**PROPUESTA DE TRABAJO FINAL o PROYECTO DE TESIS MAESTRÍA**

La propuesta o proyecto debe contener el estado del arte del tema de investigación y la evidencia clara de la posibilidad de realizar una contribución a la solución de problema(s) complejo(s) en el área del programa.

**TITULO:** Extracción de frases clave de contenido web para facturación electrónica en Colombia

DIRECTOR: Gabriel Awad

**PROGRAMA:** Maestría en INGENIERÍA DE SISTEMAS **PERFIL:** Profundización

**AUTOR**

**Pedro Javier Porras Correa 1055315078** [**pporrasc@unal.edu.co**](mailto:pporrasc@unal.edu.co) **3212961841**

1. **PLANTEAMIENTO DEL PROBLEMA**

Las palabras claves son uno de los principales elementos de SEO y seleccionar las palabras inadecuadas o ineficientes puede ser perjudicial en los resultados de búsqueda para una página web, por esto es indispensable obtener las palabras claves más relevantes. (Harikumar, 2018) esto indica que se debe realizar un proceso de investigación de palabras clave en donde se debe obtener un entendimiento claro del contenido para lograr extraer las palabras más relevantes.

Sin embargo, los usuarios de motores de búsqueda como Google tienen diferentes intenciones al escribir las consultas (Mohammadi et al., 2020) debido a esto es importante tener en cuenta los métodos y tener claro el público objetivo al que se pretende llegar, para esto se puede hacer uso de las consultas de búsquedas que se generan y los resultados que se muestran a partir de estas.

Al tomar en cuenta lo anterior se entiende porque al realizar el estudio sobre las palabras claves extrayendo las características a partir del contenido que se tiene no presenta los mismos resultados previstos a los que genera los resultados de búsqueda dado las consultas que se ingresan.

Debido a esto, si se re requiere hacer un análisis de las palabras destacadas para el tema de facturación electrónica en Colombia con el propósito de construir contenido relevante y que tenga buena visibilidad en los resultados de búsqueda, se encuentra que el proceso para generar estas palabras destacadas resulta en un proceso extenso y tedioso que depende en gran medida de la experiencia de la persona que lo realiza.

Debido a esto en la empresa Netmedik se tiene el problema donde no se conoce cuales son los temas principales que aplican para facturación electrónica en Colombia. Gracias a que no se tiene la base de referencia de los temas principales que surgen para facturación electrónica en Colombia se dificulta el poder desarrollar contenido que enfocado y efectivo que aborde los temas principales que presentan las paginas actuales.

Netmedik es una empresa de desarrollo de software a la medida, realizando implementaciones de páginas web y sitios web transacciones en plataformas como WordPress. Dentro de una de sus implementaciones se encuentra utilfy el cual es una plataforma para facturación electrónica en Colombia. En respuesta al contenido que se encuentra sobre facturación electrónica en Colombia surgió la necesidad de recolectar cuales son los temas principales para enfocar los esfuerzos en los temas más relevantes.

Como consecuencia, no se logra obtener un marco de referencia sobre los temas relevantes sobre los que se puede trabajar en facturación electrónica en Colombia, lo cual resulta en generar dificultades al momento de construir contenido sobre facturación electrónica, provocando que el contenido que se construya no se encuentre enfocado y que no sea relevante.

Esto significa que al no ser relevante y no estar claro los temas a abordar, la posición en los resultados de búsqueda se verá afectados disminuyendo la visibilidad del contenido y limitando el alcance del contenido que se construye.

1. **JUSTIFICACIÓN**

Con el objetivo de mejorar los resultados y disminuir el proceso de investigación y construcción de palabras clave para el contenido de facturación electrónica de Colombia se propone abordar el problema de extracción de frases clave para facturación electrónica de Colombia haciendo uso de modelos de extracción de frases clave haciendo uso de las consultas de búsqueda y los resultados que se obtienen para esa consulta.

Esto permitirá generar una cantidad mayor de frases claves para facturación electrónica de Colombia, el cual lograr presentar un mejor rendimiento al generar una mayor visibilidad en los resultados de búsqueda a partir de las consultas de búsqueda generadas por los usuarios del motor de búsqueda de Google.

La extracción de frases claves de las frases clave de facturación electrónica en Colombia es una tarea básica y fundamental para construir contenido relevante y enfocado, cuando se realiza una extracción de las frases claves del contenido que se encuentra mediante las consultas en el buscador se logra obtener los principales temas que están siendo buscados.

Esto permite generar contenido relevante y enfocado lo que resulta en mejorar el rendimiento del contenido que se construye al lograr identificar los temas más relevantes con base a las palabras claves que se pueden extraer a partir de las páginas web que se muestran en las primeras posiciones en los resultados de búsqueda.

Como resultado se extraerán los principales temas a partir de los resultados de búsqueda, esto permite obtener los temas más relevantes lo cual permitirá enfocar la producción de contenido enfocándolo en los temas obtenidos. De esta forma de lograr dar las pautas y el tema hacia el que se debe enfocar al momento de construir las piezas de contenido, lo cual tendrá como ventaja optimizar los recursos que se invierten para el proceso de investigación sobre el tema que se va a trabajar y mejorara la variedad, calidad y cantidad de frases claves que se pueden obtener (Thomaidou & Vazirgiannis, 2011)

1. **OBJETIVOS**
   1. **OBJETIVO GENERAL**

Desarrollar una herramienta para la extracción de frases claves de facturación electrónica en Colombia haciendo uso de los modelos disponibles.

* 1. **OBJETIVOS ESPECÍFICOS**

### 

* Construir el conjunto de datos basado en el contenido web donde se use el tema de facturación electrónica en Colombia.
* Analizar el conjunto de datos construido.
* Comparar los resultados del conjunto de datos de prueba sobre los modelos disponibles.
* Definir el modelo que se va a usar para extraer las frases claves con todo el conjunto de datos
* Seleccionar las frases claves más relevantes para facturación electrónica en Colombia.

1. **MARCO TEÓRICO**

Para el desarrollo de la tarea de extracción de frases claves se han desarrollados diversas herramientas y modelos los cuales hacen uso de las características del texto. Ya sea que hagan uso de técnicas de aprendizaje supervisado o no supervisado en donde cada una de estas técnicas tiene modelos específicos en base al método o características sobre las que se enfoca el modelo.

APRENDIZAJE SUPERVISADO

MODELOS BASADO EN EL DISEÑO DE CARACTERÍSTICAS

KEA es un algoritmo de extracción de frases clave supervisado orientado en la extracción sobre un dominio en específico, el algoritmo de KEA se plantea como un algoritmo sencillo haciendo uso de técnicas de aprendizaje supervisado y atributos como tf\*idf y la distancia, basado en los resultados que se obtienen del estudio se presenta que el rendimiento del modelo de extracción de frases claves depende de que tan relacionado al dominio se encuentran las entradas sobre las que se van a extraer las frases claves (Wu et al., 2005)

WINGNUS es un algoritmo de extracción de frases clave supervisado en el que se propone el uso de extracción de estructura lógica, esto se define como una jerarquía de componentes lógicos como son los títulos, autores etc. De esta forma se hace uso de las propiedades que se pueden extraer a partir de la jerarquía del documento junto con propiedades estadísticas (Nguyen & Luong, 2010)

APRENDIZAJE NO SUPERVISADO

MODELOS BASADOS EN ESTADÍSTICAS

El sistema de extracción de frases claves de KPMiner hace uso de reglas heurísticas para seleccionar los candidatos, para el cálculo de las propiedades de los candidatos hace uso de un algoritmo que usa se basa en tf\*idf , un factor impulsador relacionado al documento y la posición de la frase. KPMiner está orientado a ser un extractor de frases claves general en lugar de un extractor de frases claves orientado a un dominio en específico. KPMiner está orientado principalmente al idioma ingles y árabe (El-Beltagy & Rafea, 2009)

Yake presenta un algoritmo de extracción de frases claves no supervisado basado en las características estadísticas de los documentos sobre los que se aplica. A diferencia de otros modelos de extracción de características que abordan el problema como una tarea de clasificación binaria, en donde una frase puede ser o no una frase clave, YAKE obtiene una lista de frases claves y calcula la importancia de cada una (Campos et al., 2018)

MODELOS BASADOS EN GRAFOS

Dentro de los modelos de extracción de frases clave no supervisados se encuentra TextRank el cual es un modelo de clasificación de frases claves basado en grafos, el algoritmo que se muestra en este modelo define que la importancia de los nodos que esta dado por la cantidad de votos que sean emitidos a este. Esto también determina que tan importante es su voto, similar al algoritmo de pagerank de Google en donde un nodo que tenga más relevancia va a tener mayor impacto en la importancia de los nodos con los que se encuentra conectado (Mihalcea & Tarau, 2004)

CollabRank es un algoritmo de extracción de frases claves no supervisado basado en grafos, la principal característica que se propone es el uso de algoritmos de agrupación para generar unos pocos grupos de documentos, con esto se entiende que los documentos van a tener temas relacionados entre cada uno y con esto el grupo se considera como el contexto para cada documento (Wan & Xiao, 2008)

TopicRank funciona como un algoritmo de extracción de frases clave no supervisado basado en grafos, este algoritmo extrae las frases claves basado en los temas más importantes del documento, para esto hace uso del modelo de TextRank para asignar puntajes a los tópicos para posteriormente seleccionar las frases claves (Bougouin et al., 2013)

Dentro del mismo grupo de modelos de extracción no supervisados se encuentra PositionRank el cual es un algoritmo de extracción de frases clave no supervisado basado en grafos, este incorpora información de todas las posiciones de las ocurrencias de una palabra sobre un modelo de PageRank sesgado el cual va a calificar palabras claves que luego serán usados para clasificar palabras claves en trabajos de identificación. Este modelo muestra una diferencia a comparación de otros modelos al hacer uso de características que son implementadas en modelos de extracción de frases clave supervisados (Florescu & Caragea, 2017)

* 1. **ANTECEDENTES**

En esta sección se va resumir algunos de los métodos propuestos para abordar la tarea de extracción de frases clave a partir de fuentes de información como lo son los registros de consultas y los resultados generados los motores de búsqueda y la información que se puede extraer a partir de la información de la página web y la estructura de lo datos que se definen dentro del contenido web.

METODO DE APRENDIZAJE NO SUPERVISADO

Dentro de los métodos supervisados encontramos el método propuesto por (Talwar & Upadhyaya, 2017) en donde se presenta un sistema de sugerencia de palabras clave de cola larga para anuncios de búsqueda patrocinados. Sobre este trabajo propone el modelo LightLDA el cual es un modelo de temas, el cual presenta mejoras al tener una menor complejidad y disminuir el tiempo de ejecución del modelo propuesto. Este sistema se separa en dos pasos, la generación de los candidatos y la optimización del modelado de temas.

(Qiao et al., 2017) propone un método de sugerencia de palabras clave basado en registros de consultas haciendo uso de asociaciones indirectas entre palabras clave y la información oculta del tema capturada por los registros de consultas para recomendar palabras clave competitivas. El modelo propuesto aprovecha los registros de las consultas para extraer palabras clave dado una palabra, este modelo también propone dos pasos, el primer paso es hacer un análisis basado en la asociación indirecta, el segundo paso es la sugerencia de palabras clave competitivas basado en temas.

(Joshi & Motwani, 2006) propone una nueva técnica al cual llama TermsNet. Basando la utilidad de esta técnica al reconocer que las técnicas existentes no tienen en cuenta las relaciones semánticas y en los términos no comunes que no contienen el término de la consulta de entrada a menudo son ignoradas. Basado en esto presenta que TermsNet captura la relación semántica entre las palabras y que no requiere registros de consultas. Con base a lo anterior se explica el funcionamiento de TermsNet en aprovechar los motores de búsqueda para determinar la relevancia entre los términos y capturar la relación semántica.

METODO DE APRENDIZAJE SUPERVISADO

Dentro de los métodos supervisado (Mohammadi et al., 2020) presenta un método en donde se pueden identificar las palabras claves dentro del dominio general, para esta tarea propone abordarlo como un problema de detección de intención a partir de las consultas generadas en el buscador de Google. El modelo que propone seguiría el proceso de extracción de datos a partir de los resultados de búsqueda de Google y a partir de esta información se extraen las características con el propósito de agrupar las intenciones que se obtienen de este proceso. De esta forma el modelo le permite detectar automáticamente la intención de una consulta usando las palabras claves extraídas de los grupos construidos.

Dentro de las especificaciones de este modelo se especifica que se debe etiquetar y generar reglas hechas a mano para extraer las características de los resultados de búsqueda y agrupar las consultas basados en estos resultados. Con base en esta información el modelo aprende a identificar las intenciones de la nueva información basado en el entrenamiento supervisado.

Otro método supervisado basado en las características del documento es el trabajo presentado en (Yih et al., 2006) en donde también se hace necesario extraer las frases claves de las páginas que son indexadas por el buscador, para esto describe un sistema que aprende a extraer las palabras clave de las páginas web con el propósito de ser usadas en campañas de publicidad. En este sistema se comparan diferentes acercamientos para abordar la tarea de extracción de frases clave, en este se exponen métodos tradicionales como tf\*idf, sistemas como KEA y el método propuesto por ellos en donde se usan características monolíticas.

Por otro lado (Zhou et al., 2007) propone abordar la tarea como un problema de recuperación de información. La perspectiva que propone se enfoca en prestar más atención a la estructura de los datos de la página web y la publicidad, basados en esta estructura de datos realiza una evaluación de similitud entre la página web y la publicidad con lo que hace más fácil seleccionar el tipo de publicidad para la página web.

1. **METODOLOGÍA A UTILIZAR**

Para empezar el desarrollo de la propuesta se inicia con la investigación de sobre las lecturas sobre la tarea de la extracción de frases clave en el área de Procesamiento del lenguaje natural, esto permitirá organizar y presentar el problema de investigación en donde se pueda exponer cual es el problema que se quiere resolver y él porque es necesario resolverlo.

De esta forma dentro de la investigación de la literatura sobre extracción de frases claves se debe profundizar en los principales métodos de aprendizaje utilizados y los modelos disponibles para abordar la tarea de extracción de frases clave. El realizar esta investigación presentara una lista de los modelos disponibles sobre los cuales se va a investigar cuales son las ventajas y desventajas de cada uno y sobre que problemas y dominios son aplicables. Una vez se realiza la profundización sobre los tipos de aprendizaje se debe enfocar la investigación sobre la extracción de frases clave de páginas web para publicidad.

La etapa de recolección se realizará por medio del desarrollo de scripts en Python con el objetivo de automatizar y distribuir las tareas en donde lo primero será extracción de la información, para esto se va a hacer uso técnicas de web crawling para descargar el contenido de las páginas web que se obtienen a partir de los resultados del motor de búsqueda de Google sobre facturación electrónica en Colombia.

La información que se obtuvo previamente se procesara para normalizar el texto que se obtiene y eliminar el ruido que se puede descargar del contenido web como lo son la información de la página o texto no relacionado con facturación electrónica en Colombia.

La información que se tiene debe ser etiquetada, esto se va a realizar a partir de las consultas de búsquedas que fueron usadas para obtener los documentos. Con base en esta información se analizan las oraciones tokenizando las oraciones, luego se procede a limpiar las oraciones eliminando las palabras de parada, los símbolos y los números.

Una vez el texto se encuentra procesado se procede a generar los uni-gramas, bi-gramas y tri-gramas de los cuales se seleccionará el que presente mayor frecuencia entre los documentos a los que aplica la consulta de búsqueda.

Una vez que se tiene la información procesada se utilizara una fracción para probar y seleccionar el modelo que presente mejor rendimiento sobre el dominio de facturación electrónica en Colombia. Este modelo seleccionado va a continuar sobre las siguientes etapas en donde se realizará la selección de frases clave. En este proceso también se debe realizar la extracción usando el modelo de tf\*idf

el cuál es el modelo base sobre el cual se decidirá si los resultados obtenidos sobre el modelo seleccionado son positivos o negativos y que tan eficiente es sobre la línea base.

Para poder evaluar los resultados que se obtienen de extraer las frases claves del contenido web que se encuentra a partir del motor de búsqueda de Google se obtendrán los resultados obtenidos por los diferentes sistemas, a partir de estos se seleccionará el que presenta el mejor desempeño y se procederá a comparar con la línea base.

Después de que se obtiene la línea base y el modelo o sistema seleccionado se puede interpretar el rendimiento del modelo seleccionado sobre la línea base, esto apoyara o negara la teoría de que se puede tener mejores resultados al hacer uso del modelo seleccionado sobre el uso de tf\*idf.

Cuando ya se entiende el rendimiento ya sea positivo o negativo se puede exponer por qué y la magnitud de la diferencia donde se puede apoyar o negar él porque es beneficioso usar este nuevo modelo.

1. **ALCANCES DEL TRABAJO**

Como resultado del trabajo se obtendrá los siguientes elementos

• Se completará la tesis de trabajo.

• S construirá un artículo para revista.

• La base de datos utilizada durante el desarrollo del trabajo.

• Las características y los patrones encontrados a partir de la base de datos.

• El resultado del entrenamiento de los modelos sobre los datos de prueba.

• La lista de frases clave más relevantes que han sido encontradas a partir del desarrollo del trabajo.

1. **PLAN DE TEMAS Y CRONOGRAMA**

Es un plan de trabajo o un plan de actividades, que muestra en un orden lógico y secuencial la duración del proceso investigativo, en una forma gráfica o de tabla.

1. **BIBLIOGRAFÍA Y FUENTES DE INFORMACIÓN**

Bougouin, A., Boudin, F., & Daille, B. (2013). Topicrank: Graph-based topic ranking for keyphrase extraction. *International joint conference on natural language processing (IJCNLP)*, 543-551.

Campos, R., Mangaravite, V., Pasquali, A., Jorge, A. M., Nunes, C., & Jatowt, A. (2018). YAKE! collection-independent automatic keyword extractor. *European Conference on Information Retrieval*, 806-810.

El-Beltagy, S. R., & Rafea, A. (2009). KP-Miner: A keyphrase extraction system for English and Arabic documents. *Information systems*, *34*(1), 132-144.

Florescu, C., & Caragea, C. (2017). Positionrank: An unsupervised approach to keyphrase extraction from scholarly documents. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1105-1115.

Harikumar, C. (2018). Keyword analysis: Using text analysis for search engine optimization. *International Journal of Pure and Applied Mathematics*, *118*(5), 565-571.

Joshi, A., & Motwani, R. (2006). Keyword generation for search engine advertising. *Sixth IEEE International Conference on Data Mining-Workshops (ICDMW’06)*, 490-496.

Mihalcea, R., & Tarau, P. (2004). TextRank: Bringing Order into Text. *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, 404-411. https://www.aclweb.org/anthology/W04-3252

Mohammadi, S., Chapon, M., & Fremond, A. (2020). Query Intent Detection from the SEO Perspective. *arXiv:2006.09119 [cs]*. http://arxiv.org/abs/2006.09119

Nguyen, T. D., & Luong, M.-T. (2010). WINGNUS: Keyphrase extraction utilizing document logical structure. *Proceedings of the 5th international workshop on semantic evaluation*, 166-169.

Qiao, D., Zhang, J., Wei, Q., & Chen, G. (2017). Finding competitive keywords from query logs to enhance search engine advertising. *Information & Management*, *54*(4), 531-543.

Talwar, M. R., & Upadhyaya, S. (2017). *Long Tail Keyword Suggestion for Sponsored Search Advertising*.

Wan, X., & Xiao, J. (2008). CollabRank: Towards a collaborative approach to single-document keyphrase extraction. *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, 969-976.

Wu, Y. B., Li, Q., Bot, R. S., & Chen, X. (2005). Domain-specific keyphrase extraction. *Proceedings of the 14th ACM international conference on Information and knowledge management*, 283-284.

Yih, W., Goodman, J., & Carvalho, V. R. (2006). Finding advertising keywords on web pages. *Proceedings of the 15th international conference on World Wide Web*, 213-222.

Zhou, N., Wu, J., & Zhang, S. (2007). A keyword extraction based model for web advertisement. En *Integration and Innovation Orient to E-Society Volume 2* (pp. 168-175). Springer.