INSTITUTO POLITÉCNICO NACIONAL

Laboratorio de Inteligencia Artificial

CLASIFICACIÓN BIBLIOTECARIA AUTOMÁTICA USANDO IDENTIFICACIÓN SIMPLE DE TÉRMINOS CON MÉTODOS LÓGICO-COMBINATORIOS A PARTIR DE INFORMACIÓN ESCASA

TESIS

QUE PARA OBTENER EL GRADO DE
MAESTRO EN CIENCIAS DE LA COMPUTACIÓN

PRESENTA
L.I.A. RICARDO ÁVILA ARGÜELLES

DIRECTORES DE TESIS:

DR. FRANCISCO HIRAM CALVO CASTRO
DR. ALEXANDER GELBUKH KAHN

México, D.F. Diciembre 2008

SIP-14

INSTITUTO POLITECNICO NACIONAL

SECRETARIA DE INVESTIGACIÓN Y POSGRADO

ACTA DE REVISIÓN DE TESIS

En la Ciudad de México, D.F. siendo las 19:00 horas del día 10 del mes de Diciembre de 2008 se reunieron los miembros de la Comisión Revisora de Tesis designada por el Colegio de Profesores de Estudios de Posgrado e Investigación del:

Centro de Investigación en Computación

para examinar la tesis de grado titulada:

"CLASIFICACIÓN BIBLIOTECARIA AUTOMÁTICA USANDO IDENTIFICACIÓN SIMPLE DE TÉRMINOS CON MÉTODOS LÓGICO-COMBINATORIOS A PARTIR DE INFORMACIÓN ESCASA"

ÁVILA	ARGÜELLES	RICARDO
Apellido paterno	materno	nombre(s)

Con registro: B 0 6 1 0 5 8

aspirante al grado de: ***MAESTRÍA EN CIENCIAS DE LA COMPUTACIÓN***

Después de intercambiar opiniones los miembros de la Comisión manifestaron ***SU APROBACIÓN DE LA TESIS***, en virtud de que satisface los requisitos señalados por las disposiciones reglamentarias vigentes.

LA COMISIÓN REVISORA

Presidente

Dr. Adolfo Guzmán Arenas

Secretario

Dr. Edgardo Manuel Felipe Riverón

Primer vocal (Director)

Dr. Alexandre Felixovich Guelboukh Kahn

Segundo vocal

Dr. Ricardo Barrón Fernández

Suplente

Dr. Salvador Godoy Calderón

EL PRESIDENTE DEL COLEGIO

Dr. Jaime Álvarez Gallegos

INSTITUTO POLITECNICO NACIONAL
CENTRO DE INVESTIGACIÓN EN COMPUTACIÓN
DIRECCIÓN

INSTITUTO POLITECNICO NACIONAL

SECRETARÍA DE INVESTIGACIÓN Y POSGRADO

CARTA CESION DE DERECHOS

En la Ciudad de **México D.F.** el día **10** del mes de **diciembre** del año **2008**, el (la) que suscribe C. **Ricardo Ávila Argüelles** alumno (a) del Programa de **Maestría en Ciencias de la Computación** con número de registro **B061058**, adscrito al **Centro de Investigación en Computación del Instituto Politécnico Nacional**, manifiesta que es autor (a) intelectual del presente trabajo de Tesis bajo las direcciones de los doctores, **Dr. Francisco Hiram Calvo Castro y el Dr. Alexander Gelbukh,** cede los derechos del trabajo intitulado ***"Clasificación bibliotecaria usando identificación simple de términos con métodos lógico-combinatorios a partir de información escasa"***, al Instituto Politécnico Nacional para su difusión, con fines académicos y de investigación.

Los usuarios de la información no deben reproducir el contenido textual, gráficas o datos del trabajo sin el permiso expreso del autor y/o director del trabajo. Este puede ser obtenido escribiendo a la siguiente dirección de correo electrónico **ravilab06@sagitario.cic.ipn.mx**. Si el permiso se otorga, el usuario deberá dar el agradecimiento correspondiente y citar la fuente del mismo.

Ricardo Ávila Argüelles

Resumen

Muchos libros cumplen con la clasificación de la biblioteca del congreso (*LCC*) al agregar su número correspondiente de clasificación. Esto es de suma utilidad para las bibliotecas de todo el mundo porque hacen posible su búsqueda y localización por lo que la *LCC* se ha convertido en un estándar bibliotecario. Existen varias situaciones por las cuales la *LCC* es asignada a un libro, tesis, documento u obra afines: cuando un libro nuevo es preparado para su publicación: cuando un libro existente carece de este número, o artículos contenidos en una colección de documentos que no han sido previamente asignados a este número, por ejemplo los que presenta Google scholar. Como se puede ver, una tarea automática para hacer esto es altamente necesaria. Esta permitirá el desarrollo de aplicaciones que asigne automáticamente a dichas obras su clasificación correspondiente.

Clasificar dichas obras no es trivial, aun existiendo el contenido completo de ellas. Tampoco es factible escribir manualmente la obra completa o sus resúmenes, menos si se trata de un conjunto grande de estas obras. Debido a esto, en esta tesis exploramos la posibilidad de asignar la *LCC* basándose únicamente en el título de la obra. Existen sistemas que hacen esto, pero nuestra eficiencia falla al asignar la clasificación en la precisión debido a que el título provee de muy poca información para este propósito.

En esta tesis, proponemos métodos que clasifican libros basándose únicamente en el título de las obras, que en muchos casos es con lo único con lo que contamos para este propósito y proponemos nuevas medidas de comparación, donde utilizamos el esquema de votación simple de identificación de términos con ponderaciones que permiten elevar la precisión de estos algoritmos.

Nuestros experimentos demuestran mediante el algoritmo 3 que se basa en descartar los títulos que no contienen los términos suficientes en la búsqueda, nos produce un **36.13%** de precisión tratándose de una evaluación estricta y comparando la profundidad completa de esta clasificación. Adicionalmente mostramos que este algoritmo puede mostrar 5 opciones a los bibliotecarios con una precisión del **53.35%** tomando la evaluación estricta.

Abstract.

A convenient way to find publications relevant for a given topic is to look for the Library of Congress classification number (*LCC*) corresponding to this topic, and then look for works corresponding to this *LCC*. Most books published nowadays provide this number as part of their bibliographic data.

In a number of situations the *LCC* is to be assigned to a specific book, paper, or dissertation: when a new book is prepared for publication; when an existing book lacks this number, or when papers in a document collection have not been previously assigned this number—for example, papers in Google Scholar text collection. As the last example shows, automating this task is highly necessary. This leads to the development of applications that can automatically assign a publication its *LCC*.

This task is not trivial even when the full text of the publication is available to the application. However, in some scenarios providing much data is not feasible. For example, dealing with a printed book, one cannot afford typing its contents or even abstract. Dealing with a document collection provided by an editorial house, or with a great number of the snippets that gives Google Scholar, one only has access to the title of the paper or book. Consequently, in this thesis we explore the possibility to assign the *LCC* basing only on the title of the work. Existing systems that solve this task fail to provide good classification accuracy, because the title provides insufficient information for it.

Our system uses supervised learning: given a collection of titles already classified and a new title, we look for the most similar titles in the collection and assign the new one to the category to which most of those belong. In addition, we propose 5 options with a 63% of improvement with the same methods.

Contenido

1 INTRODUCCIÓN

1.1 Planteamiento del problema

La clasificación de la biblioteca del congreso (***LCC*** por sus siglas en inglés) es ampliamente aceptada y utilizada por muchas bibliotecas importantes del mundo, esto se debe a que es la clasificación más extensa de libros que existe y asegura cubrir el mayor campo del conocimiento humano en materia de clasificación de libros.

Un problema que toma mucho tiempo a los bibliotecarios, es clasificar todos aquellos trabajos como son las tesis, artículos, revistas, etc. que carecen de una clasificación formal, al hacer su clasificación manual utilizan en la mayoría de los casos la clasificación del congreso (*LCC)* para efectuar esta tarea.

La presente tesis pretende lograr un algoritmo que sea capaz de cubrir este requerimiento usando únicamente el título de la obra, y también pretendemos ver el grado de aproximación que se puede obtener, dada esta gran restricción.

Existen muchos retos por vencer para lograr una correcta clasificación, entre estos y el más importante, es la ambigüedad que existe en la clasificación de los libros en la *LCC*, ya que existen títulos demasiado semejantes en las diferentes clases que la componen, convirtiendo a nuestro problema en un algoritmo de clasificación supervisado, jerárquico-plano, duro y discreto con un espacio de representación no-lineal y con datos cualitativos.

1.2 Hipótesis

La hipótesis de nuestro trabajo se basa en que el título de un libro proporciona los suficientes elementos para ser clasificado utilizando la clasificación de la biblioteca del congreso.

En la mayoría de los casos sólo se cuenta únicamente con el título de una obra literaria en los medios digitales, existen otros elementos como el autor y la editorial, pero estos últimos

no son términos construidos para identificar el contenido de una obra, estos son los constructores de la obra.

En la presente tesis mostraremos la metodología utilizada para demostrar la hipótesis dada esta gran restricción.

1.3 Objetivos

A continuación, se resumen los objetivos alcanzados en el desarrollo del presente trabajo de tesis.

1.3.1 Objetivo General

- Crear una herramienta de apoyo bibliotecario, que auxilie en la clasificación masiva de libros, tesis y obras a fines con el propósito de colocarlas en áreas correspondientes al tema específico de la obra siguiendo el estándar de clases de la biblioteca del congreso (*LCC*).

- Explorar la clasificación de la biblioteca del congreso con el objetivo de analizar su factibilidad hacia la construcción de ontologías automáticas.

1.3.2 Objetivos Particulares

- Proponer un clasificador de libros que utilizando solo el título de la obra y la clase general a la que pertenece, sea capaz de descartar clases de la clasificación de la biblioteca del congreso y proponga únicamente 5 clases factibles a los bibliotecarios, reduciendo con esto, el tiempo que ellos invierten en esta tarea.

- Identificar que tan útil puede ser la clasificación del congreso puede ser de utilidad para clasificación de textos y para algoritmos de construcción de resúmenes de textos, utilizando los métodos que se proponen en la presente tesis.

- Proponer una medida de comparación que eleve la eficiencia de la clasificación, reduciendo el tiempo en el proceso.

1.4 Aportaciones

Esta tesis es el resultado de la investigación realizada mediante una aplicación que utiliza diversas técnicas de clasificación mostrando los resultados obtenidos en cada una de ellas, donde nos muestran el grado de aportación que da el título de una obra para su clasificación en la biblioteca del congreso (*LCC*). Con esta aplicación aportamos:

- Una medida de comparación de títulos de las obras literarias con los títulos clasificados previamente por la *LCC*.

- El grado de aportación que da un texto corto para su clasificación jerárquica, dura y discreta teóricamente, utilizando diversas medidas de comparación.

- Presentamos una utilería para los bibliotecarios que provee de 5 posibles clasificaciones para un libro utilizando sólo el título de la obra. Con una precisión del 53%.

- Presentamos dos maneras de clasificar los términos y sus relaciones, de tal forma que nos aportan información sobre el área del conocimiento a las que pertenecen. Estas formas de clasificar son:

    - Clasificación de títulos (textos cortos), usando métodos lógico-combinatorios, logrando una precisión del **93%** en la tasa de aprendizaje y un **36%** en la proyección de los títulos nuevos. Adicionalmente mostramos 5 opciones con este mismo método con una precisión del **54%**

    - Clasificación de títulos (textos cortos), usando métodos de ponderación de palabras, logrando una precisión del 35.67% en evaluación estricta y 5 opciones con una precisión del 62.88%.

1.5 Estructura del documento

El resto del documento se encuentra organizado como sigue:

El *capítulo 2* se proporciona un breve repaso de los diferentes algoritmos semejantes al nuestro mostrando en una tabla loa métodos que utilizan y también mostramos en que difieren con el nuestro dando una descripción brevemente de cada uno de ellos. En el *capítulo 3* se muestran algunos conceptos necesarios para comprender la metodología de nuestra tesis. Si el lector conoce de estos tópicos puede omitir este capítulo e ir directamente al *capítulo 4* en el que se muestra en que consiste la propuesta del presente trabajo y la forma en la que se emplean las bases teóricas del capítulo 3. En el capítulo 4 mostramos la forma en que utilizamos los algoritmos y como los modificamos para lograr nuestros propósitos de clasificación, también mostramos la forma de usar la herramienta desarrollada. Por último en el *capítulo 5*, se describen los experimentos llevados a cabo, los resultados obtenidos y la forma de evaluación propuesta. Finalmente, en el capítulo 6 se encuentran las conclusiones del desarrollo del trabajo; así como también el trabajo futuro que la presente investigación sugiere.

1.6 Alcances y limitaciones

Los algoritmos desarrollados en la presente tesis, generan una alta carga de procesamiento, por lo que se requiere un alto poder de cómputo, memoria mínima de 2gb y procesador mínimo de 2.3 GHz, un procesador lento o una memoria con poca capacidad limitaría considerablemente el desempeño de los algoritmos.

La calidad de la muestra de aprendizaje es otro de los factores que reducen la precisión de estos algoritmos. Una calidad excelente sería la obtenida directamente de la biblioteca del congreso en modo original con sus campos muy bien delimitados. En la página web que ellos muestran públicamente, duplican títulos, insertan caracteres extraños, y mueven de posición los campos de título y autor.

2. ESTADO DEL ARTE

2.1 Conceptos básicos.

Explicamos algunos conceptos para aclarar nuestra tabla de comparación.

1. LCSH (*Library of congress subject headings*). Es un compendio de sinónimos y antónimos de todos los temas que se relacionan con el contenido de la obra, este compendio es actualizado por la biblioteca del congreso. Este es ampliamente utilizado en el registro de los libros, la LCSH es una parte integral para el control bibliográfico cuya función es organizar y esparcir documentos. [6]

2. MARC (*Machine Readable Cataloging*). Es un acrónimo que se utiliza en el campo de la ciencia de los libros, es un estándar de catalogación de libros legible para las computadoras, este es utilizado en la representación y comunicación de la bibliografía e información relatada en cierto formato legible para las máquinas.[6]

Otros conceptos básicos se describirán con mayor detalle en el capítulo de marco teórico.

2.2 Análisis y descripción de los algoritmos existentes.

Los siguientes 4 algoritmos fueron seleccionados por tener mayor semejanza con el nuestro:

A. Challenges in automated classification using library classification schemes – Pharos[1]

Kwan Yi, menciona 4 algoritmos de clasificación de textos que utilizan esquemas de organización bibliotecaria como son la *LCC* y Dewey como base para la clasificación de documentos digitales.

Los algoritmos a los que se refiere en su documento son:

1) Pharos.- Derivado del proyecto de librería digital de Alejandría. Pharos es un prototipo de clasificador heterogéneo basado en la *LCC* y que implementa para el propósito de crear perfiles de información digital heterogénea clasificando grupos de noticias y catalogando registros dentro de la *LCC*. La precisión la muestra como

la media aritmética entre el 13.0 % ±3.9 y el 76% ± 19 en 4,200 clases de la biblioteca del congreso, en un conjunto de entrenamiento de 1.5 millones de registros donde se usaron el título, el lcsh.

2) Scorpion. - Fue un proyecto de investigación del centro de bibliotecas computarizadas en línea (OCLC). Que se desarrolló en el año 1996 y se terminó en 1999 con el objetivo de desarrollar un método para identificar los documentos digitales para su clasificación automática en la cual scorpion utilizó el método de clustering. Desafortunadamente los resultados obtenidos en esta investigación no fueron revelados. En conclusión mencionaron que la clasificación automática no puede reemplazar la clasificación manual, pero puede proveer una solución efectiva para darle soporte a la catalogación humana.

3) DESIRE.- desarrollado en 1996, en la unión europea con el propósito de clasificar temas en materia de ingeniería usando empate simple de términos (*STM*). En la evaluación del sistema con aproximadamente 1000 paginas web, la precisión del la clasificación automática se reportó como cerca del **60**%.

B. Predicting library of congress classification from library of congress subject headings [2]

Eibe Frank y Gordon w. Paynter (2004), abordan el problema de asignar automáticamente la clasificación de la biblioteca del congreso (*LCC*). a una obra dada por el conjunto de los encabezados de temas del congreso (*LCSH library of congress subject headings*). La *LCC* está organizada en un árbol, el nodo raíz comprende todos los temas posibles, y los nodos hoja corresponden al tema mas especializado definido por sus áreas. Ellos proponen un procedimiento que utiliza técnicas de machine learning y un modelo de entrenamiento de datos mediante un catalogo grande que se obtiene de un conjunto de *LCSH* para su clasificación desde el árbol *LCC.*

Los resultados que muestran son **50,000** relaciones en pareja (LCSH,LCC), y utilizan mas información que el título de la obra y la clase general.

Logran un **55%** de precisión usando un conjunto de entrenamiento de **800,000** registros.

C.- The utility of information extraction in the classification of books [3]

Tom Betts, Maria Milosavljevic y Jon Oberlander, describen un trabajo de asignación automática de clasificación de libros usando el esquema de la *LCC*. Esta tarea no es trivial debido al volumen y variedad de libros que existen. Ellos exploran la utilización de técnicas de extracción de información (EI) y técnicas para la categorización de textos (CT) automáticamente utilizando el contenido total del libro.

Sus experimentos los desarrollan con un conjunto de libros previamente capturados en un proyecto de la biblioteca Gutemberg en los Estados Unidos. Sus estudios demuestran que su clasificador utiliza métodos y herramientas de EI y TC mejora significativamente sobre el clasificador que usa solo el estado del arte.

En sus resultados se observa que utilizaron **19,000** obras completas para el entrenamiento de su algoritmo y tomaron un conjunto de prueba de **1,029** obras completas logrando una precisión del **80.99 %** en sus pruebas.

D. Experiments in automatic library of congress classification.

Ray R. Larson en su documento denominado "Experiments in Automatic Library of Congress Classification" [4] publicado en 1992, presenta una investigación de 60 métodos de clasificación automática de libros utilizando la *LCC*, basados en títulos, LCSH y registros MARC.

Larson menciona en su investigación que encontró el mejor método para la clasificación en un algoritmo que tiene una precisión del 86% en un conjunto de prueba de 283 registros.

Larson omite en su versión libre de costo, el conjunto de registros utilizados para la proyección de sus 60 algoritmos de clasificación, pero Kwan Yi en su documento que en seguida mostramos, menciona que Larson entrenó el clasificador con 800,000 registros del catálogo, y tomó para la prueba un conjunto de 50,000 registros logrando un rango de precisiones entre el **55% al 80%.**

2.3 Tabla comparativa de los algoritmos.

La presente tesis pretende realizar una clasificación de libros utilizando la clasificación de la biblioteca del congreso (*LCC por sus siglas en inglés*) y sólo el título de la obra.

Existen muchos clasificadores automáticos que han pretendido realizar esta tarea, solo que cada uno de ellos tiene diferentes características que dejan de ser prácticas para los bibliotecarios. Los sistemas encontrados muy parecidos al nuestro son:

Utiliza	The Utility of Information Extraction in the classification of books	Predicting Library of Congress Classification from Library of Congress Subject Headings	Experiments in Automatic Library of congress Classification	Challenges in automated classification using library classification schemes - Pharos	Discriminación por presencia de términos (Algoritmo-3)
Contenido completo del libro	Sí				
LCSH		Sí	Sí	Sí	
Catalogación legible por máquinas (*MARC)*			Sí	Sí	
Técnicas de Categorización de textos (*CT*) y extracción de información (*EI*)	Si				
Técnicas de Machine Learning		Sí			
Conjunto de entrenamiento	19,000	800,000	800,000	1,500,000	490,016
Tamaño de la muestra	1,029	50,000	50,000	7,200	117,091
Precisión	80.99%	55.00%	86.00%	16.90%	36.67%

Tabla 1: Tabla comparativa de algoritmos parecidos con el nuestro, que utilizan el título de la obra, la *lcc* y el método de identificación simple de términos.

3. MARCO TEÓRICO

3.1 Identificación simple de términos (Simple Term Match – STM)

El concepto es ampliamente utilizado para medir frecuencia de los términos idénticos en dos o más elementos de comparación, estos pueden ser frases, documentos, palabras, etc. Este concepto es simple, solo se comparan todos los términos de las frases contabilizando todos aquellos que son idénticos entre ellas. Esto es muy útil para medir la semejanza que existe entre dos o más frases basándose en los términos idénticos. Por ejemplo.

El árbol nos protege del sol y nos quita el calor.
El sol es un astro que brinda calor.

Las palabras comunes en ambas frases es a lo que llamamos *STM*, que en este ejemplo son {sol, calor}.

3.2 Esquema de votación simple en el enfoque lógico combinatorio

Este algoritmo es conocido como ALVOT. Este tuvo su origen en año de 1965 aproximadamente *[8],* y sus desarrollos se deben al especialista ruso Yu. I. Zhuravliov y su grupo, posteriormente se continúa con su aplicación en los países de Cuba y México.

El libro *[8]* del que se extrajo la metodología de este algoritmo, utiliza subconjuntos de rasgos a los que le llama grupos de apoyo o grupos omega. Para nuestro estudio, utilizamos el conjunto total de términos que compone un título, por lo que el algoritmo original se mantiene sin modificaciones, únicamente aclaramos que utilizaremos el conjunto total de rasgos que componen a un objeto, que en nuestro caso es el título de la obra.

El modelo de los algoritmos de votación, se describe en 5 etapas:

1. Conjuntos de rasgos.
2. Función de semejanza.
3. Evaluación por objeto (renglón) dado un conjunto de rasgos.
4. Evaluación por clase (columnas) para todo el conjunto de rasgos.
5. Regla de solución.

Es decir, dar un algoritmo de votación A, significa dar un conjunto de parámetros en cada una de las 5 etapas enumeradas.

Por *conjuntos de rasgos* se entiende un conjunto no vacío de rasgos en términos de los cuales se analizarán los objetos.

Criterio de comparación. Es una función que tiene como entrada dos rasgos descriptivos de un mismo dominio y define la forma en que estos deberán ser comparados, dando como resultado un valor que se encuentra en el rango del 0 al 1.

$Cc_i(A,B) \rightarrow [0,1]$, donde A y B son rasgos descriptivos del mismo dominio.

Función de semejanza, es una función que realiza cálculos utilizando los criterios de comparación definidos para cada rasgo que compone el objeto. La función de semejanza es normalizada para que resulte en el rango de 0 y 1. Su descripción formal es:

$$f = (M_1 \times M_2 \times M_3 \times ... \times M_r) \times (M_1 \times M_2 \times M_3 \times ... \times M_r) \rightarrow [0,1]$$

Donde *M* es el conjunto de todos los rasgos que componen a los objetos de un cubrimiento.

La *evaluación por objeto (renglón) dado un conjunto de rasgos fijo*, se lleva a cabo una vez que se han definido el conjunto de rasgos y la función de semejanza. En esta se inicia un proceso de "contabilización" de votación, en cuanto a la medida de semejanza entre los diferentes rasgos de los objetos ya clasificados, con los que se desean clasificar. Cada renglón correspondiente a cada objeto, se compara con el objeto a clasificar por medio de la función de semejanza.

La *evaluación por clase (columnas) para todo el conjunto de rasgos*, es totalizar las evaluaciones obtenidas para cada uno de los objetos de las clases con el objeto a clasificar, esta totalización es evidentemente función de las evaluaciones por objetos (renglones) obtenidas previamente, es decir, se calcula la pertenencia del objeto a clasificar a las diferentes clases que componen el *cubrimiento*.

La *regla de solución*, es un criterio para la toma de decisiones. En esta se define el voto final, decidiendo la clase a la que pertenece el objeto a clasificar y el grado de pertenencia a dicha clase.

3.3 Term Frequency – Inverse Document Frequency (TF·IDF)

En 1999, Chris Manning y Hinrich Schütze [0] crearon un algoritmo para ponderar los pesos en las clasificaciones de términos para la recuperación de la información y la minería de datos.

El peso TF·IDF (Term Frequency · Inverse Document Frecuency) es una medida estadística usada para evaluar la importancia de un término en un documento dentro de una colección de documentos. Esta importancia crece proporcionalmente con el número de veces que el término aparece en el documento, ponderado por su presencia en el resto de los documentos.

Es un peso frecuentemente usado en la recuperación de documentos y en la minería de datos. Este peso es una medida estadística usado para evaluar la importancia que tiene una palabra dentro de un documento que se encuentra en una colección de documentos. Este grado de importancia aumenta proporcionalmente con el número de veces que aparece una palabra en el documento, pero es ponderado por la frecuencia de la palabra en el documento donde se encuentra. Este peso se calcula de la siguiente manera.

Frecuencia del término (TF). Dado un documento, es simplemente el número de veces que aparece el término en dicho documento, el cual es usualmente normalizado para prevenir las palabras que se repiten demasiado en un documento causando ruido generando poco grado de importancia de dicho término en su documento en particular.

Frecuencia del término en los documentos df. Es el número de veces que aparece el término en el resto de los documentos, normalizado con el número de documentos totales que se están analizando.

Inversa de la frecuencia del documento (idf). Es el logaritmo base 10 o logaritmo natural, del número de documentos totales que se están analizando normalizados por el número de veces que aparece en término en el resto de los documentos.

$$idf = \log\left(\frac{|D|}{|\{d_j : t_i \in d_j\}|}\right)$$

Donde

D es el número total de documentos.

d_j es un documento que pertenece al conjunto *D*. es decir, $d_j \in D$.

t_i es el *i* término contenido en d_j.

Finalmente, el factor tf·idf, es el resultado de multiplicar *tf* con *idf* definidos previamente, donde un valor alto significa que el término es de mucha importancia para la clasificación, mientras que un valor bajo, significa que el término no es útil para este objetivo.

3.4 Clasificación de la biblioteca del congreso

Es un sistema de clasificación desarrollado desde el siglo XIX por el año de 1890, se realizó como una solución a los problemas clasificación del sistema original que fue desarrollado por Thomas Jefferson, el problema mas importante que impulsó su creación es que el sistema original había sido saturado con la cantidad de libros clasificados, la *LCC* fue inicialmente construida por el jefe de catalogación James C. M. Hanson y el jefe de clasificación Charles Martel, tomando como base el esquema de clasificación de Charles Ammi Cutter conocido como el "sistema expansivo" que combina letras con números.

La *LCC* continuamente está siendo modificada por la biblioteca del congreso y se usa en la mayoría de las bibliotecas de investigación de este país así como también en otros más como el nuestro.

El objetivo es organizar colecciones de libros en base a sus contenidos de tal forma que sea sencillo encontrarlos en áreas comúnmente relacionadas al tema de investigación.

En la mayoría de las bibliotecas públicas o académicas pequeñas, siguen utilizando el sistema *Dewey*, el cual utiliza únicamente números decimales debido a que esta clasificación es suficiente para cubrir sus necesidades de clasificación.

La *LCC* divide todo el conocimiento en las siguientes 21 clases básicas (vea ilustración1:).

Cada clase es identificada por una letra sencilla del alfabeto, y esta a su vez, se subdivide subclases, las cuales también se identifican con letras. Las clases y subclases se construyen con un máximo de 3 niveles jerárquicos utilizando el esquema de las letras.

En la *LCC* e parte de lo general a lo más específico en forma jerárquica, desde las clases más gruesas hasta las clases más delgadas, utilizando desde uno hasta cuatro dígitos en la longitud después de las letras que le preceden.

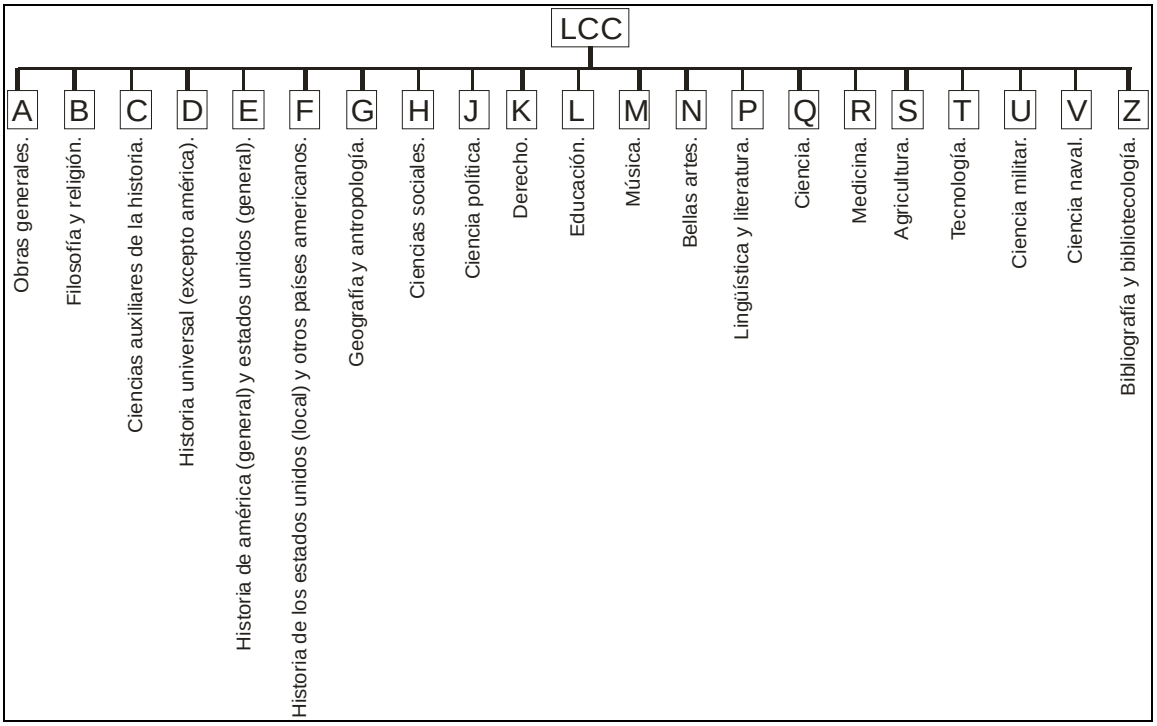


Ilustración 1: Descripción básica de la *LCC*.

La clase Q consta de las siguientes subclases:

- QA. Matemática.
- QB. Astronomía.
- QC. Física.
- QD, Química.
- QE. Geología.
- QH. Historia Natural (General).
- QK. Botánica.
- QL. Zoología.
- QM. Anatomía Humana.
- QP. Fisiología.
- QR. Microbiología.

Un ejemplo práctico de clasificación es:

LCC	Título	Autor, Edición
QA154 .W42	**First course in algebra and number theory.**	**Weiss Edwin, 1971**

Las primeras dos letras significan lo siguiente:

Q = Ciencias.

A = Matemática.

La unión de las dos letras "QA" tiene estrictamente definidos por la *LCC* los siguientes rangos para la definición del área a la que corresponde.

1 – 939	Matemática
9 – 10.3	Lógica Matemática
75.5 – 76.95	Ciencia de las computadoras. Procesamiento electrónico de datos.
101- 141.8	Matemática elemental. Aritmética.
150 – 272	**Algebra.**
273 – 274.76	Probabilidad.
276 – 280	Estadística matemática.
297 – 299.4	Análisis numérico.
299.6 – 433	Análisis (Incluye métodos analíticos conectados con problemas físicos.
440 – 699.4	Geometría.
611 – 614.97	Topología.
01 – 939	Mecánica Analítica No para teoría de la mecánica, ver QC120+

El resto de la clave del ejemplo “W42” corresponde a la primera letra del apellido paterno del autor, el número 42 es el consecutivo de la obra de los autores que comienzan con dicha letra en las obras previamente clasificadas por la biblioteca del congreso.

El alcance de nuestra tesis consiste en clasificar el título, omitiendo el resto de la clasificación, ya que este no es relevante para el uso práctico de clasificación en las bibliotecas, ya que dada la clasificación propuesta, es suficiente para agrupar las obras en base a su contenido que corresponde a un área específica del conocimiento humano.

4. METODOLOGÍA

Nuestra investigación se basa en una serie de algoritmos que aplicamos al título de un libro con el objetivo de lograr una clasificación aceptable analizando a detalle las características de los términos que lo componen buscando el mayor índice de precisión.

Debido a que contamos con una muestra inicial de aprendizaje, podemos comparar los resultados de cada algoritmo con los resultados que se consideran correctos desde la clasificación de la biblioteca del congreso.

El procedimiento aplicado a todos los algoritmos desarrollados es el siguiente:

Paso 1.) Divide el patrón de búsqueda en palabras que aportan significado para su clasificación eliminando a aquellas que no logran este objetivo *(Stopwords)* como es el caso de los artículos y preposiciones en el caso del inglés y del español.

Ejemplo:

Título patrón, es el título de búsqueda:
ANATOMY FROM THE GREEKS TO HARVEY.

Título en de la clase con el que se está comparando:
SHORT HISTORY OF ANATOMY FROM THE GREEKS TO HARVEY.

Se eliminan *stopwords.*

ANATOMY ~~FROM~~ ~~THE~~ GREEKS ~~TO~~ HARVEY.
SHORT HISTORY ~~OF~~ ANATOMY ~~FROM~~ ~~THE~~ GREEKS ~~TO~~ HARVEY.
Y las intersecciones entre los 2 títulos son los que consideramos para nuestra investigación.

Título1 ∩ Título2 = {ANATOMY, GREEKS, HARVEY}

Calculando con esto los siguientes factores:

STM = 3

Longitud del título de búsqueda = 3
Longitud del título en la clase = 5

Algunos conceptos básicos para comprender la presente tesis se explican a continuación.

Paso 2) Se crea una tabla donde se grabará la prueba indicando en el nombre de de la case que se está procesando, número del algoritmo que se aplica y finalmente la tasa utilizada para el aprendizaje y proyección.

Paso 3) Abre el archivo que contiene los títulos a clasificar y posiciona el apuntador al inicio de la lista.

Paso 4) Para todos los títulos del archivo abierto previamente, aplicamos los algoritmos que requeridos por el usuario en la pantalla.

Paso 5) Presentación de resultados. Se crea un reporte comparativo donde informa el comportamiento de la precisión a lo largo de los experimentos, la precisión final, las gráficas que informan la posición de la clave correcta y los datos generales del experimento.

Ejecutando esta serie de pasos previos a la clasificación, aplicamos los siguientes algoritmos que describimos con mayor detalle en resto de este capítulo.

4.1 Algoritmo 0 – Votación por Frecuencia de Términos (VFT)

Utilizando el concepto STM *(Explicado en el capítulo 3)*, tomamos los términos contenidos el título que se desea clasificar y medimos su frecuencia en cada una de las clases que los contienen. Al obtener las frecuencias por términos, el voto se le da a aquella clase que contenga la mayor frecuencia calculada. Un ejemplo, que puede servir para aclarar este punto es:

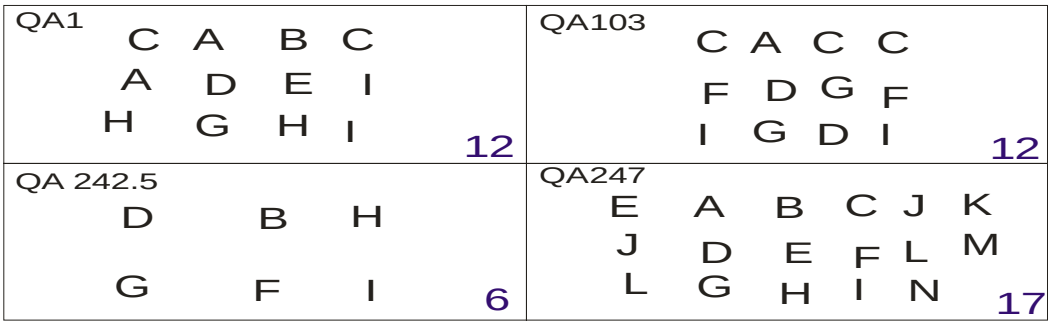
Supongamos que tenemos un título con 4 términos y 4 Subclases de QA y que A, F, D, y C, son cuatro términos contenidos en el título que se desea clasificar y existen cinco de estos términos en la clase QA1, siete en la clase QA103, dos en la clase QA242.5 y cuatro en la clase QA247, entonces se le da el voto final a aquella clase que contiene mayor

cantidad de estos términos, dado este caso, el voto corresponde a la clase QA103, como se muestra en la Ilustración 2.

Entonces tenemos:

Título = A F D C

Subclases = {QA1, QA103, QA242.5, QA247}



y queremos clasificar el título A F D C

Palabra\Clase	QA1	QA103	QA242.5	QA247
A	2	1	0	1
F	0	1	1	1
D	1	2	1	1
C	2	3	0	1
Sumas	5	7	2	4

Ilustración 2: Ejemplo de votación simple

Podemos observar que la clase que contiene la mayor cantidad de términos, es la que obtiene el voto final.

Entonces el algoritmo queda de la siguiente manera:

0. Extraemos los términos del título.
1. Eliminamos *stopwords* (artículos, preposiciones).
2. Calculamos la frecuencia de los términos en las clases
3. Aplicamos regla de solución.
4. El voto se le asigna a la clase que presenta mayor frecuencia dada por los términos.

4.2 Algoritmo 1 – Votación por Frecuencia de Términos Ponderada (VFTP)

Un concepto importante para comprender en este algoritmo es la *presencia de términos en la clase*, que es únicamente la existencia o ausencia de los términos (del título que se va a clasificar) en las clases, es decir, un término aporta 1 si existe en la case, y 0 en caso contrario. Ejemplo:

Utilizando el mismo ejemplo del algoritmo 0, tenemos una presencia de términos de la siguiente manera:

	QA1	QA103	QA242.5	QA247
A	1	1	0	1
F	0	1	1	1
D	1	1	1	1
C	1	1	0	1
Presencia	3	4	2	4

Una vez calculadas las presencias de los términos en las clases, procedemos al cálculo de la frecuencia de los términos y la ponderamos entre la cantidad de términos totales que contienen cada una de las clases. El valor obtenido de la ponderación, lo sumamos con la presencia de términos en la clase que previamente calculamos y damos el voto a aquella clase cuyo valor obtenido representa mayor cantidad. Por ejemplo:

Palabra\Clase	QA1	QA103	QA242.5	QA247
A	**2/12**	**1/12**	**0/6**	**1/17**
F	**0/12**	**1/12**	**1/6**	**1/17**
D	**1/12**	**2/12**	**1/6**	**1/17**
C	**2/12**	**3/12**	**0/6**	**1/17**
Frecuencia del término	**5/12**	**7/12**	**2/6**	**4/17**
Presencia de términos en la clase	**3**	**4**	**2**	**4**
Frecuencia del término + Presencia	**3.4167**	**4.5833**	**2.3333**	**4.2353**

Ilustración 3:Ejemplo de clasificación por algoritmo 1 - *VFTP*

En este ejemplo *(ver Ilustración 3)*, se presenta un empate entre la QA103 y la QA247 al nivel de cálculos de la presencia de los términos en las clases. Este empate lo resolvemos sumando la frecuencia de los términos y con esta ponderación damos el voto final a aquella clase que representa el mayor valor resultante de esta suma.

El algoritmo entonces, queda de la siguiente manera:

1. Extraemos los términos del título.
2. Eliminamos *stopwords* (artículos, preposiciones).
3. Calculamos la frecuencia de los términos en las clases.
4. Sumamos la cantidad de términos presentes en cada clase.
5. Calculamos el factor de votación.

$$factor = \text{presencia de términinos en las clases} + \left(\frac{\text{Frecuencia del término en la clase}}{\text{Número total de términos de búsqueda}} \right)$$

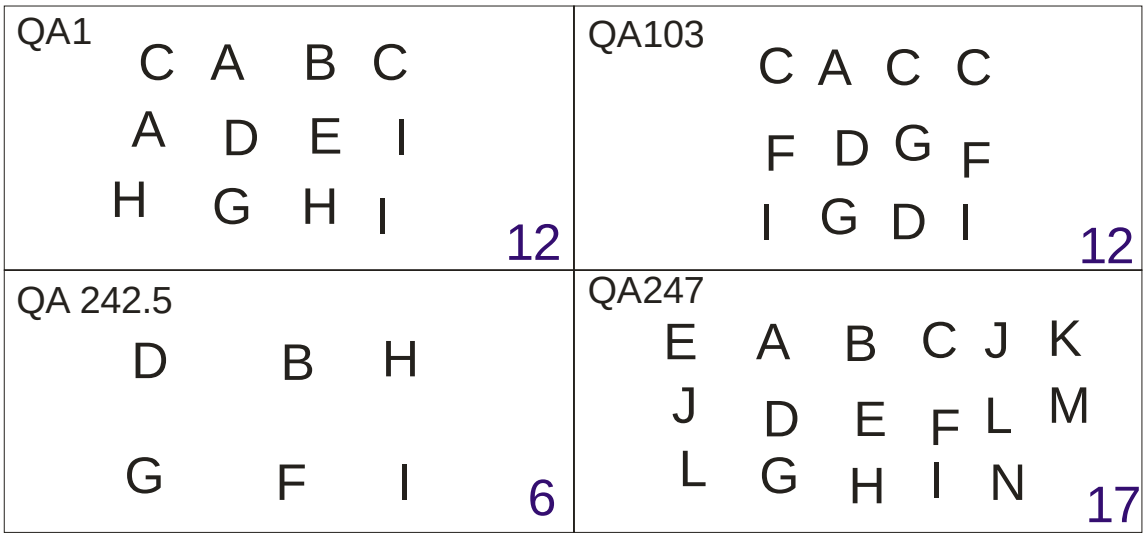
6. Aplicamos regla de solución.
    - El voto se le asigna a la clase que resulte con mayor valor calculado.

De este método, extraemos un discriminante para las clases y lo hemos denominado *DPT* (Discriminación por Presencia de Términos) que explicamos a detalle en el tema 4.4

4.3 Algoritmo 2 – Votación por Frecuencia de Términos Ponderada con factor TF·IDF (VFTP-TF·IDF)

Explicaremos brevemente como utilizamos esta medida estadística ejemplificándolo de la siguiente manera (ver Ilustración 4). Supongamos que tenemos 4 clases (QA1, QA103, QA242.5, QA247), y un patrón de búsqueda con 4 términos (A,F,D,C) que deseamos clasificar, entonces el ejemplo queda de la siguiente forma.

El contenido de las clases es:



1) Calculamos TF = $\frac{\text{Frecuencia del término en la clase}}{\text{Términos totales de la clase}}$ para todos los términos del título por clasificar.

Palabra\Clase	QA1	QA103	QA242.5	QA247
A	**2/12**	**1/12**	**0/6**	**1/17**
F	**0/12**	**1/12**	**1/6**	**1/17**
D	**1/12**	**2/12**	**1/6**	**1/17**
C	**2/12**	**3/12**	**0/6**	**1/17**
Frecuencia del término	**5/12**	**7/12**	**2/6**	**4/17**

Ilustración 4 Ejemplo TF·IDF

2) Calculamos $IDF = \log\left(\frac{\text{Número total de clases}}{\text{Clases con al menos un término}}\right)$ de la siguiente manera:

IDF en A = Log(4/3) = 0.124939
IDF en F = Log(4/3) = 0.124939
IDF en D = Log(4/4) = 0
IDF en C = Log(4/3) = 0.124939

3) Multiplicamos TF·IDF: Ejemplo del primero (2/12)(0.124939), y de la misma forma para todos los casos.

TF·IDF	QA1	QA103	QA242.5	QA247	Sumas
A	0.020823	0.010412	0.000000	0.007349	0.038584
F	0.000000	0.010412	0.020823	0.007349	0.038584
D	0.000000	0.000000	0.000000	0.000000	0.000000
C	0.020823	0.031235	0.000000	0.007349	0.059407
Sumas	0.041646	0.052058	0.020823	0.022048	

Tabla 2: Ejemplo del factor TF·IDF

4) Sumamos y damos el voto final a la clase con mayor valor, en este caso QA103.

Podemos observar claramente como el término D queda fuera de la clasificación, ya que se presenta en todas la clases y por lo tanto su aportación es cero.

La fuerza de este esquema de comparación, radica en la naturaleza del logaritmo, ya sea natural o base10, en donde se invierte el 1 en 0 indicando que si un término se encuentra en todas las clases, entonces el término no aporta nada para la clasificación. Así mismo, si se reduce o aumenta su aportación gradualmente en base a la presencia de dicho término en los documentos. Para nuestro caso, si un término se encuentra en todos los títulos que clasificaremos, entonces el término no aporta nada a la clasificación del título, en caso contrario, si un término se encuentra únicamente en una clase y no se encuentra en las demás, el grado de aportación de ese término tiende a 1 proporcionalmente a su menor presencia en el resto de las clases.

Es importante aclarar la analogía que existe con el logaritmo de cero en cualquiera de sus bases. El factor *tf·idf* únicamente se aplica cuando existen términos del título a clasificar en los títulos que componen a una clase, así mismo, si un término no aparece en alguna clase, entonces no existe empate simple de términos, por lo que no se aplica este factor y evitamos de esta manera este problema.

El algoritmo 2 (VFTP-TF·IDF), entonces queda de la siguiente manera.

1. Extraemos los términos del título.
2. Eliminamos *stopwords* (artículos, preposiciones).
3. Aplicamos Algoritmo 1 (Calculamos la frecuencia de los términos ponderada).
4. Aplicamos el factor estadístico TF·IDF
5. Aplicamos regla de solución.
    - El voto lo obtiene la clase que resulte con mayor valor calculado.

4.4 Algoritmo 3 – Discriminación por Presencia de Términos (DPT)

Esta es una aportación de nuestra tesis. Este discriminante es el producto de las observaciones hechas en los experimentos, donde consideramos que la clase correcta se encuentra entre las que contienen mayor presencia de términos.

Basado en el algoritmo 1 y tomando a aquellas clases que tienen el máximo numero de términos presentes en ellas, descartamos a aquellas que son menores a este parámetro, quedando el algoritmo de la siguiente manera.

1. Extraer los términos del título.
2. Eliminar *stopwords* (artículos, preposiciones).
3. Calcular la presencia de los términos en cada clase.
4. Retirar las clases que sean menores a la máxima presencia de términos calculada.
5. Aplicar regla de votación.
    - Si queda una clase, el voto lo obtiene esa clase con mayor presencia de términos.
6. Aplicar regla de solución.
    - En caso de 2 o más votos iguales,
        - calcular factor TF·IDF en los valores calculados.
        - Le sumar la presencia de términos calculada.
        - El voto lo obtiene la clase con el mayor valor resultante.

Siguiendo el ejemplo dado en el algoritmo 1, tenemos las siguientes clases con las siguientes frecuencias de términos en cada clase:

	QA1	QA103	QA242.5	QA247
A	1	1	0	1
F	0	1	1	1
D	1	1	1	1
C	1	1	0	1
Presencia	3	4	2	4

Descartamos las clases QA1 y QA242.5 por ser inferiores a 4 que es el máximo valor de términos presentes en las clases, resultando finalmente las clases QA103 y QA247 como posibles respuestas factibles.

4.6 Algoritmo 4 y 4´- Clasificación de Títulos con Métodos Lógico-combinatorios (CT-MLC)

Adaptando los algoritmos de este enfoque a nuestros propósitos, aclaramos los siguientes conceptos:

1. Nuestros objetos a clasificar son títulos, y cata título contiene términos que serán utilizados en la función de semejanza.
2. Comparamos con una función de semejanza cada título a clasificar, con todos los títulos en la muestra, dando como resultado una matriz que contiene todos los resultados de las comparaciones, a la que denominaremos, *matriz de semejanza*.
3. Creamos 2 matrices de semejanzas en base a las funciones de semejanza que explicaremos en este tema, mas adelante.
4. Aplicamos una *regla de solución*. Es una fórmula o conjunto de criterios que responden a dos fundamentos:
    a. Definir la clase a la que pertenece cada uno de los títulos.
    b. Que grado de pertenencia tiene cada uno de de ellos a la clase, como nuestro algoritmo en sus dos variantes es de clases duras, entonces la pertenencia del título a la clase correspondiente es 1.

Aclaramos que el costo de estos algoritmos se encuentra precisamente en el cálculo de las semejanzas del título a clasificar, con los títulos de las clases que en muchas ocasiones es muy extenso.

Para aplicar este algoritmo con enfoque lógico combinatorio, asignamos todos los títulos con su número total de términos (es decir, sin dividir los términos que lo componen) a la clase que pertenece y lo comparamos mediante funciones de semejanza con el título que se desea clasificar.

Las funciones de semejanza aplicadas a los algoritmos 4 y 4', las definimos de la siguiente manera.

a) Semejanza entre títulos. (Algoritmo 4)

$$f_{(T_i,T_j)} = \frac{\text{STM}}{\max\{\text{longitud}(T_i), \text{longitud}(T_j)\}}$$

Donde STM es la cantidad de términos idénticos entre los dos patrones

Donde

- T_i es el título a clasificar.
- T_j .es el título contenido en la clasificación aprendida.
- Max es una función que devuelve el máximo valor de los dos parámetros.
- Longitud (T_i) es una función que devuelve la cantidad de términos contenidos en el título T_i .

b) Semejanza del título a clasificar con los títulos de una clase. (Algoritmo 4')

$$f(T_i, Q_j) = \frac{\sum_{T_p \in Q_j} f(T_i, T_p)}{|Q_j|}$$

Donde:

- T_i es el título a clasificar.
- Q_j es la clase que contiene los títulos semejantes.
- T_p es el título que pertenece a la clase Q_j

En los experimentos realizados en esta tesis mostramos dos formas de evaluar el voto: *(Que en las pruebas numeramos como experimentos 4 y 4')*

1. Aquella clase que contiene un título con máxima semejanza con el título se desea clasificar.

2. Aquella clase que contiene máxima semejanza promedio con el título se desea clasificar.

Un ejemplo de clasificación utilizando este enfoque sería:

Título a clasificar: PRACTICAL MATHEMATICS

Longitud del título a clasificar = 2

Elementos en las clases

Semejanza	STM	Long. del Título	Sub-Clase	Título
0.50	1	1	QA39	MATHEMATICS
0.50	1	1	QA39	MATHEMATICS
0.50	1	1	QA43	MATHEMATICS
0.50	1	1	QA5	MATHEMATICS
0.00	0	1	QA37	BIOMATHEMATICS
1.00	2	2	QA39.2	PRACTICAL MATHEMATICS
0.50	1	2	QA103	PRACTICAL ARITHMETIC
0.50	1	2	QA39	MATHEMATICS USE
0.50	1	2	QA39	MATHEMATICS USE
0.50	1	2	QA39	NEW MATHEMATICS
0.50	1	2	QA95	FUN MATHEMATICS
0.50	1	2	QA37	TREE MATHEMATICS
0.33	1	3	QA28	MEN MATHEMATICS ET BELL
0.33	1	3	QA303	COURSE PURE MATHEMATICS
0.50	1	2	QA37	ENGINEERING MATHEMATICS
0.50	1	2	QA37	FUNDAMENTAL MATHEMATICS
0.17	1	6	QA5	DICTIONARY MATHEMATICS JA GLENN GH LITTLER
0.20	1	5	QA501	PRACTICAL DESCRIPTIVE GEOMETRY, GRANT
0.25	1	4	QA76	INTERNATIONAL JOURNAL COMPUTER MATHEMATICS
0.20	1	5	QA76.58	PRACTICAL PARALLEL COMPUTING STEPHEN MORSE
0.17	1	6	QA37	MATHEMATICS MEASUREMENTS MERRILL RASSWEILER MERLE HARRIS
0.14	1	7	QA37.2	APPLIED FINITE MATHEMATICS RICHARD COPPINS PAUL UMBERGER
0.14	1	7	QA37.2	EUCLIDEAN SPACES PREPARED LINEAR MATHEMATICS COURSE TEAM
0.17	1	6	QA37.2	FOUNDATIONS MATHEMATICS KENNETH BERNARD HENRY WELLENZOHN
0.17	1	6	QA37.2	MATHEMATICS APPLICATIONS LAURENCE HOFFMANN MICHAEL ORKIN

Tabla 3: Ejemplo práctico de clasificación con enfoque lógico combinatorio

En la Tabla 3 podemos observar que el voto lo obtiene la clase que contiene el título con mayor semejanza al título que se desea clasificar, en este caso, sería QA39.2.

En el presente trabajo de investigación mostramos también los resultados que obtuvimos aplicando la semejanza promedio del título de búsqueda con los elementos que componen a la clase.

Finalmente los algoritmos 4 y 4' tienen el siguiente procedimiento:

1. Extraer los términos del título.
2. Eliminar *stopwords* (artículos, preposiciones).
3. Calcular semejanza de los términos del título por clasificar con los términos de todos los títulos en las clases que contienen al menos un término semejante.
4. Calcular semejanza promedio por clase, aplicando las funciones de semejanza que definimos en los incisos a y b antes mencionados.
5. Aplicar reglas de votación.
   - Algoritmo 4
     1. El voto lo obtiene la clase que contiene el título más semejante al que se desea clasificar.
   - Algoritmo 4'
     2. El voto lo obtiene la clase con mayor semejanza promedio con el título que se desea clasificar.

4.7 Aplicación desarrollada

El sistema diseñado para ejecutar los algoritmos propuestos en esta tesis, fue desarrollado en *borland delphi* 7.0, utilizando el manejador de bases de datos de *microsoft sql server*, requerimos para su óptimo desempeño una máquina Pentium 4, con un monto de 2 MB de memoria RAM, y un disco duro con un monto mínimo de 10 GigaBytes disponibles.

En esta aplicación utilizamos 5 algoritmos para la clasificación experimental, los resultados obtenidos se comparan con la clasificación original dada por la *LCC*, ya que se trata de un algoritmo supervisado, y finalmente evaluamos la precisión en 3 formas:

1. *Evaluación estricta*. La clave experimental es exactamente igual a la clave original con toda la profundidad que se presentan.
2. *Evaluación hasta el punto decimal*. Las claves son idénticas hasta el punto decimal.

3. *Evaluación por posición*. Lugar de semejanza con respecto a las primeras 5 posiciones de la original. Dado el algoritmo que le corresponda en número de votos.

Los algoritmos mencionados son los que describimos con detalle en el capítulo 4 y son:

0. Algoritmo de votación por frecuencia de términos *(VFT)*.
1. Algoritmo de votación por frecuencia de términos ponderado*(VFTP)*
2. Algoritmo de votación por frecuencia de términos ponderado con factor tf idf
3. Algoritmo de discriminación por presencia de términos.
4. y 4’ - Algoritmo de clasificación de títulos con métodos lógico-combinatorios.
    a. Semejanza con el elemento más semejante a una clase.
    b. Semejanza promedio con los elementos que integran una clase.

4.7.1 Proceso del clasificador

Extracción de los registros e la página oficial de la biblioteca del congreso (1.a).

0. Fase de definición de los experimentos.
    - Preparar archivos de tipo texto con información extraída de la página oficial de la *LCC*.[1.a)]
    - Indicar en los campos de la pantalla inicial, los parámetros que requiere el experimento (Clase, nombres de los archivos de entrada, algoritmos por aplicar, utilizar solo el inglés, etc).

1. Fase de importación, filtración de los títulos y distribución de la muestra.
    - Tomar el archivo preparado en la pantalla principal que ese será la fuente de entrada de datos para el entrenamiento y distribución de las muestras de los algoritmos.
    - Eliminamos símbolos y *stopwords*.
    - Aplicamos filtro del idioma inglés si así lo requiere el usuario.
    - Creamos 2 archivos, uno para indicar los registros utilizados en el entrenamiento (train.txt) y otro para indicar los que serán utilizados en la proyección de los datos en los diferentes algoritmos (Proy.txt)

2. Fase de entrenamiento del algoritmo.
    - Recorrer todos los registros del archivo de entrenamiento preparado en la fase.
    - Por cada uno recorrido, depositar en un almacén temporal el título filtrado
    - Como los registros importados vienen en orden de clase, hacer corte en cada clase.

- En cada corte, seleccionar del almacén temporal el número de registros tomados al azar en base al porcentaje solicitado en la fase de definición.
- Grabar la tabla de fichas, que contiene los títulos de entrenamiento, en una base de datos sql
- Grabar la tabla de *keywords*, que contiene las palabras contenidas en los títulos de la tabla de fichas y la clase a la que pertenece cada uno de los *keywords* contenidos en dicho título.
- Calculamos algunas ponderaciones que se utilizarán en los algoritmos y eliminamos el tiempo de cálculo en la fase de experimentación.
  - TF·IDF por keyword (en forma global).
  - Frecuencia del keyword en la clase
  - Keywords totales en la clase

3. Fase de experimentación; proyección.
   - Toma el archivo preparado en la fase 1 que se destinó para la proyección de los algoritmos.
   - Por cada uno de ellos filtramos título y clave preasignada.
   - Colocamos el título en el algoritmo correspondiente
   - Ejecutamos el algoritmo correspondiente
   - Comparamos el resultado del algoritmo con el resultado que es el que corresponde correctamente.
     - Si son idénticos en toda la extensión de la clave, ponemos 1 a al campo “exito1” o 0 en su defecto.
     - Si son idénticos hasta el punto decimal que los divide, ponemos 1 a al campo “exito2” o 0 en su defecto.
     - Localizamos en la base de datos la posición que tomó en la tabla final de votación del algoritmo el resultado correcto que debió regresar y la guardamos en el campo “calif”.
   - Graficamos la precisión por cada uno de los registros evaluados.

4. Fase de resultados.
   - Tomamos los experimentos generados en la fase anterior que se guardaron en sus tablas correspondientes, y graficamos.
   - Actualizamos los datos del reporte
   - Presentamos reporte final con gráficas y precisiones.

4.7.2 Detalle de la aplicación

La *LCC* tiene un conjunto de registros de libros que son de dominio exclusivo de la biblioteca, así como también tiene a la venta el conjunto total de esas clasificaciones, solo

que el costo de adquisición de la clase completa llega a tener un valor hasta de 1,000 dólares por clase. Así mismo, existe un subconjunto de títulos que es de dominio público y son libres de costo, de este subconjunto extrajimos todos los existentes en la clase Q (Ciencias) para probar nuestro algoritmo y consideramos que son suficientes para los experimentos.

Es importante mencionar que estos títulos los muestra la *LCC* en páginas con 100 elementos máximo por página web. Una vez extraídos de la página web, los grabamos en archivos planos ANSI tipo texto (*.txt*), para que la aplicación los pueda leer y procesar.

Una vez grabados los archivos planos, procedemos a utilizar la herramienta de la siguiente manera

4.7.2.1 Pantalla General (inicial)

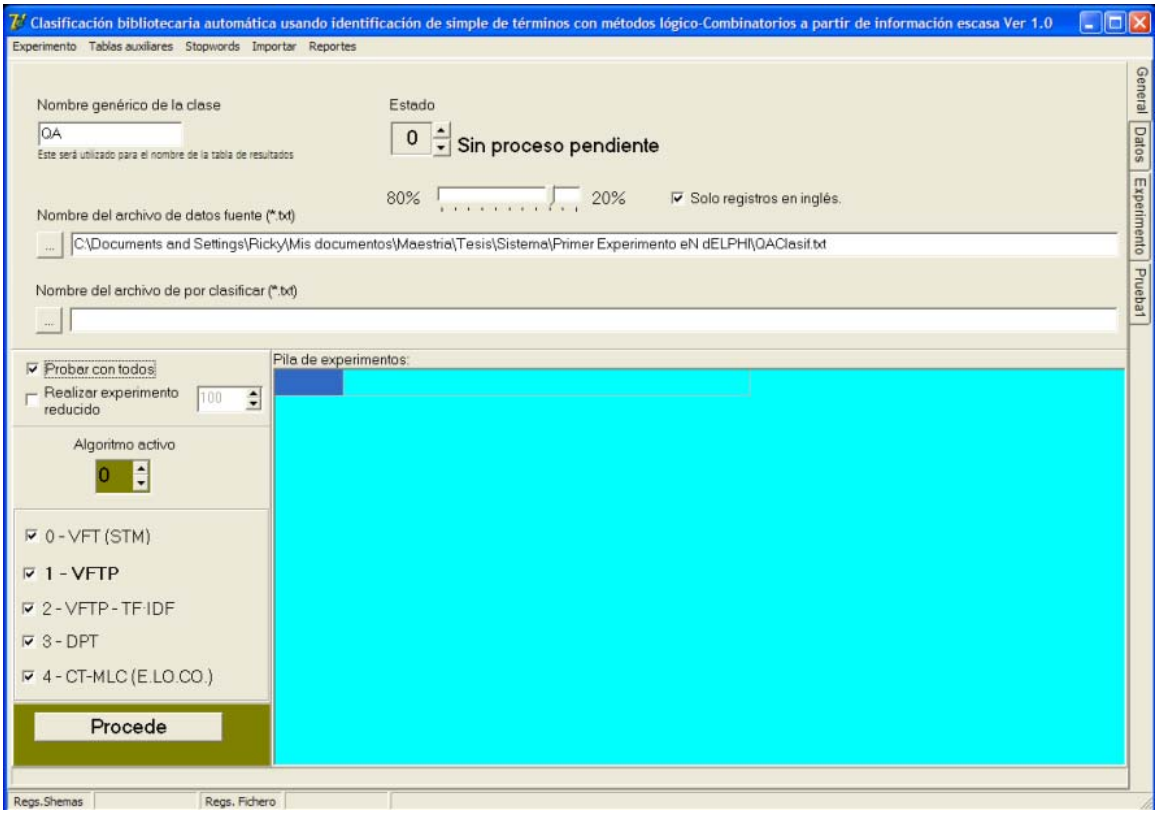


Ilustración 5: Pantalla Inicial

En la pantalla principal mostramos una interfaz gráfica que permite al usuario definir las características de su experimento. Estas especificaciones son:

- *Nombre genérico de la clase que se desea clasificar*. Funciona solo como identificador del experimento, se utiliza principalmente para nombrar los archivos de resultados.

- *Factor de entrenamiento*. Este indicador especifica el porcentaje que requiere el usuario para la tasa de entrenamiento y el resto para la tasa de proyección del algoritmo.

- *Indicador Solo ingles*. Si se encuentra activa, entonces se aplica el filtro del idioma ingles a la obtención de la muestra.

- *Nombre del archivo de texto* que contiene toda la muestra competa.

- *Nombre del archivo de texto para entrenamiento*. Este se utiliza en el caso de restauración del sistema, ya que este se genera automáticamente al terminar el proceso de importación de datos.

- *Indicador de todos*. Especifica si se desea experimentar con todos los algoritmos disponibles, generando resultados por independiente para ser analizados posteriormente.

- *Experimento reducido*. Se utiliza en el caso que se desea tomar una muestra del conjunto de prueba a un número de registros definida por el usuario y seleccionada aleatoriamente por el sistema.

- *Algoritmo Activo*. Indica el algoritmo que se aplicará al momento de ejecutarse el clasificador.

- *Casillero de algoritmos*. Es una parte de la pantalla principal que permite seleccionar el conjunto de algoritmos que serán aplicados en el clasificador durante la prueba.

- *Botón Procede*. Inicia el proceso de clasificación en base a las especificaciones antes mencionadas.

4.7.2.2 Descripción del experimento

Es procedimiento inicial para el clasificador automático, en esta fase debemos indicar:

1. El nombre de la clase o subclase con la que se desea experimentar.
2. La tasa de aprendizaje.
3. Si se desea aplicar filtro del idioma inglés.
4. La ruta y nombre del archivo fuente donde se encuentra la muestra total.
5. Los algoritmos que serán utilizados, solo dar click en las cajitas correspondientes.
6. Si se desea tomar un subconjunto de registros de la muestra de la proyección para el experimento.
7. Si se desea apilar la definición del experimento para que se ejecuten en el orden de la pila, del menú principal seleccionar la opción “experimento-a la pila” que se encuentra en el menú experimentos.
8. Y por último presionar el botón procede para iniciar el proceso de clasificación.

4.7.2.3 Pantalla de Datos

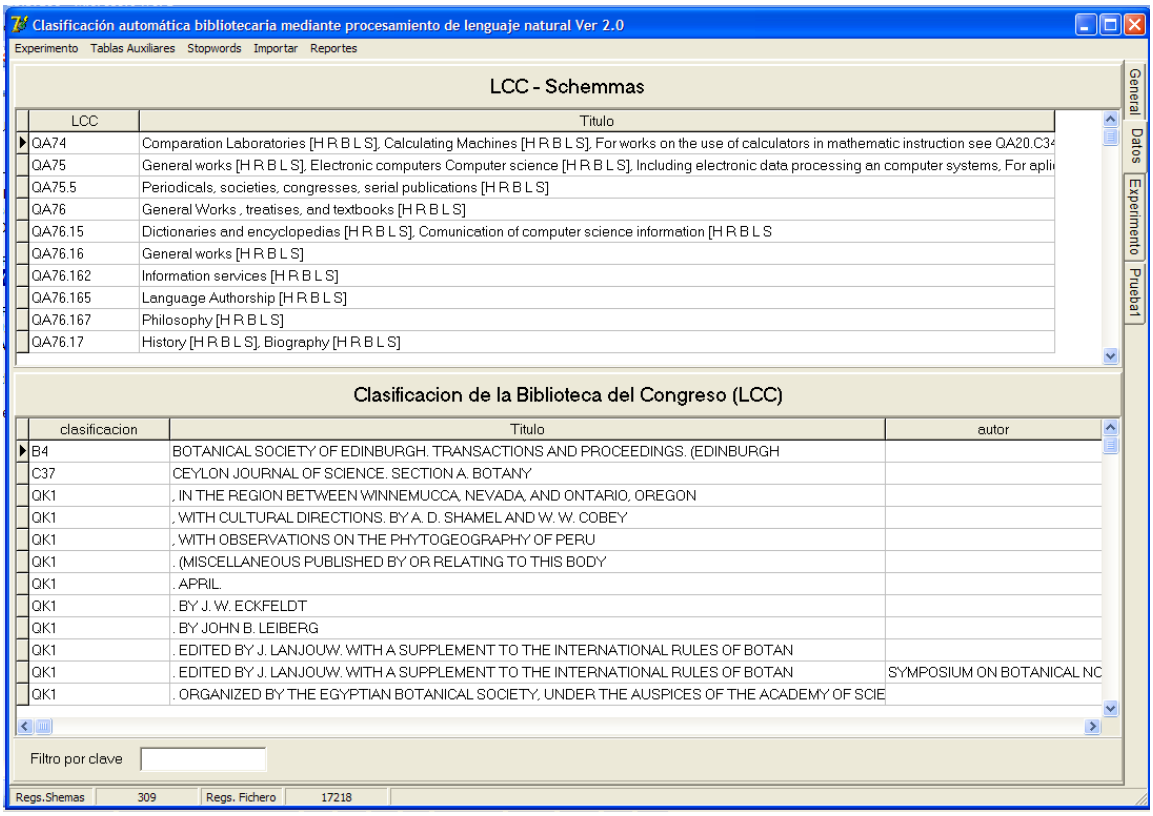


Ilustración 6 Pantalla de datos. ficha descriptiva.

Se encuentra en una de las pestañas ubicadas a la derecha de la pantalla principal, en esta pantalla mostramos los datos obtenidos de la importación de los datos. Así como también nos permite explorar los títulos con sus clases por medio del filtro de la clave.

Esta pestaña es solo de carácter informativo.

4.7.2.4 Módulo de importación y filtración de datos

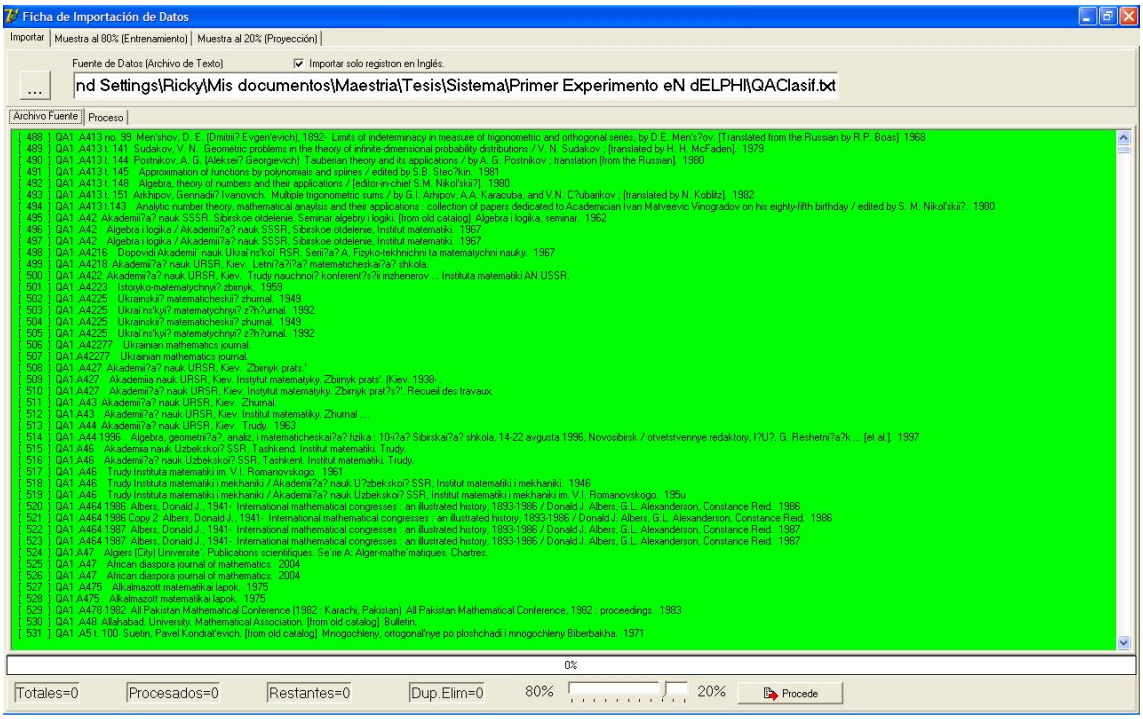


Ilustración 7 Pantalla de importación de datos.

La importación toma como entrada registros como:

[9] QA1 .A13 no. 12-13 Guichardet, A. (Alain) Tensor products of C*-algebras. Finite tensor products, by A. Guichardet. 1969

[997] QA76.76.O63 [1988 04273] Hobbes archived CDROM [electronic resource] : thousands of OS/2 programs direct from the Hobbes Internet software collection. 19uu

A estos registros se les procesa de la siguiente manera:

1. Convertimos la cadena a solo mayúscula.
2. Quitamos caracteres que pueden causar ruido en la clasificación (¿,[],?,^,”, etc).
3. Aplicamos expresiones regulares.
    a. Para la clase “([A-Z]+[A-Z]*\d+\s*\.+\s*\d+)|([A-Z]+[A-Z]*\d+)”
    b. Para la edición “\d{4}”

4. Dividimos la cadena en palabras, el divisor de campo normalmente se presenta con doble espacio, aunque también puede ser un tabulador o simplemente el carácter “/”. Esto nos trae problemas debido a que algunas veces el título viene del lado izquierdo de este carácter y algunas veces es inverso.

5. Extraemos el título, edición, autor y clasificación original para poder evaluar los algoritmos comparándolos con los resultados aportados por cada uno de ellos.

6. Tomamos sólo el título de la obra y

    a. Eliminamos *stopwords* para caso del inglés y del español.

    b. Eliminamos espacios no deseados.

    c. Aplicamos el filtro de inglés si así estuviese estipulado.

7. Grabamos en buffer intermedio y contabilizamos

8. Si la clase es diferente a la anterior, entonces calculamos los registros que serán para el entrenamiento y los que serán utilizados para la proyección del algoritmo.

9. Rompemos el título en palabras y grabamos en la tabla de *keywords* los siguientes campos.

    a. Clasificación.

    b. *Keyword.*

En el módulo de importación de información hacemos la distribución de la muestra de aprendizaje y la de proyección de los datos en base al factor ordenado por el usuario.

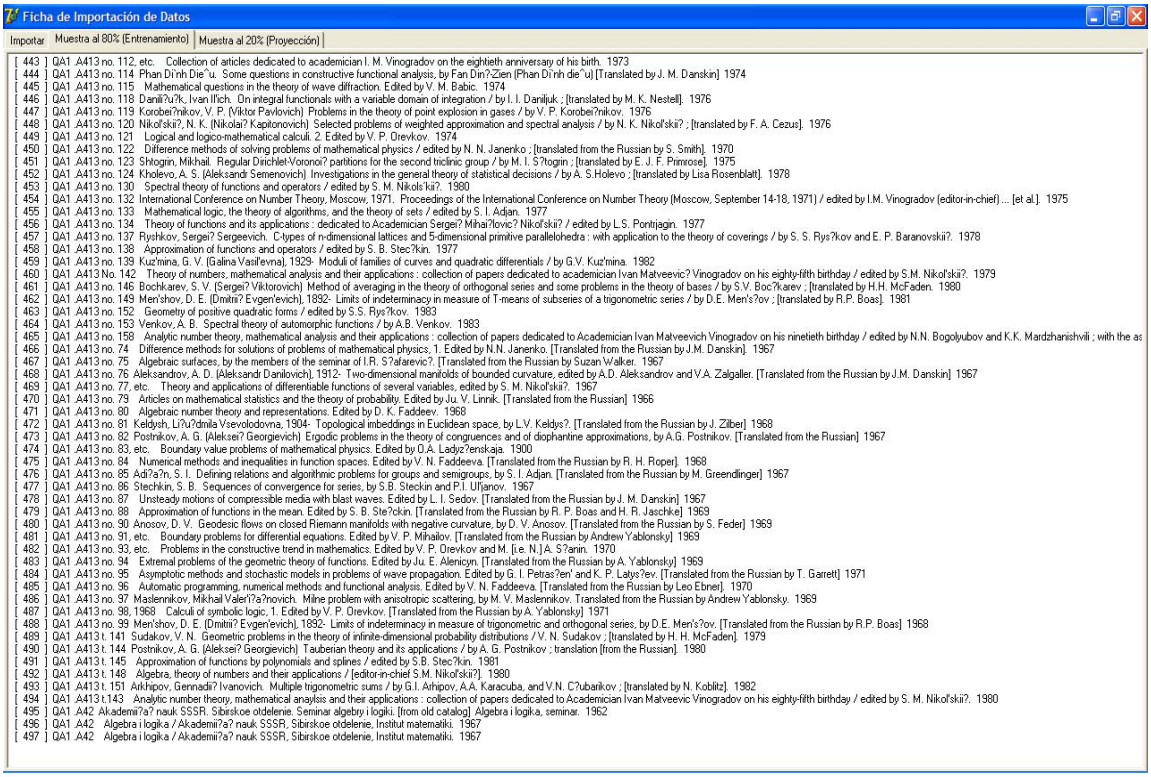


Ilustración 8 Pantalla de distribución de la muestra.

En este proceso, observamos que la clase viene ordenada, así que realizamos corte por clase y dividimos la muestra en base al factor solicitado por el usuario.

El cálculo de división de la muestra se realiza con las clases hasta el nivel más delgado, ya que se trata de una clasificación jerárquica y por lo tanto debemos dividir hasta el nivel de las hojas de este gran árbol jerárquico.

Para este propósito depositamos en un buffer los registros que pertenecen a la misma clase, calculamos los registros para entrenamiento y para proyección, tomamos al azar el número de registros calculados para el entrenamiento y los restantes se colocan para el proceso de proyección.

4.7.2.5 Descripción del módulo de experimentación en nuestra herramienta

La fase de experimentación la podemos realizar en forma manual y en forma automática. En forma manual especificamos el título de una obra, y se aplica el algoritmo sólo para ese registro, en la forma automática se ejecuta tomando como entrada todo el conjunto de datos de proyección que se especificaron previamente.

También podemos tomar de la muestra de proyección un registro aleatorio para la prueba manual, se copia y se pega usando el botón de "nuevo desde el portapapeles."

a) Clasificación Manual

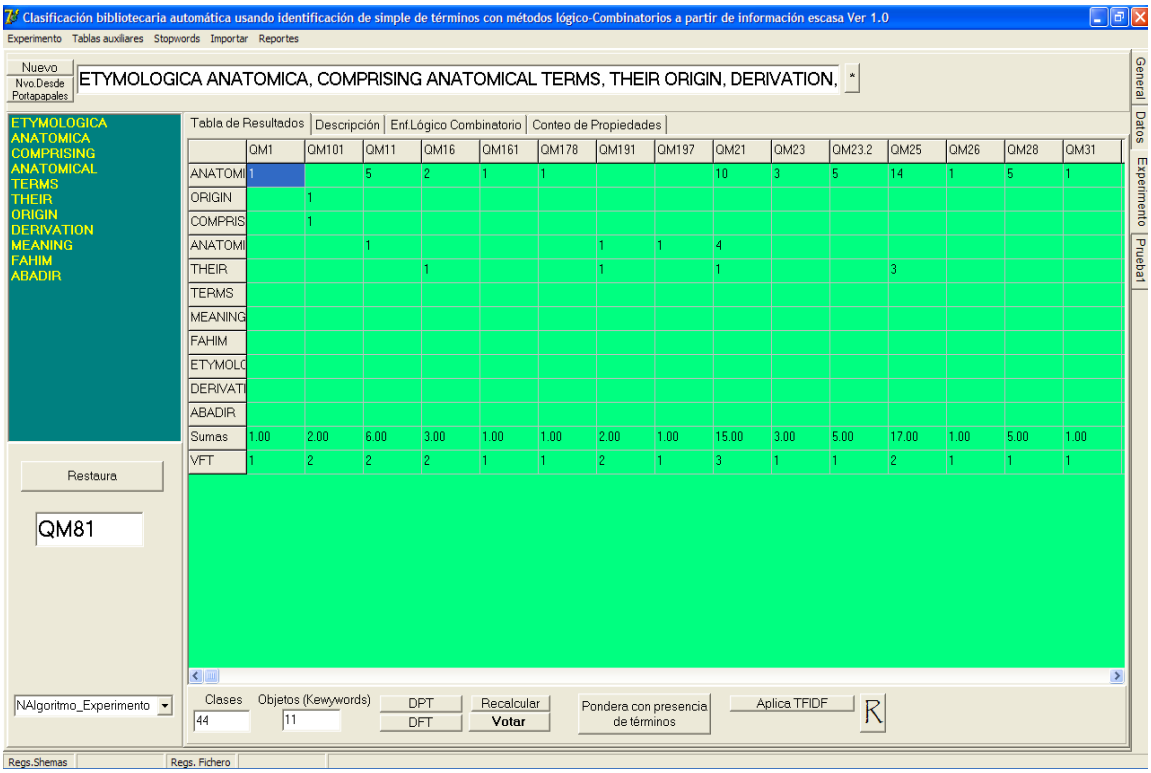


Ilustración 9 Pantalla de experimentación manual.

En la experimentación manual, contamos con los siguientes botones para poder experimentar y observar el comportamiento de la clasificación si aplicamos ciertas medidas de clasificación. (ver Ilustración 9)

Contamos también en la experimentación manual con una tabla de conteos por clase, que pueden ser alterados a gusto y conveniencia del usuario.

El conteo de los votos de cada experimento se puede apreciar claramente en la ficha de descripción, donde del lado izquierdo y de color anaranjado, presentamos la palabra clave (*keywords*) y del lado derecho el conteo final de palabras totales por clase, dando este último el voto final a aquel que presenta mayor cantidad de dichas palabras en la clase (ver Ilustración 10).

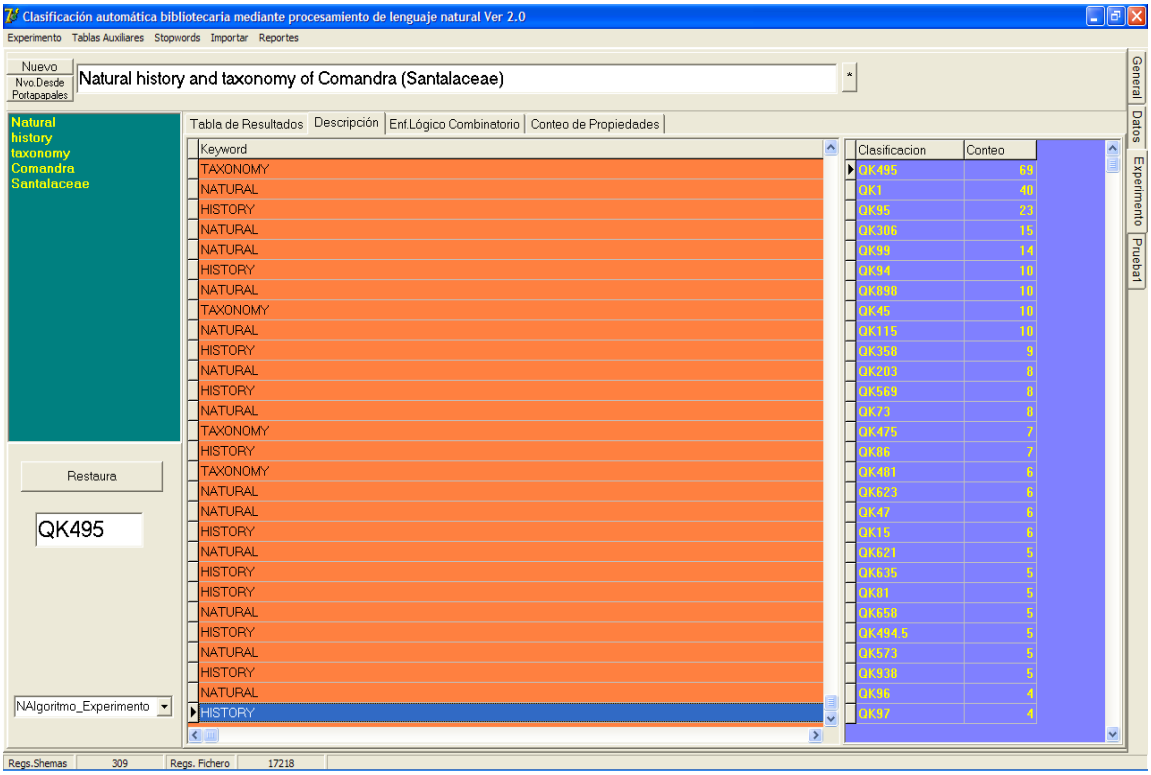


Ilustración 10 Pantalla de descripción de experimentación manual.

En la clasificación manual, también contamos con un módulo independiente de que utiliza el enfoque lógico combinatorio, con este, podemos observar el comportamiento del experimento y las semejanzas que representa con cada elemento de la muestra de entrenamiento. En la parte superior se muestra la semejanza del título por clasificar con el elemento de una clase y en la parte inferior se muestra la semejanza con el promedio de cada una de las clases, así también se muestra el voto final (Vea la Ilustración 11).

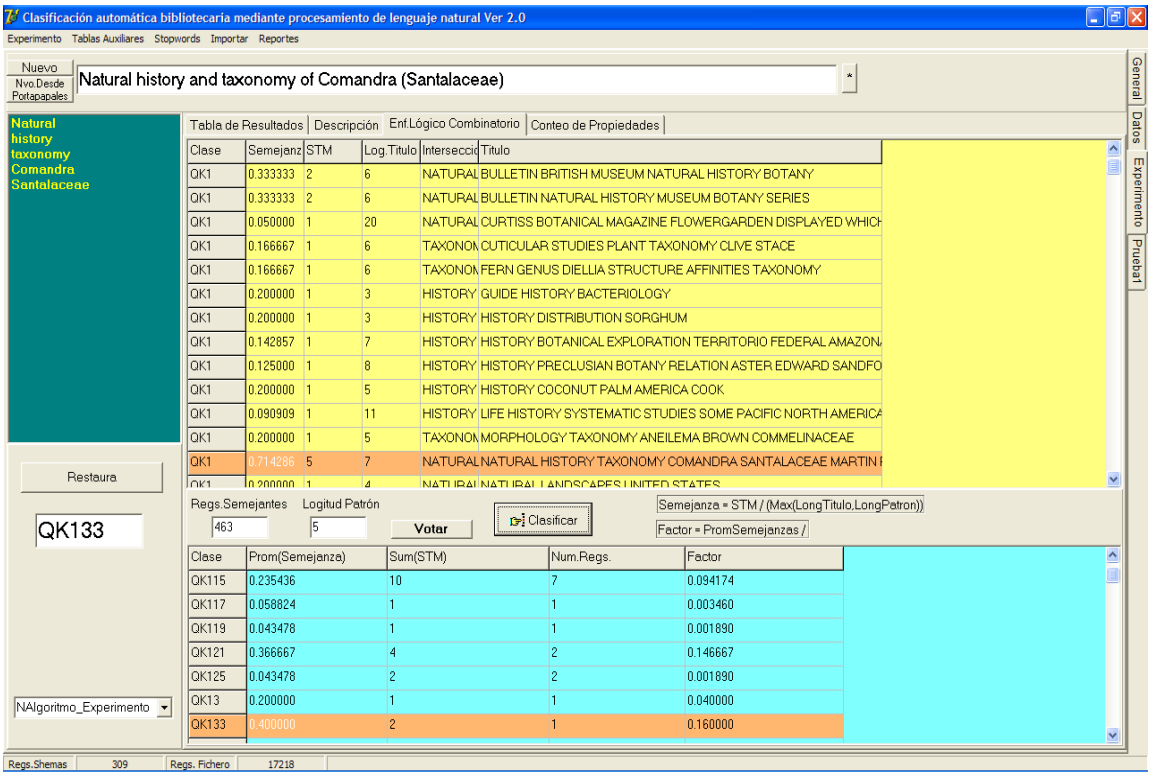


Ilustración 11 Pantalla de experimentación con enfoque lógico combinatorio.

b) Clasificación automática

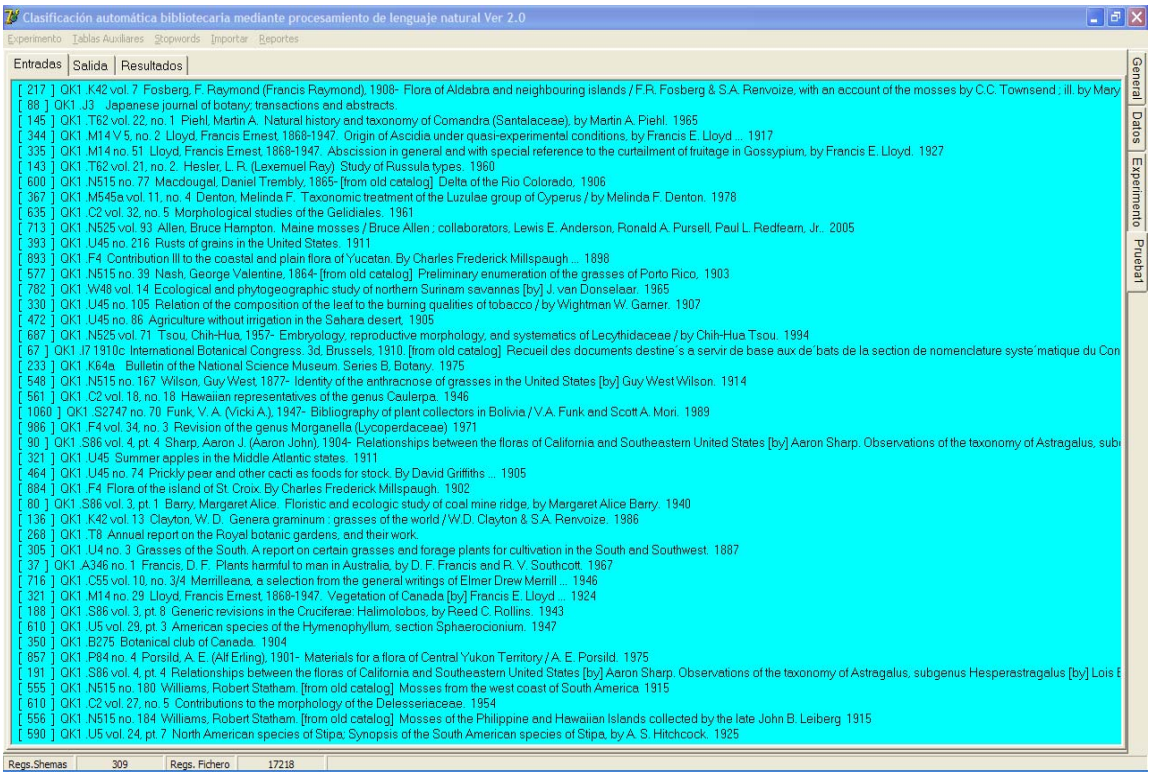


Ilustración 12 Pantalla de toma de datos para la experimentación automática.

En esta etapa, todos los elementos necesarios para proceder a la clasificación masiva de títulos de los libros, ya fueron previamente determinados por el usuario en la pantalla principal, es por esto que solo se presentan estas pantallas como carácter informativo (Ver Ilustración 12 e

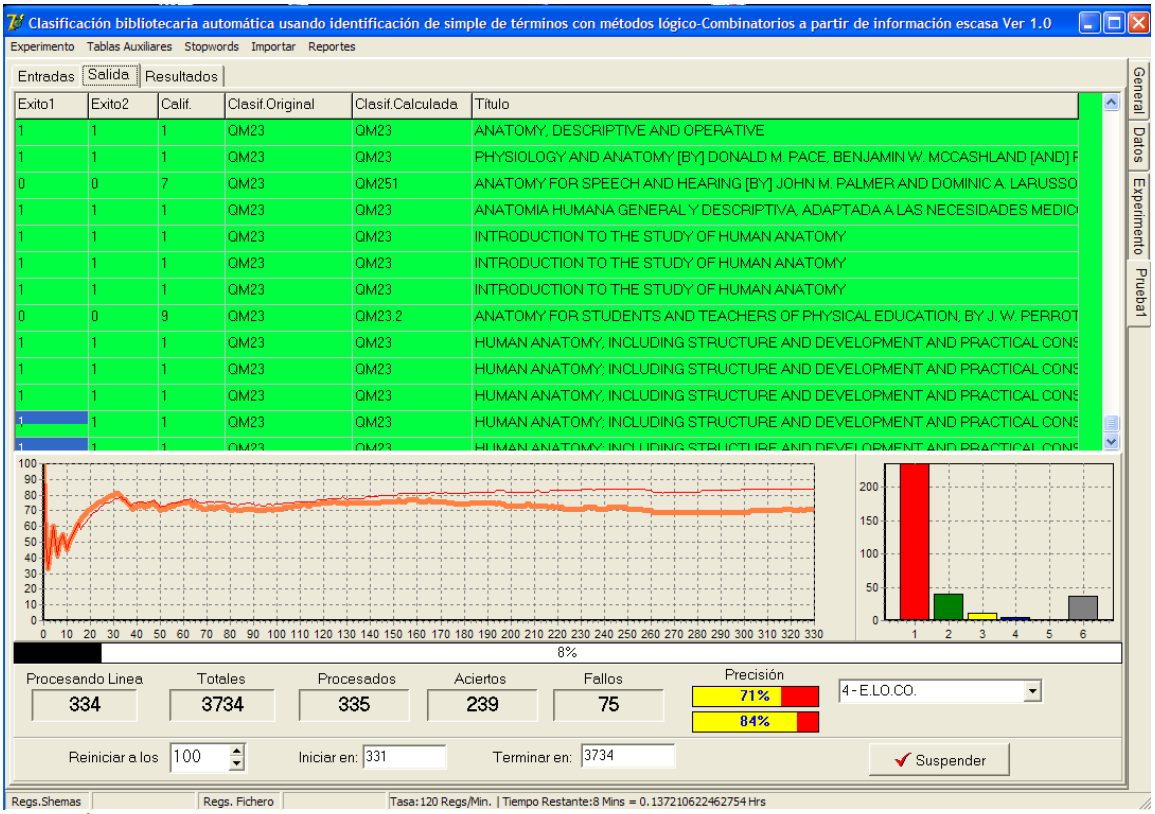


Ilustración 13).

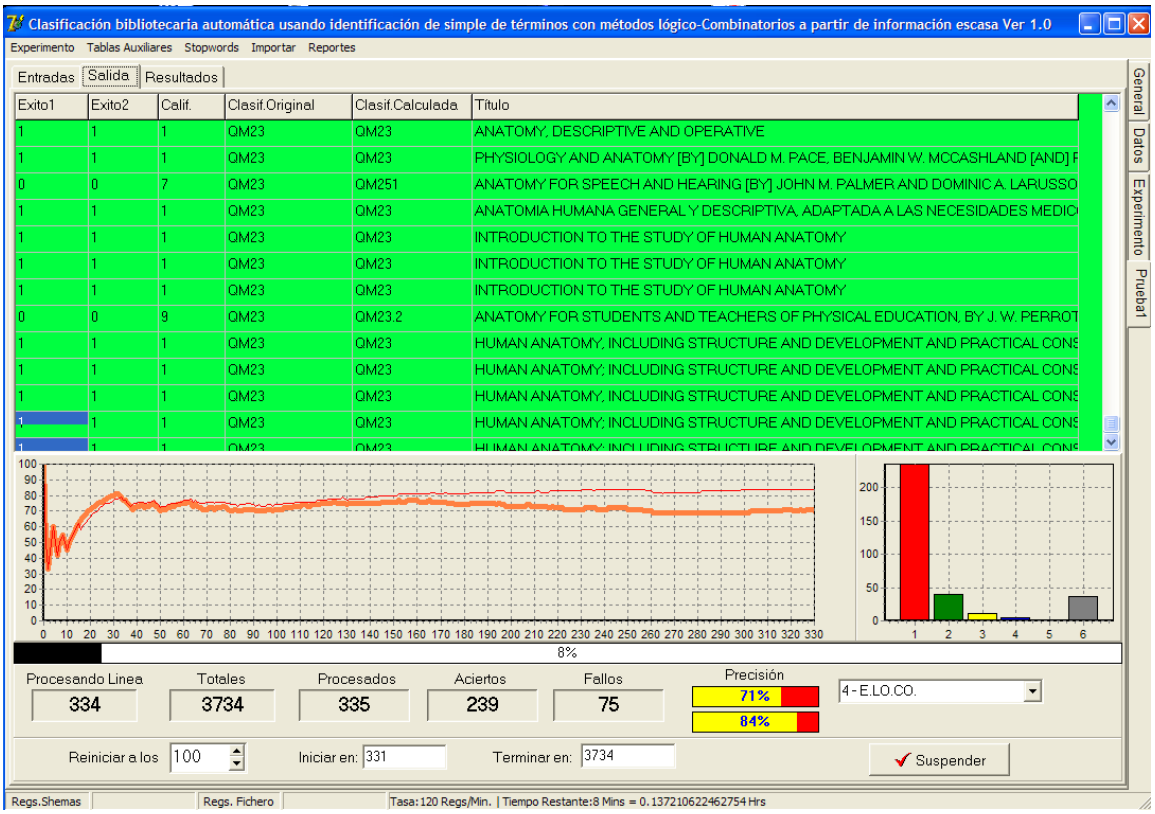


Ilustración 13 Pantalla de salida de experimentación automática

Prácticamente en esta fase del clasificador, tomamos la muestra de proyección y procedemos con la aplicación de algoritmos.

4.7.2.6 Evaluación de resultados

En la aplicación se implementó la generación automática de reportes donde se muestra detalladamente el comportamiento del experimento realizado.

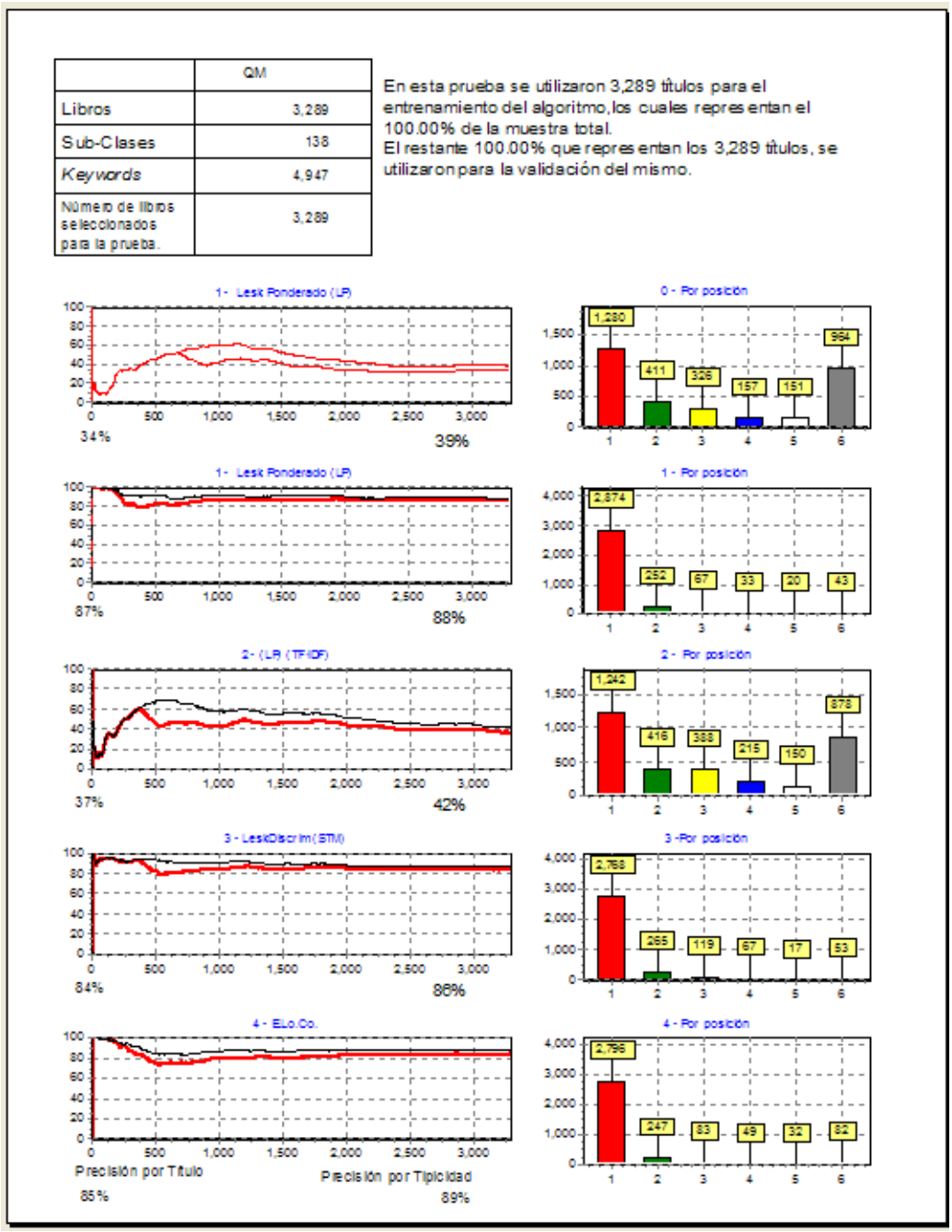


Ilustración 14: Reporte comparativo de los resultados.

Este reporte muestra la información correspondiente a la muestra de entrenamiento y a la de proyección, así como también las subclases que componen a la clase en análisis. (Ver Ilustración 14). Además muestra también los resultados obtenidos en cada experimento y la precisión lograda en cada uno de ellos. Por el lado izquierdo del reporte, se encuentran las gráficas del comportamiento de la precisión del algoritmo a lo largo de la prueba, mientras que en el lado derecho se muestran las gráficas de barras donde se muestran los registros agrupados por la posición que ocuparon al final del resultado

Al borde de cada gráfica lineal, se encuentran las precisiones finales de cada algoritmo en su forma estricta por el lado izquierdo y en su forma comparativa hasta el punto decimal por el lado derecho, en el caso de la gráfica del enfoque lógico combinatorio, se encuentra por el lado izquierdo la precisión del algoritmo utilizando máxima semejanza con un elemento de la clase y por el lado derecho la precisión del algoritmo utilizando semejanza con el promedio de la clase.

5. EXPERIMENTOS Y RESULTADOS

Los experimentos realizados en la presente tesis fueron desarrollados en la clase Q de la clasificación de la biblioteca del congreso (*LCC*) que corresponde a la clasificación de ciencias.

Se tomó esta clasificación por el hecho de que el área de cómputo corresponde a esta clase, y además por contar con una base inicial experimental que obtuvimos de la biblioteca del centro de investigación de cómputo. Finalmente logramos obtener la clase Q completa a partir de la *LCC,* quedando descartada la muestra inicial que obtuvimos.

La clase Q consta de las siguientes subclases:

- QA. Matemática.
- QB. Astronomía.
- QC. Física.
- QD, Química.
- QE. Geología.
- QH. Historia Natural (General).
- QK. Botánica.
- QL. Zoología.
- QM. Anatomía Humana.
- QP. Fisiología.
- QR. Microbiología.

Dadas las subclases, realizamos los experimentos en la clase Q aplicando los siguientes algoritmos descritos en el capítulo 2 con mayor detalle.

0. VFT, Votación por Frecuencia de Términos.
1. VFTP, Votación por Frecuencia de Términos Ponderada.
2. VFTP-TF·IDF, Votación por Frecuencia de Términos Ponderada con factor TF·IDF.
3. DPT, Discriminación por Presencia de Términos.
4. y 4’ – CT-MLC, Clasificación de Títulos con métodos Lógico-Combinatorios.

Realizamos experimentos en títulos multilingües y en títulos en inglés, donde evaluamos cada algoritmo con 3 diferentes criterios.

a) Evaluación estricta. La clasificación de prueba debe ser exactamente igual a la clasificación original, considerando todas sus subclases y numeraciones.

b) Evaluación hasta el punto decimal. La clasificación de prueba es igual al la clasificación original hasta el punto decimal de la *LCC*

c) Evaluación por posición. La clasificación de prueba está dentro de las primeras 5 propuestas como resultado final.

Quedando el conjunto de experimentos organizados de la siguiente manera:

- Tasa de aprendizaje.
    i. Multilingüe.
    ii. Solo inglés.
- Entrenamiento con 80% y 20% de proyección.
    i. Multilingüe.
    ii. Solo inglés.

5.1 Tasa de aprendizaje

Los experimentos se realizan con un conjunto de registros para el entrenamiento, y la misma muestra es utilizada en la proyección con el propósito de reafirmar que los algoritmos hayan aprendido, es decir, el 100% de los registros se utilizan para el entrenamiento y el 100% de los registros es utilizado para la proyección. De esta manera, la precisión obtenida, será el grado de aprendizaje de cada algoritmo.

Los experimentos muestran los siguientes resultados:

5.1.1 Prueba multilingüe.

Utilizando la muestra original y sin aplicar ninguna especie de filtrado de términos, tomamos la misma muestra seleccionada en el entrenamiento para la proyección del mismo y obtuvimos los siguientes resultados.

El algoritmo de método lógico-combinatorio fue evaluado únicamente en forma estricta. Los resultados mostrados en 4' son los que se obtuvieron de la semejanza promedio a la clase, mientras que los mostrados en el algoritmo 4 son aquellos con el más semejante a una clase. Las siguientes pruebas se ejecutaron con los siguientes elementos:

Registros de entrenamiento	Subclases	*Keywords*	*Conjunto de Prueba*
515,721	8,243	1,454,615	515,721

Para los experimentos 0, 1, 2 y 3

Registros de entrenamiento	Subclases	*Keywords*	*Conjunto de Prueba*
8,837	402	28,398	8,387

Para los experimentos 4 y 4'

a) Resultados con evaluación estricta.

	0	1	2	3	4'	4
	VFT	VFTP	VFTP-TF·IDF	DPT	CT-MLC.	CT-MLC.
Sin Clasificar		228		5,435		-
Cubiertos	515,721	515,493	515,721	510,286	8,837	8,837
Éxitos	178,654	433,861	177,945	396,689	7,822	8,214
Fallos	337,067	81,860	337,776	119,032	1,015	623
Precisión	34.64%	84.16%	34.50%	77.74%	88.51%	92.95%

Tabla 4: Resumen del la tasa de aprendizaje multilingüe, evaluación estricta

b) Por posición (*estricta*).

	0	1	2	3	4
	VFT	VFTP	VFTP-TF·IDF	DPT	CT-MLC.
1	34.64%	84.16%	34.50%	77.74%	88.51%
2	14.04%	7.85%	13.93%	11.27%	6.53%
3	8.92%	2.75%	8.88%	4.03%	2.24%
4	6.24%	1.48%	6.21%	2.04%	1.06%
5	4.62%	0.86%	4.60%	1.17%	0.50%
>=6	31.52%	2.90%	31.88%	3.75%	1.17%

Tabla 5: Resumen del la tasa de aprendizaje multilingüe, evaluación por posición.

c) Resultados evaluados hasta el punto decimal.

Registros de entrenamiento	Subclases	*Keywords*	*Conjunto de Prueba*
515,721	8,187	1,454,615	515,721

Para los experimentos 0, 1 , 2 y 3

	0 VFT	1 VFTP	2 VFTP-TF·IDF	3 DPT
Sin Clasificar		228	-	5,435
Cubiertos	515,721	515,493	515,721	510,286
Éxitos	205,394	451,433	204,486	413,768
Fallos	310,327	64,288	311,235	101,953
Precisión	39.83%	87.57%	39.65%	81.09%

Tabla 6: Resumen del la tasa de aprendizaje multilingüe, evaluación hasta el punto decimal

Los reportes detallados por clase se presentan en el apéndice A, y el detalle de los algoritmos se encuentra en el apéndice 1

En la muestra de aprendizaje tenemos una precisión del 92.95 con evaluación estricta mostrando que si los términos son plenamente identificados, entonces tenemos buen conocimiento de a cual clase pertenece el titulo por clasificar.

5.1.2 Prueba con el idioma inglés

Utilizando un filtro para el idioma inglés en la selección de la muestra y tomando la misma para la proyección, obtuvimos los siguientes resultados:

Registros de entrenamiento	Subclases	*Keywords*	*Conjunto de Prueba*
292,008	7,255	1,133,833	287,091

Para los experimentos 0, 1, 2 y 3

Registros de entrenamiento	Subclases	*Keywords*	*Conjunto de Prueba*
12,776	388	382	7,706

Para los experimentos 4 y 4'

a) Evaluación Estricta.

	0	1	2	3	4'	4
Estricta	VFT	VFTP	VFTP-TF·IDF	DPT	VT-MLC.	VT-MLC.
Sin Clasificar						
Cubiertos	287,091	287,091	287,091	287,091	7,706	7,706
Éxitos	77,094	229,804	76,551	76,833	6,222	6,690
Fallos	209,997	57,287	210,540	210,258	1,484	1,016
Precisión	26.85%	80.05%	26.66%	26.76%	80.74%	**86.82%**

Tabla 7: Prueba con el inglés 100% entrenamiento 100% proyección, evaluación estricta, resumen.

b) Evaluación por posición (*estricta*).

	0	1	2	3	4
	VFT	VFTP	VFTP-TF·IDF	DPT	VT-MLC.
1	26.85%	80.05%	26.66%	26.76%	80.81%
2	12.97%	9.24%	12.64%	12.77%	7.19%
3	8.27%	3.30%	8.14%	8.13%	3.31%
4	6.74%	1.71%	6.59%	6.57%	1.58%
5	5.02%	1.09%	4.94%	4.95%	1.04%
>=6	40.15%	4.61%	41.03%	40.82%	6.07%

Tabla 8: Prueba con el inglés 100% entrenamiento 100% proyección, evaluación por posición, resumen.

El algoritmo de método lógico-combinatorio fue evaluado únicamente en forma estricta y los resultados mostrados en 4' son los que se obtuvieron de la semejanza promedio a la clase.

c) Evaluación hasta el punto decimal.

Registros de entrenamiento	Subclases	Keywords	*Conjunto de Prueba*
292,008	7,207	1,133,833	287,091

Para los experimentos 0,1,2 y 3

Algoritmo	0 VFT	1 VFTP	2 VFTP-TF·IDF	3 DPT
Sin Clasificar	0	0	0	0
Cubiertos	287,091	287,091	287,091	287,091
Éxitos	77,094	229,804	76,551	76,833
Fallos	209,997	57,287	210,540	210,258
Precisión	26.85%	80.05%	26.66%	26.76%

Tabla 9: Prueba con el inglés 100% entrenamiento 100% proyección, evaluación hasta el punto decimal, resumen.

Los reportes detallados por clase se presentan en el apéndice B, y el detalle de los algoritmos se encuentra en el apéndice 2.

5.2 Proyección en los algoritmos

La muestra total es la suma del número de registros de entrenamiento con el número de registros para la proyección de cada algoritmo. La tasa de distribución de los registros son del 80% para el entrenamiento y el 20% restante lo utilizaremos para la proyección de cada algoritmo.

5.2.1 Prueba multilingüe

Utilizando el 80% de los registros de la muestra original para el entrenamiento de los algoritmos y sin aplicar ninguna especie de filtrado de términos, tomamos el 20% restante para la proyección de los mismos, obteniendo los siguientes resultados.

Registros de entrenamiento	Subclases	*Keywords*	*Conjunto de Prueba*
490,016	8,388	1,511,310	117,091

Para todos los experimentos.

a) Evaluación estricta.

	0 VFT	1 VFTP	2 VFTP-TF·IDF	3 DPT	4 VT-MLC.	4' VT-MLC.
Éxitos	32,869	41,763	32,507	42,305	41,537	30,223
Fallos	84,222	75,328	84,584	74,786	75,554	86,868
Precisión	28.07%	35.67%	27.76%	36.13%	35.47%	25.81%

Tabla 10: Prueba multilingüe, 80% entrenamiento 20% proyección, evaluación estricta, resumen.

b) Evaluación por posición (*estricta*).

	0 VFT	1 VFTP	2 VFTP-TF·IDF	3 DPT	4 VT-MLC.	4' VT-MLC.
1	28.07%	35.67%	27.76%	36.13%	35.47%	25.81%
2	12.37%	12.28%	12.08%	10.15%	8.20%	9.71%
3	7.60%	6.88%	7.51%	3.83%	4.62%	6.01%
4	5.51%	4.65%	5.42%	2.05%	3.31%	4.49%
5	4.18%	3.40%	4.11%	1.19%	2.51%	3.54%
Ente 1 y 5	57.73%	62.88%	56.88%	53.35%	54.11%	49.56%

Tabla 11: Prueba multilingüe, 80% entrenamiento 20% proyección, evaluación por posición, resumen.

c) Evaluación hasta el punto.

Registros de entrenamiento	Subclases	*Keywords*	*Conjunto de Prueba*
490,016	8,388	1,511,310	117,091

	0 VFT	1 VFTP	2 VFTP-TF·IDF	3 DPT
Éxitos	39,482	48,857	39,023	48,274
Fallos	77,609	68,234	78,068	68,817
Precisión	**33.72%**	**41.73%**	**33.33%**	**41.23%**

Tabla 12: Prueba multilingüe, 80% entrenamiento 20% proyección, evaluación hasta el punto decimal, resumen

Los reportes detallados por clase se presentan en el apéndice C, y el detalle de cada algoritmo está en el apéndice 3.

6. CONCLUSIONES

Muchas editoriales siguen la clasificación de la biblioteca del congreso (*LCC*) al incluir su número correspondiente de clasificación en cada publicación. Esto es de suma utilidad para las bibliotecas de todo el mundo porque hacen posible su búsqueda y localización en base a su contenido, lo cual se ha vuelto un estándar bibliotecario. Aun así, no todos los libros han sido previamente clasificados, particularmente en muchas universidades las tesis nuevas han de ser clasificadas manualmente.

En esta tesis hemos propuesto un método para hacer esto automáticamente. Este método clasifica libros basándose únicamente en el título de las obras, que en muchos casos es con lo único con lo que se cuenta para este propósito. Hemos propuesto nuevas aplicaciones de algoritmos de clasificación como el esquema de votación simple de identificación de términos con ponderaciones que permiten elevar la precisión de los algoritmos.

Es posible clasificar libros usando únicamente su título hasta una profundidad completa siguiendo la clasificación de la biblioteca del congreso, la cual contiene 8,377 subclases para la clase Q, sobre la cual se hicieron los experimentos. Experimentamos con 607,107 registros con títulos en cualquier idioma. Las pruebas de *tasa de aprendizaje* reportan hasta un 90% aproximado, en tanto que las pruebas de clasificación de títulos no vistos (80% de entrenamiento y 20% de prueba) reportan un 36.13% en promedio con clasificación en la presencia de los términos por cada clase.

En esta tesis dimos una aplicación novedosa a este método lógico-combinatorio, con un desempeño relativamente lento (más de una semana para obtener el resultado de clasificación mencionado anteriormente[1]).

[1] Ejecutado en una Pentium IV Core DUO, 2.33GHz, 2GB de memoria RAM

6.1 Trabajo futuro:

Los siguientes puntos pueden ser tratados como trabajo futuro:

- Aplicar técnicas de lenguaje natural.
- Aplicar métodos lógico-combinatorios más sofisticados que permitan mayor precisión.
- Preparar el algoritmo para análisis de textos largos tales como contraportadas, tablas de contenidos, resúmenes, y texto completo de la obra de ser posible. Esto con el objetivo de aportar herramientas para creación de resúmenes con aprendizaje automático.
- Agregar como base de conocimiento los siguientes elementos:
    - LCSH (library of congress subject headings).
    - MARC (Machine readable cataloging).

7. REFERENCIAS

a) Documentos científicos.

0. Manning, C. Shütze, H, **Foundations of statistical natural language Processing**, *MIT Press*, ISBN 0262133601, Cambridge, May, 620 p., 1999.

1. Kwan, Yi, **Challenges in automated classification using library classification schemes**, *Proceedings of the 97 Information Technology with Audiovisual and Multimedia and National Libraries* IFLA 2006, Seoul, Korea, 2006.

2. Frank, Ebie, Gordon W. Paynter, **Predicting Library of Congress classifications from Library of Congress subject headings**, *Journal of the American Society for Information Science and Technology*, Volume 55, Issue 3 , Pages 214-227, 2004

3. Betts, Tom, Maria Milosavljevic, and Jon Oberlander. **The utility of information extraction in the classification of books**. *In Proceedings of the 29th European Conference on Information Retrieval (ECIR 2007)*, Rome, Italy, 2004

4. Larson, Ray R., **Experiments in automatic library of congress classification**, *Journal of the American Society for Information Science and Technology*, Volume 43, Issur 2, pages 130-148, January 1999.

b) Direcciones *web*.

a) http://catalog.loc.gov/cgi-bin/Pwebrecon.cgi?DB=local&PAGE=First
b) http://www.loc.gov/catdir/cpso/LCC.html
c) http://www3.interscience.wiley.com/cgi-bin/abstract/106561106/ABSTRACT
d) http://www.ltg.ed.ac.uk/np/publications/ltg/papers/Betts2007Utility.pdf
e) http://www3.interscience.wiley.com/cgi-bin/abstract/10049642/ABSTRACT?CRETRY=1&SRETRY=0
f) http://www.ifla.org/IV/ifla72/papers/097-Yi-en.pdf
g) http://www.ifla.org/IV/ifla72/papers/097-Zhixiong_Sa_Zhengxin_Ying_trans-es.pdf
h) http://es.wikipedia.org/wiki/Aprendizaje_autom%C3%A1tico
i) http://es.wikipedia.org/wiki/Clasificaci%C3%B3n_Decimal_de_Dewey

8. BIBLIOGRAFÍA

1. Matthis, Raimund, *Adopting the library of congress classification system. A manual methods abd techniques for application or conversion*, New york : R. R. Bowker, USA, 209p.

2. Savage Helen, Droste Kathleen D., Runchock Rita, *Class Q science : Library of Congress classification schedules combined with additions and changes through 1987*, Library of Congress. Subject Cataloging Division, Detroit, Michigan : Gale research : Book Tower, USA, 862p.

3. Immroth, John Phillip, *A guide to library of congress classification*, Rochester libraries unlimited, USA, 356p.

4. Chan Lois Mai, *A guide to the Library of Congress classification*, Englewood libraries unlimited, USA, 551p.

5. Library of Congress/Decimal Classification Office, *Guide to use of dewey decimal classification. Based on the practice of the practice of the decimal classification office at the library of congress*, Forest, New york, USA, 133p.

6. Furrie, Betty, *Conociendo MARC bibliográfico: catalogación legible por máquina*, Rojas Eberhard, Bogotá, Colombia, 30p.

7. A.N. Dmitriev, Yu.I. Zhuravliov; F.P. Kredelev, *Acerca de los principios matemáticos de la clasificación de objetos y fenómenos*, Novosibirsk, Rusia, Tomo 7, 3-15p.

8. Ruiz Shulcloper, José; Guzmán Arenas, Adolfo; Martínez Trinidad J. Francisco, Enfoque lógico combinatorio al reconocimiento de patrones, Selección de variables y clasificación supervisada, IPN, México DF, México, 69-75p.

9. GLOSARIO DE TÉRMINOS

ALVOT.- Algoritmo tuvo su origen en año de 1965 aproximadamente *[8],* y sus desarrollados se deben al especialista ruso yu. I. Zhuravliov y su grupo, posteriormente se trabaja con este algoritmo en cuba. *Ver página 16*.

CLASE.- Conjunto o agrupación de objetos que contienen las mismas características.

CUBRIMIENTO.- Familia de conjuntos que contienen el universo de objetos en total, donde las clases pueden intersectar y pueden también puede contener clases vacías.

CT.- (Clasificación de Textos), es la actividad de etiquetar textos en lenguaje natural con categorías temáticas tomadas desde un conjunto previamente definido [1].

CT-MLC *(Clasificación de Títulos con métodos lógico-combinatorios)*. Algoritmo de votación con métodos lógico-combinatorios, modificado para el manejo de términos con el objetivo de clasificar títulos de libros. En esta tesis, se muestra con dos medidas de semejanza, una con el más semejante de una clase, y otro con la semejanza promedio a una clase.

DISCRIMINANTE.- Valor que utilizamos para ignorar o eliminar a aquellas clases que al comprarse con ese valor representa una cantidad inferior.

DPT.- *(Discriminación por Presencia de Términos),* Algoritmo que utiliza la existencia de los términos en las clases como un factor discriminante. *(Ver algoritmo 3 en la Pagina 31).*

DEWEY. La Clasificación Decimal de Dewey (CDD, también llamada el Sistema de Clasificación Decimal de Dewey) es un sistema de clasificación de bibliotecas, desarrollado por Melvil Dewey, bibliotecario del Amherst College en Massachusetts, EE. UU., en 1876 y desde ese momento ha sido enormemente modificado y ampliado en el curso de las veintidós principales ediciones mayores que han ocurrido hasta 2004. Durante este tiempo y desde 1894 también se han desarrollado 14 ediciones abreviadas, basadas en la Edición mayor desarrollada generalmente un año antes*.(Wikipedia [i)])*

EI *(Extracción de Información),* Extracción de Información es un término que ha comenzado a utilizarse para la actividad de extraer automáticamente tipos de información previas especificadas de los textos de lenguaje natural2. Sus objetivos son extraer conocimiento estructurado, dependiente del contexto, de la información existente, generalmente texto no estructurado, con el fin de mejorar el uso y la reutilización de esta información. Hamish define la extracción de información como un proceso que toma los textos (y a veces el habla) como entrada y que produce formatos fijos, datos no ambiguos como salida *[g)]*.

IDF. (Inverse Document Frequency). Inversa de la frecuencia del documento, elemento del factor estadístico TF·IDF que se calcula aplicando el logaritmo ya sea natural o base 10, a la división de la cantidad de términos presentes por clase entre la cantidad de clases totales del cubrimiento en cuestión *(Ver punto 3.3 en la página 18)*.

KEYWORD. Palabra con significado particular y útil para el proceso de clasificación.

LCC. (Acrónimo de Library of Congress Classification).- Clasificación de la biblioteca del congreso *(ver punto 3.4 en la página 19)*.

LCSH (Library of congress subject headings). Es un compendio de sinónimos y antónimos de todos los temas que se relacionan con el contenido de la obra, este compendio es actualizado por la biblioteca del congreso. Este es ampliamente utilizado en el registro de los libros, la LCSH es una parte integral para el control bibliográfico cuya función es organizar y esparcir documentos. [6]

MACHINE LEARNING. Se refiere al aprendizaje automático que es una rama de la Inteligencia Artificial cuyo objetivo es desarrollar técnicas que permitan a las computadoras aprender. De forma más concreta, se trata de crear programas capaces de generalizar comportamientos a partir de una información no estructurada suministrada en forma de ejemplos. Es, por lo tanto, un proceso de inducción del conocimiento. En muchas ocasiones el campo de actuación del Aprendizaje Automático se solapa con el de la Estadística, ya que las dos disciplinas se basan en el análisis de datos. Sin embargo, el Aprendizaje Automático se centra más en el estudio de la Complejidad Computacional de los problemas. Muchos problemas son de clase NP-hard, por lo que gran parte de la investigación realizada en Aprendizaje Automático está enfocada al diseño de soluciones factibles a esos problemas.[h)]

MARC. (Machine Readable Cataloging). Es un acrónimo que se utiliza en el campo de la ciencia de los libros, es un estándar para la catalogación de libros legible para las computadoras, este estándar es utilizado en la representación y comunicación de la bibliografía e información relatada en cierto formato legible para las máquinas.[6]

OBJETO. Elemento de clasificación, para nuestro caso, el título.

RASGO. Elemento que identifica a un objeto, para nuestro caso, los términos que contiene un título.

STM. *(Simple Term Match)*, Acrónimo que se refiere a la frecuencia de términos idénticos que contienen dos o más frases *(ver punto 3.1 de la página 16)*.

STOPWORD. Palabra que no aporta significado deseado en los procesos de clasificación, tales son los casos de artículos y adjetivos, que generan ruido o confusión entre clases.

TF (Term Frequency), Frecuencia del término, elemento del factor estadístico TF·IDF que se calcula dividiendo la frecuencia que presenta un término en una clase, con la cantidad de términos totales de esa misma clase. *(Ver punto 3.3 en la página 18)*.

TF·IDF. Factor estadístico propuesto por manning en su publicación “Foundations of statistical natural language Processing” *[0],* que se refiere a la multiplicación de TF con IDF *(Ver punto 3.3 en la página 18).*

VFT *(Votación por frecuencia de términos),* Algoritmo de clasificación que utiliza la frecuencia de los términos como base principal de identidad a una clase. *(Ver punto 4.1 en la página 24).*

VFTP *(Votación por Frecuencia de Términos Ponderada),* Algoritmo Semejante al VFT, la diferencia radica en el resultado que es divido por la cantidad de términos en la clase. *(vea punto 4.2 de la página 26).*

10. Descripción detallada de los experimentos.

Apéndice A

Clase	Q
Idioma	Multilingüe
Factor de Entrenamiento	100 %
Factor de proyección	100 %

Descripción de los experimentos

Número	Clase	Tasa aprendizaje	Tasa de Proyección	Lengua	Algoritmo
1	QA	100	100	Multilingüe	0-VFT
2	QB	100	100	Multilingüe	0-VFT
3	QC	100	100	Multilingüe	0-VFT
4	QD	100	100	Multilingüe	0-VFT
5	QE	100	100	Multilingüe	0-VFT
6	QH	100	100	Multilingüe	0-VFT
7	QK	100	100	Multilingüe	0-VFT
8	QL	100	100	Multilingüe	0-VFT
9	QM	100	100	Multilingüe	0-VFT
10	QP	100	100	Multilingüe	0-VFT
11	QR	100	100	Multilingüe	0-VFT
12	QA	100	100	Multilingüe	1-VFTP
13	QB	100	100	Multilingüe	1-VFTP
14	QC	100	100	Multilingüe	1-VFTP
15	QD	100	100	Multilingüe	1-VFTP
16	QE	100	100	Multilingüe	1-VFTP
17	QH	100	100	Multilingüe	1-VFTP
18	QK	100	100	Multilingüe	1-VFTP
19	QL	100	100	Multilingüe	1-VFTP
20	QM	100	100	Multilingüe	1-VFTP
21	QP	100	100	Multilingüe	1-VFTP
22	QR	100	100	Multilingüe	1-VFTP
23	QA	100	100	Multilingüe	2-VFTP-TF·IDF
24	QB	100	100	Multilingüe	2-VFTP-TF·IDF
25	QC	100	100	Multilingüe	2-VFTP-TF·IDF
26	QD	100	100	Multilingüe	2-VFTP-TF·IDF
27	QE	100	100	Multilingüe	2-VFTP-TF·IDF
28	QH	100	100	Multilingüe	2-VFTP-TF·IDF
29	QK	100	100	Multilingüe	2-VFTP-TF·IDF
30	QL	100	100	Multilingüe	2-VFTP-TF·IDF
31	QM	100	100	Multilingüe	2-VFTP-TF·IDF
32	QP	100	100	Multilingüe	2-VFTP-TF·IDF
33	QR	100	100	Multilingüe	2-VFTP-TF·IDF
34	QA	100	100	Multilingüe	3-DPT
35	QB	100	100	Multilingüe	3-DPT
36	QC	100	100	Multilingüe	3-DPT
37	QD	100	100	Multilingüe	3-DPT
38	QE	100	100	Multilingüe	3-DPT
39	QH	100	100	Multilingüe	3-DPT
40	QK	100	100	Multilingüe	3-DPT
41	QL	100	100	Multilingüe	3-DPT
42	QM	100	100	Multilingüe	3-DPT
43	QP	100	100	Multilingüe	3-DPT
44	QR	100	100	Multilingüe	3-DPT
45	QM	100	100	Multilingüe	4-CT-MLC
46	QR	100	100	Multilingüe	4-CT-MLC

Apéndice B

Clase	Q
Idioma	Inglés.
Factor de Entrenamiento	100 %
Factor de proyección	100 %

Tabla de experimentos

Número	Clase	Tasa aprendizaje	Tasa de Proyección	Lengua	Algoritmo
1	QA	100	100	Inglés	0-VFT
2	QB	100	100	Inglés	0-VFT
3	QC	100	100	Inglés	0-VFT
4	QD	100	100	Inglés	0-VFT
5	QE	100	100	Inglés	0-VFT
6	QH	100	100	Inglés	0-VFT
7	QK	100	100	Inglés	0-VFT
8	QL	100	100	Inglés	0-VFT
9	QM	100	100	Inglés	0-VFT
10	QP	100	100	Inglés	0-VFT
11	QR	100	100	Inglés	0-VFT
12	QA	100	100	Inglés	1-VFTP
13	QB	100	100	Inglés	1-VFTP
14	QC	100	100	Inglés	1-VFTP
15	QD	100	100	Inglés	1-VFTP
16	QE	100	100	Inglés	1-VFTP
17	QH	100	100	Inglés	1-VFTP
18	QK	100	100	Inglés	1-VFTP
19	QL	100	100	Inglés	1-VFTP
20	QM	100	100	Inglés	1-VFTP
21	QP	100	100	Inglés	1-VFTP
22	QR	100	100	Inglés	1-VFTP
23	QA	100	100	Inglés	2-VFTP-TF·IDF
24	QB	100	100	Inglés	2-VFTP-TF·IDF
25	QC	100	100	Inglés	2-VFTP-TF·IDF
26	QD	100	100	Inglés	2-VFTP-TF·IDF
27	QE	100	100	Inglés	2-VFTP-TF·IDF
28	QH	100	100	Inglés	2-VFTP-TF·IDF
29	QK	100	100	Inglés	2-VFTP-TF·IDF
30	QL	100	100	Inglés	2-VFTP-TF·IDF
31	QM	100	100	Inglés	2-VFTP-TF·IDF
32	QP	100	100	Inglés	2-VFTP-TF·IDF
33	QR	100	100	Inglés	2-VFTP-TF·IDF
34	QA	100	100	Inglés	3-DPT
35	QB	100	100	Inglés	3-DPT
36	QC	100	100	Inglés	3-DPT
37	QD	100	100	Inglés	3-DPT
38	QE	100	100	Inglés	3-DPT
39	QH	100	100	Inglés	3-DPT
40	QK	100	100	Inglés	3-DPT
41	QL	100	100	Inglés	3-DPT
42	QM	100	100	Inglés	3-DPT
43	QP	100	100	Inglés	3-DPT
44	QR	100	100	Inglés	3-DPT
45	QM	100	100	Inglés	4-CT-MLC
46	QR	100	100	Inglés	4-CT-MLC

Apéndice C

Clase	Q
Idioma	Multilingüe
Factor de Entrenamiento	80 %
Factor de proyección	20 %

Tabla de experimentos

Número	Clase	Tasa aprendizaje	Tasa de Proyección	Lengua	Algoritmo
1	QA	80	20	Multilingüe	0-VFT
2	QB	80	20	Multilingüe	0-VFT
3	QC	80	20	Multilingüe	0-VFT
4	QD	80	20	Multilingüe	0-VFT
5	QE	80	20	Multilingüe	0-VFT
6	QH	80	20	Multilingüe	0-VFT
7	QK	80	20	Multilingüe	0-VFT
8	QL	80	20	Multilingüe	0-VFT
9	QM	80	20	Multilingüe	0-VFT
10	QP	80	20	Multilingüe	0-VFT
11	QR	80	20	Multilingüe	0-VFT
12	QA	80	20	Multilingüe	1-VFTP
13	QB	80	20	Multilingüe	1-VFTP
14	QC	80	20	Multilingüe	1-VFTP
15	QD	80	20	Multilingüe	1-VFTP
16	QE	80	20	Multilingüe	1-VFTP
17	QH	80	20	Multilingüe	1-VFTP
18	QK	80	20	Multilingüe	1-VFTP
19	QL	80	20	Multilingüe	1-VFTP
20	QM	80	20	Multilingüe	1-VFTP
21	QP	80	20	Multilingüe	1-VFTP
22	QR	80	20	Multilingüe	1-VFTP
23	QA	80	20	Multilingüe	2-VFTP-TF·IDF
24	QB	80	20	Multilingüe	2-VFTP-TF·IDF
25	QC	80	20	Multilingüe	2-VFTP-TF·IDF
26	QD	80	20	Multilingüe	2-VFTP-TF·IDF
27	QE	80	20	Multilingüe	2-VFTP-TF·IDF
28	QH	80	20	Multilingüe	2-VFTP-TF·IDF
29	QK	80	20	Multilingüe	2-VFTP-TF·IDF
30	QL	80	20	Multilingüe	2-VFTP-TF·IDF
31	QM	80	20	Multilingüe	2-VFTP-TF·IDF
32	QP	80	20	Multilingüe	2-VFTP-TF·IDF
33	QR	80	20	Multilingüe	2-VFTP-TF·IDF
34	QA	80	20	Multilingüe	3-DPT
35	QB	80	20	Multilingüe	3-DPT
36	QC	80	20	Multilingüe	3-DPT
37	QD	80	20	Multilingüe	3-DPT
38	QE	80	20	Multilingüe	3-DPT
39	QH	80	20	Multilingüe	3-DPT
40	QK	80	20	Multilingüe	3-DPT
41	QL	80	20	Multilingüe	3-DPT
42	QM	80	20	Multilingüe	3-DPT
43	QP	80	20	Multilingüe	3-DPT
44	QR	80	20	Multilingüe	3-DPT
45	QA	80	20	Multilingüe	4-CT-MLC
46	QB	80	20	Multilingüe	4-CT-MLC
47	QC	80	20	Multilingüe	4-CT-MLC
48	QD	80	20	Multilingüe	4-CT-MLC
49	QE	80	20	Multilingüe	4-CT-MLC
50	QH	80	20	Multilingüe	4-CT-MLC
51	QK	80	20	Multilingüe	4-CT-MLC
52	QL	80	20	Multilingüe	4-CT-MLC
53	QM	80	20	Multilingüe	4-CT-MLC
54	QP	80	20	Multilingüe	4-CT-MLC
55	QR	80	20	Multilingüe	4-CT-MLC

Apéndice 1. Detalle de experimentos multilingües con 100% de entrenamiento y 100% de proyección.

Tablas comparativas detalladas de las pruebas realizadas con la muestra multilingüe y 100% de registros de entrenamiento y 100% de los registros para su proyección. El resumen de este apéndice esta en el punto 5.1.1

1.1 Resultados con algoritmo 0 – VFT

a) Evaluación estricta.

	QA	QB	QC	QD	QE	QH
Registros de entrenamiento	112,828	24,489	79,994	48,238	68,301	30,560
Subclases	956	622	1,437	517	950	637
Keywords	232,489	78,163	235,345	130,115	227,340	84,614
Conjunto de prueba	112,828	24,489	79,994	48,238	68,301	30,560
Sin Clasificar	-	-	-	-	-	-
Cubiertos	112,828	24,489	79,994	48,238	68,301	30,560
Éxitos	40,412	10,681	25,731	16,150	19,514	9,489
Fallos	72,416	13,808	54,263	32,088	48,787	21,071
Precisión	35.82%	43.62%	32.17%	33.48%	28.57%	31.05%

Tabla 13: Algoritmo VFT, prueba multilingüe 100% entrenamiento 100% proyección, evaluación estricta, Parte 1

	QK	QL	QM	QP	QR	Totales
Registros de entrenamiento	37,282	70,495	3,289	34,697	5,548	515,721
Subclases	824	1,393	138	505	264	8,243
keywords	125,914	205,579	10,612	106,658	17,786	1,454,615
Conjunto de prueba	37,282	70,495	3,289	34,697	5,548	515,721
Sin Clasificar	-	-	-	-	-	
Cubiertos	37,282	70,495	3,289	34,697	5,548	515,721
Éxitos	13,687	25,234	1,135	14,458	2,163	178,654
Fallos	23,595	45,261	2,154	20,239	3,385	337,067
Precisión	36.71%	35.80%	34.51%	41.67%	38.99%	**34.64%**

Tabla 14: Algoritmo VFT, prueba multilingüe 100% entrenamiento 100% proyección, evaluación estricta, Parte 2

b) Evaluación por posición (*estricta*)

	QA	QB	QC	QD	QE	QH	QK	QL	QP	QR	Totales
1	35.82%	43.62%	32.17%	33.49%	28.57%	31.05%	36.72%	35.80%	41.67%	39.01%	**34.64%**
2	14.80%	15.75%	13.63%	12.39%	14.80%	13.58%	14.54%	13.08%	13.01%	19.41%	**14.04%**
3	9.05%	9.71%	8.84%	9.32%	9.76%	8.71%	7.74%	8.56%	8.00%	8.71%	**8.91%**
4	6.32%	5.82%	6.65%	6.08%	6.94%	5.86%	5.69%	5.61%	6.43%	6.99%	**6.25%**
5	4.51%	4.14%	4.89%	4.57%	5.32%	4.59%	4.29%	4.49%	3.90%	5.08%	**4.62%**
>=6	29.50%	20.96%	33.82%	34.15%	34.60%	36.20%	31.01%	32.47%	26.99%	20.80%	**31.54%**

Tabla 15: Algoritmo VFT, prueba multilingüe 100% entrenamiento 100% proyección, evaluación por posición.

c) Evaluación hasta el punto decimal.

	QA	QB	QC	QD	QE	QH
Registros de entrenamiento	112,828	24,489	79,994	48,238	68,301	30,560
Subclases	937	615	1,432	515	947	631
keywords	232,489	78,163	235,345	130,115	227,340	84,614
Conjunto de prueba	112828	24489	79994	48238	68301	30560
Sin Clasificar	0	0	0	0	0	0
Cubiertos	112828	24489	79994	48238	68301	30560
Éxitos	61825	11033	28048	16391	19807	10323
Fallos	51003	13456	51946	31847	48494	20237
Precisión	54.80%	45.05%	35.06%	33.98%	29.00%	33.78%

Tabla 16: Algoritmo VFT, prueba multilingüe 100% entrenamiento 100% proyección, evaluación hasta el punto decimal, Parte 1

	QK	QL	QM	QP	QR	Totales
Registros de entrenamiento	37,282	70,495	3,289	34,697	5,548	515,721
Subclases	823	1,391	137	500	259	8,187
keywords	125,914	205,579	10,612	106,658	17,786	1,454,615
Conjunto de prueba	37282	70495	3289	34697	5548	515721
Sin Clasificar	0	0	0	0	0	
Cubiertos	37282	70495	3289	34697	5548	515721
Éxitos	13772	25750	1305	14869	2271	205394
Fallos	23510	44745	1984	19828	3277	310327
Precisión	36.94%	36.53%	39.68%	42.85%	40.93%	**39.83%**

Tabla 17: Algoritmo VFT, prueba multilingüe 100% entrenamiento 100% proyección, evaluación hasta el punto decimal, Parte 2

1.2 Resultados con algoritmo 1 – VFTP

a) Evaluación estricta.

	QA	QB	QC	QD	QE	QH
Registros de entrenamiento	112,828	24,489	79,994	48,238	68,301	30,560
Subclases	956	622	1,437	517	950	637
Keywords	232,489	78,163	235,345	130,115	227,340	84,614
Conjunto de prueba	112,828	24,489	79,994	48,238	68,301	30,560
Sin Clasificar	-	-	-	-	-	-
Cubiertos	112,828	24,489	79,994	48,238	68,301	30,560
Éxitos	83,403	21,445	67,322	41,130	60,048	26,004
Fallos	29,425	3,044	12,672	7,108	8,253	4,556
Precisión	73.92%	87.57%	84.16%	85.26%	87.92%	85.09%

Tabla 18: Algoritmo VFTP, prueba multilingüe 100% entrenamiento 100% proyección, evaluación estricta, Parte 1

	QK	QL	QM	QP	QR	Totales
Registros de entrenamiento	37,282	70,495	3,289	34,697	5,548	515,721
Subclases	824	1,393	138	505	264	8,243
Keywords	125,914	205,579	10,612	106,658	17,786	1,454,615
Conjunto de prueba	37,282	70,495	3,289	34,697	5,548	515,721
Sin Clasificar	-	-	-	228	-	228
Cubiertos	37,282	70,495	3,289	34,469	5,548	515,493
Éxitos	33,920	61,870	2,915	30,761	5,043	433,861
Fallos	3,362	8,625	374	3,936	505	81,860
Precisión	90.98%	87.77%	88.63%	89.24%	90.90%	**84.16%**

Tabla 19: Algoritmo VFTP, prueba multilingüe 100% entrenamiento 100% proyección, evaluación estricta, Parte 2

b) Evaluación por posición (e*stricta*).

	QA	QB	QC	QD	QE	QH	QK	QL	QM	QP	QR	Totales
1	73.93%	87.58%	84.16%	85.27%	87.92%	85.09%	90.99%	87.77%	88.63%	89.26%	90.92%	**84.16%**
2	11.30%	6.65%	7.59%	7.05%	7.01%	7.12%	5.51%	6.99%	6.81%	6.14%	5.41%	**7.85%**
3	4.98%	2.11%	2.63%	2.46%	1.82%	2.62%	1.25%	2.05%	2.04%	1.76%	1.98%	**2.75%**
4	2.94%	1.02%	1.40%	1.33%	0.81%	1.45%	0.73%	0.89%	0.79%	0.98%	0.83%	**1.48%**
5	1.68%	0.57%	0.91%	0.83%	0.50%	0.89%	0.31%	0.51%	0.49%	0.47%	0.22%	**0.86%**
>=6	5.17%	2.07%	3.31%	3.06%	1.94%	2.82%	1.20%	1.80%	1.25%	1.40%	0.65%	**2.90%**

Tabla 20: Algoritmo VFTP, prueba multilingüe 100% entrenamiento 100% proyección, evaluación por posición.

c) Evaluación hasta el punto decimal.

	QA	QB	QC	QD	QE	QH
Registros de entrenamiento	112,828	24,489	79,994	48,238	68,301	30,560
Subclases	937	615	1,432	515	947	631
Keywords	232,489	78,163	235,345	130,115	227,340	84,614
Conj.Prueba	112,828	24,489	79,994	48,238	68,301	30,560
Sin Clasificar	-	-	-	-	-	-
Cubiertos	112,828	24,489	79,994	48,238	68,301	30,560
Éxitos	97,383	21,683	68,929	41,650	60,307	26,276
Fallos	15,445	2,806	11,065	6,588	7,994	4,284
Precisión	86.31%	88.54%	86.17%	86.34%	88.30%	85.98%

Tabla 21: Algoritmo VFTP, prueba multilingüe 100% entrenamiento 100% proyección, evaluación hasta el punto decimal, Parte 1

	QK	QL	QM	QP	QR	Totales
Registros de entrenamiento	37,282	70,495	3,289	34,697	5,548	515,721
Subclases	823	1,391	137	500	259	8,187
Keywords	125,914	205,579	10,612	106,658	17,786	1,454,615
Conj.Prueba	37,282	70,495	3,289	34,697	5,548	515,721
Sin Clasificar	-	-	-	228	-	228
Cubiertos	37,282	70,495	3,289	34,469	5,548	515,493
Éxitos	34,011	62,192	2,959	30,960	5,083	451,433
Fallos	3,271	8,303	330	3,737	465	64,288
Precisión	91.23%	88.22%	89.97%	89.82%	91.62%	**87.57%**

Tabla 22: Algoritmo VFTP, prueba multilingüe 100% entrenamiento 100% proyección, evaluación hasta el punto decimal, Parte 2

1.3 Resultados con algoritmo 2 – VFTP-TF·IDF

a) Evaluación estricta.

	QA	QB	QC	QD	QE	QH
Registros de entrenamiento	112,828	24,489	79,994	48,238	68,301	30,560
Subclases	956	622	1,437	517	950	637
Keywords	232,489	78,163	235,345	130,115	227,340	84,614
Conjunto de prueba	112,828	24,489	79,994	48,238	68,301	30,560
Sin Clasificar	-	-	-	-	-	-
Cubiertos	112,828	24,489	79,994	48,238	68,301	30,560
Éxitos	40,089	10,694	25,513	16,155	19,080	9,521
Fallos	72,739	13,795	54,481	32,083	49,221	21,039
Precisión	35.53%	43.67%	31.89%	33.49%	27.94%	31.16%

Tabla 23: Algoritmo 2, VFTP-TF·IDF, prueba multilingüe, 100% entrenamiento 100% proyección, evaluación estricta, Parte 1

	QK	QL	QM	QP	QR	Totales
Registros de entrenamiento	37,282	70,495	3,289	34,697	5,548	515,721
Subclases	824	1,393	138	505	264	8,243
Keywords	125,914	205,579	10,612	106,658	17,786	1,454,615
Conjunto de prueba	37,282	70,495	3,289	34,697	5,548	515,721
Sin Clasificar	-	-	-	-	-	
Cubiertos	37,282	70,495	3,289	34,697	5,548	515,721
Éxitos	13,660	25,125	1,503	14,438	2,167	177,945
Fallos	23,622	45,370	1,786	20,259	3,381	337,776
Precisión	36.64%	35.64%	45.70%	41.61%	39.06%	34.50%

Tabla 24: Algoritmo 2, VFTP-TF·IDF, prueba multilingüe, 100% entrenamiento 100% proyección, evaluación estricta, Parte 2

b) Evaluación por posición (*estricta*).

	QA	QB	QC	QD	QE	QH	QK	QL	QM	QP	QR	Totales
1	35.54%	43.68%	31.90%	33.50%	27.94%	31.16%	36.65%	35.64%	45.70%	41.61%	39.08%	**34.50%**
2	14.69%	15.69%	13.38%	12.39%	14.71%	13.60%	14.55%	12.91%	12.22%	13.00%	19.38%	**13.93%**
3	9.01%	9.71%	8.80%	9.33%	9.62%	8.71%	7.75%	8.44%	10.61%	8.08%	8.67%	**8.88%**
4	6.34%	5.79%	6.47%	6.06%	6.95%	5.86%	5.69%	5.54%	5.35%	6.43%	7.01%	**6.21%**
5	4.52%	4.14%	4.83%	4.58%	5.37%	4.55%	4.28%	4.48%	2.77%	3.89%	5.08%	**4.60%**
>=6	29.91%	20.98%	34.62%	34.14%	35.41%	36.14%	31.09%	32.99%	23.35%	26.98%	20.78%	**31.88%**

Tabla 25: Algoritmo 2, VFTP-TF·IDF, prueba multilingüe, 100% entrenamiento 100% proyección, evaluación por posición

c) Evaluación hasta el punto decimal.

	QA	QB	QC	QD	QE	QH
Registros de entrenamiento	112,828	24,489	79,994	48,238	68,301	30,560
Subclases	937	615	1,432	515	947	631
Keywords	232,489	78,163	235,345	130,115	227,340	84,614
Conjunto de prueba	112,828	24,489	79,994	48,238	68,301	30,560
Sin Clasificar	-	-	-	-	-	-
Cubiertos	112,828	24,489	79,994	48,238	68,301	30,560
Éxitos	61,358	11,048	27,797	16,396	19,373	10,352
Fallos	51,470	13,441	52,197	31,842	48,928	20,208
Precisión	54.38%	45.11%	34.75%	33.99%	28.36%	33.87%

Tabla 26: Algoritmo 2, VFTP-TF·IDF, prueba multilingüe, 100% entrenamiento 100% proyección, evaluación hasta el punto decimal, Parte 1

	QK	QL	QM	QP	QR	Totales
Registros de entrenamiento	37,282	70,495	3,289	34,697	5,548	515,721
Subclases	823	1,391	137	500	259	8,187
keywords	125,914	205,579	10,612	106,658	17,786	1,454,615
Conjunto de prueba	37,282	70,495	3,289	34,697	5,548	515,721
Sin Clasificar	-	-	-	-	-	
Cubiertos	37,282	70,495	3,289	34,697	5,548	515,721
Éxitos	13,748	25,648	1,642	14,849	2,275	204,486
Fallos	23,534	44,847	1,647	19,848	3,273	311,235
Precisión	36.88%	36.38%	49.92%	42.80%	41.01%	**39.65%**

Tabla 27: Algoritmo 2, VFTP-TF·IDF, prueba multilingüe, 100% entrenamiento 100% proyección, evaluación hasta el punto decimal, Parte 2

1.4 Resultados con algoritmo 3 – DPT

a) Evaluación estricta.

	QA	QB	QC	QD	QE	QH
Registros de entrenamiento	112,828	24,489	79,994	48,238	68,301	30,560
Subclases	956	622	1,437	517	950	637
Keywords	232,489	78,163	235,345	130,115	227,340	84,614
Conjunto de prueba	112,828	24,489	79,994	48,238	68,301	30,560
Sin Clasificar	58	-	-	-	-	-
Cubiertos	112,770	24,489	79,994	48,238	68,301	30,560
Éxitos	75,901	20,840	62,863	39,238	54,156	23,888
Fallos	36,927	3,649	17,131	9,000	14,145	6,672
Precisión	67.31%	85.10%	78.58%	81.34%	79.29%	78.17%

Tabla 28: Algoritmo 3, DPT, prueba multilingüe, 100% entrenamiento 100% proyección, evaluación estricta, Parte 1

	QK	QL	QM	QP	QR	Totales
Registros de entrenamiento	37,282	70,495	3,289	34,697	5,548	515,721
Subclases	824	1,393	138	505	264	8,243
Keywords	125,914	205,579	10,612	106,658	17,786	1,454,615
Conjunto de prueba	37,282	70,495	3,289	34,697	5,548	515,721
Sin Clasificar	-	5,377	-	-	-	5,435
Cubiertos	37,282	65,118	3,289	34,697	5,548	510,286
Éxitos	31,967	50,433	2,863	29,787	4,753	396,689
Fallos	5,315	20,062	426	4,910	795	119,032
Precisión	85.74%	77.45%	87.05%	85.85%	85.67%	**77.74%**

Tabla 29: Algoritmo 3, DPT, prueba multilingüe, 100% entrenamiento 100% proyección, evaluación estricta, Parte 2

b) Evaluación por posición (*estricta*).

	QA	QB	QC	QD	QE	QH	QK	QL	QM	QP	QR	Totales
1	67.37%	85.16%	78.75%	81.43%	79.42%	78.17%	85.81%	78.11%	87.05%	85.87%	85.71%	**77.74%**
2	13.59%	8.75%	10.45%	8.39%	11.76%	10.76%	7.85%	14.17%	7.33%	7.60%	8.71%	**11.27%**
3	6.16%	2.36%	3.80%	3.77%	3.74%	3.93%	2.78%	3.43%	2.89%	2.61%	2.83%	**4.03%**
4	3.20%	1.23%	1.99%	1.97%	1.70%	2.08%	1.31%	1.75%	1.22%	1.27%	1.19%	**2.04%**
5	1.83%	0.69%	1.20%	1.03%	0.91%	1.38%	0.67%	1.09%	0.33%	0.67%	0.67%	**1.17%**
>=6	7.85%	1.81%	3.81%	3.41%	2.47%	3.67%	1.59%	1.45%	1.19%	1.97%	0.90%	**3.75%**

Tabla 30: Algoritmo 3, DPT, prueba multilingüe, 100% entrenamiento 100% proyección, evaluación por posición.

c) Evaluación hasta el punto decimal.

	QA	QB	QC	QD	QE	QH
Registros de entrenamiento	112,828	24,489	79,994	48,238	68,301	30,560
Subclases	937	615	1,432	515	947	631
keywords	232,489	78,163	235,345	130,115	227,340	84,614
Conjunto de prueba	112,828	24,489	79,994	48,238	68,301	30,560
Sin Clasificar	58	-	-	-	-	-
Cubiertos	112,770	24,489	79,994	48,238	68,301	30,560
Éxitos	90,382	21,024	63,901	39,457	54,329	24,244
Fallos	22,446	3,465	16,093	8,781	13,972	6,316
Precisión	80.15%	85.85%	79.88%	81.80%	79.54%	79.33%

Tabla 31: Algoritmo 3, DPT, prueba multilingüe, 100% entrenamiento 100% proyección, evaluación hasta el punto decimal, Parte 1

	QK	QL	QM	QP	QR	Totales
Registros de entrenamiento	37,282	70,495	3,289	34,697	5,548	515,721
Subclases	823	1,391	137	500	259	8,187
keywords	125,914	205,579	10,612	106,658	17,786	1,454,615
Conjunto de prueba	37,282	70,495	3,289	34,697	5,548	515,721
Sin Clasificar	-	5,377	-	-	-	5,435
Cubiertos	37,282	65,118	3,289	34,697	5,548	510,286
Éxitos	32,026	50,682	2,928	29,993	4,802	413,768
Fallos	5,256	19,813	361	4,704	746	101,953
Precisión	85.90%	77.83%	89.02%	86.44%	86.55%	**81.09%**

Tabla 32: Algoritmo 3, DPT, prueba multilingüe, 100% entrenamiento 100% proyección, evaluación hasta el punto decimal, Parte 2

1.5 Resultados con algoritmo 4 – CT-MLC

Esta prueba se desarrolló seleccionando al azar 2 clases (QM,QR) únicamente debido a que las pruebas requieren de mucho tiempo de procesamiento.

a) Por semejanza máxima con un elemento en la clase.

	QM	QR	Totales
Registros de entrenamiento	3,289	5,548	8,837
Subclases	138	264	402
Keywords	10,612	17,786	28,398
Conjunto de prueba	3,289	5,548	8,837
Sin Clasificar	-	-	0
Cubiertos	3,289	5,548	8,837
Éxitos	2,857	4,965	7,822
Fallos	432	583	1,015
Precisión	86.87%	89.49%	**88.51%**

Tabla 33: Algoritmo 4, CT-MLC, prueba multilingüe, 100% entrenamiento 100% proyección, máxima semejanza a un elemento de la clase.

b) Por posición en la lista de salida (*estricta*).

	QM	QR	Totales
1	86.87%	89.51%	**88.51%**
2	6.84%	6.31%	**6.53%**
3	2.52%	2.07%	**2.24%**
4	1.16%	1.01%	**1.06%**
5	0.91%	0.25%	**0.50%**
>=6	1.70%	0.85%	**1.17%**

Tabla 34: Algoritmo 4, CT-MLC, prueba multilingüe, 100% entrenamiento 100% proyección, máxima semejanza a un elemento de la clase. Evaluación por posición.

c) Por semejanza promedio máxima de una clase.

	QM	QR	Totales
Conjunto de prueba	3,289	5,548	8,837
Sin Clasificar	-	-	
Cubiertos	3,289	5,548	8,837
Éxitos	2,972	5,242	8,214
Fallos	317	306	623
Precisión	90.36%	94.48%	**92.95%**

Tabla 35: Algoritmo 4, CT-MLC, prueba multilingüe, 100% entrenamiento 100% proyección, máximo promedio semejanza con la clase.

Apéndice 2. Detalle de experimentos con el idioma inglés con 100% de entrenamiento y 100% de proyección.

Tablas comparativas detalladas de las pruebas realizadas con la muestra de registros en el idioma inglés con 100% de registros de entrenamiento y 100% de los registros para su proyección. El resumen de este apéndice esta en el punto 5.1.2

2.1 Resultados con algoritmo 0– VFT

a) Evaluación estricta.

	QA	QB	QC	QD	QE	QH
Registros de entrenamiento	56,921	14,532	43,773	23,515	34,667	28,294
SubClases	816	565	1,223	452	858	617
Keywords	170,154	65,259	172,282	79,506	151,953	114,868
Conjunto de prueba	56,921	14,532	43,773	23,515	34,667	28,294
Sin Clasificar	-	-	-	-	-	-
Cubiertos	56,921	14,532	43,773	23,515	34,667	28,294
Éxitos	17,937	4,446	10,325	6,114	6,651	6,994
Fallos	38,984	10,086	33,448	17,401	28,016	21,300
Precisión	31.51%	30.59%	23.59%	26.00%	19.19%	24.72%

Tabla 36: Algoritmo VFT, prueba con el inglés 100% entrenamiento 100% proyección, evaluación estricta, Parte 1

	QK	QL	QM	QP	QR	Totales
Registros de entrenamiento	16,879	40,422	8,648	20,229	4,128	292,008
SubClases	697	1,181	140	458	248	7,255
Keywords	77,301	158,784	32,010	90,552	21,164	1,133,833
Conjunto de prueba	16,879	40,422	3,731	20,229	4,128	287,091
Sin Clasificar	-	-	-	-	-	
Cubiertos	16,879	40,422	3,731	20,229	4,128	287,091
Éxitos	4,378	10,425	1,554	6,766	1,504	77,094
Fallos	12,501	29,997	2,177	13,463	2,624	209,997
Precisión	25.94%	25.79%	41.65%	33.45%	36.43%	**26.86%**

Tabla 37: Algoritmo VFT, prueba con el inglés 100% entrenamiento 100% proyección, evaluación estricta, Parte 2

b) evaluación por posición:

	QA	QB	QC	QD	QE	QH	QK	QL	QM	QP	QR	Totales
1	31.52%	30.59%	23.59%	26.00%	19.19%	24.72%	25.94%	25.80%	41.65%	33.45%	36.43%	**26.86%**
2	13.93%	12.98%	12.42%	11.41%	14.79%	13.03%	10.46%	12.32%	11.10%	12.77%	17.33%	**12.96%**
3	8.94%	10.87%	8.23%	6.59%	8.10%	8.64%	7.59%	7.92%	10.53%	6.99%	9.13%	**8.27%**
4	6.38%	6.60%	6.58%	8.81%	8.20%	5.61%	6.02%	5.80%	6.19%	7.26%	7.66%	**6.74%**
5	5.14%	5.22%	4.82%	4.64%	5.62%	5.08%	5.09%	5.07%	5.95%	3.98%	4.84%	**5.02%**
>=6	34.09%	33.74%	44.36%	42.55%	44.10%	42.92%	44.90%	43.09%	24.58%	35.55%	24.61%	**40.15%**

Tabla 38: Algoritmo VFT, prueba con el inglés 100% entrenamiento 100% proyección, evaluación por posición.

c) Evaluación hasta el punto decimal.

	QA	QB	QC	QD	QE	QH
Registros de entrenamiento	56,921	14,532	43,773	23,515	34,667	28,294
Subclases	802	560	1,218	449	856	610
keywords	170,154	65,259	172,282	79,506	151,953	114,868
Conjunto de prueba	56,921	14,532	43,773	23,515	34,667	28,294
Sin Clasificar	-	-	-	-	-	-
Cubiertos	56,921	14,532	43,773	23,515	34,667	28,294
Éxitos	17,937	4,446	10,325	6,114	6,651	6,994
Fallos	38,984	10,086	33,448	17,401	28,016	21,300
Precisión	31.51%	30.59%	23.59%	26.00%	19.19%	24.72%

Tabla 39: Algoritmo VFT, prueba con el inglés 100% entrenamiento 100% proyección, evaluación hasta el punto decimal, Parte 1

	QK	QL	QM	QP	QR	Totales
Registros de entrenamiento	16,879	40,422	8,648	20,229	4,128	292,008
Subclases	695	1,180	139	455	243	7,207
keywords	77,301	158,784	32,010	90,552	21,164	1,133,833
Conjunto de prueba	16,879	40,422	3,731	20,229	4,128	287,091
Sin Clasificar	-	-	-	-	-	
Cubiertos	16,879	40,422	3,731	20,229	4,128	287,091
Éxitos	4,378	10,425	1,554	6,766	1,504	77,094
Fallos	12,501	29,997	2,177	13,463	2,624	209,997
Precisión	25.94%	25.79%	41.65%	33.45%	36.43%	**26.85%**

Tabla 40: Algoritmo VFT, prueba con el inglés 100% entrenamiento 100% proyección, evaluación hasta el punto decimal, Parte 2

2.2 Resultados con algoritmo 1 – VFTP

a) Evaluación estricta.

	QA	QB	QC	QD	QE	QH
Registros de entrenamiento	56,921	14,532	43,773	23,515	34,667	28,294
SubClases	816	565	1,223	452	858	617
Keywords	170,154	65,259	172,282	79,506	151,953	114,868
Conjunto de prueba	56,921	14,532	43,773	23,515	34,667	28,294
Sin Clasificar	-	-	-	-	-	-
Cubiertos	56,921	14,532	43,773	23,515	34,667	28,294
Éxitos	43,109	12,117	33,677	17,733	28,121	23,335
Fallos	13,812	2,415	10,096	5,782	6,546	4,959
Precisión	75.73%	83.38%	76.94%	75.41%	81.12%	82.47%

Tabla 41: Algoritmo VFTP, prueba con el inglés 100% entrenamiento 100% proyección, evaluación estricta, Parte 1

	QK	QL	QM	QP	QR	Totales
Registros de entrenamiento	16,879	40,422	8,648	20,229	4,128	292,008
SubClases	697	1,181	140	458	248	7,255
Keywords	77,301	158,784	32,010	90,552	21,164	1,133,833
Conjunto de prueba	16,879	40,422	3,731	20,229	4,128	287,091
Sin Clasificar	-	-	-	-	-	
Cubiertos	16,879	40,422	3,731	20,229	4,128	287,091
Éxitos	14,285	33,162	3,314	17,283	3,668	229,804
Fallos	2,594	7,260	417	2,946	460	57,287
Precisión	84.63%	82.04%	88.82%	85.44%	88.86%	**80.05%**

Tabla 42: Algoritmo VFTP, prueba con el inglés 100% entrenamiento 100% proyección, evaluación estricta, Parte 2

b) Por posición (*estricta*).

	QA	QB	QC	QD	QE	QH	QK	QL	QM	QP	QR	Totales
1	75.74%	83.38%	76.94%	75.41%	81.12%	82.48%	84.63%	82.04%	88.82%	85.44%	88.86%	**80.05%**
2	11.25%	7.94%	9.12%	10.48%	9.63%	7.74%	8.41%	8.63%	5.79%	8.13%	5.62%	9.24%
3	4.20%	2.48%	4.08%	4.33%	2.93%	2.79%	2.68%	2.60%	2.12%	2.22%	2.18%	3.30%
4	2.20%	1.18%	2.10%	2.45%	1.30%	1.46%	1.14%	1.56%	0.94%	1.02%	1.11%	1.71%
5	1.34%	1.00%	1.46%	1.50%	0.84%	0.91%	0.48%	0.99%	0.78%	0.68%	0.48%	1.09%
>=6	5.27%	4.02%	6.30%	5.82%	4.18%	4.63%	2.67%	4.18%	1.55%	2.51%	1.74%	4.61%

Tabla 43: Algoritmo VFTP, prueba con el inglés 100% entrenamiento 100% proyección, evaluación por posición.

c) Hasta el punto decimal.

	QA	QB	QC	QD	QE	QH
Registros de entrenamiento	56,921	14,532	43,773	23,515	34,667	28,294
Subclases	802	560	1,218	449	856	610
keywords	170,154	65,259	172,282	79,506	151,953	114,868
Conjunto de prueba	56,921	14,532	43,773	23,515	34,667	28,294
Sin Clasificar	-	-	-	-	-	-
Cubiertos	56,921	14,532	43,773	23,515	34,667	28,294
Éxitos	43,109	12,117	33,677	17,733	28,121	23,335
Fallos	13,812	2,415	10,096	5,782	6,546	4,959
Precisión	75.73%	83.38%	76.94%	75.41%	81.12%	82.47%

Tabla 44: Algoritmo VFTP, prueba con el inglés 100% entrenamiento 100% proyección, evaluación hasta el punto decimal, Parte 1

	QK	QL	QM	QP	QR	Totales
Registros de entrenamiento	16,879	40,422	8,648	20,229	4,128	292,008
Subclases	695	1,180	139	455	243	7,207
keywords	77,301	158,784	32,010	90,552	21,164	1,133,833
Conjunto de prueba	16,879	40,422	3,731	20,229	4,128	287,091
Sin Clasificar	-	-	-	-	-	
Cubiertos	16,879	40,422	3,731	20,229	4,128	287,091
Éxitos	14,285	33,162	3,314	17,283	3,668	229,804
Fallos	2,594	7,260	417	2,946	460	57,287
Precisión	84.63%	82.04%	88.82%	85.44%	88.86%	**80.05%**

abla 45: Algoritmo VFTP, prueba con el inglés 100% entrenamiento 100% proyección, evaluación hasta el punto decimal, Parte 2

2.3 Resultados con algoritmo 2 – VFTP-TF·IDF

a) Evaluación estricta.

	QA	QB	QC	QD	QE	QH
Registros de entrenamiento	56,921	14,532	43,773	23,515	34,667	28,294
Subclases	816	565	1,223	452	858	617
Keywords	170,154	65,259	172,282	79,506	151,953	114,868
Conjunto de prueba	56,921	14,532	43,773	23,515	34,667	28,294
Sin Clasificar	-	-	-	-	-	-
Cubiertos	56,921	14,532	43,773	23,515	34,667	28,294
Éxitos	17,688	4,795	9,459	6,115	6,019	7,954
Fallos	39,233	9,737	34,314	17,400	28,648	20,340
Precisión	31.07%	33.00%	21.61%	26.00%	17.36%	28.11%

Tabla 46: Algoritmo 2, VFTP-TF·IDF, prueba con el inglés 100% entrenamiento 100% proyección, evaluación estricta, Parte 1

	QK	QL	QM	QP	QR	Totales
Registros de entrenamiento	16,879	40,422	8,648	20,229	4,128	292,008
Subclases	697	1,181	140	458	248	7,255
Keywords	77,301	158,784	32,010	90,552	21,164	1,133,833
Conjunto de prueba	16,879	40,422	3,731	20,229	4,128	287,091
Sin Clasificar	-	-	-	-	-	
Cubiertos	16,879	40,422	3,731	20,229	4,128	287,091
Éxitos	4,372	10,240	1,554	6,782	1,573	76,551
Fallos	12,507	30,182	2,177	13,447	2,555	210,540
Precisión	25.90%	25.33%	41.65%	33.53%	38.11%	**26.66%**

Tabla 47: Algoritmo 2, VFTP-TF·IDF, prueba con el inglés 100% entrenamiento 100% proyección, evaluación estricta, Parte 2

b) Evaluación por posición.

	QA	QB	QC	QD	QE	QH	QK	QL	QM	QP	QR	Totales
1	31.08%	33.00%	21.61%	26.00%	17.37%	28.12%	25.90%	25.34%	41.65%	33.53%	38.11%	**26.66%**
2	13.87%	13.24%	11.30%	11.37%	14.13%	12.80%	10.39%	11.99%	11.10%	12.80%	17.15%	**12.64%**
3	8.88%	10.16%	8.11%	6.67%	8.09%	8.31%	7.60%	7.69%	10.53%	6.97%	9.13%	**8.14%**
4	6.35%	6.29%	6.47%	8.82%	7.76%	5.35%	6.09%	5.59%	6.19%	7.25%	7.34%	**6.59%**
5	5.21%	5.10%	4.83%	4.67%	5.47%	4.62%	5.04%	4.89%	5.95%	4.03%	4.58%	**4.94%**
>=6	34.62%	32.22%	47.68%	42.47%	47.18%	40.80%	44.99%	44.50%	24.58%	35.42%	23.69%	**41.03%**

Tabla 48: Algoritmo 2, VFTP-TF·IDF, prueba con el inglés 100% entrenamiento 100% proyección, evaluación por posición.

c) Evaluación hasta el punto decimal.

	QA	QB	QC	QD	QE	QH
Registros de entrenamiento	56,921	14,532	43,773	23,515	34,667	28,294
Subclases	802	560	1,218	449	856	610
keywords	170,154	65,259	172,282	79,506	151,953	114,868
Conjunto de prueba	56,921	14,532	43,773	23,515	34,667	28,294
Sin Clasificar	-	-	-	-	-	-
Cubiertos	56,921	14,532	43,773	23,515	34,667	28,294
Éxitos	17,688	4,795	9,459	6,115	6,019	7,954
Fallos	39,233	9,737	34,314	17,400	28,648	20,340
Precisión	31.07%	33.00%	21.61%	26.00%	17.36%	28.11%

Tabla 49: Algoritmo 2, VFTP-TF·IDF, prueba con el inglés 100% entrenamiento 100% proyección, evaluación hasta el punto decimal, Parte 1

	QK	QL	QM	QP	QR	Totales
Registros de entrenamiento	16,879	40,422	8,648	20,229	4,128	292,008
Subclases	695	1,180	139	455	243	7,207
keywords	77,301	158,784	32,010	90,552	21,164	1,133,833
Conjunto de prueba	16,879	40,422	3,731	20,229	4,128	287,091
Sin Clasificar	-	-	-	-	-	
Cubiertos	16,879	40,422	3,731	20,229	4,128	287,091
Éxitos	4,372	10,240	1,554	6,782	1,573	76,551
Fallos	12,507	30,182	2,177	13,447	2,555	210,540
Precisión	25.90%	25.33%	41.65%	33.53%	38.11%	**26.66%**

Tabla 50: Algoritmo 2, VFTP-TF·IDF, prueba con el inglés 100% entrenamiento 100% proyección, evaluación hasta el punto decimal, Parte 2

2.4 Resultados con algoritmo 3 – DPT

a) Evaluación estricta.

	QA	QB	QC	QD	QE	QH
Registros de entrenamiento	56,921	14,532	43,773	23,515	34,667	28,294
SubClases	816	565	1,223	452	858	617
keywords	170,154	65,259	172,282	79,506	151,953	114,868
Conjunto de prueba	56,921	14,532	43,773	23,515	34,667	28,294
Sin Clasificar	-	-	-	-	-	-
Cubiertos	56,921	14,532	43,773	23,515	34,667	28,294
Éxitos	17,688	4,782	9,473	6,097	6,025	7,947
Fallos	39,233	9,750	34,300	17,418	28,642	20,347
Precisión	31.07%	32.91%	21.64%	25.93%	17.38%	28.09%

Tabla 51: Algoritmo 3, DPT, prueba con el inglés 100% entrenamiento 100% proyección, evaluación estricta, Parte 1

	QK	QL	QM	QP	QR	Totales
Registros de entrenamiento	16,879	40,422	8,648	20,229	4,128	292,008
SubClases	697	1,181	140	458	248	7,255
keywords	77,301	158,784	32,010	90,552	21,164	1,133,833
Conjunto de prueba	16,879	40,422	3,731	20,229	4,128	287,091
Sin Clasificar	-	-	-	-	-	-
Cubiertos	16,879	40,422	3,731	20,229	4,128	287,091
Éxitos	4,719	10,249	1,554	6,780	1,519	76,833
Fallos	12,160	30,173	2,177	13,449	2,609	210,258
Precisión	27.96%	25.36%	41.65%	33.52%	36.80%	**26.76%**

Tabla 52: Algoritmo 3, DPT, prueba con el inglés 100% entrenamiento 100% proyección, evaluación estricta, Parte 2

b) Evaluación por posición (*estricta*).

	QA	QB	QC	QD	QE	QH	QK	QL	QM	QP	QR	Totales
1	31.08%	32.91%	21.65%	25.93%	17.39%	28.09%	27.96%	25.36%	41.65%	33.52%	36.80%	**26.76%**
2	13.85%	13.24%	11.33%	11.40%	14.17%	12.79%	12.07%	11.98%	11.10%	12.81%	18.97%	**12.77%**
3	8.88%	10.18%	8.13%	6.62%	8.06%	8.36%	7.22%	7.69%	10.53%	6.99%	9.81%	**8.13%**
4	6.35%	6.26%	6.46%	8.81%	7.79%	5.35%	5.76%	5.57%	6.19%	7.22%	7.27%	**6.57%**
5	5.22%	5.09%	4.86%	4.69%	5.47%	4.64%	4.96%	4.90%	5.95%	3.99%	5.14%	**4.95%**
>=6	34.63%	32.32%	47.57%	42.56%	47.12%	40.76%	42.03%	44.51%	24.58%	35.46%	22.02%	**40.82%**

Tabla 53: Algoritmo 3, DPT, prueba con el inglés 100% entrenamiento 100% proyección, evaluación por posición.

b) Evaluación hasta el punto decimal.

	QA	QB	QC	QD	QE	QH
Registros de entrenamiento	56,921	14,532	43,773	23,515	34,667	28,294
Subclases	802	560	1,218	449	856	610
keywords	170,154	65,259	172,282	79,506	151,953	114,868
Conjunto de prueba	56,921	14,532	43,773	23,515	34,667	28,294
Sin Clasificar	-	-	-	-	-	-
Cubiertos	56,921	14,532	43,773	23,515	34,667	28,294
Éxitos	17,688	4,782	9,473	6,097	6,025	7,947
Fallos	39,233	9,750	34,300	17,418	28,642	20,347
Precisión	31.07%	32.91%	21.64%	25.93%	17.38%	28.09%

Tabla 54: Algoritmo 3, DPT, prueba con el inglés 100% entrenamiento 100% proyección, evaluación hasta el punto decimal, Parte 1

	QK	QL	QM	QP	QR	Totales
Registros de entrenamiento	16,879	40,422	8,648	20,229	4,128	292,008
Subclases	695	1,180	139	455	243	7,207
keywords	77,301	158,784	32,010	90,552	21,164	1,133,833
Conjunto de prueba	16,879	40,422	3,731	20,229	4,128	287,091
Sin Clasificar	-	-	-	-	-	-
Cubiertos	16,879	40,422	3,731	20,229	4,128	287,091
Éxitos	4,719	10,249	1,554	6,780	1,519	76,833
Fallos	12,160	30,173	2,177	13,449	2,609	210,258
Precisión	27.96%	25.36%	41.65%	33.52%	36.80%	**26.76%**

Tabla 55: Algoritmo 3, DPT, prueba con el inglés 100% entrenamiento 100% proyección, evaluación hasta el punto decimal, Parte 2

2.5 Resultados con algoritmo 4 – CT-MLC

Esta prueba se desarrolló seleccionando al azar 2 clases (QM,QR) únicamente debido a que las pruebas requieren de mucho tiempo de procesamiento.

a) Por semejanza máxima con un elemento en la clase.

	QM	QR	Totales
Registros de.entrenamiento	8,648	4,128	12,776
Subclases	140	248	388
keywords	32,010	21,164	53,174
Conjunto de prueba	3,578	4,128	7,706
Sin Clasificar	-	-	
Cubiertos	3,578	4,128	7,706
Éxitos	3,211	3,011	6,222
Fallos	367	1,117	1,484
Precisión	89.74%	72.94%	**80.74%**

Tabla 56: Algoritmo 4, CT-MLC, prueba con inglés, 100% entrenamiento 100% proyección, máxima semejanza a un elemento de la clase.

b) Por posición en la lista de salida (*estricta*).

	QM	QR	Totales
1	89.74%	73.06%	**80.81%**
2	6.26%	7.99%	7.19%
3	2.24%	4.24%	3.31%
4	0.48%	2.54%	1.58%
5	0.25%	1.72%	1.04%
>=6	1.03%	10.44%	6.07%

Tabla 57: Algoritmo 4, CT-MLC, prueba con inglés, 100% entrenamiento 100% proyección, máxima semejanza a un elemento de la clase. Evaluación por posición.

c) Por semejanza promedio máxima de una clase.

	QM	QR	Totales
Registros de entrenamiento	8,648	4,128	12,776
Subclases	140	248	388
keywords	32,010	21,164	53,174
Conjunto de prueba	3,578	4,128	7,706
Sin Clasificar	-	-	
Cubiertos	3,578	4,128	7,706
Éxitos	3,349	3,341	6,690
Fallos	229	787	1,016
Precisión	93.60%	80.94%	**86.82%**

Tabla 58: Algoritmo 4, CT-MLC, prueba con inglés, 100% entrenamiento 100% proyección, máximo promedio semejanza con la clase.

Apéndice 3. Detalle de experimentos multilingües con 80% de entrenamiento y 20% de proyección.

Tablas comparativas detalladas de las pruebas realizadas con la muestra multilingüe y 80% de registros de entrenamiento y 20% de los registros para su proyección. El resumen de este apéndice esta en el punto 5.2.1

3.1 Resultados con algoritmo 0 – VFT

a) Evaluación estricta.

	QA	QB	QC	QD	QE	QH
Registros de entrenamiento	100,605	43,004	70,244	40,894	56,979	45,863
Subclases	955	633	1,450	519	943	718
Keywords	225,891	140,686	238,606	123,692	198,107	151,612
Conjunto de prueba	25,151	5,376	17,561	10,224	14,245	11,466
Sin Clasificar	-	-	-	-	-	-
Cubiertos	25,151	5,376	17,561	10,224	14,245	11,466
Éxitos	8,045	1,600	4,203	2,633	3,234	3,225
Fallos	17,106	3,776	13,358	7,591	11,011	8,241
Precisión	31.99%	29.76%	23.93%	25.75%	22.70%	28.13%

Tabla 59: Algoritmo VFT, prueba multilingüe 80% entrenamiento 20% proyección, evaluación estricta, Parte 1

	QK	QL	QM	QP	QR	Totales
Registros de entrenamiento	30,932	58,625	4,305	30,142	8,423	490,016
Subclases	828	1,385	145	508	304	8,388
Keywords	115,720	178,317	16,459	95,125	27,095	1,511,310
Conjunto de prueba	7,694	14,656	1,076	7,536	2,106	117,091
Sin Clasificar	-	-	-	-	-	
Cubiertos	7,694	14,656	1,076	7,536	2,106	117,091
Éxitos	2,063	4,377	400	2,428	661	32,869
Fallos	5,631	10,279	676	5,108	1,445	84,222
Precisión	26.81%	29.86%	37.17%	32.22%	31.39%	28.07%

Tabla 60: Algoritmo VFT, prueba multilingüe 80% entrenamiento 20% proyección, evaluación estricta, Parte 2

b) Evaluación por posición (*estricta*).

	QA	QB	QC	QD	QE	QH	QK	QL	QM	QP	QR	Totales
1	31.99%	29.76%	23.93%	25.75%	22.70%	28.13%	26.81%	29.86%	37.17%	32.22%	31.39%	28.07%
2	14.27%	11.67%	11.15%	11.51%	12.13%	13.13%	11.58%	10.64%	14.50%	13.06%	14.34%	12.37%
3	8.87%	6.38%	7.14%	7.72%	7.00%	8.45%	6.91%	6.67%	6.13%	7.79%	7.31%	7.60%
4	5.77%	5.04%	5.23%	5.63%	6.15%	5.71%	5.71%	5.00%	7.43%	4.31%	6.08%	5.51%
5	4.17%	4.13%	4.29%	4.40%	4.35%	4.31%	4.11%	4.08%	5.76%	3.37%	3.85%	4.18%
1y5	**65.07%**	**56.98%**	**51.74%**	**55.01%**	**52.33%**	**59.73%**	**55.12%**	**56.25%**	**70.99%**	**60.75%**	**62.97%**	**57.73%**

Tabla 61: Algoritmo VFT, prueba multilingüe 80% entrenamiento 20% proyección, evaluación por posición.

b) Evaluación hasta el punto decimal.

	QA	QB	QC	QD	QE	QH
Registros de entrenamiento	100,605	43,004	70,244	40,894	56,979	45,863
Subclases	936	626	1446	517	940	710
Conjunto de prueba	25151	5376	17561	10224	14245	11466
Sin Clasificar	0	0	0	0	0	0
Cubiertos	25,151	5,376	17,561	10,224	14,245	11,466
Éxitos	13,279	1,704	4,708	2,672	3,314	3,539
Fallos	11,872	3,672	12,853	7,552	10,931	7,927
Precisión	52.80%	31.70%	26.81%	26.13%	23.26%	30.87%

Tabla 62: Algoritmo VFT, prueba multilingüe 80% entrenamiento 20% proyección, evaluación hasta el punto decimal, Parte 1

	QK	QL	QM	QP	QR	Totales
Registros de entrenamiento	30,932	58,625	4,305	30,142	8,423	490,016
Subclases	826	1383	144	503	298	8329
Conjunto de prueba	7694	14656	1076	7536	2106	117091
Sin Clasificar	0	0	0	0	0	
Cubiertos	7,694	14,656	1,076	7,536	2,106	117,091
Éxitos	2,097	4,506	440	2,532	691	39,482
Fallos	5,597	10,150	636	5,004	1,415	77,609
Precisión	27.26%	30.75%	40.89%	33.60%	32.81%	33.72%

Tabla 63: Algoritmo VFT, prueba multilingüe 80% entrenamiento 20% proyección, evaluación hasta el punto decimal, Parte 2

3.2 Resultados con algoritmo 1 – VFTP

a) Evaluación estricta.

	QA	QB	QC	QD	QE	QH
Registros de entrenamiento	100,605	43,004	70,244	40,894	56,979	45,863
Subclases	955	633	1,450	519	943	718
Keywords	225,891	140,686	238,606	123,692	198,107	151,612
Conjunto de prueba	25,151	5,376	17,561	10,224	14,245	11,466
Sin Clasificar	-	-	-	-	-	-
Cubiertos	25,151	5,376	17,561	10,224	14,245	11,466
Éxitos	9,511	1,789	5,443	3,393	4,687	4,125
Fallos	15,640	3,587	12,118	6,831	9,558	7,341
Precisión	37.82%	33.28%	30.99%	33.19%	32.90%	35.98%

Tabla 64: Algoritmo VFTP, prueba multilingüe 80% entrenamiento 80% proyección, evaluación estricta, Parte 1

	QK	QL	QM	QP	QR	Totales
Registros de entrenamiento	30,932	58,625	4,305	30,142	8,423	490,016
Subclases	828	1,385	145	508	304	8,388
Keywords	115,720	178,317	16,459	95,125	27,095	1,511,310
Conjunto de prueba	7,694	14,656	1,076	7,536	2,106	117,091
Sin Clasificar	-	-	-	-	-	
Cubiertos	7,694	14,656	1,076	7,536	2,106	117,091
Éxitos	2,711	5,512	483	3,336	773	41,763
Fallos	4,983	9,144	593	4,200	1,333	75,328
Precisión	35.24%	37.61%	44.89%	44.27%	36.70%	35.67%

Tabla 65: Algoritmo VFTP, prueba multilingüe 80% entrenamiento 80% proyección, evaluación estricta, Parte 2

b) Evaluación por posición (*estricta*).

	QA	QB	QC	QD	QE	QH	QK	QL	QM	QP	QR	Totales
1	37.82%	33.28%	30.99%	33.19%	32.90%	35.98%	35.24%	37.61%	44.89%	44.27%	36.70%	35.67%
2	14.12%	11.24%	10.28%	13.47%	11.82%	12.92%	12.52%	11.59%	15.52%	10.80%	11.21%	12.28%
3	7.79%	5.61%	6.05%	7.36%	7.02%	6.74%	7.23%	6.59%	7.81%	5.63%	8.36%	6.88%
4	5.44%	4.39%	4.26%	5.14%	4.81%	4.61%	4.91%	4.00%	5.30%	3.14%	4.56%	4.65%
5	3.94%	3.03%	3.55%	3.66%	3.16%	3.65%	3.16%	2.91%	2.88%	2.67%	2.75%	3.40%
1y5	**69.11%**	**57.55%**	**55.13%**	**62.82%**	**59.71%**	**63.90%**	**63.06%**	**62.70%**	**76.40%**	**66.51%**	**63.58%**	**62.88%**

Tabla 66: Algoritmo VFTP, prueba multilingüe 80% entrenamiento 80% proyección, evaluación por posición.

b) Evaluación hasta el punto decimal.

	QA	QB	QC	QD	QE	QH
Registros de entrenamiento	100,605	43,004	70,244	40,894	56,979	45,863
Subclases	936	626	1,446	517	940	710
keywords	225,891	140,686	238,606	123,692	198,107	151,612
Conjunto de prueba	25,151	5,376	17,561	10,224	14,245	11,466
Sin Clasificar	0	0	0	0	0	0
Cubiertos	25151	5376	17561	10224	14245	11466
Éxitos	14542	1922	6326	3582	4858	4377
Fallos	10609	3454	11235	6642	9387	7089
Precisión	57.82%	35.75%	36.02%	35.04%	34.10%	38.17%

Tabla 67: Algoritmo VFTP, prueba multilingüe 80% entrenamiento 80% proyección, evaluación hasta el punto decimal, Parte 1

	QK	QL	QM	QP	QR	Totales
Registros de entrenamiento	30,932	58,625	4,305	30,142	8,423	490,016
Subclases	826	1,383	144	503	298	8,329
keywords	115,720	178,317	16,459	95,125	27,095	1,511,310
Conjunto de prueba	7,694	14,656	1,076	7,536	2,106	117,091
Sin Clasificar	0	0	0	0	0	-
Cubiertos	7694	14656	1076	7536	2106	117091
Éxitos	2748	5713	500	3448	841	48857
Fallos	4946	8943	576	4088	1265	68234
Precisión	35.72%	38.98%	46.47%	45.75%	39.93%	41.73%

Tabla 68: Algoritmo VFTP, prueba multilingüe 80% entrenamiento 80% proyección, evaluación hasta el punto decimal, Parte 2

3.3 Resultados con algoritmo 2 – VFTP·(TF·IDF)

a) Evaluación estricta.

	QA	QB	QC	QD	QE	QH
Registros de entrenamiento	100,605	43,004	70,244	40,894	56,979	45,863
Subclases	955	633	1,450	519	943	718
Keywords	225,891	140,686	238,606	123,692	198,107	151,612
Conjunto de prueba	25,151	5,376	17,561	10,224	14,245	11,466
Sin Clasificar	-	-	-	-	-	-
Cubiertos	25,151	5,376	17,561	10,224	14,245	11,466
Éxitos	7,949	1,600	4,017	2,633	3,187	3,208
Fallos	17,202	3,776	13,544	7,591	11,058	8,258
Precisión	31.61%	29.76%	22.87%	25.75%	22.37%	27.98%

Tabla 69 Algoritmo 2, VFTP-TF·IDF, prueba multilingüe, 80% entrenamiento 20% proyección, evaluación estricta, Parte 1

	QK	QL	QM	QP	QR	Totales
Registros de entrenamiento	30,932	58,625	4,305	30,142	8,423	490,016
Subclases	828	1,385	145	508	304	8,388
keywords	115,720	178,317	16,459	95,125	27,095	1,511,310
Conjunto de prueba	7,694	14,656	1,076	7,536	2,106	117,091
Sin Clasificar	-	-	-	-	-	
Cubiertos	7,694	14,656	1,076	7,536	2,106	117,091
Éxitos	2,062	4,362	400	2,428	661	32,507
Fallos	5,632	10,294	676	5,108	1,445	84,584
Precisión	26.80%	29.76%	37.17%	32.22%	31.39%	27.76%

Tabla 70: Algoritmo 2, VFTP-TF·IDF, prueba multilingüe, 80% entrenamiento 20% proyección, evaluación estricta, Parte 2

b) Evaluación por posición (*estricta*).

	QA	QB	QC	QD	QE	QH	QK	QL	QM	QP	QR	Totales
1	31.61%	29.76%	22.87%	25.75%	22.37%	27.98%	26.80%	29.76%	37.17%	32.22%	31.39%	27.76%
2	13.65%	11.67%	10.37%	11.52%	12.05%	12.78%	11.58%	10.63%	14.51%	13.06%	14.33%	12.08%
3	8.68%	6.38%	6.91%	7.71%	7.07%	8.29%	6.91%	6.65%	6.13%	7.79%	7.31%	7.51%
4	5.60%	5.04%	4.99%	5.64%	6.23%	5.53%	5.68%	4.95%	7.43%	4.31%	6.08%	5.42%
5	4.06%	4.13%	4.15%	4.39%	4.28%	4.16%	4.09%	4.04%	5.76%	3.37%	3.85%	4.11%
1y5	**63.60%**	**56.98%**	**49.29%**	**55.01%**	**52.00%**	**58.74%**	**55.06%**	**56.03%**	**71.00%**	**60.75%**	**62.96%**	**56.88%**

Tabla 71: Algoritmo 2, VFTP-TF·IDF, prueba multilingüe, 80% entrenamiento 20% proyección, evaluación por posición.

c) Evaluación hasta el punto decimal.

	QA	QB	QC	QD	QE	QH
Registros de entrenamiento	100,605	43,004	70,244	40,894	56,979	45,863
Subclases	936	626	1,446	517	940	710
keywords	225,891	140,686	238,606	123,692	198,107	151,612
Conjunto de prueba	25,151	5,376	17,561	10,224	14,245	11,466
Sin Clasificar	-	-	-	-	-	-
Cubiertos	25,151	5,376	17,561	10,224	14,245	11,466
Éxitos	13,096	1,704	4,512	2,672	3,267	3,522
Fallos	12,055	3,672	13,049	7,552	10,978	7,944
Precisión	52.07%	31.70%	25.69%	26.13%	22.93%	30.72%

Tabla 72: Algoritmo 2, VFTP-TF·IDF, prueba multilingüe, 80% entrenamiento 20% proyección, evaluación hasta el punto decimal, Parte 1

	QK	QL	QM	QP	QR	Totales
Registros de entrenamiento	30,932	58,625	4,305	30,142	8,423	490,016
Subclases	826	1,383	144	503	298	8,329
keywords	115,720	178,317	16,459	95,125	27,095	1,511,310
Conjunto de prueba	7,694	14,656	1,076	7,536	2,106	117,091
Sin Clasificar	-	-	-	-	-	
Cubiertos	7,694	14,656	1,076	7,536	2,106	117,091
Éxitos	2,096	4,491	440	2,532	691	39,023
Fallos	5,598	10,165	636	5,004	1,415	78,068
Precisión	27.24%	30.64%	40.89%	33.60%	32.81%	**33.33%**

Tabla 73: Algoritmo 2, VFTP-TF·IDF, prueba multilingüe, 80% entrenamiento 20% proyección, evaluación hasta el punto decimal, Parte 2

3.4 Resultados con algoritmo 3 – DPT

a) Evaluación estricta.

	QA	QB	QC	QD	QE	QH
Registros de entrenamiento	Xº	43,004	70,244	40,894	56,979	45,863
Subclases	955	633	1,450	519	943	718
keywords	225,891	140,686	238,606	123,692	198,107	151,612
Conjunto de prueba	25,151	5,376	17,561	10,224	14,245	11,466
Sin Clasificar	-	-	-	-	-	-
Cubiertos	25,151	5,376	17,561	10,224	14,245	11,466
Éxitos	9,677	1,960	5,140	3,695	4,235	4,298
Fallos	15,474	3,416	12,421	6,529	10,010	7,168
Precisión	38.48%	36.46%	29.27%	36.14%	29.73%	37.48%

Tabla 74: Algoritmo 3, DPT, prueba multilingüe, 80% entrenamiento 20% proyección, evaluación estricta, Parte 1

	QK	QL	QM	QP	QR	Totales
Registros de entrenamiento	30,932	58,625	4,305	30,142	8,423	490,016
Subclases	828	1,385	145	508	304	8,388
keywords	115,720	178,317	16,459	95,125	27,095	1,511,310
Conjunto de prueba	7,694	14,656	1,076	7,536	2,106	117,091
Sin Clasificar	-	-	-	-	-	
Cubiertos	7,694	14,656	1,076	7,536	2,106	117,091
Éxitos	2,787	5,692	501	3,482	838	42,305
Fallos	4,907	8,964	575	4,054	1,268	74,786
Precisión	36.22%	38.84%	46.56%	46.20%	39.79%	**36.13%**

Tabla 75: Algoritmo 3, DPT, prueba multilingüe, 80% entrenamiento 20% proyección, evaluación estricta, Parte 2

b) Evaluación por posición.

	QA	QB	QC	QD	QE	QH	QK	QL	QM	QP	QR	Totales
1	38.48%	36.46%	29.27%	36.14%	29.73%	37.48%	36.22%	38.84%	46.56%	46.20%	39.79%	36.13%
2	10.91%	7.83%	6.68%	8.37%	8.11%	7.83%	31.50%	7.78%	34.85%	6.98%	8.03%	10.15%
3	5.20%	1.99%	3.01%	3.92%	3.19%	3.12%	6.78%	3.17%	5.67%	2.80%	3.32%	3.83%
4	2.70%	1.75%	1.53%	2.51%	1.72%	1.77%	3.22%	1.78%	3.53%	1.04%	1.38%	2.05%
5	1.45%	0.84%	1.09%	1.62%	0.88%	0.97%	1.86%	0.95%	2.88%	0.78%	1.14%	1.19%
1y5	58.74%	48.87%	41.58%	52.56%	43.63%	51.17%	79.58%	52.52%	93.49%	57.80%	53.66%	53.35%

Tabla 76: Algoritmo 3, DPT, prueba multilingüe, 80% entrenamiento 20% proyección, evaluación por posición

c) Evaluación hasta el punto decimal.

	QA	QB	QC	QD	QE	QH
Registros de entrenamiento	100,605	43,004	70,244	40,894	56,979	45,863
Subclases	936	626	1,446	517	940	710
keywords	225,891	140,686	238,606	123,692	198,107	151,612
Conjunto de prueba	25,151	5,376	17,561	10,224	14,245	11,466
Sin Clasificar	0	0	0	0	0	0
Cubiertos	25151	5376	17561	10224	14245	11466
Éxitos	14319	2054	5630	3786	4353	4541
Fallos	10832	3322	11931	6438	9892	6925
Precisión	56.93%	38.21%	32.06%	37.03%	30.56%	39.60%

Tabla 77: Algoritmo 3, DPT, prueba multilingüe, 80% entrenamiento 20% proyección, evaluación hasta el punto decimal, Parte 1

	QK	QL	QM	QP	QR	Totales
Registros de entrenamiento	30,932	58,625	4,305	30,142	8,423	490,016
Subclases	826	1,383	144	503	298	8,329
keywords	115,720	178,317	16,459	95,125	27,095	1,511,310
Conjunto de prueba	7,694	14,656	1,076	7,536	2,106	117,091
Sin Clasificar	0	0	0	0	0	
Cubiertos	7694	14656	1076	7536	2106	117091
Éxitos	2815	5796	527	3565	888	48274
Fallos	4879	8860	549	3971	1218	68817
Precisión	36.59%	39.55%	48.98%	47.31%	42.17%	41.23%

Tabla 78: Algoritmo 3, DPT, prueba multilingüe, 80% entrenamiento 20% proyección, evaluación hasta el punto decimal, Parte 2

3.5 Resultados con algoritmo 4 – CT-MLC

a) Por semejanza máxima con un elemento en la clase.

	QA	QB	QC	QD	QE	QH
Registros de entrenamiento	100,605	43,004	70,244	40,894	56,979	45,863
Subclases	955	633	1,450	519	943	718
Keywords	225,891	140,686	238,606	123,692	198,107	151,612
Conjunto de prueba	25,151	5,376	17,561	10,224	14,245	11,466
Sin Clasificar	-	-	-	-	-	-
Cubiertos	25,151	5,376	17,561	10,224	14,245	11,466
Éxitos	10,118	1,724	5,719	3,457	4,581	3,823
Fallos	15,033	3,652	11,842	6,767	9,664	7,643
Precisión	40.23%	32.07%	32.57%	33.81%	32.16%	33.34%

Tabla 79: Algoritmo 4, CT-MLC, prueba multilingüe, 80% entrenamiento 20% proyección, máxima semejanza a un elemento de la clase. Parte1.

	QK	QL	QM	QP	QR	Totales
Registros de entrenamiento	30,932	58,625	4,305	30,142	8,423	490,016
Subclases	828	1,385	145	508	304	8,388
keywords	115,720	178,317	16,459	95,125	27,095	1,511,310
Conjunto de prueba	7,694	14,656	1,076	7,536	2,106	117,091
Sin Clasificar	-	-	-	-	-	-
Cubiertos	7,694	14,656	1,076	7,536	2,106	117,091
Éxitos	2,462	5,140	480	3,249	784	41,537
Fallos	5,232	9,516	596	4,287	1,322	75,554
Precisión	32.00%	35.07%	44.61%	43.11%	37.23%	35.47%

Tabla 80: Algoritmo 4, CT-MLC, prueba multilingüe, 80% entrenamiento 20% proyección, máxima semejanza a un elemento de la clase. Parte 2

b) Por posición en la lista de salida.

	QA	QB	QC	QD	QE	QH	QK	QL	QM	QP	QR	Totales
1	40.23%	32.07%	32.57%	33.81%	32.16%	33.34%	32.00%	35.07%	44.61%	43.11%	37.23%	35.47%
2	8.76%	8.65%	7.34%	7.09%	7.95%	7.33%	7.54%	7.41%	54.37%	7.04%	7.55%	8.20%
3	5.01%	5.02%	4.53%	4.53%	4.97%	4.06%	4.94%	4.52%	0.00%	4.02%	5.03%	4.62%
4	3.44%	3.74%	3.19%	3.64%	3.25%	3.58%	3.18%	3.10%	0.00%	3.18%	2.99%	3.31%
5	2.66%	2.70%	2.47%	2.43%	2.79%	2.36%	2.66%	2.55%	0.00%	1.98%	2.09%	2.51%
1y5	**60.10%**	**52.18%**	**50.10%**	**51.50%**	**51.12%**	**50.67%**	**50.32%**	**52.65%**	**98.98%**	**59.33%**	**54.89%**	**54.11%**

Tabla 81: Algoritmo 4, CT-MLC, prueba multilingüe, 80% entrenamiento 20% proyección, máxima semejanza a un elemento de la clase. Evaluación por posición.

3.5 Resultados con algoritmo 4' – CT-MLC

a) Por máximo promedio ***de semejanza en la clase***.

	QA	QB	QC	QD	QE	QH
Registros de entrenamiento	100,605	43,004	70,244	40,894	56,979	45,863
Subclases	955	633	1,450	519	943	718
Keywords	225,891	140,686	238,606	123,692	198,107	151,612
Conjunto de prueba	25,151	5,376	17,561	10,224	14,245	11,466
Sin Clasificar	-	-	-	-	-	-
Cubiertos	25,151	5,376	17,561	10,224	14,245	11,466
Éxitos	7,577	1,369	3,864	2,429	3,096	2,763
Fallos	17,574	4,007	13,697	7,795	11,149	8,703
Precisión	30.13%	25.47%	22.00%	23.76%	21.73%	24.10%

Tabla 82: Algoritmo 4', CT-MLC, prueba multilingüe, 80% entrenamiento 20% proyección, máxima semejanza promedio en la clase. Parte1.

	QK	QL	QM	QP	QR	Totales
Registros de entrenamiento	30,932	58,625	4,305	30,142	8,423	490,016
Subclases	828	1,385	145	508	304	8,388
Keywords	115,720	178,317	16,459	95,125	27,095	1,511,310
Conjunto de prueba	7,694	14,656	1,076	7,536	2,106	117,091
Sin Clasificar	-	-	-	-	-	
Cubiertos	7,694	14,656	1,076	7,536	2,106	117,091
Éxitos	1,787	3,711	416	2,593	618	30,223
Fallos	5,907	10,945	660	4,943	1,488	86,868
Precisión	23.23%	25.32%	38.66%	34.41%	29.34%	25.81%

Tabla 83: Algoritmo 4', CT-MLC, prueba multilingüe, 80% entrenamiento 20% proyección, máxima semejanza promedio en la clase. Parte 2

b) Por posición en la lista de salida.

	QA	QB	QC	QD	QE	QH	QK	QL	QM	QP	QR	Totales
1	30.13%	25.47%	22.00%	23.76%	21.73%	24.10%	23.23%	25.32%	38.66%	34.41%	29.34%	25.81%
2	11.93%	9.83%	9.58%	9.76%	8.45%	9.14%	8.50%	9.33%	1.68%	8.69%	9.40%	9.71%
3	7.19%	5.75%	5.65%	6.33%	5.48%	5.72%	5.97%	5.50%	6.13%	5.07%	6.42%	6.01%
4	5.59%	4.37%	3.95%	4.48%	4.78%	4.11%	4.22%	3.86%	4.18%	3.69%	4.94%	4.49%
5	3.84%	3.85%	3.46%	4.06%	3.47%	3.21%	3.10%	3.52%	3.53%	2.90%	3.56%	3.54%
1y5	**58.68%**	**49.27%**	**44.64%**	**48.39%**	**43.91%**	**46.28%**	**45.02%**	**47.53%**	**54.18%**	**54.76%**	**53.66%**	**49.56%**

Tabla 84: Algoritmo 4', CT-MLC, prueba multilingüe, 80% entrenamiento 20% proyección, máxima semejanza promedio en la clase. Evaluación por posición.