que la clasificación exige algunos pasos para el desarrollo exitoso de la clasificación, esto constituye el pilar de futuros trabajos en el desarrollo de clasificadores de documentos o análisis de proximidad entre dos textos.

Palabras clave: *modelo de espacio de palabras, clasificación de documento.*

ABSTRACT

This article shows the basics of document classification by applying the word space model, where it is shown that classification requires some steps for the successful development of the classification, this contituye the mainstay of future work in the development classifiers anális documents or proximity between two texts.

Keywords: *Word space model, document clasification.*

INTRODUCCIÓN

En el presente artículo se presenta la mecánica de la clasificación de documentos bajo un enfoque genérico inicialmente, este será abordado considerando las deficinciones clásicas respecto a la clasificación de documentos. en este punto se resalta dos grupos fundamentales; el primero constituido por un grupo de ejemplos de documentos: preclasificados que constituyen las diferentes categorías de entrenamiento, posteriormente un segundo grupo de ejemplos de documentos preclasificados que se usaran con el propósito de probar la precisión de los algoritmos de clasificación. Enseguida se muestra la técnica del modelo de espacio de palabras, el mismo que introduce a la constitución de textos conformadas por términos que son contabilizados, que a posteriori representan un patrón de clasificación, esto conlleva a que si existiera un nuevo documentos, y este puede sometérsele a la misma técnica, también podría encontrársele el patrón individual, con lo que en forma consiguiente, posteriori se podría determinar la proximidad entre el patrón del tipo de documento con el patrón del documento individual, de ello se puede extraer en suma de varios documentos individuales, una matriz que contrasta cada documento por medio de su proximidad con cada tipo de documento establecido, a lo que se denomina matriz de confusión, de ella se puede extraer en definitiva la precisión del modelo empleado. Cabe resaltar que si se desea emplear este modelo en una aplicación real, con motivo de mejorar el acceso a los documentos clasificados, lo que en realidad se emplea es la matriz de confusión.

DEFINICION DE CLASIFICACIÒN

Consiste en colocar un documento dentro de un grupo de clases previamente definidas (Coyotl R. 2007). La mayor parte del trabajo en esta área se ha enfocado en la clasificación de textos por su tema o tópico. Sin embargo, un documento también puede ser clasificado de acuerdo a su estilo (clasificación no-temática). En la clasificación no-temática se consideran tareas tales como la clasificación de opiniones, la detección de plagio, la atribución de autoría, la clasificación por género, etc.

La clasificación automática de textos, caracteriza los documentos en referencia a un número de categorías establecidas de acuerdo a su contenido, esto debido a que un documento cualquiera puede pertenecer a una, varias o todas, y hasta incluso ninguna de las categorías establecidas con anterioridad (Joachims T. 1998). Cuando se emplea el aprendizaje automático, la razón de aprender a partir de casos definidos de documentos, es que ello nos permita hacer asignaciones a las categorías de una forma automática.
de su información, este consideró como base información de reportes anuales de aseguradoras de salud y el esquema de datos empleado por el API de geo codificación de Google. El formato propuesto consiste en un esquema XML que es presentado en forma gráfica en los siguientes párrafos.

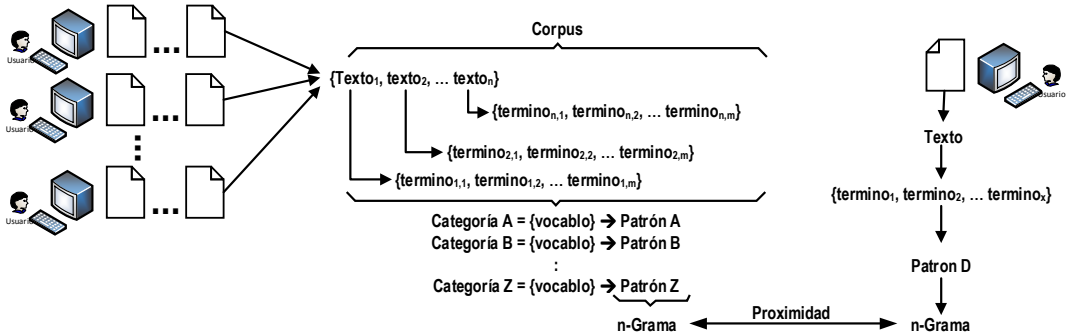


Figura 1: Esquema de la clasificación de documentos
Fuente: Cornejo V. 2013

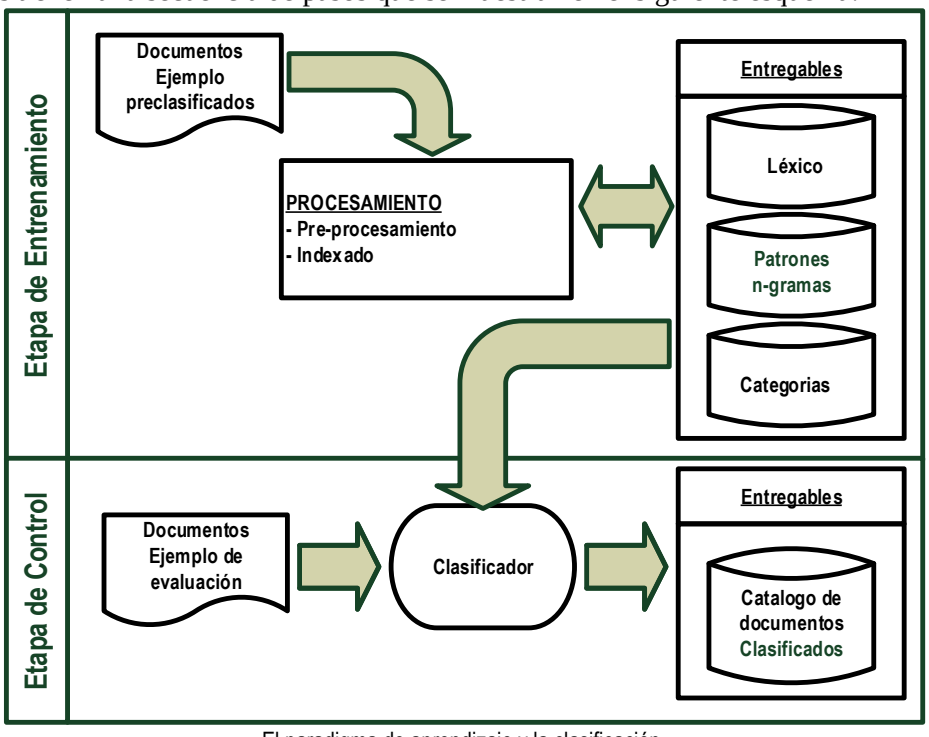
La clasificación de documentos es un proceso que consiste en coleccionar un conjunto de documentos elaborados por uno o más usuarios, los mismos que constituyen un corpus, cada documento es una colección de términos, los cuales al ser disgregados constituyen diferentes categorías, las mismas que son caracterizadas determinando un patrón que identifica a la categoría, dichas categoría son determinadas por un patrón en forma de n-grama. Los usuarios al crear un documento forman un conjunto de términos que son disgregados para de forma similar determinar su patrón en forma de n-grama. De esta manera es posible determinar la proximidad entre ambos patrones. Por lo cual, al existir una mayor proximidad entre ambos patrones, es más probable que el documento corresponda a dicha categoría.

ALGORITMO BASADO EN EL MODELO DE ESPACIO DE PALABRAS

Consiste en extraer de un conjunto de documento categorizados, los términos que estos contienen, eliminar los términos que no aportan información respecto al motivo del texto, luego en base a los términos restantes, extraer conjugaciones de uno dos o mas palabras, con lo que se construye un n-grama (monogramas, diagramas, etc.) que representa el patrón de la categoría de la cual se extrajeron los textos que lo conforman. Posteriormente se procesa de forma similar el documento a clasificar, para de forma consecutiva calcular la proximidad entre ambos patrones, la categoría con la que resulte más próxima, se considerara como la categoría a la que pertenece el referido documento.

PROCESO DE CLASIFICACIÓN DE TEXTOS

El proceso de clasificación de textos está compuesto por dos etapas (entrenamiento y prueba), las cuales tienen una secuencia de pasos que se muestran en el siguiente esquema:



El paradigma de aprendizaje y la clasificación.
Fuente: Elaboración propia.

En la etapa de entrenamiento, se debe contar con un conjunto de documentos denominados de ejemplos de entrenamiento, por medio de los cuales se efectúa una extracción de características, estas se deben representar por medio de los datos que los componen, luego de ello se hace una selección de las características relevantes, para con ello ejecutar los algoritmos de aprendizaje que permitirá pasar a la etapa de prueba.

En la etapa de prueba, por medio de un conjunto de documentos nuevos (Ejemplos nuevos), se pone a evaluación entre las características extraídas de estos documentos, y el que contienen los patrones de entrenamiento, para establecer de esta forma, el grado de asertividad que posee el modelo propuesto.

Cabe destacar que este modelo de clasificación, cumple con las etapas de la prueba experimental, donde se puede entender claramente que los ejemplos de entrenamiento son un grupo control, y los nuevos ejemplos para efectuar las pruebas, son grupos objetivo, con la salvedad de que estos últimos ya han sido clasificados con anterioridad, así que de una forma certera se puede establecer una medida de efectividad o confianza.

REPRESENTACIÓN DE UN DOCUMENTO

Cuando se desea hacer una clasificación de documentos de forma automática, el proceso de entrenamiento se lleva a cabo con un conjunto de documentos definido, al cual se denomina conjunto de entrenamiento, esto nos da una representación tipificada, esto es susceptible de que se clasifique por medio de algoritmos, una forma de hacer esto posible es por medio de un modelo vectorial, o modelado de espacio de palabras, lo cual es ampliamente usado en clasificación.

Dado un documento d_j, este es representado por un vector $\vec{d}_j = (w_{1j} \quad \cdots \quad w_{rj})$, donde w son los términos o palabras y r es el número total de palabras que se encuentran presentes en el documento, estas son parte de un diccionario particular, es usual que este número r, sea luego de haber filtrado o excluido a las palabras funcionales o vacías (Ass K. et al. 1999).

Existe otra manera de efectuar esta representación, y está dada por lo que se denomina como un lema, lo cual se efectúa con el propósito de que se contabilicen las palabras con el mismo sentido conceptual, o significado de raíz, para de manera consecuente se pueda asignarles su respectivo peso al término específico.

Existen varias formas de asignarles pesos a los términos, los mismos son:

- Ponderado booleano
- Ponderado por frecuencia de término
- Ponderado tf-idf

A. Ponderado booleano

Este tipo de ponderación acepta solamente uno de dos valores; uno (1) o cero (0), para los casos de si el término aparece o no aparece en el documento, lo cual se expresa de la forma siguiente:

$$w_{ij} = \begin{cases} 1 \; si \; t_i \; aparece \; en \; d_j \\ 0 \;\; En \; caso \; contrario \end{cases}$$

B. Ponderado por frecuencia de término

En este caso se contabiliza el número de veces que aparece un termino i en el documento d_j, lo cual se denota como f_{ij}, en este tipo de ponderación, cabe destacar que se interpreta la frecuencia, en el sentido del grado de importancia del término para el documento, esto quiere decir que a medida que el termino figura con más frecuencia, significa que el mismo es muy importante para el referido documento.

$$w_{ij} = f_{ij}$$

C. Ponderado tf-idf

Asigna el peso de la palabra i en el documento j, en proporción al número de ocurrencias de la palabra en el documento y en proporción inversa al número de documentos en la colección, para los cuales ocurre la palabra al menos una vez.

$$w_{ij} = f_{ij} * Log\left(\frac{N}{n_i}\right)$$

Donde N es el número de documentos en la colección y n_i es el número de documentos en los que el término aparece.

CORPUS

Un corpus lingüístico es una colección de elementos lingüísticos seleccionados y ordenados de acuerdo con criterios lingüísticos explícitos con la finalidad de ser usado como muestra de la lengua, un corpus lingüístico consiste en la recopilación de un conjunto de textos de materiales escritos y/o hablados, agrupados bajo un conjunto de criterios mínimos, para realizar ciertos análisis lingüísticos (Perez J. 1998).

A efectos de trabajar en tareas de procesamiento de lenguaje natural, específicamente en las tareas de clasificación de documentos, se acuña el término de corpus estandarizado; el cual es un conjunto de documentos que están al alcance de la comunidad interesada en la clasificación de textos, los mismos que serán empleados por diversos investigadores a efectos de poder comparar resultados con sus pares en otras latitudes de una forma homogénea.

Existen varios corpus estandarizados, los mismos que son de libre acceso, dentro de los cuales podemos citar a:

- Reuters21578-Apte-90Cat
- Reuters21578-Apte-115Cat
- RCV1, RCV2
- Ohsumed
- Etc.

En el ámbito de procesamiento de lenguaje natural, un corpus lingüístico es una colección de textos en soporte electrónico, normalmente amplio, que contiene ejemplos reales de uso de una lengua tal y como es utilizada por los hablantes, con sus errores, peculiaridades y excepciones (Navarro F. 2007).

LEMATIZACIÓN

La lematización es un proceso lingüístico que consiste en: dada una forma flexionada de un vocablo (es decir, en plural, en femenino, conjugada, etc), hallar el lema correspondiente es en

determinar el vocablo que de un origen pueda generar diversas variaciones de género, cantidad, etc. El lema es la forma que por convenio se acepta como representante de todas las formas flexionadas de una misma palabra. Es decir, el lema de una palabra es la palabra que nos encontraríamos como entrada en un diccionario tradicional: singular para sustantivos, masculino singular para adjetivos, infinitivo para verbos.

STEMMING

El *Stemming* es un método para reducir una palabra a su raíz o a un *stem* (en inglés) o lema. Hay algunos algoritmos de *stemming* que ayudan en sistemas de recuperación de información.

ALGORITMO DE PORTER

Es un algoritmo que por medio de un análisis morfológico de las palabras, se reducen a una base común o raíz, sobre la cual se puede implementar programas que requieran de la lematización de textos. El algoritmo de Porter permite hacer stemming, esto es extraer los sufijos y prefijos comunes de palabras literalmente diferentes pero con una raíz común que pueden ser consideradas como un sólo término

CALCÙLO DE PROXIMIDAD

El patrón de un tipo determinado de documento está dado por la conjunción de los vocablos relevantes que constituyen su corpus de documentos, esto ensamblado en la forma de algún n-grama, de la misma manera, el patrón de un documento por clasificar, debe estar definida por el mismo nivel de n-grama del patrón del tipo de documento con el cual se quiere determinar su proximidad.

En la medida que el patrón de documento a clasificar se aproxime más al patrón del tipo de documento, se puede afirmar que es muy probable que el documento pertenezca a dicho tipo clasificado.

Para estimar la proximidad de estos patrones, se puede aplicar la fórmula del coseno de dos vectores (Sahlgren M. 2006), la misma que presentamos a continuación:

$$\cos\_measure(\vec{w},\vec{r}) = \frac{\vec{w},\vec{r}}{|\vec{w}|\quad|\vec{r}|} = \frac{\sum_{j=1}^{n} w_j \times r_{i,j}}{\sqrt{\sum_{j=1}^{n}(w_j)^2} \times \sqrt{\sum_{j=1}^{n}(r_{i,j})^2}}$$

Donde $\vec{w}$ corresponde al vector patrón del documento a clasificar y $\vec{r}$ es el vector del patrón del tipo de documento i con el cual se quiere determinar su proximidad

PREMISAS DE LA INVESTIGACIÓN

Premisa 1: Los documentos tienen una naturaleza y estructura, los mismos que a su vez están constituidos por textos que son un conjunto de vocablos que son regularmente empleados en documentos de similar categoría.

Premisa 2: Los vocablos individualmente constituyen información, y estos a la vez que se asocian entre sí, incrementan el volumen de información, la misma que podría caracterizar en mejor manera a los documentos que los contengan.

Premisa 3: Cuando se emplean más de un vocablo, en un proceso de clasificación automática (n-gramas), puede darse el caso que una conjunción de vocablos (A, B), pueda presentarse como (B, A) en el mismo documento o uno similar del mismo tipo, lo cual en términos prácticos,

constituiría una dispersión de las frecuencias asociadas a la categoría definida, para lo cual, dado el caso se debería indexar horizontalmente los vocablos, y de esta forma evitar la dispersión de las frecuencias.

Premisa 4: Al constituirse los vocablos asociados de uno, dos o más, estos se deberán catalogar asociados al tipo de documento que les dio origen, una vez constituida la asociación y elaborado la concentración de frecuencias, estos vocablos se asumirán como únicos a efectos de desarrollar los cálculos requeridos para la determinación de las proximidades entre los vocablos y el tipo de documento asociado.

Premisa 5: Los corpus de entrenamiento no serán modificados o reevaluados a efectos de construir los n-gramas que configuren el patrón, esto debido a que al desarrollar una aplicación real, los corpus de entrenamiento ya pasaron filtros diversos que permiten superar esta fase, y no es dable evaluar constantemente cada vez que se requiera hacer una clasificación. Cabe aclarar que esta premisa contradice los postulados de la evaluación de un corpus, que muchas veces requiere de un ajuste del contenido a efectos de mejorar el proceso de entrenamiento de los algoritmos.

PRE-PROCESAMIENTO

Esta fase tiene el propósito de eliminar elementos textuales que no contienen información relevante. Esta información haría que se eleve los costos de procesamiento, así como la calidad de información, debido a que los patrones se harían más semejantes, por tanto dichos no serían muy efectivos al momento de ejecutar las rutinas de clasificación, en este caso se efectuara las siguientes acciones:

Etiquetas. Los documentos suelen contener elementos textuales en forma de cabeceras o pies de página, etiquetas en formatos html o xml, todos ellos serán retirados del texto para proseguir el procesamiento, dicho de otra forma, se debe borrar todos estos elementos y dejar el texto de forma continua.

Palabras Vacías. Las preposiciones, artículos, conjunciones y otros, son elementos textuales que abundan en un texto normal. Estas tampoco suelen aportar información, sus conjunciones son muy reiteradas, y de incluirlos en el proceso de construcción de los digramas y más, confundirían el propósito de estas combinaciones para una clasificación efectiva, por lo tanto estas palabras deben eliminarse.

Extracción de Verbos y Sustantivos. Las palabras cuando conforman un texto, en esencia comprenden su contenido en función a verbos y sustantivos. La particularidad de un texto está definido por estos, y su patrón de clasificación se perfecciona al incluir únicamente estos elementos. Por ello se debe reconstituir el texto con solamente estas palabras.

Lematización de Palabras. Las palabras empleadas en una redacción de texto normal, son usadas haciendo uso de sufijos para darles sentidos de acción, tiempo, etc, los mismos que en el contexto de la redacción original, hace un texto entendible fonéticamente, pero todos estos términos tienen una raíz, la misma que es general a todas y que encierra el significado base de los diversos términos. Por ello la lematización consiste en reducir las palabras a su lema o raíz, para de esta forma las conjunciones sean concentradas de una manera uniforme, por ejemplo hablar, hablará, hablando, hablo, etc., tiene su raíz en habl.

Los procesos a desarrollar podrían efectuarse tomando como base el conjunto de procesos que se ejemplifican en el siguiente módulo de procesos:

```
…
… …
… … …
sTexto = ExtraerTexto(“d:\Reuter21578\test\corn\0009622”)
sTexto = PreProcesamiento(sTexto)
… … …
… …

Funcion PreProcesamiento( texto)
sTextoProcesado = ProcesarTexto(texto)
sTextoProcesado = LematizarTexto(sTextoProcesado)
Retornar sTextoProcesado
Fin Funcion

Funcion ProcesarTexto( texto)
_texto = EliminaLinks(_texto)
_texto = EliminarSignos(_texto)
_texto = EliminarCaracteresEspeciales(_texto)
_texto = EliminarApostrofe(_texto)
_texto = EliminarVacias(_texto)
_texto = EliminarSignosPuntuacion(_texto)
_texto = EliminarNumeros(_texto)
_texto = QuitarPalabrasConNumeros(_texto)
_texto = ExtraerSustantivosVerbos(_texto)
Retornar _texto
Fin Funcion
```

INDEXADO

Es un proceso de ordenamiento de las palabras existentes en todos los documentos, por ello esto puede ser representado por una matriz en la cual se incluyan todos los términos existentes en todos los documentos de la forma siguiente:

$$A = \begin{Bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{1n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{Bmatrix}$$

Figura 2: Representación de un n-grama
Fuente: Cornejo V. 2013

Si A es un n-grama de grado n, significa que cada término está compuesto por n conjugaciones de vocablos, de acuerdo a la ubicación dentro de los documentos que constituyen el corpus, esto significaría que este n-grama está compuesto por m términos, donde cada línea de la matriz representaría un término.

Es simple concebir la idea de un ordenamiento vertical, lo que dicho de otro modo es el ordenamiento entre líneas diferentes de una matriz, esto de acuerdo a un criterio generalmente alfabético, pero lo que vale la pena aclarar es el ordenamiento horizontal, esto implica que se debe ordenar dos o más vocablos conjugados de acuerdo al nivel de n-grama que se esté trabajando, esto significa que un termino t_{ij} contiene un conjunto de vocablos $\{v_1, \quad v_2 \ldots \quad v_n\}$ donde v_x debe ser menor que v_{x+1}. Por ejemplo si tenemos el término t cuyo digrama que contiene los vocablos {"aaa", "bbb"}, en alguna parte del texto, lo cual tendría una frecuencia igual a uno, y en algún otro lugar tiene los vocablos {"bbb", "aaa"} con una frecuencia igual a uno, lo que se debería hacer es ordenar alfabéticamente estos vocablos quedándonos con {"aaa", "bbb"} y una frecuencia igual a dos, lo cual evitaría la dispersión de las frecuencias. Debido a que $a^2 + b^2$ es menor que $(a + b)^2$ al aplicar la ley de los cosenos en el cálculo de proximidad. Esto se puede apreciar de mejor manera en el gráfico siguiente:

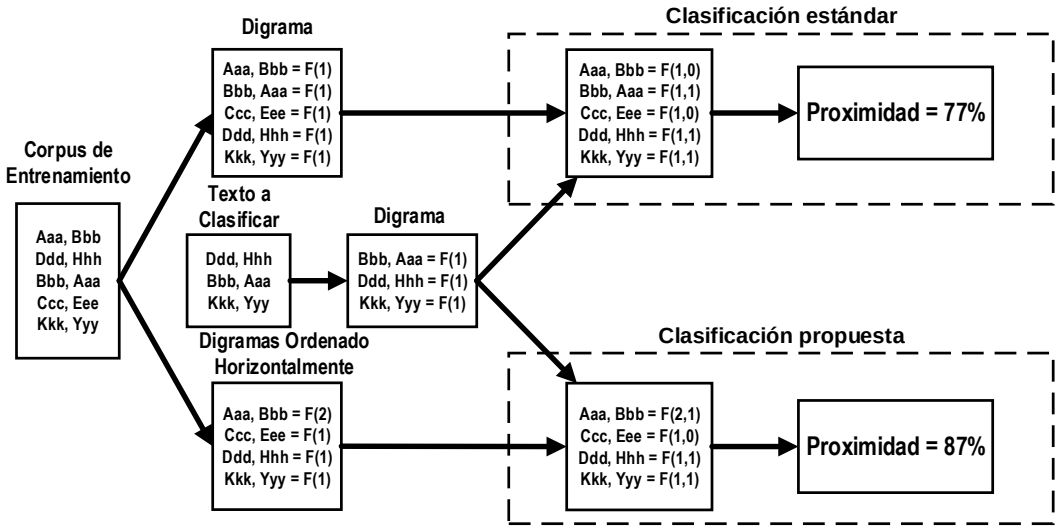


Figura 3: Ejemplo de clasificación
Fuente: Cornejo V. 2013

Posterior a la indexación vertical y horizontal se debe crear tantas tablas como tipos de documentos puedan existir los mismos que deben contener la frecuencia de los términos que contenga el texto pre-procesado, el mismo que debe estar organizado de la forma siguiente:

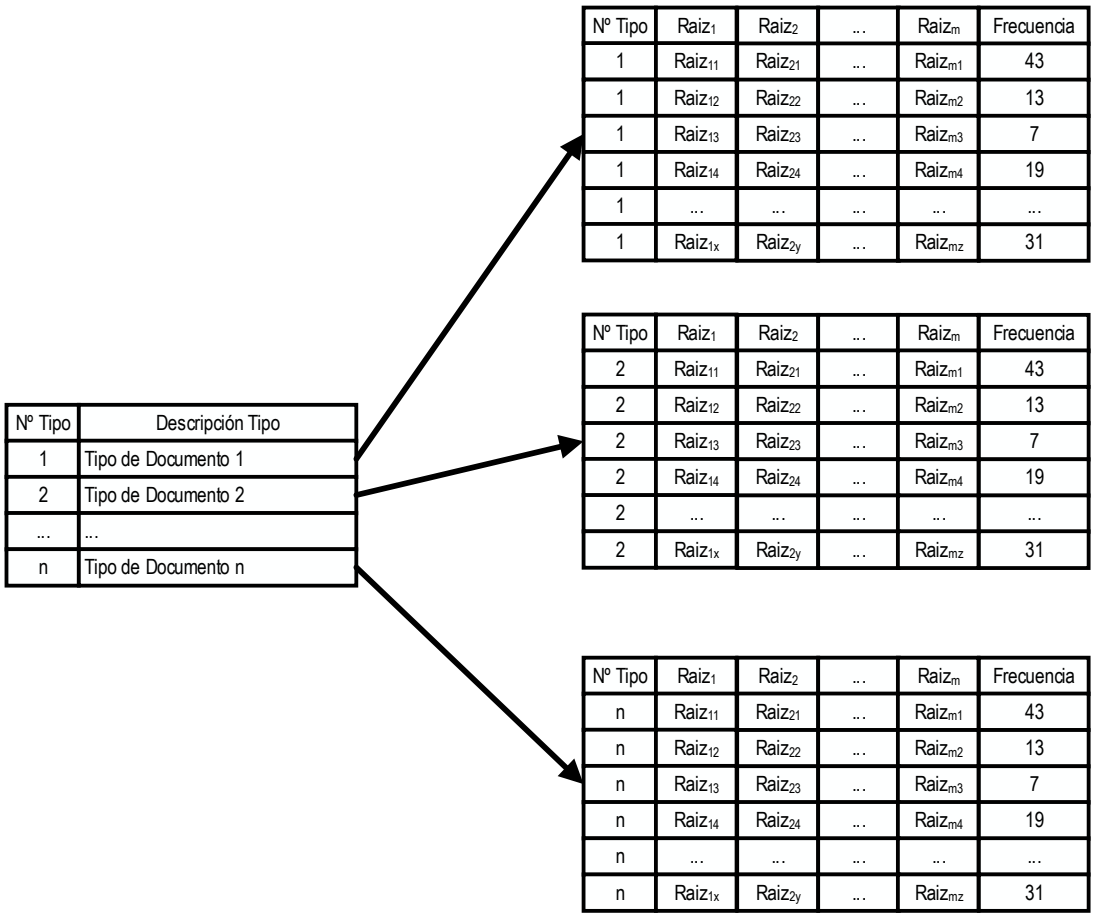


Figura 4: Esquema de la creación de tablas patrón por tipo de documento
Fuente: Cornejo V. 2013

ENTREGABLES

Al término de esta etapa, se tendrá un catálogo de documentos pre clasificados, comparados a la par con la clasificación automática que haya efectuado la mecánica de la propuesta, que en términos totales nos darán un grado de certeza sobre la precisión del método. De forma tangible se debe presentar una matriz de doble entrada (matriz de confusión) que además de presentar la cantidad de documentos clasificados, debe presentar la cantidad de documentos por tipo que fueron clasificados como los diferentes tipos definidos, esta matriz puede elaborarse como se muestra en la tabla siguiente:

	Total	Doc. Tipo 1	Doc. Tipo 2	Doc. Tipo 3	…	Doc. Tipo n
Doc. Tipo 1	A	a_1	a_2	a_3	…	a_n
Doc. Tipo 2	B	b_1	b_2	b_3	…	b_n
Doc. Tipo 3	C	c_1	c_2	c_3	…	c_n
…	…	…	…	…	…	…
Doc. Tipo n	N	n_1	n_2	n_3	…	n_n

Tabla 1: Matriz de confusión
Fuente: Elaboración propia

Posteriormente se debe sacar la proporción de la clasificación efectuada, para ello se debe calcular la proporción de la cantidad de documentos clasificados por tipo, dividido entre el total de documentos de ese tipo que se emplearon para probar el método de clasificación

	Doc. Tipo 1	Doc. Tipo 2	Doc. Tipo 3	…	Doc. Tipo n
Doc. Tipo 1	**a_1/A**	a_2/A	a_3/A	…	a_n/A
Doc. Tipo 2	b_1/B	**b_2/B**	b_3/B	…	b_n/B
Doc. Tipo 3	c_1/C	c_2/C	**c_3/C**	…	c_n/C
…	…	…	…	…	…
Doc. Tipo n	n_1/N	n_2/N	n_3/N	…	**n_n/N**

Tabla 2: Matriz de confusión porcentual
Fuente: Elaboración propia

De la tabla anterior podemos extraer los resultados de la intersección de los tipos definidos de documentos que fueron clasificados correctamente, su promedio determinaría el grado de precisión del método, sin embargo, ello no impediría sacar conclusiones de los resultados individuales que presenten los demás casilleros.

En una aplicación real, la clasificación se emplea para tener un acceso más rápido a la información requerida, permite hacer una búsqueda segmentada por las categorías definidas, en el caso de la clasificación de documentos, un documento puede pertenecer a más de una categoría, esto se determinará por un umbral de pertenencia que se determine de acuerdo a la exhaustividad con las que se desee hacer dicha búsqueda, en tal sentido se empleará el porcentaje de proximidad del tipo de documento en el cual se desee efectuar la búsqueda, en un sentido práctico se puede decir que se empleara la columna del tipo de documento para segmentar los documentos con los que se efectuara la búsqueda, como se muestra en la tabla siguiente:

	Doc. Tipo 1	Doc. Tipo 2	Doc. Tipo 3	…	Doc. Tipo n
Documento 1	$P_{1,1}$	$P_{1,2}$	$P_{1,3}$	…	$P_{1,n}$
Documento 2	$P_{2,1}$	$P_{2,2}$	$P_{2,3}$	…	$P_{2,n}$
Documento 3	$P_{3,1}$	$P_{3,2}$	$P_{3,3}$	…	$P_{3,n}$
…	…	…	…	…	…
Documento m	$P_{m,1}$	$P_{m,2}$	$P_{m,3}$	…	$P_{m,n}$

Tabla 3: Catálogo de Documentos
Fuente: Elaboración propia

RESULTADOS

Se empleó el corpus “corpora Reuters21578-Apte-90Cat” (Moschitti A.). El núcleo de cualquier experimentación de categorización de texto es la precisión final y la posibilidad de compararlo con trabajos anteriores. El corpus Reuters ofrece esta posibilidad, ya que se ha usado en gran medida en el trabajo de la categorización de textos. Las categorías se expresan en diferentes directorios. En cada directorio se almacenan el conjunto de archivos asociados a la categoría de destino (tipo de documento). La división de entrenamiento entre prueba es proporcionada por medio de dos directorios principales diferentes (training y test).

Según David Lewis Reuters es actualmente una de las colecciones de prueba más ampliamente utilizadas para la investigación en categorización de textos. Los datos fueron recogidos inicialmente y etiquetados por Carnegie Group, Inc. y de Reuters, Ltd. en el curso del desarrollo del sistema de categorización e interpretación de textos (Lewis D.).

En este corpus y de acuerdo a las recomendaciones que presenta la fuente, se ha tomado como unidad experimental ocho categorías de documentos los mismos que presentamos a continuación.

Documento Tipo	Cantidad de Archivos	
	Entrenamiento	Prueba
• Acq	1650	719
• Crude	389	189
• Earn	2877	1087
• Grain	433	149
• Interest	347	131
• money-fx	538	179
• trade	369	117
• unknown	1830	280
Total =	8433	2851
	11284	

Tabla 4: Cantidad de documentos por tipo para entrenamiento y clasificación
Fuente: Elaboración propia

A continuación se presenta una tabla que resume las frecuencias de los archivos de prueba clasificados empleando el método de clasificación.

Tipo de documento	Archivos procesados	Clasificación con la propuesta		
		Monograma	Digrama	Digrama Ordenado
Acq	719	661	652	659
crude	189	175	145	144
Earn	1087	995	957	961
Grain	149	127	119	125
interest	131	101	91	94
money-fx	179	125	108	109
trade	117	105	104	103
unknown	280	211	193	189

Tabla 5: Resumen en frecuencias de clasificación
Fuente: Elaboración propia

De la tabla presentada anteriormente, se puede calcular los porcentajes de categorización por tipos definidos, los mismos que se muestran en la tabla siguiente:

Tipo de documento	Clasificación con la propuesta		
	Monograma	Digrama	Digrama Ordenado
acq	91.93%	90.68%	91.66%
crude	92.59%	76.72%	76.19%
earn	91.54%	88.04%	88.41%
grain	85.23%	79.87%	83.89%
interest	77.10%	69.47%	71.76%
money-fx	69.83%	60.34%	60.89%
trade	89.74%	88.89%	88.03%
unknown	75.36%	68.93%	67.50%
Promedio	84.17%	77.87%	78.54%

Tabla 6: Resumen en porcentajes de clasificación
Fuente: Elaboración propia

De la tabla anterior se puede apreciar que la clasificación empleando la propuesta que ofrece en promedio la mayor precisión, es la que se efectúa sin la conjunción de vocablos, dicho de otra forma, empleando monogramas, la misma que obtuvo un nivel de precisión del orden del 84.17%.

ANÁLISIS Y DISCUSIÓN DE RESULTADOS

El corpus empleado cuenta con un total de 11284 archivos correspondientes a un total de 15437 archivos, lo quiere decir que se está empleando un total de 73.10% del cuerpo total de archivos propuestos por Reuters. De una forma más analítica se puede ver en el cuadro siguiente:

Categorías	Numero de Archivos Entrenamiento		Numero de Archivos Prueba	
	Cantidad	%	Cantidad	%
Empleadas	8433	73.89%	2851	70.85%
No Empleadas	2980	26.11%	1173	29.15%
Total	11413		4024	

Tabla 1: Constitución de archivos del Corpus Reuters 21578-90Cat por segmentos de entrenamiento y prueba
Fuente: Elaboración propia

Del cuadro precedente podemos ver que de los 11413 archivos de entrenamiento existentes en las 91 categorías, se está empleando 8433, lo que constituye una representatividad del orden del 73.89% de los archivos de entrenamiento. Del mismo modo se está empleando 2851 archivos de prueba de un total de 4024 archivos, lo que constituye un 70.85% de los archivos de prueba.

De todo lo anteriormente expueto, se puede extrater entender que con el uso de un corpus estandarizado, las pruebas a las que se somete el método, son prácticas y contrastables con los esfuerzos que otros hayan desarrollado en el área de clasificación de texto, y siendo los resultados de la clasificación superiores a un 80% de precisión, estos pueden asumirse como muy buenos.

CONCLUSIONES

Hoy en día es innegable que el procesamiento de textos en forma de documentos guard gran cantidad de información a la que se debe aaceder de alguna manera, esta información al ser diversa y de volúmenes muy considerables, debe ser segmentada por temas o categorías. En este sentido los modelos de espacio de palabras en la calsificación de documentos aporta mecanismos que contribuyen a la accesibilidad a dicha información, permite además generar elementos que permiten encontrar elementos de búsqueda por lo cual su uso es de carácter necesario, cuando se trata de efectuar sistemas que en su contenido tengan documentos con volúmenes considerables de información.

El empleo de un corpus estandarizado, permite a los investigadores en materia de clasificación de documentos, verificar los resultados planteados por os diversos métodos existentes, y probar los propios, para que de una manera consecuente, proponer nuavas alternativas de clasificació, por consiguiente, "Reuters21578" provee esa posibilidad, no obstante, no es necesario utilizar el total de las categorías existentes en ese corpus, sino las más representativas y que provean un volumen lo suficientemente basto para efectuar los experimentos requeridos.

REFERENCIAS BIBLIOGRÁFICAS

- Ass K. y Eikvil L. (1999), "*Text categorization: a survey*", Technical Report 941, Norwegian Computing Center, Noruega.

- Cornejo A., Victor M. (2013), "*Construcción automática y análisis de Modelos de Espacios de Palabras de n-gramas y su aplicación a tareas de procesamiento de lenguaje natural*", Tesis doctoral, Universidad Nacional de San Agustin de Arequipa

- Coyotl Morales; Rosa Maria (2007), "*Clasificación Automática de Textos considerando el Estilo de Redacción*", Tesis de Maestría, Instituto Nacional de Astrofísica, Óptica y Electrónica – INAOE, Tonantzintla, Pue.

- Joachims T. (1998), "*Text Categorization with Support Vector Machines: Learning with many relevant features*", 10th European Conference on Machine Learning, Edición 1298, pp 137-142, Dorint-Parkhotel, Chemnitz, Germany.

- Lewis; David D., Test Collections - Reuters-21578, Disponible en: http://www.daviddlewis.com/resources/testcollections/reuters21578/

- Moschitti; Alessandro, "*Text Categorization Corpora*", Disponible en: http://disi.unitn.it/moschitti/corpora.htm,

- Navarro Colorado; Francisco de Borja (2007), "*Metodología, construcción y explotación de corpus anotados semántica y anafóricamente*", Ph.D. Tesis, Universidad de Alicante, España.

- Pérez Guerra; J. (1998), "*Introducción a la lingüística de corpus. Un ejercicio con herramientas informáticas aplicadas al análisis textual*", Santiago de Compostela: Tercera Edición.

- Sahlgren; Magnus (2006), "*The Word-Space Model - Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces*", Tesis doctoral, Stockholm University Department of Linguistics Computational Linguistics Stockholm, Sweden - National Graduate School of Language Technology Gothenburg University,

Gothenburg, Sweden - Swedish Institute of Computer Science Userware Laboratory Kista, Sweden.

- Sahlgren; Magnus (2006), "*Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces*", Ph.D. dissertation, Stockholm University, Sweden.