

4 RESULTADOS Y EVALUACIÓN

Habiendo explicado la metodología empleada para la obtención automática de términos y su validación en el capítulo anterior, es necesario pasar a los resultados obtenidos y su evaluación.

Este capítulo estará conformado por las siguientes partes, la primera de ellas se refiere a la extracción de candidatos a término del COCIEM; la segunda es la selección de candidatos a término a validar. La tercera parte está conformada por la validación de los candidatos a término empleando Wikipedia. Y finalmente, la cuarta parte es la evaluación de los resultados obtenidos y algunas observaciones de ésta.

4.1 Extracción de candidatos a término del COCIEM

Para llevar a cabo la extracción de candidatos a términos, primeramente, se establecieron dos parámetros. El primero de ellos era la división del COCIEM en tres subcorpus, cada uno determinado por el nivel escolar, es decir, primaria, secundaria y bachillerato. La razón de realizar esta división fue para que los candidatos a término se quedaran en cada uno de los niveles educativos y no se perdieran o disminuyera su importancia por el avance del conocimiento a través de los años. El segundo parámetro era considerar cada libro de los subcorpus como un documento, es decir, los libros no se concatenaron por materia o grado escolar. Lo que se buscó con esto último es que los términos no se vieran sobre o subcalificados al hacer el análisis de TF-IDF y además se tuvieran listas de candidatos a término por libro.

Posteriormente, ya con los parámetros establecidos se procedió a preprocesar cada uno de los subcorpus y a extraer los candidatos a término siguiendo las dos primeras partes de la arquitectura mostrada en la sección 3.5. De estas dos tareas se obtuvieron tres listas de candidatos a término por subcorpus que pertenecen a unigramas, bigramas y trigramas, respectivamente; asimismo estas se encuentran divididas por libro siguiendo el formato de salida mostrado en la Tabla 13 de la sección 3.3.3. En la Tabla 16 se muestra el número de

candidatos a término con repetición que se obtuvieron de la extracción de candidatos a término.

Nivel educativo	Materia	Número de libros de texto	Unigramas	Bigramas	Trigramas
Primaria	Ciencias naturales	6	6,606	6,658	14,536
	Matemáticas	9	4,891	3,458	7,618
Total		15	11,497	10,116	22,154
Secundaria	Biología	8	13,968	17,636	32,036
	Matemáticas	24	10,728	17,945	35,127
	Física	9	7,837	19,823	38,296
	Química	8	10,812	16,583	31,745
	Educación ambiental	1	1,625	3,738	7,370
Total		50	44,970	75,725	144,574
Bachillerato	Biología	3	7,202	7,057	11,484
	Matemáticas	11	8,450	9,924	18,405
	Física	3	3,296	8,286	15,817
	Química	5	7,785	12,700	23,393
	Educación para la salud	3	3,958	8,570	11,689
	Ecología	2	4,817	5,999	10,690
Total		27	35,508	52,536	91,478
Gran total		92	91,975	138,377	258,206

Tabla 16. Número de candidatos a término con repetición por nivel educativo, materia y tipo de n-grama

Como se puede observar en la Tabla 16, en los casos de matemáticas de los tres niveles educativos el número de candidatos a término es menor a la proporción de libros que conforman a las materias. La razón de ello es que la materia de matemáticas se encuentra

constituida por varios libros de ejercicios y actividades y no tanto de teoría como en otras materias.

Los pesos de TF-IDF para la mayoría de los candidatos a término son medianos y bajos, pocos son los que obtuvieron un peso alto. Posiblemente porque al ser libros con un enfoque educativo se incluye una gran cantidad de información o ejemplos que pueden estar altamente relacionados con la vida cotidiana, como lo son las compras en el mercado o los medios de transporte; asimismo se incluye información que está frecuentemente relacionada con otras materias para comprender de mejor forma los temas explicados.

4.2 Selección de los candidatos a término a validar

Una vez finalizada la extracción de candidatos a término se procedió a continuar con la tercera parte de arquitectura del sistema de extracción terminológica mostrada en la sección 3.5. Esta tercera parte es la selección de los candidatos a término a validar, el cual es una tarea que se lleva a cabo a partir de las listas de candidatos a término obtenidas en los dos pasos anteriores dados a conocer en la sección 4.1.

A pesar de haber obtenido listas de candidatos a término para cada uno de los subcorpus del COCIEM y sus libros, solamente se eligieron tres materias para proseguir con la selección de los candidatos a término a validar. Las razones de realizar esto es que solamente se pudieron obtener 3 listas de términos creados por expertos para realizar una evaluación, de las cuales se hablarán más adelante en la sección 4.4; también se eligieron éstas porque son materias que tienen una categoría bien definida dentro de Wikipedia. Las materias seleccionadas son las siguientes:

- Matemáticas de bachillerato
- Ecología de bachillerato
- Matemáticas de primaria

Para llevar a cabo la selección de los candidatos a término a validar de estas tres materias se establecieron dos parámetros. El primero de ellos consiste en la generación de distintas listas de candidatos a término a validar usando distintos umbrales de pesos de TF-

IDF, esto para observar los distintos resultados que se obtienen al modificar el límite de lo que se considera un candidato a término a validar o no. Con base en que los pesos de TF-IDF que se obtuvieron no fueron altos sino bajos y medianos, se instituyeron los siguientes tres umbrales:

- $TF-IDF \geq 0.03$
- $TF-IDF \geq 0.01$
- $TF-IDF > 0$

El segundo parámetro que se estableció fue el de unir las listas de candidatos a término a validar por materia, en otras palabras, se unieron las listas de candidatos a término de unigramas, bigramas y trigramas de cada libro según la materia a la que perteneciera. Esto para disminuir el número de candidatos a término comunes entre libros a validar y tener al final una lista de términos validados con todas las posibles construcciones analizadas.

En la Tabla 17 se da a conocer un extracto de los candidatos a término para la materia de matemáticas de bachillerato; en la Tabla 18 se muestra un extracto para ecología de bachillerato y finalmente en la Tabla 19 se expone un extracto de las listas obtenidas por el extractor para matemáticas de primaria. Cabe aclarar que en estas tablas se mezclaron unigramas, bigramas y trigramas de los libros que conforman cada una de las materias analizadas.

Matemáticas de Bachillerato			
cuadrado	0.652319739	codominio	0.073741693
método de simpson	0.450862706	isomorfismo	0.057477441
cónica	0.226610224	ley de senos	0.036520337
triángulo oblicuángulo	0.185921002	escondite	0.030788647
binomios	0.153943209	reducir el contaminación	0.01194413
función primitivo	0.150593012	mauricio gómez	0.011891407
función exponencial natural	0.125413295	incluir el fórmula	0.008840447
ecuación trigonométrico	0.092960501	velocidad uniforme	0.007931875
construcción alternativa	0.092859351	voluntario	0.006376907
lado y ángulo	0.09130083	proyectil	0.002365373

Tabla 17. Extracto de las listas de candidatos de los libros de matemáticas de bachillerato

Ecología de bachillerato			
ecosistema	0.363877584	fuentes de contaminación	0.037250487
selva tropical	0.182599668	diclorodifeniltricloroetano	0.03127365
bosque	0.146871039	bosque de conífero	0.029398861
ciclo de vida	0.118631532	cambiar radicalmente	0.02102087
ppm	0.107039546	región tropical	0.012179076
índice de natalidad	0.10192139	inagotable	0.009825895
biodiversidad	0.084841923	quemar el hidrocarburo	0.009706804
bosque de coníferas	0.058240796	gota de agua	0.006204729
organismo descomponedor	0.042041732	biogeografía	0.004912947
lechuza	0.040531815	cálculo	0.000179262

Tabla 18. Extracto de las listas de candidatos de los libros de ecología de bachillerato

Matemáticas de primaria			
números romano	0.395719045	línea recto	0.009294068
itzel	0.388469382	diagrama de árbol	0.077376202
decámetro	0.257031928	expropiación petrolero	0.073713457
multiplicación	0.182761216	porcentaje	0.065171074
mónica e itzel	0.178761075	encerrar el número	0.050622666
probabilidad	0.141134807	tabla de frecuencia	0.038387955
juanita	0.118100483	rombo	0.035094508
fracción impropio	0.113304833	derecha exclusivo	0.028265639
mínimo común múltiplo	0.104221928	desgraciadamente	0.017478482
eje de simetría	0.102802493	aproximación	0.006075622

Tabla 19. Extracto de las listas de candidatos de los libros de matemáticas de primaria

Como se puede observar en las tablas anteriores existen algunos errores que se deben a la lematización por parte de FreeLing o por problemas del texto original. Por ejemplo, en la Tabla 17 existe el caso de “construcción alternativa”, el cual no concuerda con su lema, el cual debería ser “construcción alternativo”; la razón de ello es que FreeLing consideró la palabra “alternativa” como un sustantivo porque el bigrama era un renglón del texto original y no tenía más contexto para determinar que era un adjetivo. En cambio, en la Tabla 18 se muestra el caso de “bosque de conífero” y “bosque de conifera”; el primero de ellos sí está bien lematizado, el segundo no porque le faltó un acento a la palabra “conífera” y por tanto FreeLing no pudo encontrar el lema correcto.

Además de los errores de lematización, en las tablas mostradas anteriormente existen casos en los que los candidatos a término no son en realidad posibles términos, como “desgraciadamente”, el cual es un adverbio, o “lado y ángulo” que son dos sustantivos unidos por una conjunción. Este tipo de error se debe a que se usó conocimiento estadístico para extraer los candidatos a término y no uno lingüístico o híbrido que permitiera filtrar las construcciones que posiblemente no eran candidatos a término.

Se puede observar de igual forma en las tablas que algunos de los candidatos a término no pertenecen a las materias donde se encontraron, como “juanita”, “expropiación petrolera” en matemáticas de primaria, o “reducir el contaminación” en matemáticas de bachillerato. Esto se debe por que los libros empleados en el COCIEM son de carácter educativo y frecuentemente se emplean ejemplos en el cual se emplean nombres comunes de personas o se tienen cuadros del estilo de ¿sabías qué? que incluyen información adicional o relacionada con otras materias.

En la Tabla 20 se muestra el número de candidatos a término sin repetición para cada umbral determinado para los pesos de TF-IDF. A partir de la tabla se puede verificar que la gran mayoría de los candidatos término tiene un peso de TF-IDF menor a 0.03, es decir, un peso bajo en una escala del 0 al 1. Además se puede observar que una gran parte de los candidatos a término a validar sin repetición de matemáticas de primaria tiene un peso de TF-IDF mayor o igual a 0.01 a diferencia de ecología y matemáticas de bachillerato donde poco más de la mitad de candidatos a término tiene un peso de TF-IDF menor a 0.01.

Candidatos a términos a validar sin repetición			
Materia	TF-IDF > 0	TF-IDF ≥ 0.01	TF-IDF ≥ 0.03
Matemáticas de bachillerato	33,944	19,582	4,561
Ecología de bachillerato	19,304	9,428	583
Matemáticas de primaria	13,851	13,062	4,081

Tabla 20. Número de candidatos a término a validar sin repetición por materia para cada uno de los umbrales de TF-IDF determinados

4.3 Obtención de términos validados por Wikipedia

Para llevar a cabo la validación de los candidatos a término por parte de Wikipedia, fue necesario en una primera instancia la selección de las fronteras de dominio para cada una de las materias que se analizaron, es decir, para matemáticas de primaria y ecología y matemáticas de bachillerato.

El procedimiento de selección de las fronteras de dominio se llevó a cabo de manera manual, buscando en la base de datos de Wikipedia las posibles categorías que serían equivalentes o que contendrían las materias analizadas. En la Tabla 21 se muestran las fronteras de dominio empleadas para cada una de las materias que se validaron.

Materia	Fronteras de dominio
Matemáticas de primaria	Matemáticas, Álgebra, Aritmética y Geometría
Matemáticas de bachillerato	Matemáticas, Álgebra, Cálculo y Geometría
Ecología de bachillerato	Ecología, Medio ambiente y Climatología

Tabla 21. Fronteras de dominio para cada una de las materias analizadas

Aunque la mayoría de las fronteras de dominio pertenecen a la primera frontera de dominio que aparece en cada renglón de la Tabla 21, se decidió indicar de manera más específica las fronteras de dominio, porque conforme se realizó el procedimiento de validación se observó que en determinados casos era necesario aumentar el límite de caminos a analizar³⁵ para encontrar la relación entre el candidato a término y la frontera de dominio; este aumento en el límite causaba que en ocasiones no se pudiera terminar la validación por el exceso de información que se generaba. Por consiguiente, para no aumentar el número máximo de caminos se decidió aumentar el número de categorías, que aunque son redundante no afectan el desempeño del programa.

Con las fronteras de dominio establecidas se procedió a validar cada una de las listas de candidatos a término dadas a conocer en la sección 4.2. El proceso de validación consiste básicamente en el cálculo del coeficiente de dominio, el cual como se vio en la sección 3.4.2, puede ser calculado de tres formas que son las siguientes:

- Número de caminos (CDnc)
- Longitud de caminos (CDlc)
- Longitud media de caminos (CDlmc)

³⁵ El límite de caminos a analizar previene que la validación de un término caiga en un ciclo infinito. En el caso de esta tesis el límite es de 500 caminos antes de detener la validación y proseguir con el siguiente candidato a término.

La información para el cálculo de estos coeficientes de dominio, la cual se explicó de igual forma en la sección 3.4.2, consiste a grandes rasgos en buscar primero la existencia del candidato a término en Wikipedia, y posteriormente en determinar si éste está ligado a una página de desambiguación o no. Si es este último caso se procede a buscar las categorías con las que está ligado el candidato a término y con cuáles están sus categorías relacionadas a la vez; con la información obtenida de esta búsqueda se calculan los tres coeficientes de dominio. De encontrarse ligado a una página de desambiguación se analizan cada una de las entradas polisémicas de forma normal, es decir, como si fuera otro candidato a término, y el que obtenga el mayor valor del coeficiente de dominio, sin importar la forma de ser calculado, se considera como la acepción buscada.

Además cabe recordar que el coeficiente de dominio (CD) tiene cuatro rangos, los cuales son los siguientes:

- **$CD = 1.00$** : El candidato a término es un término de la frontera de dominio.
- **$0.00 < CD < 1.00$** : El candidato a término está medianamente relacionado con la frontera de dominio pero se le considera como un término de la frontera de dominio.
- **$-1.00 < CD \leq 0.00$** : El candidato a término no está relacionado con la frontera de dominio y por tanto no es un término.
- **$CD = -1.00$** : El candidato a término no se encontró en Wikipedia y no se sabe si es un término o no de la frontera de dominio.

En la Tabla 22 se muestra el número de términos validados para cada una de las materias analizadas según los distintos umbrales establecidos para los pesos de TF-IDF. Se puede observar que por lo general los coeficientes CD_{nc} y CD_{lc} obtienen un número similar de términos validados, mientras que CD_{mc} obtiene aproximadamente un tercio menos de términos validados. Además, el número de términos validados aumenta conforme se disminuye el umbral de TF-IDF, esto por el aumento de candidatos a término a validar; cabe observar que estos aumentos no son tan grandes como los mostrados en la Tabla 20, la cual muestra los candidatos a término a validar sin repetición.

Términos validados									
Peso	TF-IDF > 0			TF-IDF ≥ 0.01			TF-IDF ≥ 0.03		
Métrica Materia	CDnc	CDlc	CDlmc	CDnc	CDlc	CDlmc	CDnc	CDlc	CDlmc
Matemáticas de bachillerato	2,909	2,830	1,806	1,575	1,524	1,099	591	577	454
Ecología de bachillerato	644	707	638	325	337	309	96	93	88
Matemáticas de primaria	1,748	1,698	901	1,368	1,334	736	535	517	310

Tabla 22. Número de términos validados por materia para cada uno de los umbrales de TF-IDF determinados

En la Tabla 23 se muestra el número de candidatos a término que tenían su respectivo artículo en Wikipedia pero no se les pudo calcular el coeficiente de dominio porque las páginas de los artículos no estaban ligados a ninguna categoría. A estos casos se les asigna un coeficiente de dominio igual a cero, el cual indica que fueron encontrados en Wikipedia pero no están relacionados con la frontera de dominio, en este caso porque no se pudo saber si pertenecían o no a ella.

Candidatos a término encontrados en Wikipedia pero sin categoría asignada									
Peso	TF-IDF > 0			TF-IDF ≥ 0.01			TF-IDF ≥ 0.03		
Tipo Materia	Unigrama	Bigrama	Trigrama	Unigrama	Bigrama	Trigrama	Unigrama	Bigrama	Trigrama
Matemáticas de bachillerato	289	80	28	128	60	19	46	28	5
Ecología de bachillerato	227	32	27	98	24	14	12	6	3
Matemáticas de primaria	172	30	11	132	30	11	37	19	6

Tabla 23. Número de candidatos a término que fueron encontrados como artículos en Wikipedia pero que no estaban asignados a ninguna categoría

Como se puede observar en la Tabla 23, la mayor cantidad de páginas de Wikipedia no asignadas a ninguna categoría están relacionadas con unigramas. Algunos ejemplos de candidatos a términos sin una categoría asignada son: “determinante”, “módulo”, “culiacán”, “ecuación lineal”, “señal de tránsito” o “integración por sustitución” en matemáticas de bachillerato; en ecología de bachillerato “pirámides”, “transgénico”, “control poblacional”, “cuenca hidrológico”, “valor monetario”, “campo de estudio” o “grano de polen”; mientras que en matemáticas de primaria “jugo”, “área”, “imperio inca”, “fracción equivalente”, “paleta de agua” o “suma de fracción”. Entre los ejemplos dados anteriormente existen algunos candidatos a término que realmente son términos de la materia a la que están adjudicados.

A continuación en la Tabla 24 se muestran los primeros 20 términos validados de matemáticas de bachillerato para cada umbral establecido de TF-IDF, donde cabe aclarar que se obtuvieron los mismos términos para cada uno de los métodos del cálculo del coeficiente de dominio. En cambio, en la Tabla 25 se muestran los primeros 20 términos validados de ecología de bachillerato para cada umbral determinado de TF-IDF, donde se observa que para $TF-IDF \geq 0.03$ se obtuvieron distintos términos validados a partir de la posición 16 para cada uno de los métodos del cálculo del coeficiente de dominio. En la Tabla 26 se muestran los primeros 20 términos validados para matemáticas de primaria para cada umbral establecido de TF-IDF; en este caso para $TF-IDF \geq 0.01$ varía para el cálculo de CDlc la lista de términos en la posición 19 y 20.

En general los primeros lugares de las listas de términos validados muestran términos que están altamente relacionados con la materia a la que se les adjudica. Sin embargo, existen casos en los que se aceptan términos con faltas de ortografía o que tienen una cierta ambigüedad.

Matemáticas de bachillerato		
TF-IDF>0	TF-IDF $\geq$ 0.01	TF-IDF $\geq$ 0.03
adición	adición	aritmética
afinidad	álgebra	asociatividad
álgebra	aproximación lineal	binomio
ancho	aritmética	binomio conjugado
aproximación lineal	asociatividad	binomio de newton
aritmética	binomio	bisectriz
asociatividad	binomio conjugado	cardioides
astroide	binomio de newton	caso de factorización
bidimensional	bisectriz	cifra
binomio	cálculo diferencial	circunferencia
binomio conjugado	caracol de pascal	circunferencia inscribir
binomio de newton	cardioides	circunferencia unitario
bisectriz	caso de factorización	cociente
cálculo diferencial	centena	combinaciones
caracol de pascal	centro	conjunto
característica	cicloide	construcción geométrica
cardioides	ciencia matemática	corona circular
caso de factorización	cifra	cuantificador existencial
casquete esférico	circunferencia	cubo
cateto	circunferencia inscribir	cuerpo geométrico

Tabla 24. Los 20 primeros términos de matemáticas de bachillerato para cada umbral empleado de TF-IDF

Ecología de bachillerato				
TF-IDF>0	TF-IDF ≥ 0.01	TF-IDF ≥ 0.03		
abiótico	abiótico	abiótico		
acaricida	acaricida	acaricida		
amensalismo	amensalismo	amensalismo		
biodegradabilidad	biodegradabilidad	biodegradación		
biodegradación	biodegradación	biótico		
biótico	biótico	clima		
clima	clima	contaminación		
clima árido	clima árido	ddt		
clordano	clordano	descomponedor		
clímax	clímax	diclorodifeniltricloroetano		
comensal	comensal	explotación		
condición climático	condición climático	hábitat		
conservación biológico	conservación biológico	neutralismo		
contaminación	contaminación	protooperación		
contaminante	corriente oceánico	semiárido		
corriente de humboldt	crisis ecológico	comunidad	nutriente	territorial
corriente marino	cálido	continental	vertido	vertido
corriente oceánico	ddt	ecológico	migratorio	naturaleza
crisis ecológico	descomponedor	entropía	naturaleza	migratorio
cálido	desecho tóxico	selva	territorial	ambiental
		CDnc	CDlc	CDlmc

Tabla 25. Los 20 primeros términos de ecología de bachillerato para cada umbral empleado de TF-IDF. En el caso TF-IDF≥0.03 se muestran a partir de la posición 16 los distintos términos validados ordenados según el cálculo del coeficiente de dominio.

Matemáticas de primaria			
TF-IDF>0	TF-IDF ≥ 0.01		TF-IDF ≥ 0.03
ancho	ancho		ancho
antecesor	antecesor		antecesor
arista	arista		cardinalidad
aritmética	aritmética		centena
billón	billón		cifra
cara	cara		cúbico
característica	característica		decena
cardinalidad	cardinalidad		diámetro
centena	centena		dígito
centro	cifra		elementos
cifra	compacto		figura
combinaciones	completar el cuadrado		figura geométrico
compacto	cuerpo geométrico		fracción impropio
completar el cuadrado	cuerpos geométricos		heptágono
cubo	curva		hexágono regular
cuerpo geométrico	cúbico		lado
cuerpos geométricos	decena		longitud
curva	decágono		líneas
cúbico	denominador	dodecaedro	matemática
decena	diámetro	denominador	minuendo
CDnc y CDlmc		CDlc	

Tabla 26. Los 20 primeros términos de matemáticas de primaria para cada umbral empleado de TF-IDF. En el caso TF-IDF≥0.01 se muestran a partir de la posición 19 los distintos términos validados ordenados según el cálculo del coeficiente de dominio.

Se puede observar en la Tabla 24 que uno de los primeros términos validados es “algebra” sin acento, esto puede ser porque hubo un error en la digitalización del libro o en el libro mismo, pero Wikipedia tiene una redirección a “álgebra”; cabe aclarar que “álgebra” también aparece, solamente que algunos lugares más abajo pero también con un coeficiente de dominio igual a 1. Con respecto a la Tabla 25 no muestra errores como los anteriores, pero tiene la peculiaridad de que cambia el orden de algunos términos validados para el umbral $TF-IDF \geq 0.03$ dependiendo de la forma en que se calculó el coeficiente de dominio. En la Tabla 26 se muestra que existe un error, posiblemente por la digitalización del libro o su escritura, el cual es “denominador” el cual fue encontrado en Wikipedia porque tiene un redireccionamiento hacia “denominador”.

De igual forma, se pueden observar que existen casos en el cual la lematización no fue la correcta, como en “aritmética” o “cuerpos geométricos”, aunque sí fueron localizados en Wikipedia gracias a su estructura.

4.4 Evaluación de resultados

La evaluación de los sistemas extractores de términos es uno de los procesos más importantes que se deben llevar siempre a cabo. En esta tesis, la evaluación del sistema extractor desarrollado no es una excepción; para ello se emplean las métricas más empleadas en los proyectos relacionados con el PLN, que son precisión y cobertura.

Una de las razones por las cuales sólo se eligieron tres materias para llevar a cabo la obtención de los términos fue la falta de listas validadas por expertos para la evaluación de los resultados. Para matemáticas de primaria se obtuvo una lista de términos obtenida por el método de lista de referencia (sección 2.3.1), es decir, se analizó de manera manual el corpus, más específicamente los libros de matemáticas de primaria, para la extracción de los términos que se encontraban en esta sección del COCIEM.

En el caso de matemáticas de bachillerato, se empleó una lista de términos que fue obtenida a través del método de validación (sección 2.3.2); esta lista es la misma que fue empleada en el proyecto descrito en el artículo de Cabrera-Diego et al. (2011), donde estudiantes de ingeniería y lingüística analizaron los candidatos a término que se obtuvieron

empleando el extractor terminológico del programa WordSmith Tools 5.0 (Scott, 1996) que se basa en el logaritmo de la verosimilitud³⁶.

Finalmente, en el caso de ecología de bachillerato, se empleó igualmente el método de validación para la obtención de una lista de términos; a diferencia de la lista de matemáticas de bachillerato, en este caso se empleó la salida generada por el extractor terminológico desarrollado en esta tesis, donde un biólogo validó cada uno de los n-gramas existentes en los libros. Cabe aclarar que para ello no se le dio ninguna información sobre los pesos obtenidos en el proceso de extracción, esto para no sesgar los posibles resultados.

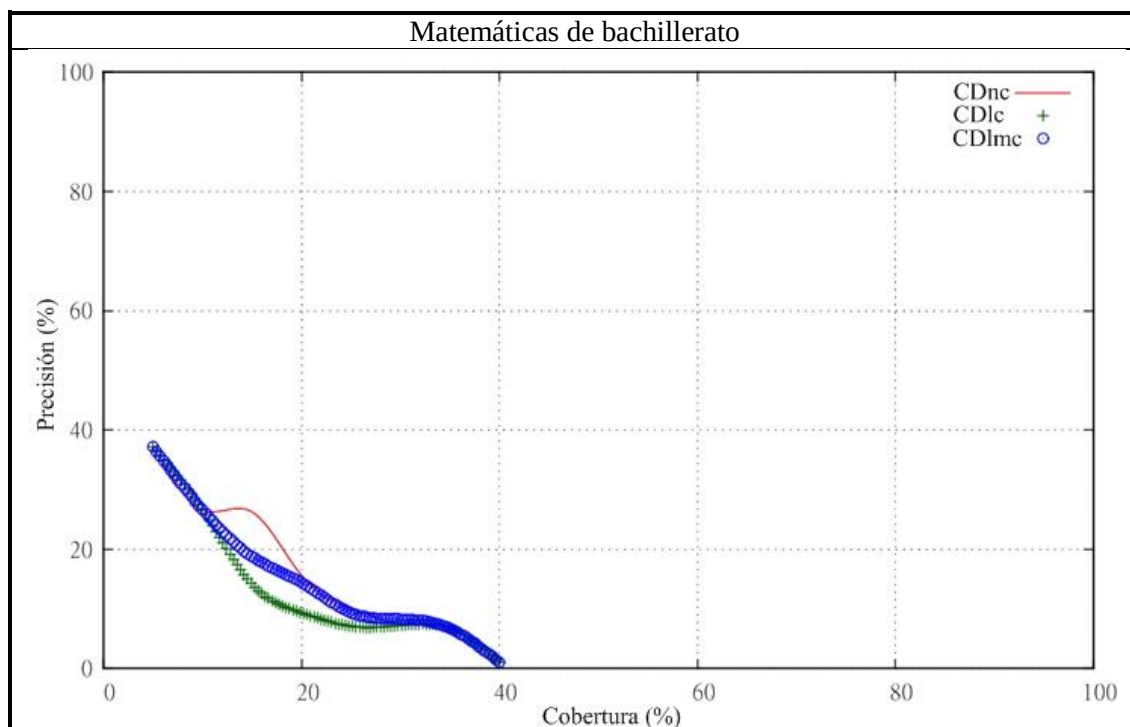
La evaluación de los resultados obtenidos por la validación de Wikipedia se llevó a cabo para unigramas, bigramas, trigramas y los tres n-gramas unidos. Para llevar a cabo esto, primero se lematizaron las listas de términos creadas por los expertos de matemáticas de bachillerato y primaria siguiendo un método similar al empleado para lematizar la base de datos de Wikipedia (sección 3.4.1.2), la única diferencia es que en este caso no existe un identificador; en el caso de la lista de términos de ecología de bachillerato esta no fue lematizada, ya que ya se encontraba en ese estado por provenir de la salida del análisis de TF-IDF. Posteriormente, las listas de términos de los expertos se dividieron en tres partes, en unigramas, bigramas y trigramas, en esta última parte se agregaron también los n-gramas que tuvieran un tamaño mayor que 3 en el caso de matemáticas de bachillerato y de primaria. Finalmente, se comparan las listas de términos validados y las listas de términos creados por los expertos; cabe aclarar que los términos validados con coeficientes de dominio mayores a cero se consideran términos y en teoría deberían ser encontrados en las listas de términos de los expertos. Con base en la información obtenida se calculan los valores de precisión y cobertura con base en las Ecuaciones (17) y (18), las cuales fueron mostradas en la sección 1.2.3.

Esta evaluación se llevó a cabo para verificar la precisión y cobertura de las listas de términos validados por Wikipedia para n-gramas de manera conjunta, es decir, los tres tipos

³⁶ El logaritmo de la verosimilitud consiste en comparar dos corpus o documentos, el primero de ellos es de donde se extraerán los términos y el otro de referencia. Para obtenerlo se emplean las frecuencias relativas de cada palabra en ambos corpus y el número total de palabras en ellos (Cabrera-Diego et al., 2011) .

de n-gramas unidos, y para cada tipo de ellos (unigramas, bigramas y trigramas). Asimismo, la evaluación se llevó a cabo para conocer cómo cambiaban los resultados con base en los diferentes umbrales de pesos de TF-IDF designados (sección 4.2).

En la Figura 21 se muestran las gráficas de cobertura contra precisión de las materias analizadas para la unión de los tres n-gramas con pesos de TF-IDF mayores a cero; en ellas se muestran tres líneas que representan las diferentes formas de calcular el coeficientes de dominio (sección 3.4.2). En los Anexos B, C y D se pueden encontrar todas las gráficas de cobertura versus precisión que se generaron para la evaluación. En estas gráficas, el eje de las abscisas representa la cobertura, es decir, el porcentaje de términos que se encuentran tanto en la lista de términos validados como en la lista de términos de los expertos; en cambio, el eje de las ordenadas representa la precisión, en otras palabras, el porcentaje de términos validados que son realmente términos según la lista de términos de los expertos.



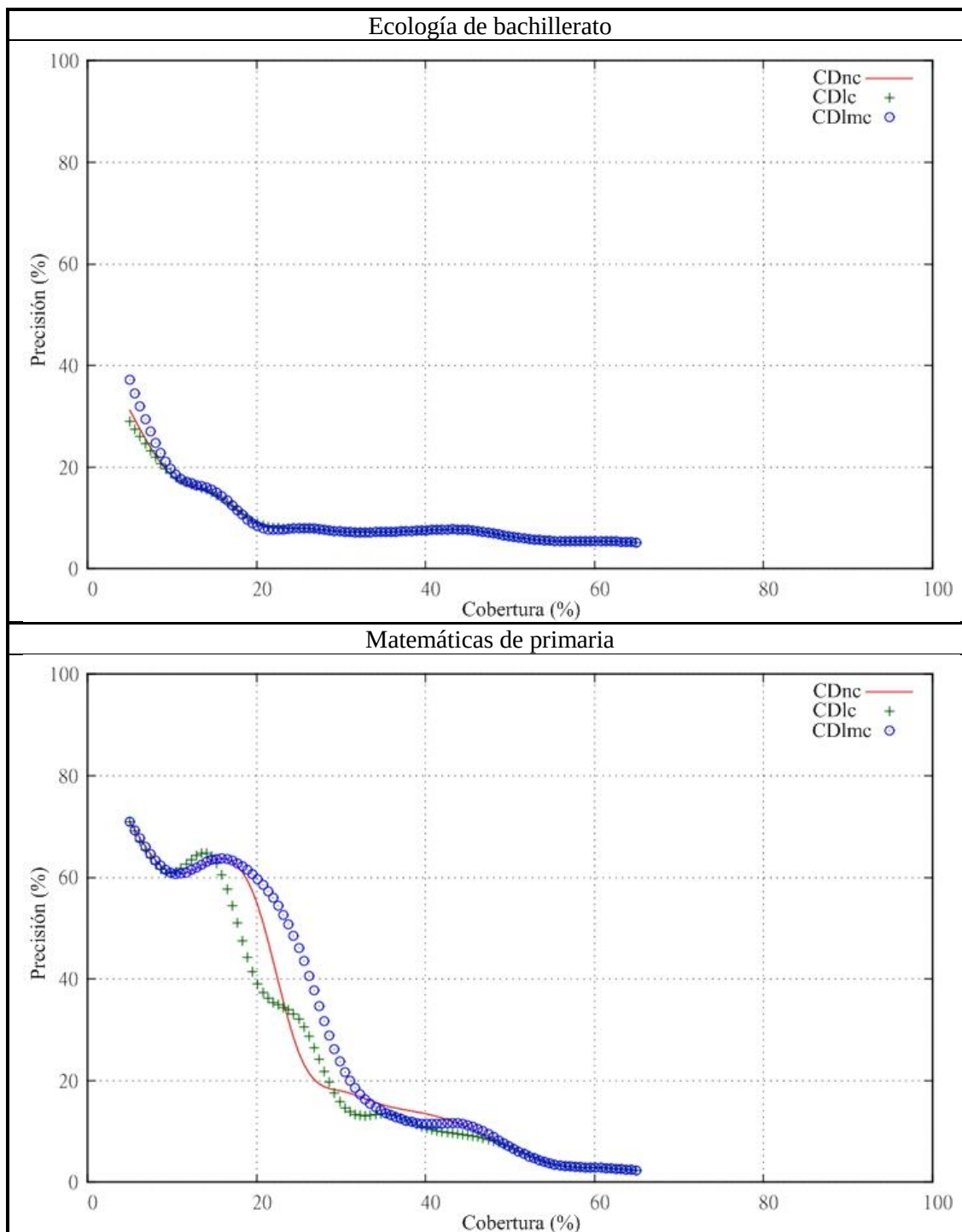


Figura 21. Gráficas de precisión y cobertura para todos los n-gramas generados para pesos de TF-IDF mayores a cero

Como se puede observar en la Figura 21, hay tres gráficas de cobertura versus precisión. La primera de ellas corresponde a matemáticas de bachillerato, donde se puede ver que los coeficientes de dominio CDnc y CDlmc empleados para el cálculo del coeficiente de dominio tuvieron mejores resultados que el coeficiente de dominio CDlc entre un 10% y 30%

de cobertura; para los tres coeficientes de dominio el máximo de precisión fue 37.14% para una cobertura del 5%, en cambio la cobertura máxima fue de 40% con una precisión de 1.05%. La segunda gráfica corresponde a ecología de bachillerato, en donde los tres métricas para el cálculo del coeficiente de dominio tuvieron un comportamiento similar; CDlc tuvo una precisión máxima de 29.02%, CDlmc de 37.19% y CDnc de 31.22% para una cobertura del 5%, mientras que los tres coeficientes tuvieron una cobertura máxima del 65% con una precisión de 5.21%. La tercera gráfica es de la evaluación de matemáticas de primaria, en la cual se puede observar que los tres cálculos del coeficiente de dominio tuvieron un comportamiento distinto entre un 10% y 30% de cobertura, siendo la métrica CDlmc el de mejor desempeño; de manera más específica las tres medidas obtuvieron una precisión máxima del 70.97% para un 5% de cobertura, mientras que se obtuvo una máximo de cobertura en 65% con una precisión del 2.33%. En todos los casos la métrica con peor rendimiento en general es la basada en la longitud de caminos (CDlc).

Finalmente, observando la Figura 21 se puede decir que la validación con Wikipedia con peor rendimiento fue la de matemáticas de bachillerato y la mejor la de matemáticas de primaria, al menos para la unión de unigramas, bigramas y trigramas y pesos de TF-IDF mayores a cero. Además, todas las gráficas muestran en un mayor o menor grado una tendencia similar a la gráfica típica de precisión y cobertura, que fue mostrada en la sección 1.2.3.

A pesar de lo anterior, es necesario analizar las razones de los resultados obtenidos y por ello en la siguiente sección se presentará una serie de observaciones sobre la evaluación realizada en esta tesis.

4.4.1 Observaciones de la evaluación de resultados

A partir de los resultados mostrados en la sección anterior, se establecieron ciertos comentarios, los cuales serán expuestos a continuación. No sólo se abordarán las gráficas de cobertura versus precisión, sino también las listas de validación y las listas de términos que se emplearon en la evaluación.

Como se pudo observar en la Figura 21, el mejor resultado de precisión se obtuvo en matemáticas de educación primaria (70.97%). En cambio, en cobertura los mejores

resultados fueron para ecología de bachillerato y matemáticas de primaria, ambos con una cobertura del 65%.

Sin embargo, debido a los resultados no tan buenos en el caso de matemáticas de bachillerato, especialmente, se decidió analizar la lista validada de términos que se empleó en la evaluación de los resultados. En este análisis se pudo observar que faltan términos en la lista validada por expertos, por ejemplo: “campana de Gauss”, “ecuación punto-pendiente”, “matriz triangular superior”, “antidiferencial”, entre otros. Esto ciertamente afecta a la precisión, debido a que la aparición de términos en la lista validada por Wikipedia pero no en la lista de evaluación (creada por los expertos) es como si se hubieran aceptado muchos no términos que lo eran en realidad. Asimismo se observó que la lista de los expertos tenía plurales que al ser lematizados en ciertas ocasiones no se volvían singulares por problemas de FreeLing, creando términos que no se encontraban de ninguna manera en las listas de candidatos de término o en las validadas por Wikipedia, provocando una falta de cobertura. Ejemplo de este caso es “ángulos interno”, el cual aparece en la lista de expertos pero no en la de candidatos a término; el más cercano, y correcto, es el de “ángulo interno”. También en la lista validada por expertos no se tenían en todos los casos los sinónimos de un término, aun cuando estos estaban en los documentos de matemáticas de bachillerato; por ejemplo, en las listas de los expertos se encontraba ecuación punto-pendiente y lado recto, pero no sus sinónimos fórmula punto-pendiente y ancho focal.

Por los problemas vistos en la lista de términos de matemáticas de expertos para bachillerato, se decidió hacer una validación de las listas obtenidas por el extractor terminológico descrito en esta tesis; para llevar esto a cabo no se utilizó de ninguna forma la información provista por la lista de términos validados por Wikipedia o de la lista de pesos de TF-IDF, esto para no sesgar la lista de validación. En la Figura 22 se muestra la gráfica de cobertura versus precisión empleando la segunda lista validada de matemáticas de bachillerato y la lista de candidatos con pesos de TF-IDF mayores a cero para la unión de los tres n-gramas empleados en esta tesis. En el Anexo E se pueden encontrar las demás gráficas generadas para pesos de TF-IDF mayores a cero.

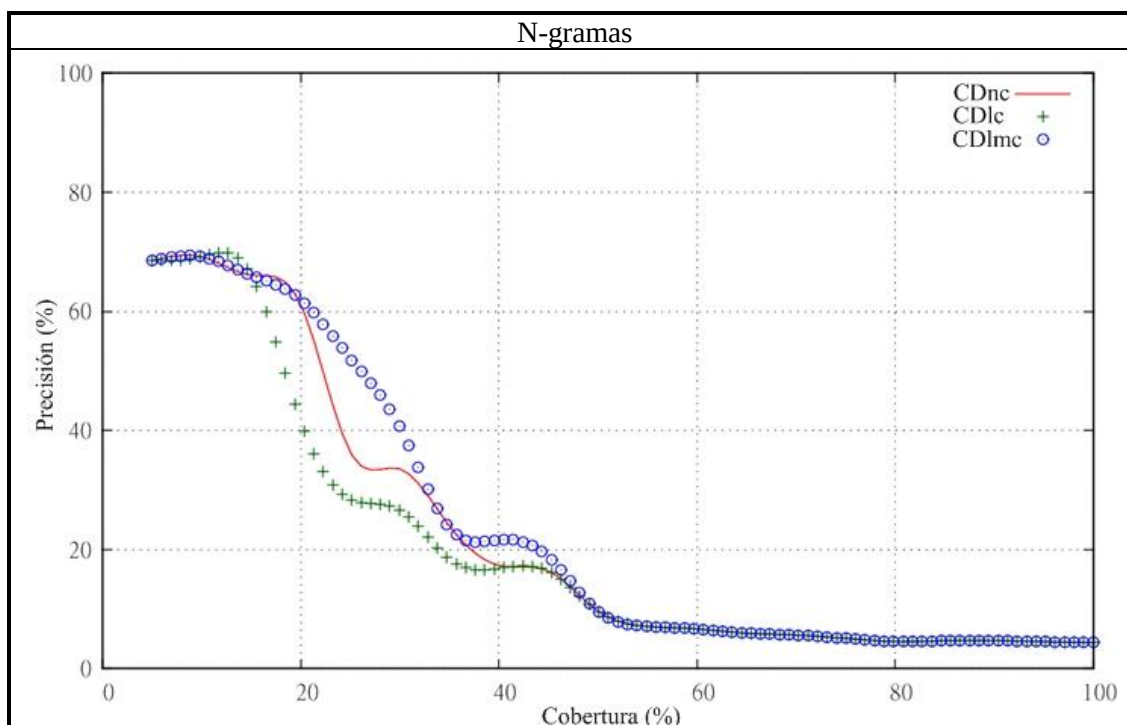


Figura 22. Gráfica de cobertura versus precisión de matemáticas de bachillerato usando la lista validada de candidatos a término obtenida del extractor terminológico desarrollado en esta tesis para n-gramas con pesos de TF-IDF mayores a cero.

Aunque, los resultados mostrados en la Figura 22 no se consideran como resultados, debido a que no se puede ser juez y parte, ciertamente la falta de términos y los errores en la lista validada por expertos afectó seriamente los resultados obtenidos en matemáticas de bachillerato. Asimismo se puede observar que al igual que en la gráfica de matemáticas de bachillerato de la Figura 21, la métrica con peor rendimiento es la que se basa en la longitud de caminos (CDlc).

Con respecto a la lista de términos de Ecología validada por un experto, existen algunos errores que son intrínsecos al trabajo manual que se realizó y algunas variaciones entre lo que se considera término ciertas veces. Sin embargo, no se observaron casos que pudieran afectar el cálculo de la precisión y cobertura en gran medida.

Sobre la lista de referencia de términos de matemáticas de primaria, la cual fue realizada por un terminólogo con conocimientos de matemáticas, sufrió de igual manera algunos problemas al llevarse a cabo la lematización; por ejemplo, la palabra “muestra” pasó a “mostrar”. Sin embargo, la lista de términos otorgada por el experto tiene algunos casos que desde mi punto de vista no son términos, por ejemplo, “al azar”, “altura de un triángulo”

o “descomposición de un número en productos de dos o más factores”. De igual forma, contiene términos que, por su estructura, no pueden ser detectados por el extractor terminológico o son eliminados en el proceso anterior al cálculo del IDF (sección 3.3.2), como los signos “<” y “%”, o como “mayor que” y “a escala”.

Además de analizar las listas que se emplearon para evaluar el sistema de extracción terminológica y su validación usando Wikipedia, se procedió a revisar las listas validadas por la enciclopedia. En esta revisión se encontraron diversos problemas que son propios de la estructura de Wikipedia; por ejemplo, se observó que había términos que se indicaban que eran de matemáticas cuando claramente no lo eran, como “mandarina” o “fiesta de disfraz” para matemáticas de primaria. Por medio de la generación de sus grafos, como el mostrado en la sección 3.4.2, se observó que la unión de ambos términos a la categoría de “matemáticas” era la categoría de “dimensión” que estaba conectada a su vez con la categoría “tiempo”. En la Figura 23 se muestra una sección de los grafos generados para “fiesta de disfraz” y “mandarina”, donde se puede apreciar la conexión de estos a la categoría tiempo. También en algunos casos las páginas de desambiguación de Wikipedia ayudaban, como en “razón” y “dividendo” para matemáticas o “residuo” para ecología, en algunos otros no lo hacía. Por ejemplo, en el caso de matemáticas la palabra “nulo” se desambiguaba como “voto nulo”, “rango” como “Rango (película)”; mientras que en el caso de ecología “reproducción” se desambiguaba como “reproducción (economía)”.

De igual manera, dentro de las listas de términos validados por Wikipedia, existen términos validados que, a pesar de formar parte de la materia y tener un coeficiente de dominio mayor a cero, en realidad fueron encontrados como otros términos, pero que de algún modo están relacionados con la materia. Por ejemplo, en matemáticas “ft” y “arreglo” son marcados como términos³⁷, los cuales sí los son en teoría, pero en realidad estos fueron encontrados como “arreglo musical” y “featuring” que de alguna manera, posiblemente por casos similares a los mostrados en la Figura 23, están relacionados con matemáticas.

³⁷ Ambos términos tienen un coeficiente de dominio igual a 0.75 y son de la lista de matemáticas de primaria para $TF-IDF > 0$ usando la métrica CDlc.

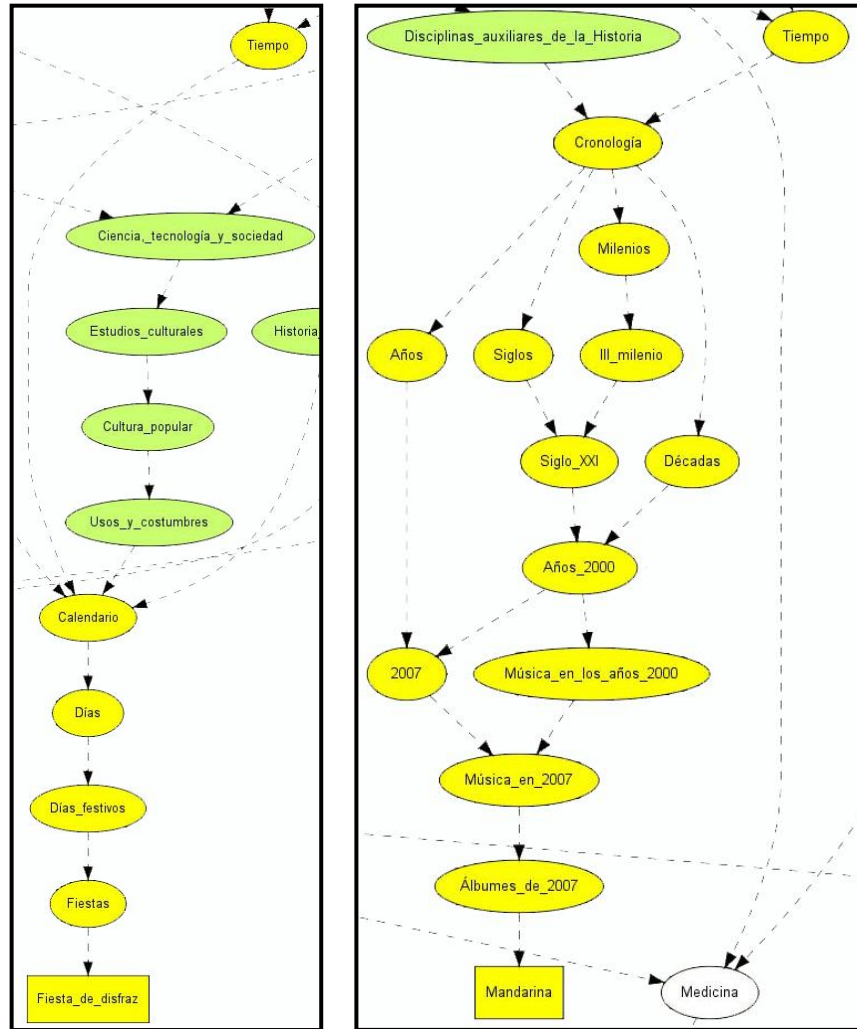


Figura 23. Secciones de los grafos de “Fiesta de cumpleaños” (izquierda) y “Mandarina” (derecha)

También como se dio a conocer en la sección 4.3 existen casos en el cual las páginas de artículos de Wikipedia no están asignadas a ninguna categoría, ya sean por errores u omisiones al momento de crear el artículo. Esto afecta la cobertura, pues hay términos que pertenecen realmente a la frontera de dominio pero que al no estar asignados a ninguna categoría no pueden tomarse como términos al momento de realizar la validación con Wikipedia, pues se les asigna un coeficiente de dominio igual a cero.