Universidad Federal del Amazonas

Instituto de Computación

Programa de Pos-Graduación en Informática

Minería de Términos Frasales aplicada en tareas de Recuperación de Información.

Zulema Sánchez Vera

Manaos

2019

Zulema Sánchez Vera

Minería de Términos Frasales aplicada en tareas de Recuperación de Información.

Disertación presentada para el Curso de Pos-Graduación en Informática de la Universidad Federal del Amazonas como requisito parcial para la obtención del grado de Máster en Informática.

Universidad Federal del Amazonas
Instituto de Computación
Programa de Pos-Graduación en Informática

Orientador: Prof. Edleno Silva de Moura, D.Sc

Manaos
2019

Ficha Catalográfica

Ficha catalográfica elaborada automaticamente de acordo com os dados fornecidos pelo(a) autor(a).

Sánchez Vera, Zulema
S999m Minería de Términos Frasales aplicada en tareas de Recuperación de Información. / Zulema Sánchez Vera. 2019
69 f.: il. color; 31 cm.

Orientador: Edleno Silva de Moura
Dissertação (Mestrado em Informática) - Universidade Federal do Amazonas.

1. Términos frasales. 2. Clasificación. 3. Clusterización. 4. Búsqueda ad hoc. I. Moura, Edleno Silva de II. Universidade Federal do Amazonas III. Título

FOLHA DE APROVAÇÃO

"Minería de Términos Frasales aplicada en tareas de Recuperación de Información"

ZULEMA SÁNCHEZ VERA

Dissertação de Mestrado defendida e aprovada pela banca examinadora constituída pelos Professores:

Prof. Edleno Silva de Moura - PRESIDENTE

Prof. Altigran Soares da Silva - MEMBRO INTERNO

Prof. André Luiz da Costa Carvalho - MEMBRO EXTERNO

Manaus, 29 de Abril de 2019

A ustedes, queridos padres.

Agradecimentos

- Mis más profundos agradecimientos a todos los profesores del programa, principalmente al profesor Edleno Silva de Moura, D.Sc (orientador).
- A los especiales colegas que gané durante el proceso.
- Al coordinador Eduardo L. Feitosa, sin su ayuda nada hubiera sido posible.

Resumen

El crecimiento espectacular y constante de la *web* con el consiguiente aumento en el número de documentos digitales disponibles y el uso cada día más frecuente de sistemas que lidian con información textual, han motivado constantes esfuerzos en el desarrollo de sistemas eficaces para el tratamiento de la información que realizan tareas como busca, clasificación y clusterización en bases de datos textuales. Conocida la relevancia de la representación del texto en los resultados de la recuperación de información, este trabajo investiga el impacto de la adición de términos frasales como unidades, debido a su interpretabilidad superior, con el objetivo de enriquecer la representación tradicional del modelo *BoW*. La idea es que con el uso de términos frasales el ruido y ambigüedad inherente de la representación del texto basada solo en palabras individuales sea reducida, traduciéndose en mayor calidad en los resultados obtenidos.

Para la minería de términos frasales se utilizó el método *Autophrase* que integra los enfoques de segmentación y evaluación de la calidad para la extracción de secuencias de palabras, que constituyen unidades semánticas completas, no precisa de expertos humanos, es independiente del idioma, dominio e incorpora información sintáctica en forma de etiquetas *POS* siempre que esté disponible. En la búsqueda *ad hoc* se utilizó el modelo vectorial en los conjuntos de datos: *OHSUMED, Cystic Fibrosis* y *Glasgow Herald 1995*, los experimentos realizados muestran ganancias en el orden de 34,97 % utilizando la métrica de *MAP*. Observándose que la adición de información semántica en forma de términos frasales en las consultas, favorece la identificación de los documentos relevantes.

En las tareas de clasificación y clusterización se comparó la mejora de rendimiento en términos de precisión, cuando los términos frasales mejor evaluados por las técnicas *Chi2* y *Mutual information* son adicionados para ampliar la representación de los documentos, basadas en palabras individuales en las colecciones *20 newsgroups, DBpedia ontology classification* y *AG'news corpus* respectivamente. Para esta comparación fueron empleados los clasificadores *Naive Bayes* y *Support vector machine* en la clasificación y *K-means* en la clusterización. Los resultados no mostraron ganancias significativas con la incorporación de los términos frasales. La conclusión, en este caso, es que los documentos ya de por si contienen suficiente información en forma de unigramas que aportan mayor peso que los términos frasales que aumentan la dispersión de los datos.

Palabras claves: Términos frasales, Clasificación, Clusterización, Búsqueda *ad hoc*

Abstract

The spectacular and constant growth of the web with the consequent increase in the number of digital documents available and the increasingly frequent use of systems that deal with textual information, have motivated constant efforts in the development of effective systems for the treatment of information. who perform tasks such as search, classification and clustering in textual databases. Well-known relevance of the representation of the text in the results of the retrieval of information, this research investigates the impact of the addition of frasal terms as units, due to its superior interpretability, with the aim of enriching the traditional representation of the BoW model. The idea is that with the use of phrasal terms the inherent noise and ambiguity of the representation of the text based only on individual words is reduced, resulting in higher quality in the results obtained.

For the mining of phrasal terms the method was used Autophrase that integrates the segmentation and quality evaluation approaches for the extraction of word sequences, which constitute complete semantic units, does not require human experts, is independent of the language, domain and incorporates syntactic information in the form of POS labels provided it is available. In the ad hoc search the vector model was used in the data sets: OHSUMED, Cystic Fibrosis and Glasgow Herald 1995, the experiments performed show gains in the order of 34.97 % using the MAP metric. Observing that the addition of semantic information in the form of phrasal terms in the queries, favors the identification of the relevant documents.

In the tasks of classification and clustering, performance improvement in terms of precision was compared, when the best phrasal terms evaluated by the techniques Chi2 and mutual information were added to extend the representation of the documents, based in individual words in the collections 20 newsgroups, DBpedia ontological classification and AG'news corpus respectively. For this comparison, the classifiers Naive Bayes, Support vector machines were used in classification and K-means in the clustering. The results did not show significant advances with the incorporation of the phrasal terms. The conclusion, in this case, is that the documents already contain enough information in the form of unigrams that contribute more weight than the phrasal terms that increase the dispersion of the data.

Key-words: Phrasal terms, Classifications, Clustering, Ad hoc search.

Índice de figuras

Índice de tablas

Índice general

Página

1 Introducción

Aunque desde mediados del siglo XX se viene trabajando en el área de la Recuperación de información, en los últimos diez años su relevancia ha aumentado notablemente. Entre otros posibles factores desencadenantes de este efecto se encuentra: en primer lugar, el crecimiento espectacular y constante de la web, con el consiguiente aumento en el número de documentos digitales a disposición de los usuarios de la red; en segundo lugar, el cambio producido en los hábitos de los usuarios, a raíz de la preponderancia de internet entre las diversas modalidades de acceso a la información, lo que ha traído consigo una modificación paralela en los servicios que demanda, incrementando la necesidad de una recuperación de alto rendimiento.

Son muchos los enfoques que se han experimentado para abordar el objetivo esencial de la Recuperación de información (facilitar la tarea de discernimiento de los escasos documentos relevantes que puedan existir en la red, frente a los millones de documentos irrelevantes en relación a cada consulta formulada), desde el modelo booleano hasta la aplicación de técnicas de Inteligencia artificial, entre las que podemos citar las redes neuronales, los algoritmos genéticos, el procesamiento del lenguaje natural, etc. Sin embargo, el resultado de la recuperación depende en gran medida de la representación de los documentos.

En este contexto, donde la mayor parte de los datos disponibles están almacenados en documentos escritos en lenguaje natural, la literatura sobre minería de texto describe como un problema fundamental durante el proceso de análisis: la representación del texto de manera efectiva, no solo desde la perspectiva de rendimiento del algoritmo, sino también para una mejor interpretación y análisis de los resultados. Un enfoque común es utilizar n-gramas, es decir, una secuencia continua de n unigramas como unidades básicas. Sin embargo, tal representación plantea preocupaciones de crecimiento exponencial del vocabulario cuando n crece; así como falta de interpretabilidad (LIU; SHANG; HAN, 2017). Como alternativa: la minería de términos frasales[1] utiliza solo un subconjunto compacto

[1] Término frasal: secuencia ordenada de palabras con un significado específico, que puede ser totalmente distinto del significado de las palabras que lo componen. (CARVALHO; MOURA; CALADO, 2010)

de n-gramas, que generan una representación explicable dado un documento de texto (CARVALHO; MOURA; CALADO, 2010; LIU et al., 2015; SHANG et al., 2018).

La minería de términos frasales, es una tarea fundamental para el análisis de texto en varios dominios (ciencias, noticias, media social, etc.), porque muchos conceptos, entidades y relaciones se expresan mediante secuencias de palabras que constituyen una unidad semántica propia. Ejemplo de aplicaciones incluyen: detección y seguimiento de temas (*topic detection and tracking*) (DANILEVSKY et al., 2014; LINDSEY; III; STIPICEVIC, 2012), OLAP (*Online Analitycal processing*) en colecciones de texto multidimensionales, categorización de documentos (LEWIS, 1992; EL-KISHKY et al., 2014; TUYET; HANH, 2016), busca de palabras claves (*keyword search*), descubrimiento de eventos sociales (*social event discovery*), resúmenes de documentos (*document summarization*), recuperación *ad-hoc* (ZHANG et al., 2007; CARVALHO; MOURA; CALADO, 2010), además del enriquecimiento de los modelos tradicionales de recuperación de información.

Los modelos clásicos de recuperación de información textual, representan los documentos como "bolsas de palabras"(*Bag of words model*) BoW, referido a conjuntos de términos completamente independientes (BAEZA-YATES; RIBEIRO-NETO et al., 1999), sin tener en cuenta el orden en que aparecen las palabras en el documento (cada término se considera como una información independiente), ni su ubicación, ni los términos a su alrededor. A pesar de estos modelos BoW haber demostrado ser efectivos en las tareas de recuperación de información, la transformación del texto no estructurado en unidades estructuradas (términos frasales), reduce sustancialmente la ambigüedad semántica y mejora la potencia y eficiencia en la manipulación de dichos datos.

La clasificación y clusterización de documentos de texto, herramientas poderosas para transformar repositorios no estructurados, son tareas importantes en el procesamiento del lenguaje natural, han sido ampliamente estudiadas y son componentes esenciales en muchas aplicaciones como: el filtrado de información, el análisis de sentimientos y la búsqueda web, facilitando y mejorando los resultados de la recuperación (ABDULLAH; ZAMIL, 2018). Donde no solo un clasificador o clusterizador de texto idóneo, sino también una representación de documento adecuada tiene influencia en la precisión resultante. El enfoque básico y con buenos resultados es representar documentos por palabras individuales (modelo BoW). Sin embargo, a menudo se utilizan otras características para lograr mejores resultados, donde diagramas, bigramas, n-gramas o algunos patrones diseñados se suelen

extraer como características. Además de varias técnicas de selección de atributos como: la frecuencia, Información Mutua (MI) (COVER; THOMAS, 2012), X^2, que se aplican para seleccionar las características más discriminatorias, mejorando el desempeño predictivo a través de modelos más eficientes.

Estos métodos tradicionales de representación de texto que ignoran la información contextual o el orden de las palabras, siguen siendo insatisfactorios para capturar la semántica de las palabras. Por lo que aunque existen diferentes aproximaciones y enfoques propuestos, la clasificación y clusterización de documentos continúa centrando la atención de los investigadores, dado que su eficacia aún necesita ser mejorada.

La recuperación *ad hoc*, probablemente la tarea más representativa por ser aquella en la que se basan los buscadores *web*, además del gran volumen de consultas y la calidad de las respuesta provistas a los usuarios, un factor importante es el tamaño creciente de los conjuntos de datos, donde una simple tarea puede acarrear costos computacionales grandes sino se optimiza adecuadamente. El rendimiento de un sistema de Recuperación de información puede verse afectado por muchos factores: la ambigüedad de los términos de consulta, la falta de familiaridad con las características del sistema y los factores relacionados con la representación de los documentos de la colección.

Dada la relevancia de la representación del texto en la recuperación, entre los diversos métodos estudiados para la minería de términos frasales en grandes colecciones de texto, tanto de la corriente de procesamiento lingüístico como estadístico del lenguaje natural, se implementó el método de los autores(LIU et al., 2015) perteneciente a la corriente de procesamiento estadístico, en la familia de los métodos semi supervisados. Este método integra los enfoques de segmentación y evaluación de la calidad para la extracción de secuencias de palabras, que constituyen unidades semánticas completas, requiere de un conjunto pequeño de ejemplos rotulados para entrenamiento y alcanza altos niveles de precisión, superando en calidad y eficiencia enfoques anteriores (SHANG et al., 2018), además de poseer una complejidad lineal en tiempo y espacio. Fue empleado en los experimentos iniciales, pero su grande desventaja radica en la necesidad de emplear expertos humanos para crear la base de entrenamiento, haciéndose aun más difícil en colecciones especializadas como la *OHSUMED* y *Cystic Fibrosis*, utilizadas en este trabajo.

En el articulo "*Automated Phrase Mining from Massive Text Corpora*" (SHANG

et al., 2018), extensión del anterior, los autores automatizan el proceso de minería de términos frasales, obteniendo un mejor rendimiento al utilizar técnicas que lo diferencian del método anterior como: *Robust Positive-Only Distant Training*, metodología para reducir la dependencia de expertos humanos, que constituye un impedimento para el análisis oportuno de grandes colecciones de texto en dominios emergentes, utilizando para ello el gran cúmulo de términos frasales disponibles en bases de conocimiento externas. *POS-Guided Phrasal Segmentation* que incorpora un etiquetador *POS* pre entrenado (disponible en muchos idiomas), consiguiendo mayor precisión en la localización de los limites de los n-gramas e independencia del idioma. Los autores refieren que el método tiene un alto impacto en una variedad de aplicaciones relacionadas con texto, que incluyen representación e indización de documentos, búsqueda de relevancia, clasificación, resumen y recomendaciones. El código disponibilizado por los autores, fue escogido y utilizado en los experimentos como el método de minería de términos frasales.

Por todo lo anterior, en este trabajo estamos interesados en el estudio del impacto de la minería de términos frasales, aplicados a tareas de recuperación de información de texto como: recuperación *ad hoc*, clusterización y clasificación de documentos, esperando mejorar la calidad final de los resultados obtenidos.

Los experimentos realizados se subdividieron en tres partes, de acuerdo a las tareas de recuperación (recuperación *ad-hoc*, clusterización y clasificación). En todos los casos se utilizaron términos frasales de hasta 6-gramas de longitud, con el objetivo de enriquecer la representación de los documentos utilizada en los modelos tradicionales.

Para medir el impacto del uso de los términos frasales en la tarea de recuperación *ad-hoc*, se utilizó el modelo vectorial (BAEZA-YATES; RIBEIRO-NETO et al., 1999) y los resultados obtenidos muestran mejoras en la calidad de los resultados de hasta un 34,97 % y 8.93 % cuando son utilizadas las métricas *MAP* y *P@10*. Lo que muestra un impacto positivo en la inclusión de los términos frasales, enriqueciendo la representación de los documentos utilizada por el modelo *BoW*. Estos resultados se presentan en el Capitulo 4 sección 1.

En el caso de la clasificación fueron utilizados diferentes métodos de aprendizaje automático como: *Support Vector Machine, Descenso de Gradiente estocástico, Naive Bayes, PassiveAgresive, Ramdon Forest* y *clasificador Ridge* reconocidos por tratar matrices

dispersas de forma eficiente, así como técnicas filtro de selección de características con el objetivo de seleccionar los términos frasales más discriminatorios y adicionarlos al modelo enriqueciendo la representación de los documentos. Para medir el impacto de la adición de los términos frasales se utilizaron las colecciones *20 Newsgroups* y *DBpedia ontology classification*; los experimentos en la primera colección mostraron deterioro en los resultados de la clasificación en todos los casos, mientras en la segunda se obtuvo una ganancia del orden del 1,3 %, utilizando el clasificador *SVM* con el mayor *baseline* y de 2,3 % en *Naive Bayes* con un *baseline* menor, estos resultados pueden ser consecuencia de las características de las colecciones utilizadas y propias de la clasificación de documentos de texto, como se discute en el capítulo 4 sección 2. La clusterización muy similar a la clasificación no mostró ganancias cuando fueron adicionados los términos frasales a la representación de los documentos, se utilizó la colección *AG's news corpus*, los resultados se presentan en el capítulo 4 sección 3.

Este trabajo está estructurado de la siguiente forma. En el capítulo 2 presentamos métodos de minería de texto para la extracción de términos frasales, así como su utilización en las tareas de clasificación, clusterización y busca encontrados en la literatura. En el capítulo 3 describimos el método de minería de términos frasales utilizado. En el capítulo 4 son presentados los experimentos realizados. El capítulo 5 cierra con las conclusiones.

2 Estado del arte

La minería de términos frasales ha sido estudiada por dos corrientes fundamentales: la corriente de procesamiento lingüístico del lenguaje natural *(PLN)*, se basa en la aplicación de técnicas y reglas que codifican de forma explícita el conocimiento lingüístico. Los documentos son analizados a partir de los diferentes niveles lingüísticos, por herramientas que incorporan al texto etiquetas propias de cada nivel. El enfoque más común se basa en la identificación de frases[1] nominales utilizando documentos etiquetados (*POS Tagged*)(ZHANG et al., 2007). Obviamente, estos métodos basados en reglas carecen de suficiente flexibilidad para manejar varios idiomas y colecciones heterogéneas.

Con el fin de mejorar la precisión, métodos *PLN* introducen modelos de aprendizaje supervisado o modelos estocásticos. Los métodos supervisados de fragmentación (*parsing*) de frases, toman una serie de textos etiquetados como datos de entrenamiento y aprenden las reglas de clasificación basadas en las características *POS* (*Part-Of-Speech*). Otros métodos utilizan funciones *PLN* más sofisticadas como: analizadores sintácticos de dependencias (MCDONALD; CRAMMER; PEREIRA, 2005), uso de atributos distribucionales basados en n-gramas de la *web* (BERGSMA; PITLER; LIN, 2010), modelos basados en redes neuronales (LAI et al., 2015).

No obstante, la mayoría de los métodos *PLN* necesitan entrenamiento pesado y rotulado complejo, su dependencia de diversos tipos de analizadores lingüísticos con reglas de lenguaje dependientes del dominio, dificulta la aplicación de estas técnicas en grandes colecciones emergentes (Dominios científicos, Media social, *Yelp, Twitter*, etc.), que se desvían de las reglas rigurosas del lenguaje, haciendo que estos enfoques sean difíciles de generalizar.

Con el objetivo de superar estas limitaciones, la corriente de recuperación de información que utiliza el procesamiento estadístico del lenguaje natural basado en la frecuencia, considera cada n-grama minerado como un patrón frecuente (AHONEN, 1999; CAROPRESO; MATWIN; SEBASTIANI, 2001), que se extrae si su ocurrencia es mayor

[1] Frase (Phrase): Secuencia de palabras que constituyen una unidad conceptual, no necesariamente tiene un significado semántico completo.

que un determinado limiar conocido. Estos métodos no depende de las reglas estrictas del lenguaje; están sujetos para alcanzar altos niveles de eficiencia a grandes colecciones de documentos, su simplicidad y eficacia los han convertido hoy en los modelos más utilizados en los sistemas de recuperación de información textual. Sin embargo, ranquear n-gramas según la frecuencia generará muchas secuencias de palabras que no constituyen términos frasales. Recientemente, diversos enfoques estadísticos basados en la frecuencia (EL-KISHKY et al., 2014; LIU et al., 2015; SHANG et al., 2018) para estimar la calidad y ranquear los n-gramas candidatos, integran la segmentación con la estimación de calidad, para de esta forma rectificar la calidad inexacta estimada inicialmente utilizando el contexto de ocurrencia local.

Existen tres familias de métodos que abarcan estas dos corrientes: Métodos supervisados, aquellos a los que pertenecen las técnicas de procesamiento lingüístico del lenguaje natural; métodos no supervisados, basados solamente en estadísticas extraídas de corpus y no requieren de documentos rotulados y por último, los métodos semi supervisados que emplean estadísticas y aprendizaje de máquina.

2.1. Métodos supervisados.

(ZHANG et al., 2007) propuso identificar frases nominales en consultas cortas que incluían nombres propios, frases de diccionario simples y complejas. Los nombres propios comprendían entidades con nombre de personas, ubicaciones y organizaciones, derivadas principalmente de *Wikipédia*, mientras que las frases del diccionario se extraían principalmente de *WordNet*[2]. Cualquier otro n-grama gramaticalmente válido se identificó como frases simples (para bigramas) o complejas (para n-gramas de orden superior), utilizando un analizador probabilístico (Collins Parser), con estadísticas obtenidas de *Google* para fines de verificación y selección de frases.

El método de Zhang que combina diferentes herramientas, mostró que cuando utilizado en tareas de recuperación de documentos *ad hoc*, el reconocimiento y utilización de un subconjunto de términos frasales (frases nominales) en las consultas, mejora sustancialmente la efectividad de la recuperación.

[2] WordNet: diccionario electrónico. Se utiliza para reconocer frases de diccionario y ciertos nombres propios.

A pesar de que el objetivo de este trabajo tiene aspectos similares (además del espectro más amplio del uso de términos frasales, que constituyen un superconjunto de las frases nominales), el uso de bases de datos externas y técnicas de procesamiento de lenguaje natural, hacen que el método no sea adecuado para muchos escenarios de recuperación de información, debido a su alto costo computacional y dependencia de fuentes externas, además solo explora su impacto en las tareas de recuperación *ad hoc*. Por el contrario, el método de minería utilizado, aprovecha la gran cantidad de términos frasales disponibles en bases de conocimiento públicas y solo incorpora información sintáctica en forma de etiquetas *POS* para la mejora del rendimiento, si existe un tokenizador y etiquetador para el idioma de la colección.

Con el objetivo de aprovechar el recurso lingüístico proporcionado por la *web*, aproximaciones un poco más recientes utilizan atributos distribucionales (*distributional features*) basados en n-gramas de la *web*. El trabajo de (BERGSMA; PITLER; LIN, 2010) presentó resultados en diversas tareas de investigación *PLN*: generación, desambiguación[3], análisis sintáctico y etiquetado. Creando clasificadores robustos a través de datos de n-gramas a escala *web* para el ordenamiento de adjetivos, la corrección ortográfica, el agrupamiento de sustantivos propios compuestos y desambiguación semántica verbal. Utilizando para ello una colección de n-gramas (hasta 5 gramas) y la distribución de las etiquetas *POS* para cada n-grama. Mostrando que los n-gramas funcionan consistentemente bien en diversas tareas de *PLN*. Pero como técnica perteneciente a la vertiente de procesamiento lingüístico del lenguaje natural, tiene un elevado costo de rotulaje, no es fácilmente adaptable a otros idiomas y/o géneros, ni a las nuevas y dinámicas aplicaciones emergentes.

2.2. Métodos no supervisados.

Muchos estudios de minería de términos frasales están vinculados con el modelado de temas (*Topic modeling*), existiendo tres estrategias fundamentales:

La primera estrategia combina la detección de los limites de los n-gramas con la inferencia del tema *(topics)*, en un modelo unificado (inferencia simultánea de n-gramas y tema). Ejemplos incluyen los modelos: *Topic N-Grams (TNG)* (WANG; MCCALLUM;

[3] Desambiguación semántica: El proceso de identificación del sentido activado por una palabra ambigua

WEI, 2007), que genera palabras en orden textual y crea n-gramas representativos (términos frasales), mediante la concatenación de bigramas sucesivos y *Phrase discovery LDA (PDLDA)* propuesto por (LINDSEY; III; STIPICEVIC, 2012) que amplía *TNG* incorporando la jerarquía *Pitman-Yor Processes* en el proceso de formación de n-gramas; donde simultáneamente se divide un corpus en secuencias de diferentes longitudes y les asigna temas basándose en el supuesto de que el tema de una secuencia de palabras cambia periódicamente, las palabras entre estos puntos de cambio comprenden una frase.

A pesar de que en términos de resultados cuantitativos han demostrado que mejoran muchas aplicaciones como: la clasificación de documentos, la recuperación de información, etc. son modelos computacionalmente complejos y los n-gramas detectados a menudo tienen una calidad inferior, no constituyen términos frasales.

La segunda estrategia consiste en extraer n-gramas que representan temas (*topicals phrases*) como un paso de post-procesamiento del modelado de temas basados en unigramas. Tales enfoques suponen que las palabras que se etiquetan simultáneamente con el mismo tema, muchas veces se pueden agrupar como una frase.

En líneas generales, dos métodos que pertenecen a esta estrategia son: *TurboTopic* y *KERT* propuestos por (BLEI; LAFFERTY, 2009) y (DANILEVSKY et al., 2014) respectivamente, ambos consideran la construcción de la frase como un paso posterior al procesamiento estadístico generativo *LDA (Latent Dirichlet Allocation)* (BLEI; NG; JORDAN, 2003), por lo que luego de ejecutar *LDA* (hipótesis de bolsa de palabra) en la colección, en el primer método se mezclan recursivamente los unigramas adyacentes pertenecientes al mismo tema hasta que todos los unigramas adyacentes significativos ya fueron mezclados y en el segundo luego de aplicado *LDA* se particiona cada documento en k documentos (k= No. temas); para cada tema se minera patrones frecuentes y se realiza el ranqueamiento basado en cuatro criterios que solo utilizan la frecuencia de los datos.

Sin embargo, en el modelado de temas basado en unigramas, las palabras dentro de la misma frase pueden asignarse a más de un tema. Además, como estos modelos se ejecutan típicamente en datos con *stopwords* eliminadas, los procesos posteriores a la ejecución de *LDA* tendrían dificultades para reconocer frases que contengan *stopwords*.

Por su parte, la última estrategia realiza pre-procesamiento de los documentos para la extracción de términos frasales (utilizando métodos estadísticos o lingüísticos) y luego

ejecuta el modelado de temas. (EL-KISHKY et al., 2014) propuso el método conocido como *ToPMine* que extrae n-gramas realizando minería de patrones frecuentes, ranquea utilizando los mismos criterios del método *KERT* y como último paso esos n-gramas son pasados como entrada para *PhraseLDA* (extensión de *LDA*), que restringe que todas las palabras de los n-gramas sean asignada al mismo tema. En este método como cada frase extraída se considera como un término único, el vocabulario resultante se amplía de manera significativa, lo que lleva a datos más dispersos.

Estos métodos basados en medidas estadísticas no dependen de las características lingüísticas específicas del idioma y, por lo tanto, pueden lograr una mayor escalabilidad en comparación con los métodos antes mencionados, pero tienden a producir secuencias de palabras que no son términos frasales. El método empleado utiliza procesamiento estadístico en conjunción con información sintáctica si está disponible y es compatible con cualquier idioma siempre que exista el acceso a alguna base de conocimiento general en el idioma de la colección.

2.3. Métodos semi supervisados.

Recientemente se han desarrollado con éxito algunos métodos basados en datos *(data-driven)*, para la minería de términos frasales que utilizan aprendizaje de máquina, dentro de estos métodos semi supervisados tenemos:

En el trabajo de (CARVALHO; MOURA; CALADO, 2010) se estudia el impacto de la adición de términos frasales (bigramas) detectados automáticamente, en un sistema de recuperación de información *ad hoc* utilizando diversas colecciones de texto. Primero, son extraídos todos los bigramas de la colección de documentos y recolectadas sus estadísticas. Luego de eso, es construido un modelo de clasificación utilizando el método *SVM* (*Support Vector Machine*) en una base pequeña de bigramas rotulados manualmente como válidos o no válidos. Después de la obtención del modelo, son procesados todos los bigramas obtenidos en la primera etapa y seleccionados aquellos que son términos frasales.

La técnica no utiliza procesamiento lingüístico del lenguaje natural, se basa en aprendizaje de máquina, usa las estadísticas extraídas de los bigramas para su identificación como términos frasales. Aunque no es necesario la utilización de base de datos externas, el método necesita de la intervención de especialistas para criar la base de entrenamiento con

bigramas rotulados. Además existe un costo adicional de crear un modelo de clasificación del algoritmo *SVM*, pero es importante resaltar que esta clasificación no es compleja computacionalmente, ya que el método utiliza un número reducido de atributos, también el procesamiento del corpus y el entrenamiento del clasificador es realizado *offline* en el proceso de indexación de la colección.

Los experimentos mostraron que esta abordaje puede ser eficaz para determinar los bigramas que representan términos frasales, además de poder mejorar la recuperación de la información, aumentando la precisión en algunas situaciones hasta un 36 %. Sin embargo, solo detecta términos frasales de 2-gramas. Mientras que el método de minería de términos frasales empleado, consigue determinar términos de mayor longitud, además se estudia el impacto de la adición de estos términos en la clasificación y clusterización de documentos.

La segmentación de secuencias de palabras, es otra estrategia de extracción de términos frasales, que divide una secuencia de palabras en subsecuencias disjuntas. En el trabajo de (LIU et al., 2015) se propone un método para la minería de *"quality phrase"* de un corpus de texto. Los autores definen *"quality phrase"* como secuencia de palabras que aparecen contiguamente en la colección y forman una unidad semántica completa (no composible), es decir, una secuencia de n-gramas que representan términos frasales.

La idea clave es que la utilización de estadísticas basadas en la frecuencia de los datos, tienden a producir una evaluación engañosa de la calidad de los n-gramas (posibilidad de que la secuencia de n-gramas sea considerada término frasal) y por tanto un resultado no satisfactorio, por lo que propone rectificar esa estimación inicial e inexacta de calidad, en función del contexto de ocurrencia.

El método integra los enfoques de segmentación frasal y evaluación de la calidad. Con un esfuerzo de etiquetado limitado, el modelo creado puede segmentar de forma iterativa el corpus en palabras o secuencias de palabras no superpuestas de tal manera que: la calidad de secuencia de palabras estimada en la iteración anterior guía la segmentación y los resultados de segmentación rectifican la frecuencia y mejoran el proceso de estimación de calidad de una secuencia. Dicho marco integrado se beneficia de la mejora mutua y logra alta calidad y eficiencia. Este método fue utilizado en los experimentos iniciales de minería con resultados de calidad.

Casi todos los métodos de vanguardia de cualquier corriente requieren expertos

humanos en ciertos niveles, para el diseño de reglas o etiquetado de n-gramas. Tal dependencia se convierte en un impedimento para el análisis oportuno de colecciones de textos masivos y emergentes. (SHANG et al., 2018) propone un método que automatiza la minería de términos frasales utilizando bases de conocimiento externas, para eliminar los esfuerzos humanos y minimizar la dependencia del idioma, incorporando un análisis lingüístico superficial y limitado. La diferencia con el trabajo anterior está dada en la utilización de las técnicas:

Robust Positive-Only Distant Training, la cual aprovecha el hecho de que muchas términos frasales están disponibles en bases generales de conocimiento como *Wikipédia* y *Freebase* y pueden ser obtenidas en una escala mayor que las producidas por expertos humanos. Esta técnica se utiliza para generar etiquetas positivas y negativas con ruido desde una base de conocimiento. Siendo innecesario el etiquetado por expertos humanos.

POS-Guided Phrasal Segmentation, que propone incorporar un etiquetador *POS* pre entrenado para mejorar aún más el rendimiento, cuando esté disponible para el idioma de la colección de documentos. Esta técnica utiliza la información de contexto incorporada en las etiquetas *POS* y localiza con mayor precisión los límites de los n-gramas en la colección dada, lo que mejora la precisión. De esta manera la información sintáctica guía el modelo de segmentación.

Además, la dependencia del idioma se minimiza de manera que es compatible con cualquier idioma siempre que exista una base de conocimiento (por ejemplo, *Wikipédia*), un tokenizador y un etiquetador *POS* para el idioma de la colección. Experimentos realizados mostraron que este método además de tener características deseables como ser independiente del dominio e idioma de la colección, tuvo un mejor rendimiento y mayor calidad de los términos frasales extraídos.

2.4. Términos frasales en la clasificación y clusterización de documentos.

Existen muchos estudios relacionados con la clasificación y clusterización de documentos. La combinación entre atributos y técnicas de aprendizaje de máquina es uno de los enfoque más notables. El uso de n-gramas fue visto durante mucho tiempo como una

forma natural de mejorar el rendimiento de la recuperación, sobre los modelos tradicionales que ignoran el aspecto secuencial de las ocurrencias de palabras. (D'HONDT et al., 2013) dá una visión general de los principales hallazgos en investigaciones previas sobre la utilización de n-gramas estadísticos o sintácticos (generados a través de métodos sintácticos o estadísticos). Mostrando que la mayoría de resultados positivos no eran significativamente superiores a los resultados anteriores que utilizan el modelo *BoW* y cuando lo eran, los resultados iniciales tendían a ser muy bajos.

Trabajos de investigación sobre el tema se enfocaron en el estudio de dos aproximaciones para la incorporación de n-gramas estadísticos (fundamentalmente bigramas) y/o sintácticos (frases nominales, verbales, dependencias triples, etc.) en la representación de los documentos. La primera combinaba unigramas y n-gramas, mientras que la segunda excluía los unigramas. Los experimentos mostraron que el uso solo de los n-gramas conlleva al decrecimiento en los resultados de la categorización, cuando comparado al modelo *BoW* (LEWIS, 1992; APTÉ; DAMERAU; WEISS, 1994; BEKKERMAN; ALLAN, 2004).

(LEWIS, 1992) examinó el efecto de diferentes tipos de atributos, así como su tamaño en la eficacia de la categorización, la selección de atributos estaba basadas en análisis *POS* de la colección de documentos. El conjunto de n-gramas sintácticos eran creados por una secuencia de palabras que cumplían con las reglas gramaticales predefinidas. Los experimentos revelaron que los n-gramas sintácticos no mostraron mejoras significativas en la clasificación de documentos.

Mientras (MLADENIC; GROBELNIK, 1998) mostraron que clasificadores entrenados en una combinación de unigramas y n-gramas (hasta 3-gramas) tuvieron un rendimiento mejor que los clasificadores solo entrenados con unigramas, pero fue utilizada solo una base de datos en los experimentos y el clasificador utilizado como *baseline* basado en el modelo *BoW* produjo resultados da categorización base bajos; diversas investigaciones basadas en la primera aproximación (BRAGA; MONARD; MATSUBARA, 2009; BEKKERMAN; ALLAN, 2004) mostraron que a pesar de que en general los bigramas pueden predecir mejor las categorías que unigramas (CAROPRESO; MATWIN; SEBASTIANI, 2001), con el uso de n-gramas no se reportaron mejoras significativas cuando comparado al modelo *BoW*.

En (TAN; WANG; LEE, 2002) los autores utilizan bigramas y n-gramas en con-

junción, adoptan rigurosos criterios para la selección de los n-gramas, con el objetivo de utilizar solo los más discriminatorios, como resultado solo el 2 % de los bigramas fueron utilizados en los experimentos, obteniendo ganancias significativas aunque el *baseline* no exhibía los mejores resultados alcanzados en la literatura. En (CRAWFORD; KOPRINSKA; PATRICK, 2004) n-gramas fueron utilizados para clasificar e-mails, pero sin ganancias consistentes a pesar de resultados *baseline* bajos.

La técnica *Acquaintance* desarrollada por (DOUCET; AHONEN-MYKA, 2004) basada en n-gramas, es una representación que convierte documentos en vectores de atributos de alta dimensión, donde cada atributo es el equivalente a una sub cadena continua. Esta técnica funciona de manera efectiva, pero la dimensión del vector de atributos suele ser muy alta. (TESAR et al., 2006) emplea bigramas para extender el modelo *BoW* los resultados no superaron los mejores resultados alcanzados *BoW* en las colecciones Reuter-21578 y *20 Newsgroups*, utilizando el clasificador *Naive Bayes*.

(FORMAN, 2007) reportó buenos resultados empleando bigramas en la clasificación de artículos de ciencias en computación, pero los experimentos fueron enfocados en clases que comprendían el 2,8 % de los documentos de la colección, incluso el autor señaló que este estudio no era lo suficientemente amplio como para sacar conclusiones generales. (FIGUEIREDO et al., 2011) presentan una metodología para generar nuevas características discriminatorias llamadas *c-features* (2-itemsets), que a diferencia de los n-gramas, son términos que co-ocurren pero sin ninguna restricción en el orden o distancia entre los términos en un documento, con ganancias utilizando los clasificadores *SVM* y *Knn*. (ÖZGÜR; GÜNGÖR, 2012) logra pequeñas pero pequeñas significativas mejoras al utilizar una combinación de unigramas con un subconjunto de dependencias léxicas y podas en tres conjuntos de datos diferentes, incluido el conjunto Reuters-21578.

El trabajo de investigación (D'HONDT et al., 2013), describe como para la clasificación de documentos de patentes, la combinación de representaciones de n-gramas (generados estadística o sintácticamente) y unigramas conduce a resultados de clasificación significativamente mejores, porque los n-gramas son más adecuados para capturar los términos de múltiples palabras, abundantes en los textos de patentes técnicas ricos en terminología. Recientemente, se han propuesto muchos modelos neuronales para aprender representaciones de palabras *(word embedding)* (LAI et al., 2015), así como diferentes estrategias para generar características discriminatorias.

(TUYET; HANH, 2016) proponen un clasificador basado en un atributo llamado *maximal frequent sequences (MFSs)*, que consiste en la extracción de una secuencia de palabras frecuentes en la colección y que no constituyen una sub secuencia de otra secuencia más larga. La extracción de este atributo es independiente del idioma e ignora el impacto de la mayoría del pre procesamiento lingüístico como: tokenización y lematización. Pero los resultados obtenidos no superaron los resultados del modelo *BoW* cuando aplicados a los mismos conjuntos de datos.

En resumen, en varios de los trabajos mostrados anteriormente, aunque se reportan ganancias con el uso de n-gramas como características, las ganancias son marginales o sujetas a circunstancias especificas, no siendo reportados ganancias en bases conocidas por altos resultados cuando se utiliza el modelo *BoW* (por ejemplo: *Reuter21578, 20newsgroups*) e incluso algunas veces se deteriora el rendimiento comparado con el modelo tradicional. En el presente trabajo se experimenta con varias técnicas de selección de características, clasificadores y con la adición de términos frasales desde bigramas hasta 6-gramas, la diferencia con métodos anteriores consiste en el uso del método de minería empleado.

3 Minería de Términos Frasales

Este capítulo describe el método de minería términos frasales *(AutoPhrase)*, utilizado en los experimentos, aún no siendo propuesta de esta pesquisa es básico para su comprensión. *AutoPhrase* perteneciente a la corriente de procesamiento estadístico del lenguaje natural, se fundamenta en la idea de que la utilización de estadísticas basadas en la frecuencia de los datos, tiende a producir una evaluación engañosa de la calidad de los n-gramas y por tanto un resultado no satisfactorio, como se expone en el Ejemplo 1.

AutoPhrase es un método que combina el uso de estadísticas basadas en la frecuencia de los datos con la segmentación de frases, a diferencia de la mayor parte de métodos existentes que utilizan complejos y entrenados analizadores lingüísticos, resultando en un rendimiento insatisfactorio en colecciones de nuevos dominios y géneros, sin una costosa adaptación extra. Este método utiliza la metodología *Robust Positive-Only Distant Training* que reduce la dependencia a los expertos humanos, utilizando para ello el gran cumulo de términos frasales disponibles en bases de conocimiento externas, la dependencia de expertos humanos constituía un impedimento para el análisis oportuno de grandes colecciones de textos en dominios emergentes, con el uso de esta metodología se logra un mejor rendimiento en comparación con el limitado número de secuencias etiquetadas por humanos. Además la técnica de *POS-Guided Phrasal Segmentation* que incorpora información sintáctica en forma de etiquetas *POS* cuando un etiquetador *POS* está disponible, consiguiendo mayor precisión en la localización de los limites de los n-gramas e independencia del idioma. Por tanto, este método soporta cualquier idioma siempre que esté disponible una base de conocimiento general en ese idioma y se beneficia de la información sintáctica obtenida a través de un etiquetador de *POS*, pero no lo requiere.

3.1. Minería de n-gramas basada en la frecuencia

Ejemplo 1. ***Minería de n-gramas basada en la frecuencia:*** *Considerando un conjunto de publicaciones científicas y la frecuencia de dos n-gramas: "relational database system" y "support vector machine" y sus sub partes, podemos observar: i) la frecuencia decrece con el tamaño de la secuencia, ii) tanto n-gramas que podemos considerar como*

termos frasales y aquellos que no lo son, pueden tener una frecuencia alta como: "support vector machine" y "vector machine"; y iii) la frecuencia de un n-grama ("relational database system") y sus sub partes pueden tener similar escala a otros n-gramas ("support vector machine") y sus contra partes.

Secuencias.	**Frecuencia**	**Término Frasal**	**Frecuencia rectificada**
relational database system	100	si	70
relational database	150	si	40
database system	160	si	35
relational	500	-	20
database	1000	-	20
system	10000	-	1000
Secuencias.	**Frecuencia**	**Término Frasal**	**Frecuencia rectificada**
support vector machine	100	si	80
support vector	160	si	50
vector machine	150	no	6
support	500	-	150
vector	1000	-	200
machine	10000	-	50

Tabla 3.1.1 –
Ejemplo de minería de n-gramas basada en la frecuencia.

Por lo que un método que ranquea n-gramas de acuerdo a la frecuencia, puede producir falsos términos frasales como "*vector machine*". Suponiendo que mediante alguna heurística se pudiera diferenciar entre "*support vector*" y "*vector machine*" discriminando sus frecuencias (150 y 160) , la misma heurística podría fallar con los n-gramas "*relational database*" y "*database system*". Además, utilizando las frecuencias del ejemplo, todas las heurísticas podrían producir igual predicción para "*relational database*" y "*vector machine*", garantizando que una de ellas está errada. Este ejemplo argumenta las limitaciones del uso de la frecuencia, especialmente para juzgar si una secuencia es muy larga (mayor que una unidad semántica mínima) o muy corta (incompleta) o con una longitud adecuada. Por lo que los autores proponen rectificar esa estimación inicial e inexacta de la calidad en función del contexto de ocurrencia. El objetivo de la rectificación es estimar cuántas veces cada n-grama puede ser interpretado como un término frasal en su contexto de ocurrencia. El Ejemplo 2 ilustra la idea.

3.2. Segmentación Frasal

Ejemplo 2. *Rectificación de la frecuencia: Considerando la siguiente ocurrencia de las secuencias mostradas en la Tabla 3.1.1.*

1. *A ⌈ relational database system ⌋ for images ...*
2. *⌈ Database system ⌋ empowers everyone in your organizations ...*
3. *More formally, a ⌈ support vector machine ⌋ constructs a hyperplane ...*
4. *The ⌈ support vector ⌋ method is a new general method of ⌈ function estimation ⌋ ...*
5. *A standard ⌈ feature vector ⌋ ⌈ machine learning ⌋ setup is used to describe ...*
6. *⌈ Relevance vector machine ⌋ has an identical ⌈ functional form ⌋ to the ⌈ support vector machine⌋ ...*
7. *The basic goal for ⌈object-oriented relational database ⌋ is to ⌈ bridge the gap⌋ ...*

Las primeras cuatro instancias contarían en la frecuencia de estos n-gramas mientras las últimas tres no contarían en la frecuencia de los n-gramas "*vector machine*" o "*relational database*", porque no son interpretados como una unidad en su contexto de ocurrencia. Suponiendo que podamos colectar la frecuencia rectificada de cada n-grama identificado, como se muestra en la Tabla 3.1.1. La frecuencia rectificada ahora distingue que "*vector machine*" raramente ocurre como una unidad completa.

Para recuperar la frecuencia rectificada con el mejor esfuerzo, la estrategia es examinar el contexto de aparición de cada secuencia de palabras y decidir si contarla como una unidad. El examen de una ocurrencia puede implicar la enumeración de posibilidades alternativas, como: extender la secuencia o romper la secuencia, y la comparación entre ellas. Esta prueba podría ser costosa, perdiendo la ventaja en la eficiencia de los enfoques de minería de patrones frecuentes. Con este objetivo en mente los autores proponen un enfoque de segmentación frasal, integrado con el proceso de evaluación de la calidad de los n-gramas en un marco unificado.

La segmentación frasal busca dividir una secuencia de palabras en sub secuencias disjuntas, cada una de las cuales se correlaciona con una unidad semántica, a diferencia

de enfoques anteriores donde una misma palabra puede pertenecer a más de una secuencia (solapamiento), lo que puede conducir a un crecimiento exponencial del vocabulario, así como falta de interpretabilidad. En la primera instancia del Ejemplo 2 "*Relational Database System*" se considera un único n-grama; la segmentación de la quinta instancia aumenta la frecuencia de los n-gramas "*feature vector*" y "*machine learning*" no siendo así para "*feature vector machine*". Con la segmentación frasal, se condensan las pruebas individuales para cada secuencia de palabras y se reduce la complejidad general. Además aunque existe un número exponencial de posibles particiones de los documentos, solo interesan aquellas relevantes para la extracción de términos frasales de modo que: i) solo se tengan en cuenta aquellos n-gramas frecuentes de calidad razonable al enumerar las particiones; y ii) la calidad del n-grama determinada en un principio guíe la segmentación y la segmentación rectifique esa estimación inicial de la calidad. El recuento total de una secuencia que aparece en la colección segmentada se denomina frecuencia rectificada. Con dicha integración se logra una mejora mutua de ambos procesos, lográndose mayor calidad y eficiencia.

Definition 1. ***Segmentación frasal.*** *Dada una secuencia de palabras* $C = \omega_1\omega_2...\omega_n$*, una segmentación* $S = s_1s_2...s_m$ *es inducida por una secuencia de índices limites* $B = b_1, b_2, ..., b_{m+1}$ *satisfaciendo* $1 = b_1 < b_2 < ... < b_{m+1} = n + 1$*, donde un segmento* $s_i = (\omega_{b_i}\omega_{b_i+1}...\omega_{b_i+|s_i|-1})$ *. s/i/ se refiere al número de palabras en segmento* s_i*. Desde que* $b_i + |s_i| = b_{i+1}$*, para mayor claridad se utiliza* $\omega_{[b_i,b_{i+1})}$ *para denotar la secuencia de palabras* $\omega_{b_i}\omega_{b_i+1}...\omega_{b_i+|s_i|-1}$ *.*

Ejemplo 3. *La segmentación de la primera instancia del Ejemplo 2 seria:*
C= a relational database system for images
S = / a / relational database system / for / images /
B = {1,2,5,6,7}

Aunque no existe una definición universalmente aceptada para cuantificar la calidad que debe tener un n-grama para ser considerado término frasal, los autores utilizan los siguientes criterios:

1. Popularidad: Para ser considerado un término frasal un n-grama debe aparecer con suficiente frecuencia en los documentos de la colección.

2. Concordancia: Se refiere a la colocación de las palabras en una frecuencia que es significativamente más alta que lo esperado debido al acaso.

3. Informatividad: Un n-grama es informativo si es indicativo de un tema o concepto específico.

4. Completud: Un n-grama debe interpretarse como una unidad semántica en su contexto de ocurrencia.

3.3. Método automático de minería de Términos Frasales (*AutoPhrase*)

AutoPhrase incluye varios módulos: Minería de n-gramas frecuentes, generación de rótulos, entrenamiento positivo distante (*Robust positive-only distant training*), estimación de la calidad, segmentación frasal guiada por etiquetas *POS* y re estimación de la calidad de los n-gramas, como se presenta en la Figura 3.3.1.

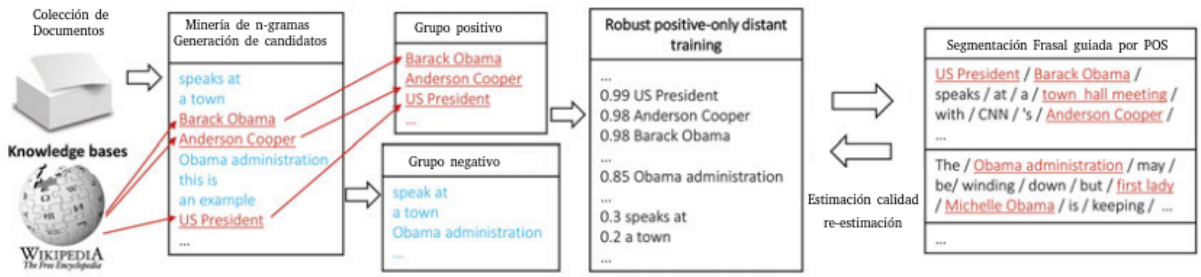

Figura 3.3.1 –
Marco general para la minería automática de términos frasales.

3.3.1. Minería de n-gramas frecuentes (Generación de candidatos).

Para la generación de candidatos a través de la minería de n-gramas frecuentes se recopilan todos los n-gramas que satisfacen cierto limiar de soporte mínimo τ, de acuerdo con el requisito de popularidad y una determinada longitud máxima ω. Para la minería eficiente, se recurre a dos propiedades:

Propiedad de cierre descendente: si una secuencia no es frecuente, se garantiza que no será frecuente su súper secuencia. Por lo tanto, esas secuencias más largas se filtrarán y nunca se expandirán.

Propiedad de prefijo: si una secuencia es frecuente, cualquiera de sus unidades de prefijo también debería ser frecuente. De esta forma, todas las secuencias frecuentes pueden generarse al expandir sus prefijos.

El algoritmo 1, muestra el proceso de generación de candidatos. Donde C [.] es utilizado para indexar una palabra de la colección y |C| para denotar su tamaño. El operador $\oplus$ para concatenar dos palabras o n-gramas. El resultado es un diccionario de clave-valor f. Las claves son el vocabulario U que contiene todas las secuencias frecuentes P y las palabras U \P. Los valores es la frecuencia absoluta.

Entrada: Colección de documentos, limiar de soporte mínimo τ
Salida : Diccionario de palabras y n-gramas frecuentes
inicio
  $f \leftarrow \emptyset$ //diccionario vacío.
  $index \leftarrow \emptyset$ //diccionario vacío
  para $i \leftarrow 1$ ***to*** $|C|$ **hacer**
    $index[C[i]] \leftarrow index[C[i]] \cup i$
  fin
  mientras $index \neq \emptyset$ **hacer**
    $index' \leftarrow \emptyset$ //diccionario vacío
    para $u \in index.keys$ **hacer**
      si $|index[u]| \geq \tau$ **entonces**
        $f|u| \leftarrow |index[u]|$
        para $j \in index[u]$ **hacer**
          $u' \leftarrow u \oplus C|j+1|$
          $index'[u'] \leftarrow index'[u'] \cup \{j+1\}$
        fin
      en otro caso
    fin
    $index \leftarrow index'$
  fin
  devolver f
fin

Algoritmo 1: Generación de candidatos

3.3.2. Generación de etiquetas.

Los autores consiguen superar la dependencia de los expertos humanos y con ello conferir una mayor autonomía al método, utilizando las bases de conocimiento público (por ejemplo, *Wikipédia*) que generalmente codifican un número considerable de términos frasales en los títulos, palabras claves y enlaces internos de las páginas[1]. Utilizando el

[1] WikipediaEntities

análisis de los enlaces internos y sinónimos en la *Wikipédia* en inglés, fueron extraídos términos frasales y colocados en un grupo positivo; aquellos candidatos extraídos del corpus pero que no coinciden con ningún término frasal derivado de la base de conocimiento, se utilizan para poblar un grande y ruidoso grupo negativo (pueden existir secuencias candidatas que sean términos frasales y que no estén presentes en la base de conocimiento). Con estos grupos conformados se utiliza el clasificador *Random Forest* por ser una técnica *ensemble* que promedia el resultado de clasificadores bases independientes, con el objetivo de aliviar el efecto del ruido en el grupo negativo.

3.3.3. Estimación de la calidad de los n-gramas candidatos.

Para estimar la calidad de los candidatos, como se mencionó anteriormente, fueron utilizado cuatro criterios: popularidad, concordancia, informatividad y completud. Las características calculadas son divididas en dos categorías, según los criterios de concordancia e informatividad.

El conjunto de características relativas al criterio de concordancia está diseñado, para medir la concordancia entre sub unidades. Para ello se particionan las secuencias en dos partes disjuntas de todas las maneras posibles, con el objetivo de comprobar si la co-ocurrencia es significativamente más alta que la aleatoriedad. De cada palabra o secuencia $u \, \epsilon \, U$ se tiene su frecuencia $f[u]$. Su probabilidad es definida como:

$$p(u) = \frac{f[u]}{\sum_{u' \epsilon U} f[u']} \tag{3.1}$$

Dada una secuencia $v \, \epsilon$ P, se particiona en dos sub unidades $< u_l, u_r >$, tal que la información mutua puntual (*Pointwise mutual information*) *PMI* es minimizada. La información mutua puntual cuantifica la diferencia entre la probabilidad de su coincidencia dada su distribución conjunta y sus distribuciones individuales, bajo el supuesto de independencia.

$$< u_l, u_r >= argmin u_l \oplus u_r = v log \frac{p(v)}{p(u_l)p(u_r)} \tag{3.2}$$

Con $< u_l, u_r >$, se utiliza *PMI* como una característica para medir la concordancia.

$$PMI(u_l, u_r) = log \frac{p(v)}{p(u_l)p(u_r)} \tag{3.3}$$

Otra característica utilizada por los autores es la divergencia puntual de Kullback-Leibler (*pointwise Kullback-Leiber divergence*):

$$PKL(v \parallel < u_l, u_r >) = p(v) log \frac{p(v)}{p(u_l)p(u_r)} \tag{3.4}$$

El $p(v)$ adicional se multiplica con la información mutua puntual, lo que conduce a un menor *bias* hacia aquellas secuencias de rara ocurrencia.

En cuanto al requisito de informatividad, algunos candidatos son funcionales o *stopwords*. Los autores incorporan las siguientes características.

- Candidatos que comienzan o terminan con *stopwords*, a menudo son funcionales más que informativos.
- Utilizar IDF promedio sobre palabras que conforman el n-grama para medir la semántica.
- Puntuación: Por lo general, las probabilidades de un n-grama entre comillas, corchetes o conectado por guión ser un término frasal es mayor.
- Proporción entre la frecuencia del n-grama y la frecuencia mínima entre sus sub partes. Una proporción baja generalmente indica que la secuencia se puede acortar. Por ejemplo: "*classifier SVM*" la proporción debe ser baja porque tanto "*classifier*" como *"SVM"* son palabras frecuentes.
- Proporción entre la frecuencia máxima entre sus supersecuencia y la frecuencia del n-grama. Una proporción alta implica que la secuencia no está completa. En el caso de: *"NP-complete in the strong"* tiende a tener una alta proporción porque siempre ocurre en *"NP-complete in the strong sense"*.

3.3.4. Segmentación frasal guiada por etiquetas *POS*.

La segmentación frasal guiada por etiquetas *POS* es un componente crucial para la minería automática de términos frasales, que a diferencia de la propuesta anterior (LIU et al., 2015), propone abordar el desafío de medir la completud y la independencia al ubicar cada n-grama candidato en el corpus y rectificar su frecuencia previa obtenida mediante la comparación de cadenas.

Para lograr un alto rendimiento con una dependencia mínima del lenguaje, los autores utilizan los resultados de: tokenización y etiquetado *POS*, herramientas disponibles en diferentes idiomas, con el objetivo de combinar la información de contexto de las etiquetas *POS* para identificar los limites de los n-gramas en complemento con la estadística basada en la frecuencia. El Ejemplo 4 sugiere como la información de contexto brindada por la etiquetas *POS* es un complemento para una mejor identificación de los limites de las secuencias.

Ejemplo 4. *Combinación de la frecuencia con la información de las etiquetas POS*

1	⌈ Sopia NNP	Smith ⌋ NNP	was VBN	born VBN	in IN	England NNP
2	... ...	the DT	⌈ Great NNP	Firewall ⌋ NNP	is VBZ	... ..
3	This DT	is VBZ	a DT	great JJ	⌈ firewall NN	software ⌋ NN

Métodos basados solo en la frecuencia pueden cometer errores como: fraccionar una combinación de palabras frecuentes *sophia* y *smith* que son nombres comunes, porque para formar la secuencia *sophia smith* esta también debería ser frecuente, lo que no ocurre necesariamente; con la información *POS* dos sustantivos continuos son un indicador de una probable secuencia completa. Por otra parte secuencias frecuentes tienen a ser menos fraccionable como el caso de *great firewall* ocasionando que secuencias como *firewall software* no sean consideradas; con la información de contexto la instancia 2 la transición de *firewall* (sustantivo) a *is* (verbo) implica un limite en la secuencia y en 3 *great* es un adjetivo en cambio *firewall* y *software* son sustantivos, haciendo que "*firewall software*" sea más probable.

Definition 2. ***Segmentación frasal guiada por etiquetas POS.*** *Dado un corpus C (secuencia de palabras con etiquetas POS* $(\omega_1\omega_2..., t_1t_2...t_n)$*), una segmentación* $S = s_1s_2...s_m$ *es inducida por una secuencia de índices limites* $B = b_1, b_2, ..., b_{m+1}$ *satisfaciendo* $1 = b_1 < b_2 < ... < b_{m+1} = n + 1$*, donde el i-ésimo segmento* $s_i = (\omega_{b_i}\omega_{b_i+1}...\omega_{b_i+|s_i|-1}, t_{b_i}t_{bi+1}...t_{b_i+|s_i|-1})$ *. s/i/ se refiere al número de palabras/etiquetas en el segmento* s_i*. Desde que* $b_i + |s_i| = b_{i+1}$*, para mayor claridad se utiliza* $\omega_{[b_i,b_{i+1})}$ *para denotar la secuencia de palabras* $\omega_{b_i}\omega_{b_i+1}...\omega_{b_i+|s_i|-1}$ *y* $t_{[b_i,b_{i+1})}$ *para denotar la secuencia de etiquetas POS* $t_{b_i}t_{bi+1}...t_{b_i+|s_i|-1}$*. Por tanto, el i-ésimo segmento* $s_i = (\omega_{[b_i,b_{i+1})}, t_{[b_i,b_{i+1})})$*.*

Definition 3. *Calidad de una secuencia POS es la probabilidad de que una secuencia de palabras sea una unidad semántica, dada su correspondiente secuencia de etiquetas POS, de acuerdo con el criterio anterior: dada una secuencia de etiquetas POS de tamaño k* $t_1t_2...t_k$*, su calidad es calculada:*

$$T(t_1...t_k) = p(\lceil v_1...v_k \rfloor \, | tag(v_1...v_k) = t_1...t_k) \epsilon [0,1]$$

donde $tag(v_1...v_k)$ *es la secuencia de etiquetas POS de la secuencia* $v_1...v_k$

El puntaje de calidad de una secuencia POS T(.) está diseñado para recompensar las secuencias con patrones *POS* significativos. La forma es:

$$T(t_{[b_i,b_{i+1})}) = (1 - \delta(t_{b_{i+1}-1}, t_{b_{i+1})}) x \prod_{j=b_i+1}^{b_{i+1}-1} \delta(t_{j-1}, t_j)$$

donde $\delta(t_1, t_2)$ es la probabilidad que la etiqueta *POS* tag t_2 esté exactamente después que la etiqueta *POS* tag t_1 dentro de una secuencia en la colección de documentos dada. El primer término indica que hay un índice limite entre b_{i+1} y bi, mientras que el producto indica que todas las etiquetas *POS* entre $t[b_i, b_{i+1})$ están en la misma secuencia. Este puntaje de calidad dado en las etiquetas *POS* puede contrarrestar naturalmente el *bias* a segmentos más largos porque exactamente uno de $\delta(t_1, t_2)$ y $(1 - \delta(t_1, t2))$ siempre se multiplica, no importa como se segmenta la colección. Matemáticamente $\delta(t_1, t_2)$ fue definido como:

$$\delta(t_1, t_2) = p(... \lceil ...\omega_1\omega_2... \rfloor ... | C, tag(\omega_1) = t_1 \wedge tag(\omega_2) = t_2)$$

Como depende de como se segmentan los documentos en n-gramas, $\delta(t_1, t_2)$ es aprendido durante la segmentación frasal. Una vez detallados Q (calidad de la secuencia) y T (calidad de la secuencia de etiquetas *POS*, se define formalmente el modelo se segmentación frasal guiado por *POS*. La probabilidad conjunta del corpus C y la segmentación $S = s_1...s_m$ es factorizada de la siguiente forma:

$$p(S, C) = \prod_{t=1}^{m} p(b_{i+1}, \left\lceil \omega_{[b_i,b_{i+1})} \right\rfloor \, | b_i, t_{[b_i,b_{i+1})}) \tag{3.5}$$

donde $p(b_{i+1}, \left\lceil \omega_{[b_i,b_{i+1})} \right\rfloor | b_i, t_{[b_i,b_{i+1})}$ es la probabilidad de observar la secuencia $\omega_{[b_t,b_{t+1})}$ como el i-ésimo segmento de calidad dado el índice limite anterior b_i y sus correspondientes secuencia de etiquetas *POS* $t_{b_i,b_{i+1}}$.

Para cada segmento, dada la secuencia de etiquetas *POS* t y el índice de inicio b_i de un segmento s_i, el proceso generativo se define de la siguiente manera.

Modelo generativo para cada segmento:

1. Generar el índice b_{t+1}, de acuerdo a la calidad de la secuencia *POS*.

$$p(b_{i+1}|b_i, t_{[b_i,b_{i+1})}) = p(\left\lceil \omega \right\rfloor | t_{[b_i,b_{i+1})}) = T(t_{[b_i,b_{i+1})})$$

2. Dado los índices b_i, b_{i+1}, se genera la secuencia $\omega_{[b_i,b_{i+1})}$ de acuerdo a una distribución multinomial sobre todos los segmentos de longitud $(b_{i+1} - b_i)$.

$$p(\omega_{[b_i,b_{i+1})}|b_i, b_{i+1}) = p(\omega_{[b_i,b_{i+1})}||s_i| = b_{i+1} - b_i)$$

3. Finalmente se generará un indicador si $\omega_{[b_i,b_{i+1})}$ forma un segmento de calidad, de acuerdo a la calidad de la secuencia.

$$p(\left\lceil \omega_{[b_i,b_{i+1})} \right\rfloor | \omega_{[b_i,b_{i+1})}) = Q(\omega_{[b_i,b_{i+1})})$$

Integrando los tres pasos del modelo, se tiene la siguiente factorización probabilística:
$p(b_{i+1}, \left\lceil \omega_{[b_i,b_{i+1})} \right\rfloor | b_i, t_{[b_i,b_{i+1})})$
$= p(b_{i+1}|b_i, t_{[b_i,b_{i+1})}) p(\omega_{[b_i,b_{i+1})}|b_i, b_{i+1}) p(\left\lceil \omega_{[b_i,b_{i+1})} \right\rfloor | \omega_{[b_i,b_{i+1})})$
$= T(t_{[b_i,b_{i+1})}) p(\omega_{[b_i,b_{i+1})}||s_i| = b_{i+1} - b_i) Q(\omega_{[b_i,b_{i+1})})$

Por lo tanto, para una determinada colección C con documentos D, hay tres subproblemas:

- Aprender $p(u||u|)$ de cada palabra y secuencia $u \epsilon P$. Se denota $p(u||u|)$ como θ_u.
- Aprender $\delta(t_1, t_2)$ de cada par de etiquetas *POS*.
- Inferir la segmentación S cuando θ y δ son fijas.

Utilizando el principio de máximo a posteriori y maximizando la probabilidad conjunta:

$$\sum_{d=1}^{D} \log p(S_d, C_d) = \sum_{d=1}^{D} \sum_{i=1}^{m_d} \log p(b_{i+1}^{(d)}, \omega_{b_i, b_{i+1})}^{(d)} \rfloor | b_t^{(d)}, t_{[b_i, b_{i+1})}) \tag{3.6}$$

Para encontrar la mejor segmentación, maximizando ecuación (3.6) se utiliza programación dinámica, como se muestra en el Algoritmo 2.

Entrada: Corpus $C = w_1, w_2, ..., w_n, t_1 t_2 ... T_n$, Calidad Q, parámetros θ y δ.
Salida : Segmentación óptima S
inicio
  $h_1 \leftarrow 1$, $h_i \leftarrow 0$ $(0 < i \leq n+1)$
  para $i = 1$ ***to*** n **hacer**
    para $j = i+1$ ***to*** $n+1$ **hacer**
      //implementado vía Trie
      si *No existe secuencia comenzando* $\omega_{i,j}$ **entonces**
        break
      fin
      //log y adición son usados para evitar *underflow*
      si h_i x $p(j, \lceil \omega_{[i.j)} \rfloor | i, t_{[i.j)}) > h_j$ **entonces**
        $h_j \leftarrow h_i$ x $p(j, \lceil \omega_{[i,j)} \rfloor | i, t_{[i,j)})$
        $g_j \leftarrow i$
      fin
    fin
  fin
  $j \leftarrow n+1$
  $m \leftarrow 0$
  mientras $j > 1$ **hacer**
    $m \leftarrow m+1$
    $s_m \leftarrow \omega_{[g_j, j)}, t_{[g_j, j)}$
    $j \leftarrow g_j$
  fin
  devolver $S \leftarrow s_m s_{m-1} ... s_1$
fin

Algoritmo 2: Segmentación Frasal guiada por etiquetas *POS*

Cuando la segmentación y el parámetro θ son fijos, la forma cerrada para $\delta(t_1, t_2)$ es:

$$\delta(t_1, t_2) = \frac{\sum_{d=1}^{D} \sum_{i=1}^{m_d} \sum_{j=b_i^{(d)}}^{b_{i+1}^{(d)}-2} 1(t_j^{(d)} = t_1 \wedge t_{j+1}^{(d)} = t_2)}{\sum_{d=1}^{D} \sum_{i=1}^{n_d-1} 1(t_i^{(d)} \wedge t_{i+1}^{(d)} = t_2)} \tag{3.7}$$

donde 1(.) denota el indicador de identidad y $\delta(t_1.t_2)$ es la relación no segmentada entre todos los pares $t_1 t_2$ en el corpus dado.

De forma similar teniendo la segmentación S y el parámetro δ fijos, la forma cerrada de θ_u, es derivada como:

$$\theta_u = \frac{\sum_{d=1}^{D} \sum_{i=1}^{m_d} 1(\omega_{[b_i, b_{i+1})} = u)}{\sum_{d=1}^{D} \sum_{i=1}^{m_d} 1(|s_i^{(d)}| = |u|)} \tag{3.8}$$

Podemos observar que θ_u es el números de veces que u es un segmento completo, normalizado por el número de segmentos de longitud $|u|$

Los autores utilizan *Viterbi Training* para iterativamente optimizar parámetros, dado su rápida convergencia y resultados en modelos sencillos y dispersos para tareas similares.

3.3.5. Re - estimación de la Calidad.

Con la frecuencia rectificada se reconstruye todo el espacio de características de la siguiente manera: Cuando se generan características relacionadas con la frecuencia, como la concordancia y la completud, la frecuencia sin procesar se reemplaza por la frecuencia rectificada. Cuando se calculan las características relacionadas con la ocurrencia, como la informatividad, solo se consideran las apariciones coincidentes de segmentos completos. La reconstrucción explota la frecuencia rectificada de una manera más completa y por lo tanto produce un mejor ganancia de rendimiento.

4 Experimentos

En este capítulo se presentan los experimentos realizados para evaluar el impacto de la adición de términos frasales, extraídos por el método *Autophrase*, en la representación de los documentos con el consiguiente enriquecimiento de los modelos tradicionales basados en el modelo *BoW*, en las tareas de recuperación de información *ad hoc*, clasificación y clusterización de documentos.

Para evaluar el impacto de la adición de los términos frasales en todas las tareas de recuperación de información, se utilizó el tradicional modelo vectorial con ponderación de los términos *TF - IDF* (BAEZA-YATES; RIBEIRO-NETO et al., 1999), donde cada documento fue representado como un conjunto de palabras claves, o sea, términos de un índice. Un término de indización es generalmente una palabra o una agrupación de palabras que representa un concepto o significado presente en el documento (FERNEDA, 2003). En el caso de la clasificación y clusterización se utilizaron además la aplicación de técnicas filtro de selección de atributos para utilizar en el índice. Los términos frasales detectados por el método fueron utilizados para expandir la representación de los documentos. En este caso, todos los términos frasales detectados en la colección se tratan como términos de una sola palabra y son indexados por el motor de búsqueda como tal como adición a todas las palabras individuales indexadas.

En la práctica, estamos ampliando los documentos de la colección para que contengan, además de todas sus palabras individuales, también todos sus términos frasales. Las normas de los documentos se vuelven a calcular teniendo en cuenta esto. El mismo proceso se aplica a la consulta. Por ejemplo, en un escenario en el que en una consulta se detecta *"hormone replacement therapy"* como un término frasal, la consulta *"60 year old menopausal woman with hormone replacement therapy"* se expandiría en una consulta con nueve términos distintos: "60", "*year*", "*old*", "*menopausal*", "*woman*", "*with*", "*hormone*", "*replacement*", "*therapy*", y *"hormone replacement therapy"*. Lo mismo se aplicaría a todos los documentos.

4.1. Recuperación *ad hoc*

Para evaluar este impacto en la tarea de búsqueda fueron utilizadas las colecciones *OHSUMED*[1], *Glasgow Herald 1995 (GH95)* y *Cystic Fibrosis (CF)*. La colección de pruebas *OHSUMED* (*TREC-9 Filtering Track*) es un conjunto de 348,566 referencias de *MEDLINE*, la base de datos de información médica en línea que consta de títulos y / o resúmenes de 270 revistas médicas durante un período de cinco años (1987-1991). Los campos disponibles son: título, resumen, términos de indexación de *Medical Subject Headings (MeSH)*, autor, fuente y tipo de publicación. Para este trabajo se indexó solo el título y el resumen cuando están disponibles, ya que algunas referencias no tienen ningún resumen (solo títulos).

GH95 es un compendio de 56,472 noticias del periódico *The Herald* de la colección *British-English data collection* del año 1995, que forma parte de *CLEF test collections*. Los campos disponibles son: título, autor, texto. Fueron indexados los campos título y texto. Por su vez la colección *CF* también subconjunto de la base de información médica *MEDLINE*, tiene un total de 1239 documentos publicados entre los años 1974 y 1979 sobre aspectos de la Fibrosis Cística, cuenta con 11 campos disponibles: autor, título, citación bibliográfica, términos de indexación principal y secundarios, resumen o extracto del documento, referencias y citaciones. En este caso fueron utilizados los campos título, resumen o extracto, términos de indexación principal y secundarios.

La tabla 4.1.1 muestra algunas estadísticas sobre las colecciones de documentos donde podemos observar que, aunque contiene más documentos que *GH95, OHSUMED* tiene un vocabulario que no es significativamente mayor (léxico limitado), pero sus documentos ricos en terminología médica tienen alta probabilidad de construir n-gramas significativos, adecuados para ser capturados como términos frasales.

Colección.	**Tam. vocabulario**	**No. términos Frasales**
OHSUMED	189773	100843
GH95	186609	39866
CF	11237	430

Tabla 4.1.1 –
Estadísticas de las colecciones utilizadas.

[1] https://trec.nist.gov/data/t9$_filtering.html$

4.1.1. Configuración del método de minería de términos frasales.

Luego de realizar algunos ajustes el limiar de suporte mínimo establecido fue de 10, es decir, se mineraron términos frasales con una frecuencia mayor o igual a 10 en las colecciones de documentos; en los experimentos se utilizaron términos de hasta 3-gramas y de hasta 6-gramas de longitud, con el objetivo de estudiar el impacto de los términos frasales de mayor orden; se utilizó tokenizador y etiquetador *POS* disponible para el idioma de las colecciones y se utilizaron aquellos términos frasales extraídos con una calidad igual o superior a 0,5 (parámetro utilizado por los autores) en las colecciones *OHSUMED* y *GH95*, en cambio dado el tamaño pequeño de la colección *CF* se utilizaron términos con una calidad igual o superior a 0.48 (aún estando por debajo del limiar utilizado en (SHANG et al., 2018) constituyen términos frasales). Las siguientes tablas 4.1.2 y 4.1.3 muestran las características de los términos frasales detectados en las colecciones utilizadas.

	OHSUMED	**GH95**	**CF**
2-gramas	72105	32590	353
3-gramas	21559	5558	67
4-gramas	5356	1291	9
5-gramas	1336	350	1
6-gramas	487	77	0
Total	100843	39866	430

Tabla 4.1.2 –
Términos frasales detectados en las colecciones utilizadas.

Términos frasales x documentos	**OHSUMED**	**GH95**	**CF**
0	30141	396	0
1-3	72877	4788	220
4-6	29189	6139	449
7-9	24562	5606	321
>= 10	191797	39543	249
TOTAL	91,35 %	99,30 %	100 %

Tabla 4.1.3 –
Términos frasales por documentos.

Como se puede observar, en todas colecciones los brigramas y trigramas representan más del 93 % del total, mientras los términos más largos (4 - 6 gramas) están por debajo del 8 %, por lo que se realizaron varios experimentos con diferentes tamaños de términos frasales para estudiar su aporte en la tarea de recuperación, observándose que el uso

de estos términos de mayor orden aporta ganancias ínfimas en las métricas utilizadas. También se pudo observar que los términos frasales en el tope del *ranking* son en su mayoría nombres de entidades (Tabla 4.1.4), lo que es compatible con los artículos de la *Wikipédia*. La tabla 4.1.5 muestra algunos cálculos para medir de alguna forma la precisión del método en las colecciones utilizadas, mostrándose alta calidad en la minería de los términos frasales.

OHSUMED	**GH95**	**CF**
ewing's sarcoma	stefan edberg	diabetes mellitus
pseudomonas aeruginosa	myra hindley	haemophilus influenzae
colorectal cancer	johan cruyff	staphylococcus aureus
streptococcus pyogenes	vladimir zhirinovsky	escherichia coli

Tabla 4.1.4 –
Términos frasales en el Top del *ranking*.

Métricas	**OHSUMED**	**GH95**	**CF**
Primeros 30 n-gramas considerados TF	100 %	100 %	96,66 %
Últimos 30 n-gramas considerados TF	60,00 %	73,33 %	80,00 %
Primeros 30 n-gramas descartados como TF	46,67 %	60 %	36,66 %
30 n-gramas aleatorios	96,67 %	90,00 %	86.67 %

Tabla 4.1.5 –
Precisión del método *AutoPhrase*.

4.1.2. Impacto de la adición de los Términos frasales para la búsqueda de documentos.

En esta sección presentamos los resultados obtenidos de la adición de términos frasales en la búsqueda de documentos. Los resultados de la búsqueda en las colecciones antes y después de la inclusión de los términos frasales, se evaluaron en términos de *Mean Average Presition (MAP)* y precisión en los 10 primeros resultados proporcionados para cada consulta *(Pg @ 10)* (BAEZA-YATES; RIBEIRO-NETO et al., 1999). Las consultas utilizadas en los experimentos fueron de la 1 - 63 de *OHSUMED*, 251 - 300 de la *GH95* y 1 - 100 en la *CF*, donde solo se consideró la porción de títulos de las consultas.

Se adicionaron los términos frasales detectados a la maquina de busca implementada con el modelo vectorial, como términos adicionales en la representación de los documentos y se detectaron los términos frasales de cada consulta, las consultas de la colección *GH95*

son consultas de texto cortos (2 - 5 palabras), por tanto tienen una menor aparición de términos frasales cuando comparado a *OHSUMED*, en el caso de *CF* es una colección con un número reducido de términos frasales, lo que se refleja en la cantidad de consultas que los contiene, como se muestra en la tabla 4.1.6:

	Número de Consultas		
	Total	**con Término frasal**	**% con Término frasal**
OHSUMED	63	56	88 %
GH95	46	20	43,47 %
CF	100	35	35 %

Tabla 4.1.6 –
Número de consultas modificadas.

La tabla 4.1.7 muestra los resultados obtenidos de todas las consultas de prueba. La columna *Baseline* se refiere a los resultados obtenidos utilizando el modelo vectorial con el modelo *BoW*. La columna +TF se refiere al uso de los términos frasales en las consultas y a la adición de los términos frasales en la representación de los documentos. Como podemos notar, el uso de los términos frasales para expandir la representación de los documentos, dió lugar a ganancias en todas las colecciones de documentos, tanto en términos de *MAP* como en *P@10*. Para verificar la importancia estadística de los resultados, se utilizó la prueba de *McNemar* (considerando un valor de p de 0.01) en las colecciones *OHSUMED* y *GH95* (marcado en negrita).

	MAP				
	Hasta 6-gramas			**Hasta 3-gramas**	
	Baseline	**+TF**	**Ganancia**	**+TF**	**Ganancia**
OHSUMED	0,1547	0,1680	**8,60** %	0,1683	8,79 %
GH95	0,2060	0,2456	**19,23** %	0,2463	19,56 %
CFC	0,2936	0,2950	0,48 %	0,2944	0,27 %
	P@10				
	Hasta 6-gramas			**Hasta 3-gramas**	
	Baseline	**+TF**	**Ganancia**	**+TF**	**Ganancia**
OHSUMED	0,2460	0,2667	**8,41** %	0,2680	8,94 %
GH95	0,2217	0,2326	**4,92** %	0,2326	4,92 %
CF	0,4709	0,4759	1,06 %	0,4750	0,87 %

Tabla 4.1.7 –
Resultados obtenidos considerando todas las consultas.

Al analizar los resultados, podemos pensar que la dimensión de los documentos tuvo un papel destacado en este comportamiento. La colección GH95 está compuesta por documentos más largos comparados con el resto de las colecciones utilizadas, lo que podría traer como resultado una mayor probabilidad de ocurrencia; también podemos analizar en la colección *OHSUMED* que a pesar de que existe un gran número de términos frasales, los mismos se repiten en gran número de documentos, lo que podría traer consigo poco poder discrimitativo. Aún en el caso de la colección *CF* se obtuvo una ganancia mínima, a pesar del número reducido de términos frasales y que la evaluación de los documentos relevantes de cada consulta de esta colección está incompleta, por lo que se puede considerar que los resultados obtenidos en esta colección están condicionados también a factores independientes del poder discriminativo de los términos frasales. Podemos observar que la ganancia (o pérdida) en las métricas al utilizar términos frasales de 6-gramas no es significativa, debido a que no existe ninguna consulta que tenga algún término frasal mayor que 3-gramas, por tanto lo que está influenciando los resultados son otros factores como la norma de los documentos. De forma general, se puede observar que se obtuvieron ganancias al usar los términos frasales en la recuperación de documentos.

Cuando se consideran solo las consultas que contienen términos frasales, los resultados incrementan en hasta un 33,26 %. Presentamos en la Tabla 4.1.8 los valores obtenidos al considerar solo las consultas que contienen términos frasales. Como puede verse, las ganancias obtenidas fueron significativas, con hasta un 34,97 % en términos de *MAP* y 8,93 % de *P@10* en la colección *GH95* donde se obtuvieron los mayores valores. Lo que demuestra que con la adición de los términos frasales en la recuperación de documentos, se consiguen ganancias expresivas. Resultados estadísticamente significativos de acuerdo con la prueba de *McNemar* en las colecciones *OHSUMED* y *GH95* (marcado negrita).

Como se muestra en la Tabla 4.1.9 se realizó un análisis de cada consulta que posee términos frasales, para determinar si existe algún deterioro en las métricas utilizadas como consecuencia de su adición. Los resultados no fueron estadísticamente significativos cuando comparados con las consultas sin la adición de los términos frasales. Los resultados nos dan la idea de que las colecciones con consultas cortas, como es el caso de *GH95* son favorecidas al aumentar su información semántica adicionando términos frasales. De cualquier forma aunque existen consultas con términos frasales que deterioran las métricas su influencia es menor si la comparamos con los resultados globales.

	MAP		
	Baseline	**+TF**	**Ganancia**
OHSUMED	0,1535	0,1691	**10,16** %
GH95	0,2425	0,3273	**34,97** %
CF	0,3421	0,3454	0,96 %
	P@10		
	Baseline	**+TF**	**Ganancia**
OHSUMED	0,2446	0,2661	**8,79** %
GH95	0,2800	0,3050	**8,93** %
CF	0,5314	0,5514	3,76 %

Tabla 4.1.8 –
Resultados obtenidos considerando las consultas con términos frasales.

	No. Consultas que deterioran las métricas			
	MAP	**%**	**P@10**	**%**
OHSUMED	17	30,35	15	26,78
GH95	2	10	2	10
CF	14	14	7	7

Tabla 4.1.9 –
Número de consultas con términos frasales que deterioran las métricas.

Podemos notar que, al expandir un documento con la adición de términos frasales, hay un pequeño cambio en el modelo vectorial, en la norma de los documentos, ya que los términos frasales se agregan como términos adicionales con el consiguiente incremento de la norma. En la tabla 4.1.10 se cuantifica como el incremento de la norma afectó los resultados, obtenidos a partir de las consultas que no tienen términos frasales, los resultados en estas consultas solo se ven afectados por el cambio en la norma de los documentos.

	MAP			**P@10**		
	Baseline	**+TF**	**Ganancia**	**Baseline**	**+TF**	**Ganancia**
OHSUMED	0,1650	0,1593	-3,45 %	0,2571	0,2614	1,67 %
GH95	0,1779	0,1826	2,64 %	0,1769	0,1769	0 %
CF	0,2661	0,2665	0,15 %	0,4385	0,4354	-0,71 %

Tabla 4.1.10 –
Resultados obtenidos considerando las consultas sin términos frasales.

Los resultados en la Tabla 4.1.10 muestran que la adición de términos frasales tuvo un pequeño impacto en todas las bases de datos en términos de MAP, con una ganancia (o pérdida) de un máximo del 3,45 %, mientras que en términos de P @ 10 el impacto fue

aún menor. Lo que indica que, si bien la información adicional tuvo algún impacto en los resultados, la ganancia obtenida por el método se debió principalmente a los términos frasales existentes en algunas consultas y no a la información agregada a los documentos, afirmando la hipótesis del impacto positivo de la adición de términos frasales en la tarea de recuperación *ad hoc*.

4.2. Clasificación.

La clasificación de texto permite el agrupamiento de documentos semánticamente significativos, ayudando tanto a los usuarios como a las herramientas de recuperación de información, a localizarlos con mayor precisión. En esta sección se compara el rendimiento en términos de precisión de la clasificación. La idea es que cuando los términos frasales son utilizados en conjunción con características simples (palabras), la ambigüedad y ruido inherente de la representación *BoW* sean reducidos. Se utilizaron las colecciones de texto: *20 Newsgroups*, *DBpedia ontology classification*.

La colección *20 Newsgroups* ampliamente utilizada, reconocida por alta precisión cuando se utiliza el modelo *BoW* (BEKKERMAN; ALLAN, 2004), está compuesta por documentos que son mensajes enviados a grupos de noticias, acerca de temas como ciencia, religión, política, entre otros. La versión utilizada de la colección fue *bydate*[2], con un total de 18846 documentos ordenados por fecha, distribuidos entre 20 categorías. El número de documentos por clase varia de 628 a 999. Por otra parte no tiene documentos duplicados ni multi-etiquetados y los encabezamiento que identifican los grupos de noticias (*Xref, Newsgroup, Followup-to, Date*) fueron eliminados.

El conjunto de datos *DBpedia ontology classification*, (ZHANG; ZHAO; LECUN, 2015) fue construido seleccionando 14 clases no superpuestas de *DBpedia* 2014 (ontología de varios dominios derivada de la *Wikipédia*). De cada una de las 14 clases de la ontología, fueron elegidos al azar 40,000 muestras de entrenamiento y 5,000 muestras de prueba. El tamaño total del conjunto de datos de entrenamiento es de 560,000 y el conjunto de datos de prueba de 70,000. Los campos de esta colección contienen el título y el resumen de cada artículo de la *Wikipédia*, fue utilizado el resumen de cada artículo.

4.2.1. Técnicas de selección.

La selección de características es una técnica que se utiliza cuando solo se selecciona un subconjunto de características (en este caso, términos frasales) disponibles de la colección de datos y se especifican cuales son más relevantes para la tarea de clasificación. Debido a que la cantidad de términos frasales minerados a partir de los conjuntos de datos de

[2] http://qwone.com/ jason/20Newsgroups/

entrada puede ser potencialmente grande, se utilizaron técnicas filtro, para la selección de los términos frasales más discriminatorios, que trataran estas nuevas características como atributos individuales, sin ninguna transformación:

1. chi2 (X^2), para determinar la dependencia entre un término frasal f y una clase c;
$$X^2 = \frac{NX(AD - CB)^2}{(A + C)X(B + D)X(A + B)X(C + D)} \tag{4.1}$$
2. Información mutua (MI), mide el número de bits de información obtenidos para una predicción de clase c al conocer la presencia o ausencia de una característica f en un documento.
$$MI(f, c) = \frac{P(f)}{P(f)XP(c)} \approx \frac{AXN}{(A + C)X(A + B)} \tag{4.2}$$

Donde:

A es el número de veces en que f y c co-ocurren.

B es el número de veces que f ocurre sin c.

C es el número de veces que c ocurre sin f.

D es el número de veces que ni c ni f ocurren.

N es el número de documentos.

k es el número de clases.

4.2.2. Clasificadores empleados.

Para la comparación del rendimiento se utilizaron diferentes clasificadores lineales, aplicados exitosamente en problemas de alta dimensión con datos dispersos, como es el caso de la clasificación de texto.

1. **Descenso de Gradientes Estocástico *(SGD)***: enfoque simple pero muy eficiente para el aprendizaje discriminatorio de clasificadores lineales bajo funciones de pérdida convexa, en el esquema "*one versus all*" (OVA).
2. **Naive Bayes**: familia de algoritmos estadísticos, basados en el Teorema de Bayes, que calcula las probabilidades condicionales de la ocurrencia de dos eventos en función de las probabilidades de ocurrencia de cada evento individual.

3. **Ramdon Forest**: meta estimador que se ajusta a una serie de clasificadores de árboles de decisión en varias sub muestras del conjunto de datos y utiliza el promedio para mejorar la precisión predictiva y control del ajuste excesivo (*over-fitting*).

4. **SVM**: algoritmo de vanguardia, se basa en el principio de minimización del riesgo estructural de la teoría del aprendizaje computacional. La idea es seleccionar un hiper plano de separación que equidista de los ejemplos más cercanos de cada clase para, de esta forma, conseguir lo que se denomina un margen máximo a cada lado del hiper plano.

5. **NearestCentroid**: cada clase está representada por un centroide, las etiquetas de los ejemplos de prueba son asignadas a la clase con el centroide más cercano. Cuando se utiliza los vectores tf * idf para representar los documentos, se conoce como el clasificador de Rocchio.

6. **PassiveAgresive**: familia de algoritmos de aprendizaje en línea. Después de cada observación, el algoritmo predice un resultado. Una vez que el algoritmo ha hecho una predicción, recibe retroalimentación que indica el resultado correcto. Entonces, el algoritmo en línea puede modificar su mecanismo de predicción, probablemente mejorando las posibilidades de hacer una predicción precisa en rondas subsiguientes

7. **Ridge**: Utiliza la regresión de Ridge sobre el problema de mínimos cuadrados, imponiendo una penalización cuadrática al tamaño de los coeficientes.

4.2.3. Configuración del método de minería de términos frasales.

El limiar de soporte mínimo fue establecido en 10, aunque se mineraron términos con hasta 6 gramas para los experimentos (Tabla 4.2.1); se relacionan los resultados empleando solo términos frasales de 2-gramas en la colección *20 neswgroups*, porque en términos de orden superior se experimentó un deterioro mayor en los resultados de la clasificación.

Términos Frasales	**20 Newsgroups**	**DBpedia**
2-gramas	2494	40864
3-gramas	523	9500
4-gramas	107	4146
5-gramas	72	1756
6-gramas	43	1622
Total	3239	57888

Tabla 4.2.1 –
Términos frasales detectados en las colecciones.

4.2.4. Impacto de la adición de términos frasales en la clasificación de documentos.

En los experimentos en todas las colecciones se utilizó el modelo *BoW* como *baseline* y ninguna técnica de selección de características fue aplicada. Se utilizó además *cross validation = 5* en la colección *20 newsgroups*; luego se fueron adicionaron determinado número de los términos frasales mejor evaluados en conjunción a todos los unigramas (excepto *stop-words*) y fue observado su influencia sobre el rendimiento de la clasificación.

La Tabla 4.2.2 muestra los resultados de los clasificadores utilizando los parámetros con el mejor rendimiento; sin la adición de los términos frasales con el objetivo de identificar el mejor *baseline*, utilizando como métricas la precisión *micro-average F1* calculada sobre la precisión y revocación global y *macro-average F1* que es el promedio de los valores F1 para cada clase. *micro-average F1* tiende a estar dominado por el desempeño del clasificador en categorías comunes y *macro-average F1* están más influenciada por el desempeño en categorías raras (YANG; LIU et al., 1999). Ambas métricas ampliamente utilizadas para evaluar el rendimiento de los clasificadores.

Clasificadores	**20 newsgroups**		**DBpedia**	
	micro F1	**macro F1**	**micro F1**	**macro F1**
RidgeClassifier	0.78802	0.77801	0.96669	0.96662
PassiveAgressive	0.76514	0.75671	0.96794	0.96792
RandomForest	0.69437	0.67559	0.95920	0.95906
LinearSVC (l2)	0.78480	0.77563	0.97104	0.97102
SGDClassifier	0.78278	0.76918	0.95783	0.95766
NearestCentroid	0.68143	0.68702	0.88274	0.88699
Naive Bayes (ComplementNB)	0.79299	0.77669	0.91390	0.92055

Tabla 4.2.2 –
Precisión de clasificadores, usando modelo *BoW*

Como lo muestran los resultados, la mayor precisión se obtuvo en la colección *20 newsgroups* utilizando el clasificador *Naive Bayes*, empleado también en (MLADENIC; GROBELNIK, 1998; TESAR et al., 2006; ADI; ÇELEBI, 2014) en su versión *ComplementNB*. Mientras en la colección *DBpedia* el mayor resultado fue alcanzado por el clasificador *SVM* algoritmo estado del arte, ampliamente utilizado para la clasificación de documentos de texto. En la colección *20 newsgroups* algunos de los grupos de noticias (categorías) están fuertemente correlacionados (Tabla 4.2.3), lo que no ocurre en *DBpedia*, que con mayor cantidad de características discriminatorias, consigue diferenciar con mayor precisión los documentos de cada categoría, por lo que el clasificador consigue altos resultados en la clasificación.

comp.graphics comp.os.ms-windows.misc comp.sys.ibm.pc.hardware comp.sys.mac.hardware comp.windows.x	rec.autos rec.autos rec.sport.baseball rec.sport.hockey	sci.crypt sci.electronics sci.med sci.space
misc.forsale	talk.politics.misc talk.politics.guns talk.politics.mideast	talk.religion.misc alt.atheism soc.religion.christian

Tabla 4.2.3 – Categorías de la colección *20 newsgroups* agrupadas (aproximadamente) por temas

Es interesante entender como la adición de los términos frasales podrían ayudar a mejorar los modelos de diferentes clasificadores. Mediante la adición de estos términos discriminatorios se puede, por ejemplo, (i) incrementar la probabilidad de que un documento de prueba pertenezca a su verdadera clase, en el caso de *Naive Bayes* (ii) confirmar que algunas características compuestas son consideradas más discriminatorias que sus términos individuales, de acuerdo al peso asignada a ellas por *SVM*. En el caso de la colección *20 Newsgroups* el unigrama *screen* es común en las clases *comp.graphics* y *comp.os.ms.windows.mics*; en la clase *comp.graphics* aparece formando parte de términos frasales como: *LCD screen, virtual screen*, mientras en la clase *comp.os.ms.windows.mics* tenemos *windows screen, screen saver*, lo que puede ayudar a determinar la categoría del documento. El unigrama *screen* por si solo no tiene mucho significado, ya el caso de los términos frasales anteriores brindan mayor información discriminatoria.

4.2.4.1. *20 Newsgroups*.

La Figura 4.2.1 muestra los valores de *micro-F1* de las técnicas de selección de características en términos frasales de 2-gramas, utilizando el clasificador que obtuvo el mayor *baseline* (*Naive bayes*). El valor 0 en el eje X representa la situación cuando solo se utilizan los unigramas para la clasificación y este punto se entiende como *baseline* cuyo valor de referencia para este conjunto de datos es de 79,299 % , así como el último valor del eje de los términos frasales es el total de bigramas, por tanto sin ninguna técnica de selección (consulte la Tabla 4.2.4).

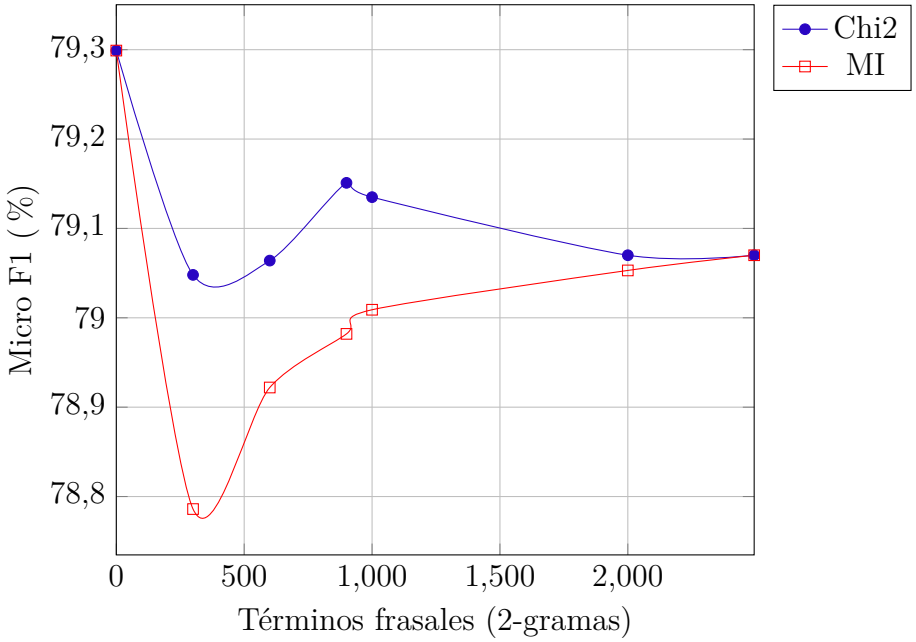

Figura 4.2.1 – Dependencia de micro-F1 y Términos Frasales de 2-gramas utilizados para la colección *20 Newsgroups*

Las técnicas de selección de características utilizadas muestran una disminución en el rendimiento de la clasificación, especialmente cuando se utiliza una cantidad menor de características mejor evaluadas. Una mirada más cercana indica que *MI* favorece, es decir, principalmente selecciona bigramas menos frecuentes, lo que no puede influir significativamente en el rendimiento de la clasificación y el caso de esta colección la aplicación de *MI* no fue significativa pues el mejor rendimiento se obtuvo con la adición de todos los términos frasales. Por otro lado *Chi2* asigna una evaluación alta a las características comunes, (YANG; PEDERSEN, 1997) que no parecen ser adecuadas para este conjunto de datos. Se realizaron experimentos extrayendo de los bigramas adicionados cierta cantidad de términos frasales mejor evaluados como en (TESAR et al., 2006),

consiguiéndose mejorías no significativas, y en ninguno de los casos se superó el *baseline*. La Tabla 4.2.4 resume los resultados de la Figura 4.2.1 para cada enfoque de selección de características.

Técnica selección	**Bigramas adicionados**	**Micro F1**	**Pérdida** %	**Macro F1**	**Pérdida** %
chi2	900	0.79151	-0,1866	0.77595	-0,0953
MI	2494	0.79070	-0,2888	0.77524	-0,1867
baseline	0	0.79299	0	0.77669	0

Tabla 4.2.4 –
Mejores resultados para términos frasales de 2-gramas con *Naive Bayes*.

Esta colección está formada por mensajes enviados a grupos de noticias, en su mayoría son textos cortos, por lo que aunque la dimensionalidad de la representación del texto es grande, los datos subyacentes son escasos. En otras palabras, el léxico del cual se extraen los documentos puede ser del orden de 10^5, pero un documento dado puede contener solo unos pocos cientos de palabras. Se caracteriza además porque cada palabra del corpus tiene importancia en la clasificación, los mejores resultados (*baseline*) se obtuvieron utilizando todos los unigramas con excepción de los *stop words*; por lo que aunque existen términos frasales de 2-gramas altamente discriminatorios (Tabla 4.2.5) capaces de mejorar los resultados de la clasificación, su contribución es débil en comparación con lo que cientos de unigramas pueden contribuir y su adición aumenta la varianza. Incluso los términos frasales que pueden tener una frecuencia más alta que sus componentes a menudo no ocurren con la frecuencia suficiente para hacer una gran diferencia. (KOSTER; SEUTTER, 2003) escriben: Incluso la selección de atributos más cuidadosa no puede superar las diferencias en la Frecuencia del Documento entre n-gramas y palabras. (LIU et al., 2003) consideran que aunque la mayoría de términos en este conjunto tienen valor discriminatorio para la clasificación, pocos términos por clase son suficientes para lograr una precisión aceptable, existiendo pocos términos ruidosos.

(BEKKERMAN; ALLAN, 2004) considera que en colecciones de referencia conocidas, como *20 Newsgroups*, el empleo de bigramas en las representación de documentos no ha reportado mejoras estadísticamente significativas; lo que puede estar dado porque los resultados alcanzados en esta colección son altos y probablemente no pueden mejorarse con ninguna técnica, debido a que todos los elementos clasificados incorrectamente están básicamente mal etiquetados.

comp.graphics	**alt.atheism**	**sci.space**
graphics library	strong atheism	space agency
image quality	religious beliefs	lunar surface
virtual reality	anti semitism	space center
rec.sport.hockey	**sci.crypt**	**sci.electronics**
hockey players	brute force	pc board
bob gainey	strong encryption	radio shack
eric lindros	private key	electrical engineering
rec.sport.baseball	**sci.med**	**soc.religion.christian**
white sox	infectious diseases	roman catholic
pitching staff	public health	christian faith
american league	health service	lord jesus
rec.autos	**comp.os.ms-windows.misc**	**talk.politics.guns**
sports car	dos window	gun control
fuel injection	ms windows	self defense
high speeds	windows nt	second amendment
rec.motorcycles	**comp.sys.ibm.pc.hardware**	**talk.politics.mideast**
speed limit	hard drive	armenian genocide
shaft drive	video card	turkish government
high performance	controller card	peace talks
comp.sys.mac.hardware	**comp.windows.x**	**mis.forsale**
mac lc	multi screen	money order
mac portable	windows manager	external modem
mac ii	share memory	buyer paids
talk.politics.misc	**talk.religion.misc**	
civil rigth	eternal life	
sales tax	moral code	
white house	jesus christ	

Tabla 4.2.5 – Ejemplo de términos frasales de 2-gramas extraídos por categorías de la colección *20 Newsgroups*.

En el caso de los términos frasales de mayor orden (hasta 6-gramas) a pesar de que están diseñados para capturar más información contextual y órdenes de palabras, aún tienen el problema de la escasez de datos, lo que afecta considerablemente la precisión de la clasificación. Los experimentos realizados con estos términos de mayor orden fue observado mayor decrecimiento en la precisión de la clasificación.

4.2.4.2. *DBpedia.*

La Figura 4.2.2 muestra los valores de *micro-F1* de las técnicas de selección en términos frasales de 2-gramas, utilizando el clasificador *SVM* con kernel lineal, ambas métricas tienen un comportamiento bien similar, a pesar de que evalúan las características

de modos diferentes. En esta colección a diferencia de la colección *20 newsgroups*, la precisión de la clasificación no decrece cuando los términos frasales son adicionados a la representación de los documentos, mejorando los resultados de la clasificación en menos de un 2 %. La Tabla 4.2.6 y 4.2.7 incluyen los mejores resultados del uso de bigramas y términos frasales de mayor orden. Los resultados de la Tabla 4.2.7 se construyeron utilizando los n-1 gramas mejor evaluados de la técnica con mayor rendimiento en la clasificación,(ejemplo: los términos de 3-gramas se adicionaron a los 30000 bigramas mejor evaluados por la técnica *Chi2*), la columna Ganancia n-1 gramas se refiere a la ganancia cuando se compara la adición de los términos frasales de orden superior con respecto a los términos del orden estrictamente inferior y se comparó el rendimiento del clasificador con estas características, pero como se puede notar a pesar de que la inclusión de términos frasales de mayor orden no deterioró los resultados, el incremento de la precisión a consecuencia de su adición es despreciable, menor que 0,02 %.

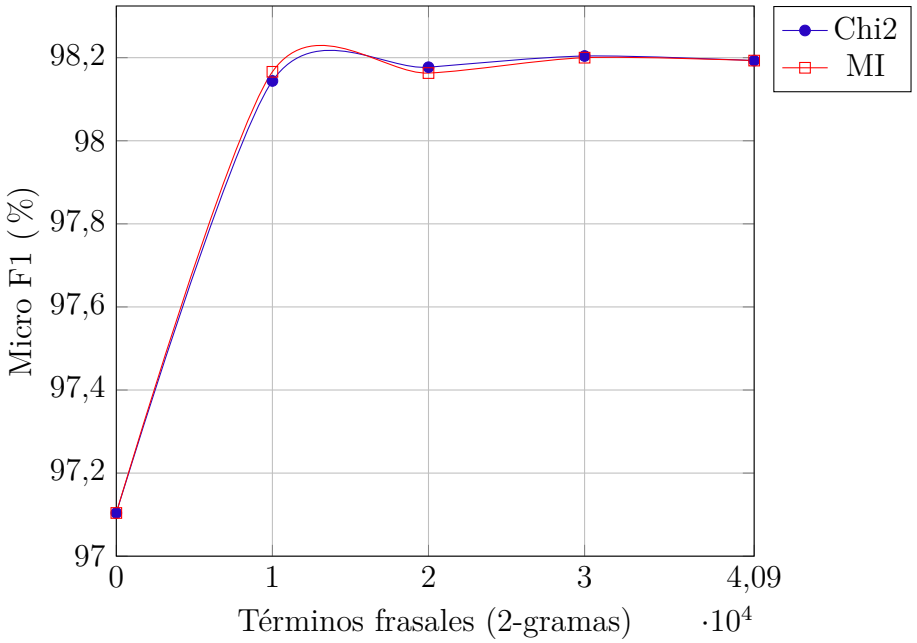

Figura 4.2.2 – Dependencia de micro-F1 y Términos frasales de 2-gramas utilizados para la colección *DBpedia*

Esta colección diferente del conjunto de datos *20 newsgroups* está 100 % balanceada, no existen categorías fuertemente relacionadas, la extensión de sus documentos es mayor y de similar longitud,por lo que existe una mayor probabilidad de distinguir la clase a la que pertenece cada documento, los términos frasales extraídos por categorías (Tabla 4.2.8)

Técnica selección	**Bigramas adicionados**	**Micro F1**	**Ganancia** %	**Macro F1**	**Ganancia** %
chi2	30000	0.98204	1.133	0.98203	1,132
MI	30000	0.98200	1,129	0.98198	1,128
baseline	0	0.97104	0	0.97102	0

Tabla 4.2.6 –
Mejores resultados para términos de 2-gramas, utilizando *SVM*

Técnica selección	**TF adicionados**	**Micro F1**	**Ganancia n-1 grama**	**Macro F1**	**Ganancia n-1 grama**
Trigramas					
chi2	3000	0.98221	0,0173	0.98220	0,0173
MI	6000	0.98216	0,0122	0.98214	0,0112
4gramas					
chi2	4146	0.98233	0,0122	0.98231	0,0111
MI	4146	0.98233	0,0122	0.98231	0,0111
5 -6gramas					
chi2	3378	0.98236	0,0030	0.98234	0,0030
MI	3378	0.98236	0,0030	0.98234	0,0030

Tabla 4.2.7 –
Mejores resultados para términos frasales de mayor orden, utilizando *SVM*

son altamente discriminatorios, lo que permite una mayor contribución en los resultados de la clasificación. Una dificultad podría radicar en que colecciones con *baseline* tan altos, es difícil de experimentar un alta ganancia, hasta porque pueden existir instancias mal rotuladas. En el caso de clasificadores con un *baseline* menor como el clasificador *Naive Bayes* la ganancia del uso de los términos frasales es solo un poco mayor (menor que un 2,3 % en todos los casos), como se muestra en la Tabla 4.2.9. Lo que puede indicar que la contribución de los términos frasales adicionados al modelo es baja cuando comparada ya a la información que ya posee cada documento en forma de unigramas.

SVM provee un método general que asigna pesos para cada característica, que se utilizó para investigar acerca de los términos frasales discriminatorios. Mediante estos pesos se puede ranquear las características y permite evaluar la importancia de los mismos. La Tabla 4.2.10 muestra algunos pesos de las características para la clase: *Company.*

Como lo muestran los resultados de la Tabla 4.2.10, entre las características más importantes para discriminar la clase, se encuentran tres términos frasales con mayor peso que las palabras que los componen, lo que nos indica que los términos frasales pueden

Company	**Educational Institution**	**Artist**
manufacturing company	private university	film actor
commercial bank	public university	visual artist
technology company	college preparatory	pop singer
Athlete	**Office Holder**	**Mean Of Transportation**
tennis player	national congress	cargo ship
baseball pitcher	republican politician	steam locomotive
basketball player	democratic politician	class submarine
Building	**NaturalPlace**	**Village**
pennsylvania house	national forest	southeastern brazil
supreme court	national park	baden württemberg
presbyterian church	mare river	bavaria germany
Animal	**Plant**	**Album**
coleophoridae family	bromeliad family	country music
rock snails	flowering plants	death metal
finned fish	family orchidaceae	hip hop
Film	**Written Work**	
crime film	romance novels	
american comedy	manga magazine	
bollywood film	comic book	

Tabla 4.2.8 – Ejemplo de términos frasales de 2-gramas extraídos de la colección *DBpedia* por categoría.

Técnica selección	**Bigramas adicionados**	**Micro F1**	**Ganancia %**	**Macro F1**	**Ganancia %**
chi2	20000	0.93369	2,165	0.93102	2,248
MI	30000	0.93323	2,115	0.93052	2,193
baseline	0	0.91390	0	0.91055	0

Tabla 4.2.9 – Mejores resultados para términos frasales de 2-gramas, utilizando *Naive Bayes*.

Característica	**Peso**	**Característica**	**Peso**
company	5.9784	retailer	3.1576
manufacturer	4.7581	brewery	2.9424
airline	4.4907	corporation	2.8275
label	3.9068	chain	2.7949
publisher	3.4503	publishing house	2.7686
record label	3.2201	film studio	2.6932

Tabla 4.2.10 – Características con los mayores pesos de la clase de entrenamiento *Company*

ser características discriminatorias. Pero en el caso de la clasificación no se obtuvieron resultados significativos y puede estar dado además de las características de la colección, del hecho que la adición de términos frasales aumenta la dispersión de los datos ya

naturalmente dispersos en clasificación de documentos de texto; también está el hecho de que a diferencia de la tarea de búsqueda, los documentos ya de por si tienen suficiente información en forma de unigramas que aportan un mayor peso en la clasificación.

4.3. Clusterización.

Una forma alternativa de presentar los resultados de sistemas de recuperación de información es organizarlos en grupos de interés. La clusterización de documentos de texto, proceso de agrupar de forma no supervisada un conjunto de documentos en clases semánticamente similares, es una vía eficaz de ayudar tanto a las personas como a sistemas automatizados a descubrir los documentos relevantes. (LIU; CROFT, 2004) demostraron que la recuperación de documentos basadas en clúster supera la efectividad de la recuperación basada en documentos tradicionales. La representación *BoW* utilizada para estos métodos de clusterización es a menudo insatisfactoria, ya que ignora las relaciones entre términos importantes, lo que generalmente entra en conflicto con la realidad. En orden a lidiar con este problema, de forma similar a la sección anterior proponemos adicionar a la representación tradicional un conjunto de términos frasales, con la expectativa que el uso de conceptos para retener la semántica entre palabras tenga como resultado que las categorías sean más distinguibles.

El método elegido para agrupar documentos fue *K-means* uno de los algoritmos de agrupamiento más populares; conocido por ser más eficiente que los algoritmos jerárquicos (AGGARWAL; ZHAI, 2012) en la clusterización de grandes conjuntos de datos. *K-means* almacena k centroides que utiliza para definir agrupamientos; se considera que un documento está en un clúster particular si está más cerca del centroide de ese clúster que cualquier otro centroide. Se utilizó la colección: *AG's news corpus*.

La colección *AG's news corpus* (ZHANG; ZHAO; LECUN, 2015) fue construida seleccionando las 4 clases más grande de la colección original *AG's corpus*: colección de más de un millón de noticias colectadas de diferentes fuentes. Cada clase contiene 30,000 muestras de entrenamiento y 1,900 muestras de prueba. El tamaño del conjunto de datos de entrenamiento es de 120,000 y el conjunto de datos de prueba de 7,600. Los campos de la colección contienen título y descripción.

4.3.1. Configuración del método de minería de Términos frasales y *K-means*

La Tabla 4.3.1 muestra el número de términos frasales minerados estableciendo el limiar de suporte mínimo en 10 y de hasta 6-gramas de longitud. (LIU et al., 2003) investigó el uso de la selección de características en el problema de la clusterización de

texto, demostrando que la selección de características puede mejorar su rendimiento y eficiencia, por lo que se utilizaron las técnicas supervisadas *Chi2* y *MI* empleadas en la clasificación ya que son conocidos los rótulos de cada instancia. Dado que la selección de los centroides iniciales influye grandemente en el algoritmo de agrupación *K-means*, se utilizó *K-means++* para seleccionar 20 conjuntos de centroides iniciales para el conjunto de datos y se promedio 20 veces el rendimiento de la agrupación final.

Términos Frasales	**AG's news corpus**
2-gramas	8231
3-gramas	1907
4-gramas	508
5-gramas	128
6-gramas	36
Total	10810

Tabla 4.3.1 –
Términos frasales detectados en la colección.

4.3.2. Impacto de la adición de términos frasales en la clusterización de documentos

Debido a que la representación del texto mediante el modelo *BoW* plantea el problema de alta dimensionalidad y dispersión inherente de los datos, que puede traer como consecuencia la disminución drástica del rendimiento de los algoritmos de clusterización; se utilizó el modelo *BoW* con la aplicación de las técnicas de selección para determinar el mejor *baseline*. Los resultados obtenidos de la Tabla 4.3.2 fueron resultado de la aplicación de la técnica *Chi2* con 50000 características mejor evaluadas. Luego se fueron adicionando los términos frasales mejor evaluados para determinar su influencia sobre el rendimiento de la clusterización. Se utilizó como métricas la precisión *micro-average F1*, *macro-average F1*, *completeness-score* que mide si todos los puntos de datos que son miembros de una clase determinada son elementos del mismo clúster y *homogeneity-score* si todos sus clústeres contienen solo puntos de datos que son miembros de una sola clase.

4.3.2.1. *AG's news corpus*

La Figura 4.3.1 muestra los valores *micro-F1* de las técnicas de selección aplicadas, usando términos frasales de hasta 2-gramas adicionados a los unigramas, en todos los casos

Colección	Micro F1	Macro F1	Homogeneidad	Completud
AG's news corpus	0,23214	0,30518	0,20892	0,37752

Tabla 4.3.2 –
Baseline de la colección utilizada.

se observó una disminución en las métricas empleadas. Al igual que en la colección *20 newsgroups* en la sección anterior, se extrajeron determinado número de términos frasales mejor evaluados, dado que el descenso en la precisión es mayor en los primeros términos mejor evaluados por ambas métricas, pero no se obtuvo mejorías cuando comprados al *baseline*. La Tabla 4.3.3 muestra todos los resultados obtenidos en la clusterización cuando se adicionaron los términos frasales mejor evaluados por las técnicas *Chi2* y *MI* a la representación tradicional de los documentos; donde Hom: Homogeneidad y Comp: Completud.

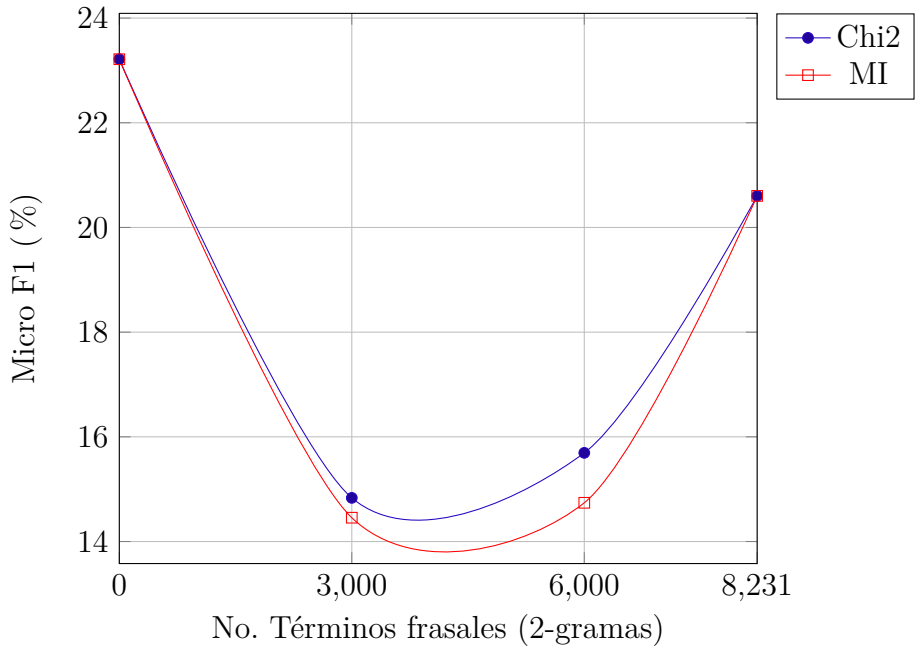

Figura 4.3.1 – Dependencia de micro-F1 y Términos Frasales de 2-gramas utilizados para la colección *AG's news corpus*

Los documentos de esta colección están formados por textos cortos desde 8 palabras hasta 196, con un promedio de 37.15 palabras por documentos. El mayor desafío en el manejo de documentos de texto cortos es precisamente lidiar con la escasez de palabras que contienen, que proporcionan muy pocas pistas contextuales para aplicar las técnicas tradicionales de extracción de datos . En este caso, las técnicas tradicionales para calcular

Técnica	**TF 2-gram**	**Micro F1**	**Gan %**	**Macro F1**	**Gan %**	**Hom**	**Gan %**	**Com**	**Gan %**
chi2	3000	0,1483	-36,10	0,2020	-33,79	0,1957	-6,32	0,3305	-12,45
MI		0,1445	-37,74	0,1907	-37,49	0,2413	15,51	0,3717	-1,54
chi2	6000	0,1569	-32,40	0,2113	-30,74	0,1904	-8,85	0,3239	-14,20
MI		0,1474	-36,49	0,1898	-37,79	0,2065	-1,15	0,3421	-9,38
TF	8231	0,2060	-11,24	0,2555	-16,25	0,2042	-2,25	0,3350	-11,26
baseline	0	0,2321	0	0,3051	0	0,2089	0	0,3775	0

Tabla 4.3.3 –
Mejores resultados para términos frasales de 2-gramas, utilizando *K-means*

la similitud del texto dan como resultado medidas que están muy cerca de cero, ya que los documentos, incluso los más parecidos, tienen muy pocos o casi ningún término en común. En investigaciones recientes con el objetivo de resolver el problema de la dispersión de los vectores de características se utilizan diferentes técnicas para expandir textos cortos a textos más largos, ya sea mediante el uso de fuentes de conocimiento externas como *Wikipédia, WordNet, HowNet*, resultados de búsqueda en la Web, otros bases de conocimiento construidas por el usuario; como el uso de métodos cuya idea fundamental es hacer uso de las relaciones entre términos para compensar la escasez de datos (SEIFZADEH et al., 2015; JIA et al., 2018).

Los términos frasales minerados guardan una relación semántica entre los términos que los componen trayendo información contextual. Pero contrario a la idea que la adición de información en forma de términos frasales discriminatorios (Tabla 4.3.4) debería contribuir a la mejor identificación del agrupamiento, no agregaron valor que contribuyera a mejorar el rendimiento final de la clusterización y aumentaron la dispersión de los datos.

World	**Sports**	**Business**	**Sci/Tech**
president bush	red sox	crude oil	chief executive
john kerry	world cup	economic growth	open source
united states	sports network	wall street	e mail

Tabla 4.3.4 – Ejemplo de términos frasales (2-gramas) extraídos por categorías de la colección *AG's news corpus*.

5 Conclusiones

Los experimentos se enfocaron en medir el impacto de la adición de términos frasales en las tareas de *Búsqueda ad hoc*, clasificación y clusterización. En la tarea de *Búsqueda ad hoc* se realizó una comparación con el tradicional modelo vectorial, obteniéndose una ganancia de hasta un 19,56 % y 8,60 % en la métrica *MAP* cuando son consideradas todas las consultas y de hasta un 34,97 % y 10.16 % cuando solo se utilizan las consultas que poseen términos frasales, en las colecciones *GH95* y *OHSUMED* respectivamente. De los experimentos realizados se pudo conocer que:

La extensión de los documentos con la adición de los términos frasales tuvo un pequeño impacto en términos de *MAP*, lo que indica que si bien la información adicionada a los documentos tuvo un impacto en los resultados, la ganancia obtenida se debió principalmente al enriquecimiento semántico de las consultas a través de términos frasales, que favoreció la identificación de los documentos relevantes a dicha consulta; Por lo que las colecciones con consultas cortas son mayormente beneficiadas con la adición de los términos frasales. De igual forma, entonces la adición de términos frasales de mayor orden (4 a 6 gramas) no resulta beneficioso cuando se trabaja con consultas de extensión media a corta. De forma general, podemos considerar que el uso de términos frasales es una solución sólida para mejorar la calidad de las tareas de búsqueda sin agregar una gran sobrecarga computacional, ya que la mayoría del procesamiento se realiza en el momento de la indexación.

En las tareas de clusterización y clasificación, a pesar de haber utilizado términos frasales mejor evaluados por las técnicas *Chi2* y *MI* extendiendo la representación del modelo *BoW*, los resultados no superaron los mejores resultados de los *baselines* obtenidos por el simple modelo *BoW* para las colecciones *20 newsgroups, AG's news corpus* y la ganancia obtenida en la colección *DBpedia* no fue significativa. Las colecciones utilizadas en su mayoría están formadas por documentos cortos, por lo que a pesar de que la dimensionalidad de la representación ser grande, los datos subyacentes son escasos. A pesar de que los términos frasales minerados guardan una relación semántica incorporando información contextual a los documentos capaces de mejorar la precisión, su contribución es débil en comparación con lo que cientos de unigramas pueden contribuir y su adición incrementa la varianza y dispersión de los datos.

Bibliografía

ABDULLAH, M.; ZAMIL, M. G. The effectiveness of classification on information retrieval system (case study). *arXiv preprint arXiv:1804.00566*, 2018. Citado en la página 2.

ADI, A. O.; ÇELEBI, E. Classification of 20 news group with naïve bayes classifier. In: IEEE. *2014 22nd Signal Processing and Communications Applications Conference (SIU)*. [S.l.], 2014. p. 2150–2153. Citado en la página 41.

AGGARWAL, C. C.; ZHAI, C. A survey of text clustering algorithms. In: *Mining text data*. [S.l.]: Springer, 2012. p. 77–128. Citado en la página 49.

AHONEN, H. Knowledge discovery in documents by extracting frequent word sequences. Graduate School of Library and Information Science. University of Illinois . . . , 1999. Citado en la página 6.

APTÉ, C.; DAMERAU, F.; WEISS, S. M. Automated learning of decision rules for text categorization. *ACM Transactions on Information Systems (TOIS)*, ACM, v. 12, n. 3, p. 233–251, 1994. Citado en la página 13.

BAEZA-YATES, R.; RIBEIRO-NETO, B. et al. *Modern information retrieval*. [S.l.]: ACM press New York, 1999. v. 463. Citado 4 vez(es) en la(s) página(s) 2, 4, 29 y 32.

BEKKERMAN, R.; ALLAN, J. *Using bigrams in text categorization*. [S.l.], 2004. Citado 3 vez(es) en la(s) página(s) 13, 37 y 43.

BERGSMA, S.; PITLER, E.; LIN, D. Creating robust supervised classifiers via web-scale n-gram data. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. [S.l.], 2010. p. 865–874. Citado 2 vez(es) en la(s) página(s) 6 y 8.

BLEI, D. M.; LAFFERTY, J. D. Visualizing topics with multi-word expressions. *arXiv preprint arXiv:0907.1013*, 2009. Citado en la página 9.

BLEI, D. M.; NG, A. Y.; JORDAN, M. I. Latent dirichlet allocation. *Journal of machine Learning research*, v. 3, n. Jan, p. 993–1022, 2003. Citado en la página 9.

BRAGA, I.; MONARD, M.; MATSUBARA, E. Combining unigrams and bigrams in semi-supervised text classification. In: CITESEER. *Proceedings of Progress in Artificial Intelligence, 14th Portuguese Conference on Artificial Intelligence (EPIA 2009), Aveiro*. [S.l.], 2009. p. 489–500. Citado en la página 13.

CAROPRESO, M. F.; MATWIN, S.; SEBASTIANI, F. A learner-independent evaluation of the usefulness of statistical phrases for automated text categorization. *Text databases and document management: Theory and practice*, Citeseer, v. 5478, p. 78–102, 2001. Citado 2 vez(es) en la(s) página(s) 6 y 13.

CARVALHO, A. L. da C.; MOURA, E. S. de; CALADO, P. Using statistical features to find phrasal terms in text collections. *Journal of Information and Data Management*, v. 1, n. 3, p. 583, 2010. Citado 3 vez(es) en la(s) página(s) 1, 2 y 10.

COVER, T. M.; THOMAS, J. A. *Elements of information theory*. [S.l.]: John Wiley & Sons, 2012. Citado en la página 3.

CRAWFORD, E.; KOPRINSKA, I.; PATRICK, J. Phrases and feature selection in e-mail classification. In: *ADCS*. [S.l.: s.n.], 2004. p. 59–62. Citado en la página 14.

DANILEVSKY, M. et al. Automatic construction and ranking of topical keyphrases on collections of short documents. In: SIAM. *Proceedings of the 2014 SIAM International Conference on Data Mining*. [S.l.], 2014. p. 398–406. Citado 2 vez(es) en la(s) página(s) 2 y 9.

D'HONDT, E. et al. Text representations for patent classification. *Computational Linguistics*, MIT Press, v. 39, n. 3, p. 755–775, 2013. Citado 2 vez(es) en la(s) página(s) 13 y 14.

DOUCET, A.; AHONEN-MYKA, H. Non-contiguous word sequences for information retrieval. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the Workshop on Multiword Expressions: Integrating Processing*. [S.l.], 2004. p. 88–95. Citado en la página 14.

EL-KISHKY, A. et al. Scalable topical phrase mining from text corpora. *Proceedings of the VLDB Endowment*, VLDB Endowment, v. 8, n. 3, p. 305–316, 2014. Citado 3 vez(es) en la(s) página(s) 2, 7 y 10.

FERNEDA, E. *Recuperação de informação: análise sobre a contribuição da ciência da computação para a ciência da informação*. Tese (Doutorado) — São Paulo: USP, 2003. Citado en la página 29.

FIGUEIREDO, F. et al. Word co-occurrence features for text classification. *Information Systems*, Elsevier, v. 36, n. 5, p. 843–858, 2011. Citado en la página 14.

FORMAN, G. Feature selection for text classification. *Computational methods of feature selection*, Chapman and Hall/CRC Press, v. 16, p. 257–274, 2007. Citado en la página 14.

JIA, C. et al. Concept decompositions for short text clustering by identifying word communities. *Pattern Recognition*, Elsevier, v. 76, p. 691–703, 2018. Citado en la página 52.

KOSTER, C. H.; SEUTTER, M. Taming wild phrases. In: SPRINGER. *European Conference on Information Retrieval*. [S.l.], 2003. p. 161–176. Citado en la página 43.

LAI, S. et al. Recurrent convolutional neural networks for text classification. In: *AAAI*. [S.l.: s.n.], 2015. v. 333, p. 2267–2273. Citado 2 vez(es) en la(s) página(s) 6 y 14.

LEWIS, D. D. An evaluation of phrasal and clustered representations on a text categorization task. In: ACM. *Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval*. [S.l.], 1992. p. 37–50. Citado 2 vez(es) en la(s) página(s) 2 y 13.

LINDSEY, R. V.; III, W. P. H.; STIPICEVIC, M. J. A phrase-discovering topic model using hierarchical pitman-yor processes. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. [S.l.], 2012. p. 214–222. Citado 2 vez(es) en la(s) página(s) 2 y 9.

LIU, J.; SHANG, J.; HAN, J. Phrase mining from massive text and its applications. *Synthesis Lectures on Data Mining and Knowledge Discovery*, Morgan & Claypool Publishers, v. 9, n. 1, p. 1–89, 2017. Citado en la página 1.

LIU, J. et al. Mining quality phrases from massive text corpora. In: ACM. *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*. [S.l.], 2015. p. 1729–1744. Citado 5 vez(es) en la(s) página(s) 2, 3, 7, 11 y 23.

LIU, T. et al. An evaluation on feature selection for text clustering. In: *Proceedings of the 20th international conference on machine learning (ICML-03)*. [S.l.: s.n.], 2003. p. 488–495. Citado 2 vez(es) en la(s) página(s) 43 y 49.

LIU, X.; CROFT, W. B. Cluster-based retrieval using language models. In: ACM. *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*. [S.l.], 2004. p. 186–193. Citado en la página 49.

MCDONALD, R.; CRAMMER, K.; PEREIRA, F. Online large-margin training of dependency parsers. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the 43rd annual meeting on association for computational linguistics*. [S.l.], 2005. p. 91–98. Citado en la página 6.

MLADENIC, D.; GROBELNIK, M. Word sequences as features in text-learning. In: CITESEER. *In Proceedings of the 17th Electrotechnical and Computer Science Conference (ERK98*. [S.l.], 1998. Citado 2 vez(es) en la(s) página(s) 13 y 41.

ÖZGÜR, L.; GÜNGÖR, T. Optimization of dependency and pruning usage in text classification. *Pattern analysis and applications*, Springer, v. 15, n. 1, p. 45–58, 2012. Citado en la página 14.

SEIFZADEH, S. et al. Short-text clustering using statistical semantics. In: ACM. *Proceedings of the 24th International Conference on World Wide Web*. [S.l.], 2015. p. 805–810. Citado en la página 52.

SHANG, J. et al. Automated phrase mining from massive text corpora. *IEEE Transactions on Knowledge and Data Engineering*, IEEE, v. 30, n. 10, p. 1825–1837, 2018. Citado 6 vez(es) en la(s) página(s) 2, 3, 4, 7, 12 y 31.

TAN, C.-M.; WANG, Y.-F.; LEE, C.-D. The use of bigrams to enhance text categorization. *Information processing & management*, Elsevier, v. 38, n. 4, p. 529–546, 2002. Citado en la página 13.

TESAR, R. et al. Extending the single words-based document model: a comparison of bigrams and 2-itemsets. In: ACM. *Proceedings of the 2006 ACM symposium on Document engineering*. [S.l.], 2006. p. 138–146. Citado 3 vez(es) en la(s) página(s) 14, 41 y 42.

TUYET, H. N. T.; HANH, T. Maximal frequent sequences for document classification. In: IEEE. *Advanced Technologies for Communications (ATC), 2016 International Conference on*. [S.l.], 2016. p. 152–157. Citado 2 vez(es) en la(s) página(s) 2 y 15.

WANG, X.; MCCALLUM, A.; WEI, X. Topical n-grams: Phrase and topic discovery, with an application to information retrieval. In: IEEE. *Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on*. [S.l.], 2007. p. 697–702. Citado en la página 9.

YANG, Y.; LIU, X. et al. A re-examination of text categorization methods. In: *Sigir*. [S.l.: s.n.], 1999. v. 99, n. 8, p. 99. Citado en la página 40.

YANG, Y.; PEDERSEN, J. O. A comparative study on feature selection in text categorization. In: *Icml*. [S.l.: s.n.], 1997. v. 97, n. 412-420, p. 35. Citado en la página 42.

ZHANG, W. et al. Recognition and classification of noun phrases in queries for effective retrieval. In: ACM. *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*. [S.l.], 2007. p. 711–720. Citado 3 vez(es) en la(s) página(s) 2, 6 y 7.

ZHANG, X.; ZHAO, J.; LECUN, Y. Character-level convolutional networks for text classification. In: *Advances in neural information processing systems*. [S.l.: s.n.], 2015. p. 649–657. Citado 2 vez(es) en la(s) página(s) 37 y 49.