Procesamiento del Lenguaje Natural
ISSN: 1135-5948
secretaria.sepln@ujaen.es
Sociedad Española para el Procesamiento del Lenguaje Natural
España

Montejo Ráez, Arturo; Perea Ortega, José Manuel; Martín Valdivia, María Teresa; Ureña López, L. Alfonso

Uso de la detección de bigramas para categorización de texto en un dominio científico

Procesamiento del Lenguaje Natural, núm. 44, marzo, 2010, pp. 91-98
Sociedad Española para el Procesamiento del Lenguaje Natural
Jaén, España

Disponible en: http://www.redalyc.org/articulo.oa?id=515751744017

Cómo citar el artículo
Número completo
Más información del artículo
Página de la revista en redalyc.org

Sistema de Información Científica
Red de Revistas Científicas de América Latina, el Caribe, España y Portugal
Proyecto académico sin fines de lucro, desarrollado bajo la iniciativa de acceso abierto

Procesamiento del Lenguaje Natural, Revista nº 44, marzo de 2010, pp 91-98 recibido 20-01-10 revisado 18-02-10 aceptado 04-03-10

Uso de la detección de bigramas para categorización de texto en un dominio científico*

Using bigrams detection for text categorization in scientific domain

Arturo Montejo Ráez **José Manuel Perea Ortega**
María Teresa Martín Valdivia **L. Alfonso Ureña López**

Departamento de Informática, Escuela Politécnica Superior
Universidad de Jaén, E-23071 - Jaén
{amontejo,jmperea,maite,laurena}@ujaen.es

Resumen: En este artículo se presentan una serie de experimentos aplicando la técnica de detección de *multi-palabras* para categorización de texto en un dominio científico. Para ello, se ha utilizado parte de la colección de artículos científicos de Física de Altas Energías (HEP) proporcionada por el Laboratorio Europeo de Física de Partículas (CERN). Los algoritmos de aprendizaje supervisado empleados para la experimentación han sido Rocchio y PLAUM. La técnica de detección de *multi-palabras* utilizada se ha limitado a secuencias fijas de dos términos como máximo, es decir, lo que se conoce como bigramas. El objetivo de este trabajo ha sido comprobar si el uso de bigramas frecuentes como términos característicos puede ser una mejora para la tarea de categorización de textos en este dominio específico, llegando a la conclusión de que la detección de *multi-palabras* no merece la pena ser usada para esta tarea en el dominio HEP.
Palabras clave: Bigramas, Categorización de Texto, Multi-palabras, colección HEP

Abstract: This paper presents some experiments using the technique of multi-words detection for text categorization in scientific domain. We have used part of the collection of scientific papers of High Energy Physics (HEP) provided by the European Laboratory for Particle Physics (CERN). The supervised machine learning algorithms employed have been Rocchio and PLAUM. The technique of multi-words detection used has been limited to fixed sequences of maximum two terms, known as bigrams. The aim of this study is to determine whether the use of frequent bigrams as unique features may be an improvement for text categorization task in this specific domain. Our conclusion is that multi-words detection should not be used for this task in the HEP domain.
Keywords: Bigrams, Text Categorization, Multi-words, HEP collection

1. Introducción

La asignación automática de palabras clave a documentos de texto se ha enmarcado dentro de la creciente área de la categorización de texto (Text Categorization, TC), un área donde las técnicas de recuperación de información y los algoritmos de aprendizaje automático están relacionados, ofreciendo soluciones a problemas que utilizan colecciones de documentos del mundo real (Sebastiani, 2002).

La categorización de texto es la disciplina encargada de la clasificación automática de documentos de texto bajo categorías o clases predefinidas (puede verse una introducción en (Buenaga, Gómez, y Díaz, 1997)). La tarea de la categorización de texto se enmarca dentro del problema de la clasificación automática (Automatic Classification), también cono-

* Esta investigación ha sido parcialmente financiada por el Gobierno Español, proyecto TEXT-COOL 2.0 (TIN2009-13391-C04-02), por la Junta de Andalucía, proyecto GeOasis (P08-TIC-41999) y por la Universidad de Jaén, proyecto RRFC/PP2008/UJA-08-16-14

 © 2010 Sociedad Española para el Procesamiento del Lenguaje Natural

cido como reconocimiento de patrones (Pattern Recognition). Es una tarea que se engloba también dentro del aprendizaje automático. Se consideran principalmente dos tipos de clasificación automática: supervisada y no supervisada. Cuando se opta por aplicar una clasificación no supervisada, no existe posibilidad de conocer las clases o categorías finales a priori. Precisamente, uno de los objetivos de las técnicas de aprendizaje no supervisado como *clustering* de documentos es descubrir las posibles categorías o clases finales en un problema de clasificación. Sin embargo, si estas categorías o clases ya se encuentran predefinidas, es recomendable utilizar técnicas supervisadas para lograr una aproximación a la solución del problema. El uso de aprendizaje supervisado también es ideal cuando se dispone de una colección de entrenamiento. Esto último es lo que comúnmente se conoce como categorización de texto, considerando el *clustering* de documentos como una disciplina diferente.

Normalmente, el significado de una palabra no tiene sentido sin las palabras adyacentes que le acompañan en cualquier texto. En algunos casos, un concepto queda mejor representado mediante la combinación de las palabras que rodean al término principal, es decir, utilizando lo que se conoce como *multi-palabras* o *n-gramas*. Dentro de la categorización de texto, la identificación de características se puede mejorar detectando estas *multi-palabras* y formando un único término a partir de ellas. Las *multi-palabras* pueden estar formadas por dos o más términos.

En este trabajo vamos a analizar la conveniencia que la detección de *multi-palabras* puede tener en la categorización de documentos en un dominio científico específico. El artículo se organiza de la siguiente manera: en primer lugar, se explica la identificación de características mediante el reconocimiento de *multi-palabras*. Seguidamente, se describe la colección de documentos utilizada para los experimentos de este trabajo. A continuación, se muestran los experimentos y resultados obtenidos, haciendo un análisis comparativo. Finalmente, se comentan las conclusiones.

2. Identificación de características mediante el reconocimiento de multi-palabras

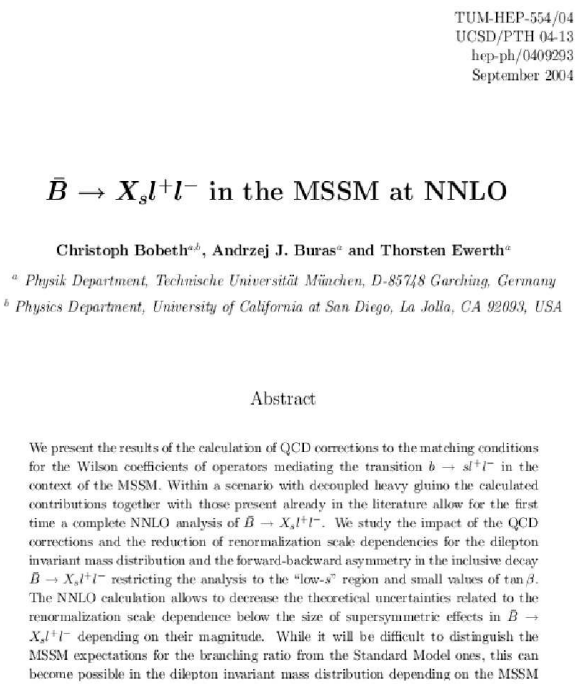
El reconocimiento de *multi-palabras* o la detección de *n-gramas* es una técnica que tiene en cuenta la posición entre ciertas palabras del texto, de forma que una secuencia fija de éstas puede representar un concepto único (Church y Hanks, 1990; Kilgarriff y Tugwell, 2001). Por tanto, se podría utilizar está técnica como método de identificación de características (términos o conceptos) en una colección empleada para categorización de texto. Por ejemplo, los pares "*top quark*" o "*Higgs boson*" se refieren a nombres de partícula (hacen referencia al mismo concepto pero están formados por más de un término). En algunos trabajos se han alcanzado buenos resultados utilizando la información proporcionada por el *Part Of Speech* (POS) para seleccionar las *multi-palabras* candidatas (Cavnar y Trenkle, 1994; Peng y Schuurmans, 2003). Un ejemplo de estas técnicas hace uso de ciertos patrones POS (un verbo, un nombre más un adjetivo, etc.) junto con un índice de co-ocurrencia (*Información Mutua*), que debe ser superior a un umbral preestablecido.

Bigrama	POS	Decisión
offer-some	VB-DT	No válido
potential-advantages	JJ-NNS	Válido
in-the	IN-DT	No válido
nuclear-power	NP-NP	Válido
be-placed	VB-VBN	No válido
power-station	NP-NP	Válido

Tabla 1: Algunos bigramas candidatos con sus etiquetas POS y la decisión tomada

Para la experimentación llevada a cabo en este trabajo hemos considerado únicamente *multi-palabras* de dos componentes, es decir, bigramas, utilizando dos tipos de patrones basados en el POS: *nombre-nombre* y *adjetivo-nombre*. Estas dos posibles combinaciones de términos son detectadas después de aplicar un etiquetado POS a la colección, marcando las posibles *multi-palabras* candidatas. La razón para explorar sólo los bigramas es debido a que los modelos de más de dos palabras son más complejos y menos frecuentes. Un resultado satisfactorio, no obstante, animaría a explorar cadenas más largas. La Tabla 1 muestra algunos ejemplos de posibles bigramas junto con la decisión tomada (válido o no válido), dependiendo del patrón POS identificado.

También ha sido necesario el cálculo de la *Información Mutua* (IM) (MacKay, 2003) entre cada par de términos de toda la colección de documentos. La *Información Mutua* entre



TUM-HEP-554/04
UCSD/PTH 04-13
hep-ph/0409293
September 2004

$\bar{B} \to X_s l^+ l^-$ in the MSSM at NNLO

Christoph Bobeth[a,b], Andrzej J. Buras[a] and Thorsten Ewerth[a]

[a] Physik Department, Technische Universität München, D-85748 Garching, Germany

[b] Physics Department, University of California at San Diego, La Jolla, CA 92093, USA

Abstract

We present the results of the calculation of QCD corrections to the matching conditions for the Wilson coefficients of operators mediating the transition $b \to s l^+ l^-$ in the context of the MSSM. Within a scenario with decoupled heavy gluino the calculated contributions together with those present already in the literature allow for the first time a complete NNLO analysis of $\bar{B} \to X_s l^+ l^-$. We study the impact of the QCD corrections and the reduction of renormalization scale dependencies for the dilepton invariant mass distribution and the forward-backward asymmetry in the inclusive decay $\bar{B} \to X_s l^+ l^-$ restricting the analysis to the "low-s" region and small values of $\tan\beta$. The NNLO calculation allows to decrease the theoretical uncertainties related to the renormalization scale dependence below the size of supersymmetric effects in $\bar{B} \to X_s l^+ l^-$ depending on their magnitude. While it will be difficult to distinguish the MSSM expectations for the branching ratio from the Standard Model ones, this can become possible in the dilepton invariant mass distribution depending on the MSSM

Figura 1: Ejemplo de la primera página de un documento HEP

dos términos x e y se puede calcular mediante la siguiente fórmula:

$$IM(x,y) = \log_2 \frac{p(x,y)}{p(x)p(y)} \qquad (1)$$

donde $p(x,y)$ es la probabilidad de que los términos x e y aparezcan juntos. Esto puede ser computado como la frecuencia de ambos términos seguidos en toda la colección dividido por el número total de posibles pares (bigramas):

$$p(x,y) = \frac{freq(x,y)}{N_{bigramas}}$$

$p(x)$ y $p(y)$ son las probabilidades de los términos x e y respectivamente. La probabilidad de un término será la frecuencia de ese término dividido por el número total de términos de la colección:

$$p(x) = \frac{freq(x)}{N_{palabras}} \qquad p(y) = \frac{freq(y)}{N_{palabras}}$$

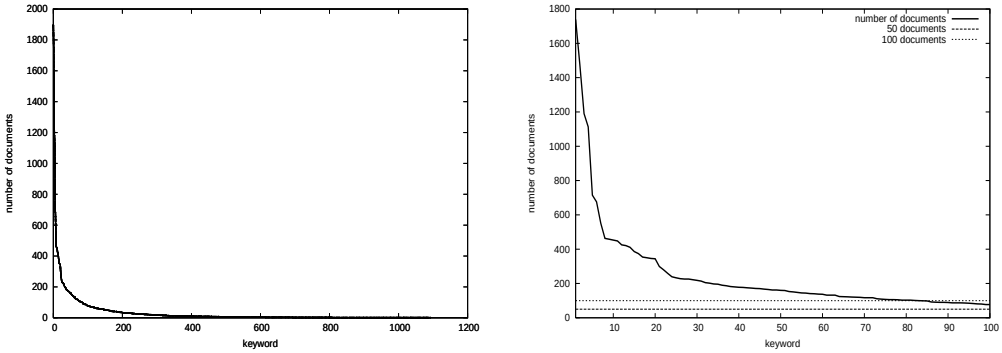
Cada bigrama se ha clasificado en dos listas: una ordenada por el valor de *Información Mutua* y otra por el valor de la frecuencia de aparición. El objetivo de este trabajo es comprobar si el uso de bigramas frecuentes como características únicas puede ser una mejora para el sistema de categorización de textos en el dominio científico.

3. La colección HEP

Para este trabajo se ha utilizado la colección de artículos de Física de Alta Energía[1] (High Energy Physics, HEP) (Vassilevskaya, 2002). Esta colección trata de recoger la mayoría de artículos o documentos publicados y prepublicados en revistas, conferencias, comités científicos, etc., que tienen que ver con la física de partículas y tecnologías relacionadas. La colección contiene alrededor de 400.000 documentos, de los cuales el 50 % son accesibles electrónicamente. Los documentos HEP son artículos técnicos que normalmente contienen abundantes gráficas, diagramas, fórmulas y otras notaciones científicas. Su vocabulario debería tener, en nuestra opinión, un bajo grado de ambigüedad, debido a la especialización de los textos en cada área (principalmente Astrofísica, Física Teórica y Física Experimental).

La colección usada en el presente trabajo consiste en 22.903 documentos etiquetados (6,1 gigabytes), convertidos a texto plano

[1] Estos documentos están libremente disponibles en el Servidor de Documentos del CERN (CDS), en la dirección http://cds.cern.ch



(a) Todas las clases (b) 100 más frecuentes

Figura 2: Distribución de clases por documentos en la partición HEP-EX

usando la herramienta *pdftotext* incluida en el paquete *xpdf*[2] (versión 2.01). Los ficheros que contienen las palabras clave (etiquetas) para ese conjunto de artículos representa 112 megabytes de ficheros. En la Figura 1 se muestra un ejemplo de la primera página de un artículo HEP típico.

3.1. La partición HEP-EX

Es importante destacar que, debido al tamaño de la colección descrita anteriormente, el tiempo necesario para realizar los diferentes cálculos podría ser prohibitivo. Es por ello que hemos considerado la posibilidad de trabajar con una partición generada a partir de la colección principal. A esta partición la hemos denominado HEP-EX, y ha sido obtenida manteniendo únicamente algunos de los artículos experimentales de la colección principal HEP. Los artículos completos de esta partición producen más de 300.000 características después de aplicar el proceso de extracción de raíces (*stemming*) y de eliminar las palabras vacías. Por tanto, esta alta dimensionalidad hace necesario aplicar un método que decremente este número de características si nuestra intención es utilizar algoritmos de aprendizaje supervisado en un tiempo de computación razonable.

La partición HEP-EX está formada por 2.839 documentos relacionados con la Física experimental de alta energía, que están indexados con 1.093 palabras clave (las categorías). Las Figuras 2a y 2b muestran la distribución de las palabras clave en la partición HEP-EX. Como se puede observar, esta partición está muy desbalanceada: únicamente 84 categorías están representadas por más de 100 documentos y sólo cinco clases por más de 1.000 documentos.

Otra posibilidad para acelerar el tiempo de experimentación es considerar, en lugar del texto completo de los artículos, solamente el *abstract* de los mismos. Cualquier documento HEP incluye una sección (*abstract*) donde los autores resumen el contenido del texto del artículo y su tamaño varía desde unas pocas líneas hasta una página entera. En este trabajo, los experimentos se han llevado a cabo utilizando tanto el texto completo de los artículos, como el *abstract* de forma aislada.

4. Experimentos y resultados

Los algoritmos de aprendizaje supervisado utilizados han sido el algoritmo de clasificación de Rocchio y el algoritmo de PLAUM (Li et al., 2002). A partir de todos los posibles pares bien formados que aparecen en la partición HEP-EX, se han generado dos listas de bigramas clasificados de mayor a menor valor de *Información Mutua* y de mayor a menor frecuencia de aparición. Por otro lado, es necesario determinar cuáles de esos bigramas serán detectados como *multipalabras* en el corpus, considerando únicamente los primeros clasificados (*top-ranked*) como bigramas relevantes. Durante la experimentación, hemos probado diferentes valores de *top-ranked*: 50, 100, 500 y 1.000. Por tanto, se han ejecutado un total de 16 experimentos como combinación de los dos clasificadores base utilizados (Rocchio y PLAUM), los cuatro valores de *top-ranked* seleccionados y las dos listas de bigramas candidatos (por frecuencia y por valor de *Información Mutua*).

[2] El paquete *xpdf* está disponible en http://www.foolabs.com/xpdf

Precisión	Recall	F1	BEP	Accuracy	% clases	*n*-top	Valor	Algoritmo
0.6961	0.4122	0.4910	0.5541	0.9819	53.28	0	ninguno	PLAUM
0.6921	0.4167	0.4931	0.5544	0.9820	52.58	50	IM	PLAUM
0.6925	0.4111	0.4880	0.5518	0.9818	52.50	100	IM	PLAUM
0.6860	0.4121	0.4883	0.5490	0.9817	52.15	500	IM	PLAUM
0.6890	0.4082	0.4862	0.5486	0.9818	51.50	1000	IM	PLAUM
0.6850	0.4101	0.4869	0.5475	0.9818	52.89	50	Frec.	PLAUM
0.6911	0.4173	0.4937	0.5542	0.9819	51.94	100	Frec.	PLAUM
0.6935	0.4028	0.4821	0.5481	0.9819	51.16	500	Frec.	PLAUM
0.6923	0.3952	0.4762	0.5437	0.9818	52.41	1000	Frec.	PLAUM
0.4668	0.5410	0.4570	0.5039	0.9720	88.15	50	IM	Rocchio
0.4683	0.5420	0.4581	0.5051	0.9719	87.30	100	IM	Rocchio
0.4656	0.5367	0.4545	0.5012	0.9717	87.53	500	IM	Rocchio
0.4647	0.5390	0.4543	0.5018	0.9716	87.60	1000	IM	Rocchio
0.4764	0.5454	0.4648	0.5109	0.9727	88.19	50	Frec.	Rocchio
0.4767	0.5487	0.4683	0.5127	0.9727	88.12	100	Frec.	Rocchio
0.4721	0.5484	0.4678	0.5103	0.9728	87.84	500	Frec.	Rocchio
0.4813	0.5469	0.4706	0.5141	0.9730	89.12	1000	Frec.	Rocchio

Tabla 2: Resultados de los experimentos usando los *abstracts* como corpus

Precisión	Recall	F1	BEP	Accuracy	% clases	*n*-top	Valor	Algoritmo
0.6961	0.4122	0.4910	0.5541	0.9819	53.28	0	ninguno	PLAUM
0.7114	0.4387	0.5147	0.5751	0.9828	58.90	50	IM	PLAUM
0.7081	0.4382	0.5140	0.5732	0.9827	59.80	100	IM	PLAUM
0.5565	0.1857	0.2633	0.3711	0.9261	31.35	500	IM	PLAUM
0.5723	0.2045	0.2820	0.3884	0.9261	34.14	1000	IM	PLAUM
0.7122	0.4395	0.5172	0.5759	0.9828	59.04	50	Frec.	PLAUM
0.7051	0.4325	0.5091	0.5688	0.9826	57.82	100	Frec.	PLAUM
0.7081	0.4365	0.5123	0.5723	0.9828	58.47	500	Frec.	PLAUM
0.7028	0.4347	0.5099	0.5687	0.9825	56.75	1000	Frec.	PLAUM
0.4242	0.5223	0.4268	0.4733	0.9699	86.92	50	IM	Rocchio
0.4216	0.5199	0.4236	0.4708	0.9693	86.92	100	IM	Rocchio
0.3257	0.5128	0.3695	0.4192	0.8750	91.04	500	IM	Rocchio
0.3464	0.5143	0.3853	0.4304	0.8783	91.78	1000	IM	Rocchio
0.4256	0.5204	0.4261	0.4730	0.9698	85.74	50	Frec.	Rocchio
0.4232	0.5188	0.4244	0.4710	0.9696	86.89	100	Frec.	Rocchio
0.4310	0.5240	0.4321	0.4775	0.9703	87.44	500	Frec.	Rocchio
0.4227	0.5235	0.4307	0.4731	0.9700	86.55	1000	Frec.	Rocchio

Tabla 3: Resultados de los experimentos usando los textos completos como corpus

Además, como se ha comentado anteriormente, haremos uso de dos corpus a partir de la partición HEP-EX: uno con los textos completos de los artículos y otro con los *abstracts* solamente. Por un lado, cuanto mayor sea el tamaño del corpus, la *Información Mutua* llega a ser más representativa, pero el objetivo de este trabajo no es establecer conclusiones acerca del uso de *abstracts* o de textos completos, sino determinar si merece la pena considerar la identificación de *multi-palabras* durante el proceso de clasificación en el dominio de la Física de altas energías.

En las Tablas 2 y 3 se muestran los resultados obtenidos utilizando el corpus de *abstracts* y el de textos completos respectivamente. Para ambas tablas, las medidas dadas son:

- Precisión (media por documento del cociente "clases acertadas" entre "clases propuestas" por el sistema)
- Cobertura (*Recall*) (media por documento del cociente "clases acertadas" entre "clases esperadas")
- F1

- Break-Even-Point (BEP), computado como la media entre precisión y *recall*
- *Accuracy*
- El porcentaje de clases entrenadas satisfactoriamente
- El número de bigramas *top-ranked* considerados
- El valor de clasificación empleado (IM para *Información Mutua* y *Frec.* para frecuencia)
- El algoritmo de clasificación utilizado

Existe una fila en cada tabla de resultados en la que no se ha considerado el uso de bigramas, es decir, no se ha llevado a cabo la detección de *multi-palabras*. Esos resultados nos permiten realizar un análisis completo sobre la técnica empleada en este trabajo para la colección HEP.

5. Análisis y conclusiones de los resultados

En las Figuras 3a y 3b se muestra una representación gráfica de la precisión, *recall* y medida F1 para el algoritmo de clasificación Rocchio utilizando el corpus de *abstracts* de la partición HEP-EX. Las Figuras 4a y 4b muestran el mismo tipo de gráficas calculadas para el otro algoritmo de aprendizaje utilizado (PLAUM), usando el mismo corpus (*abstracts*). Por último, las Figuras 5a y 5b muestran las gráficas de los resultados obtenidos empleando PLAUM como algoritmo de clasificación y usando el corpus de los textos completos.

Analizando estos gráficos, se puede obtener una conclusión simple: el uso de la técnica de detección de *multi-palabras* descrita con anterioridad, no reporta mejoras significativas para la categorización de textos en el dominio HEP. Además, los experimentos que utilizan el corpus de documentos completos obtienen peores resultados, como puede comprobarse en la Figura 5. Estos resultados son coherentes con las conclusiones obtenidas por Lewis (Lewis, 1992) en su estudio sobre la selección de características en categorización de texto. La utilización de frases y la selección de características con un alto ratio de valor de *Información Mutua* no es un argumento lo suficientemente fuerte como para justificar su uso. Por tanto, para categorización de texto en el dominio HEP, la detección de *multi-palabras* no parece recomendable.

Bibliografía

Buenaga, M., J.M. Gómez, y B. Díaz. 1997. Using WORDNET to complement training information in text categorization. En *Proceedings of Second International Conference on Recent Advances in Natural Language Processing (RANLP)*.

Cavnar, W.B. y J.M. Trenkle. 1994. N-gram-based text categorization. En *Symposium On Document Analysis and Information Retrieval*, páginas 161–175, Las Vegas.

Church, K. W. y P. Hanks. 1990. Word association norms, mutual information and lexicography. *Computational Linguistics*, 16(1):22–29.

Kilgarriff, A. y D. Tugwell. 2001. WORD SKETCH: Extraction and display of significant collocations for lexicography. En *Proc. Collocations Workshop, ACL 2001*, páginas 32–38.

Lewis, D. D. 1992. Feature Selection and Feature Extraction for Text Categorization. En *Proceedings of Speech and Natural Language Workshop*, páginas 212–217, San Mateo, California. Morgan Kaufmann.

Li, Y., H. Zaragoza, R. Herbrich, J. Shawe-Taylor, y J. Kandola. 2002. The perceptron algorithm with uneven margins. En *Proceedings of the International Conference of Machine Learning (ICML'2002)*.

MacKay, David J. C. 2003. *Information theory, inference, and learning algorithms?* Cambridge.

Peng, F. y D. Schuurmans. 2003. Combining naive bayes and n-gram language models for text classification. En Fabrizio Sebastiani, editor, *ECIR*, volumen 2633 de *Lecture Notes in Computer Science*, páginas 335–350. Springer.

Sebastiani, F. 2002. Machine learning in automated text categorization. *ACM Comput. Surv.*, 34(1):1–47.

Vassilevskaya, Lyubov A. 2002. An approach to automatic indexing of scientific publications in high energy physics for database spires-hep. Master's thesis, Fachhochsule Potsdam, Institut für Information und Dokumentation, September.

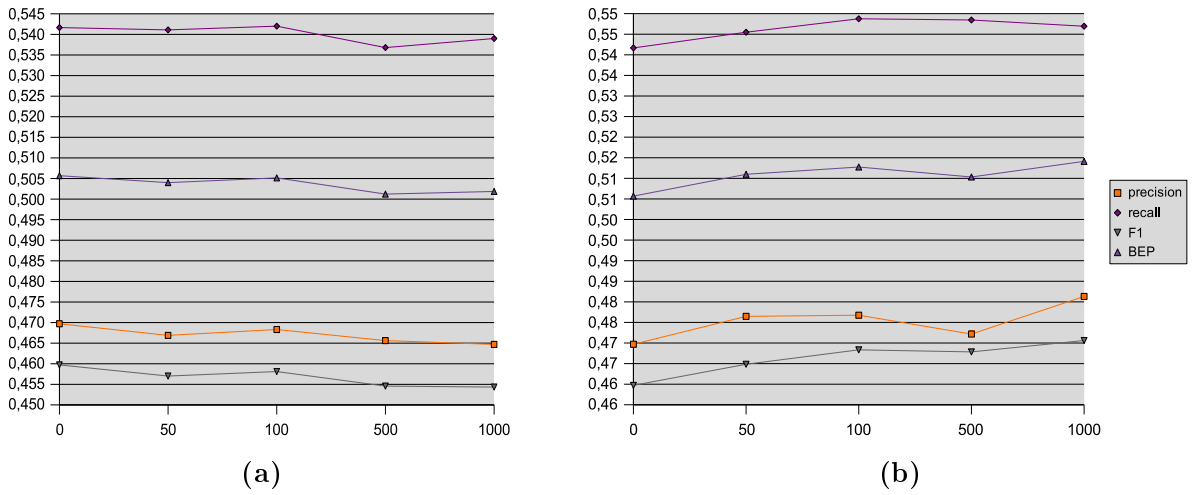


Figura 3: Influencia de la detección de *multi-palabras* sobre el corpus de *abstracts* para el algoritmo **Rocchio** usando la lista ordenada por (a) *Información Mutua*, o (b) Frecuencia

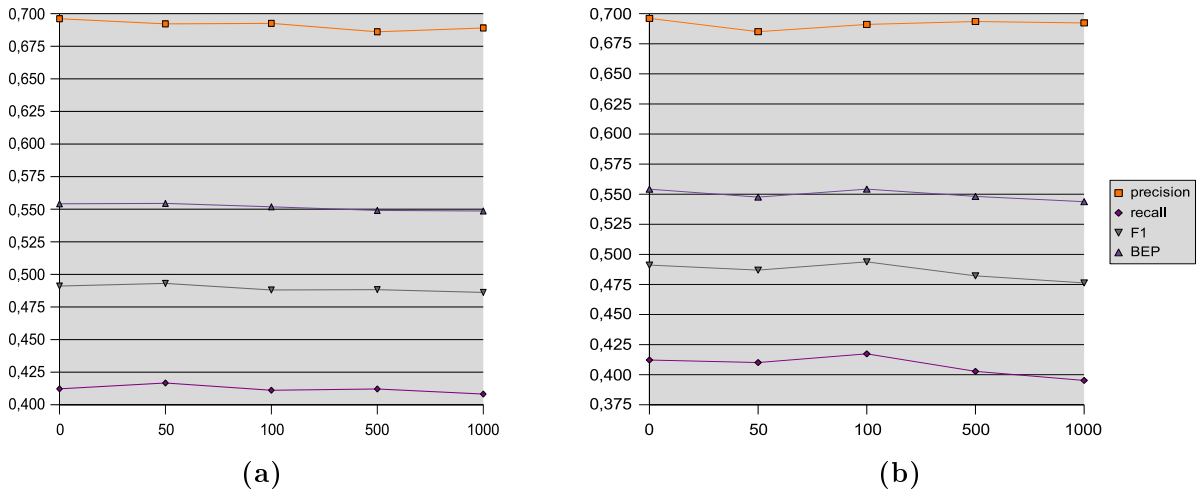


Figura 4: Influencia de la detección de *multi-palabras* sobre el corpus de *abstracts* para el algoritmo **PLAUM** usando la lista ordenada por (a) *Información Mutua*, o (b) Frecuencia

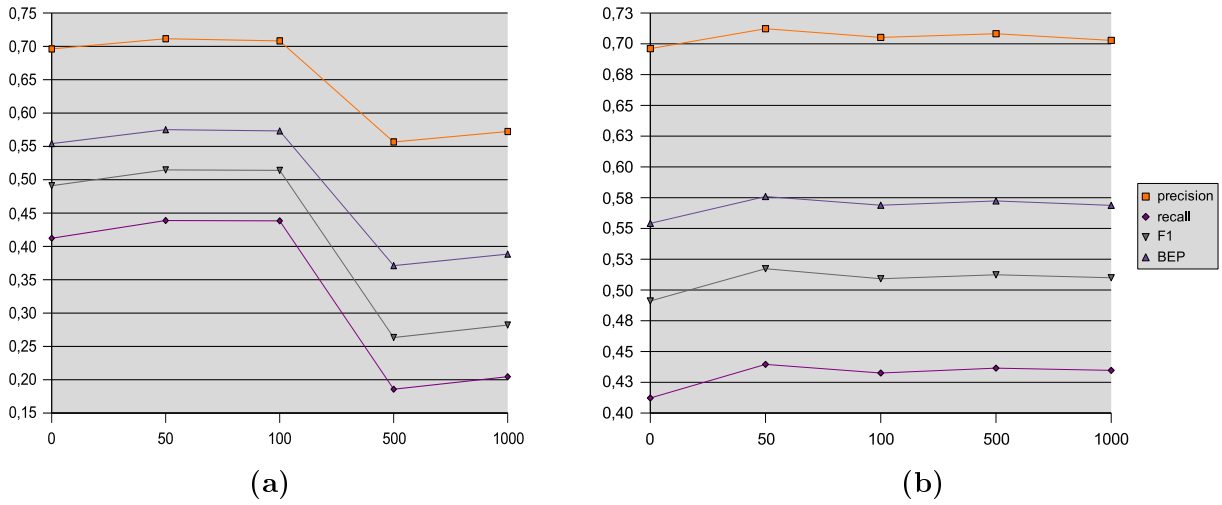


Figura 5: Influencia de la detección de *multi-palabras* sobre el corpus de textos completos para el algoritmo **PLAUM** usando la lista ordenada por (a) *Información Mutua*, o (b) Frecuencia