ADAPTACIÓN DE UN MODELO DE ESPACIO VECTORIAL DE RECUPERACIÓN DE INFORMACIÓN A TEXTOS ESCRITOS EN NASA YUWE

Tesis de Maestría
Ciclo de Formación I Doctorado en Ingeniería Telemática

Mag. Luz Marina Sierra Martínez

Director: Ph.D. Juan Carlos Corrales Muñoz
Codirector: Ph.D. Carlos Alberto Cobos
Asesor: Ph.D. Tulio Rojas Curieux

Universidad del Cauca
Facultad de Ingeniería Electrónica y Telecomunicaciones
Maestría en Ingeniería Telemática
Popayán, 4 de febrero de 2016

ADAPTACIÓN DE UN MODELO DE ESPACIO VECTORIAL DE RECUPERACIÓN DE INFORMACIÓN A TEXTOS ESCRITOS EN NASA YUWE

Mag. Luz Marina Sierra Martínez

Tesis presentada a la Facultad de Ingeniería Electrónica y Telecomunicaciones de la Universidad del Cauca para la obtención del título de Magister en Ingeniería Telemática
Ciclo de Formación I Doctorado en Ingeniería Telemática

Director: Ph.D. Juan Carlos Corrales Muñoz
Codirector: Ph.D. Carlos Alberto Cobos
Asesor: Ph.D. Tulio Rojas Curieux

Universidad del Cauca
Facultad de Ingeniería Electrónica y Telecomunicaciones
Maestría en Ingeniería Telemática
Popayán, 4 de febrero de 2016

Nota de aceptación

_____

_____

_____

_____
Tutor

_____
Presidente del Jurado

_____
Jurado

Popayán, 4 de febrero de 2016

A mi madre,
Mis hermanos,
Luis,
Muñeco,
Familia y amigos

AGRADECIMIENTOS

La autora expresa sus agradecimientos a:

- Mi madre, hermanos y familia quienes me han apoyado en todos mis proyectos.

- Los profesores Juan Carlos Corrales, Carlos Alberto Cobos y Tulio Enrique Rojas, por todo el apoyo, enseñanzas y sus valiosos aportes para el desarrollo de este proyecto que es un paso muy importante para mi crecimiento y formación profesional.

- Departamento de Sistemas y la Universidad del Cauca por la oportunidad que me han brindado para participar de este programa.

- Los profesores y personas de la Comunidad Nasa de los diferentes resguardos que han participado en esta experiencia; gracias a su colaboración este trabajo se ha podido concluir adecuadamente.

CONTENIDO

LISTA DE ECUACIONES

Pág.

LISTA DE TABLAS

Pág.

LISTA DE FIGURAS

Pág.

LISTA DE ANEXOS

Pág.

GLOSARIO

ANALIZADOR LÉXICO (TOKENIZER): Es el algoritmo que se encarga de dividir el texto en unidades pequeñas o tokens [1], la opción más fácil para obtener una lista de tokens de textos en lenguajes europeos es segmentar utilizando caracteres de espacio (espacios, tabulador, entre otros) [2].

COLECCIÓN DE PRUEBA: Es una herramienta para evaluar la efectividad de la recuperación de un sistema de recuperación de información consta de tres componentes: 1-Una colección de documentos; 2- Un conjunto de prueba de necesidades de información, expresadas como consultas; 3- Un conjunto de juicios de relevancias, evaluados generalmente de manera binaria de si el par documento-consulta es relevante o no [3], estas colecciones se utilizan para.

COMUNIDAD NASA (PÁEZ): Es un pueblo indígena de Colombia, es el segundo del país en cuanto a su población absoluta y número de hablantes de su lengua el nasa yuwe, son alrededor de 200.000 personas distribuidas en 7 departamentos [4].

FUNCIÓN DE RANKING: Es el corazón de cualquier motor de búsqueda enfocado a determinar qué tan relevante un documento es para una consulta, obteniendo una lista ordenada de todos relevantes para una consulta específica [5]

LUCENE: Es una API de código abierto para recuperación de información, que contiene diversas funciones para indexar documentos, realizar su búsqueda, entre otras cosas [6].

MEDIDA F: es una medida de calidad usada en recuperación de información para evaluar los modelos y algoritmos de clasificación de documentos y se define como la media armónica de las medidas de precisión y recuerdo [7].

MODELO ESPACIO VECTORIAL: Es un modelo de recuperación de información el cual los documentos y consultas se representan como vectores de términos mediante la asignación de pesos a cada término en el documento, estos pesos son usados para computar un grado de similitud entre cada consulta y documento, la cual está dada por el coseno del ángulo entre los vectores que los representan, es decir la función de similitud obtendrá valores entre 1 (cuando el documento es igual a la consulta) y 0 (cuando no existe ningún término común entre la consulta y el documento) [5].

NASA: gente, ser humano, Páez [8].

NASA YUWE: es la lengua de comunidad indígena Nasa o Páez, esta lengua étnica es una de las más importantes que se habla en el territorio colombiano, dado el número de hablantes [4], a pesar de esto y de la zona que ocupan, múltiples hechos vienen atentando desde la época colonial contra la supervivencia de la lengua [4], por tanto, hoy día es considerada una lengua en peligro de extinción [9]. Gran parte del conocimiento de esta lengua se basa en la tradición oral, de tal forma que la necesidad de preservar y lograr que éste trascienda los ha llevado a utilizar varias representaciones gráficas y desde el siglo XVIII la escritura del nasa yuwe ha estado presente en su vida, pero es en el siglo XX

cuando se da la unificación del alfabeto, lo cual representa un importante paso para encontrar caminos que coadyuven en el desarrollo y revitalización de la lengua nasa [10].

PRECISIÓN (PRECISION): Es la proporción de material relevante recuperado, del total de los documentos que son relevantes en la base de datos, independientemente de que éstos, se recuperen o no [11] [3].

PRECISIÓN R (R-PRECISION): es la proporción de los top R documentos recuperados que son relevantes, donde R es el número de documentos relevantes para la consulta actual [12].

PRINCIPALES MODELOS DE REPRESENTACIÓN PARA RECUPERACIÓN DE INFORMACIÓN EN TEXTOS NO ESTRUCTURADOS: 1- El modelo Booleano, 2- Modelo probabilístico, 3- El modelo de espacio vectorial [5] [3].

PRECISIÓN MEDIA EN DOCUMENTOS RELEVANTES OBSERVADOS (AVERAGE PRECISION AT SEEN RELEVANT DOCUMENTS - AP): es la media de rankings de precisión después de que cada documento es recuperado [13].

RECUERDO (RECALL): Mide la capacidad del sistema para recuperar los documentos que son relevantes, en otras palabras, representa la fracción de todos los documentos relevantes que son recuperados [14].

RECUPERACIÓN DE INFORMACIÓN: Es un proceso complejo que busca entre otras tareas, producir una función de ranking (orden) de documentos con respecto a una consulta, lo cual implica dos tareas principales [5]: 1- La concepción de un forma lógica para la representación de los documentos y las consultas; 2- la definición de la función de ranking que permite cuantificar la similitud entre los documentos y las consultas.

REMOCIÓN DE PALABRAS VACÍAS (STOPWORDS REMOVAL LIST): Es el proceso de eliminar las palabras que más se repiten en la escritura de un idioma y que no aportan significado a la búsqueda por sí solas, esta tarea se basa en la definición de aquellas palabras que son las más frecuentemente usadas, en la mayoría de los lenguajes contienen: determinantes, conjunciones, preposiciones e interjecciones. Generalmente, el uso de análisis de frecuencias basado en esas listas ha demostrado ser efectivo y es un proceso útil, dado que aparte de reducir el número de tokens a ser indexados, ha mostrado mejorar la efectividad de la recuperación de información [2].

SISTEMA DE RECUPERACIÓN DE INFORMACIÓN (SRI): son aquellos sistemas que realizan las operaciones (tokenizer, stopwords removal list, stemming, entre otras) que permiten: la representación de documentos y consultas, y la definición de la función de ranking para evaluar la similitud entre los documentos y las consultas [5].

TOKENS: son unidades pequeñas de caracteres, en la mayoría de los casos con significado (como por ejemplo las palabras), resultantes de aplicar el análisis léxico (tokenizer) a un documento o consulta [1].

YUWE: boca, palabra, lengua [8].

RESUMEN

El nasa yuwe es una lengua oficial de Colombia, actualmente se encuentra en peligro de extinción, desde diferentes instancias nacionales e indígenas se vienen adelantando estrategias en pro de revitalizar la lengua entre las que se encuentran las tecnologías de la información que buscan apoyar la visibilización de la lengua y su uso a través de herramientas computacionales.

Este documento describe el desarrollo y los resultados obtenidos en la adaptación de un modelo de espacio vectorial para la recuperación de información de textos escritos en nasa yuwe mediante:

La construcción de una colección cerrada de prueba de textos escritos en nasa yuwe, la cual involucró:

- Trabajo de campo con profesores de la comunidad nasa de varios resguardos cercanos al municipio de Popayán
- La conformación de 97 documentos escritos en nasa yuwe
- La definición de 8 consultas
- El registro del juicio de expertos sobre la relevancia de los documentos para cada consulta.

Un prototipo de sistema de recuperación de información de textos escritos en nasa yuwe el cual se ha desarrollado teniendo en cuenta:

- La adaptación de un analizador léxico (tokenizer) para nasa yuwe basado en el analizador léxico de Lucene .NET (versión 2.9.4)
- La definición de una lista de palabras vacías para remover de los documentos de la colección y las consultas (Stopwords Removal list)
- La evaluación del desempeño del prototipo a través de medidas tradicionales del área de investigación como la Curva Precisión – Recuerdo.

En el desarrollo de este trabajo, se pudo observar que a pesar de que el nasa yuwe, es una lengua que está en proceso de descripción fue posible hacer la adaptación del análisis léxico y la definición de palabras vacías para esta lengua, para finalmente obtener un prototipo de sistemas de recuperación de información para textos escritos en nasa yuwe, y a través de la medición del desempeño de este prototipo fue posible apreciar que la adaptación del analizador léxico es tarea crucial en la recuperación y se muestran resultados prometedores con relación a la línea base, a diferencia de los resultados obtenidos con la lista de palabras vacías, con la cual no se muestran mejoras sustanciales en el desempeño del prototipo de esta lengua.

Entre las palabras clave de este trabajo se encuentran:

- Recuperación de información para textos escritos en nasa yuwe
- Colección de prueba de textos escritos en nasa yuwe

- Prototipo de Sistema de Recuperación de Información de textos escritos en nasa yuwe
- Adaptación de un analizador léxico para nasa yuwe
- Tokenizador nasa
- Lista de palabras vacías para nasa yuwe.

ABSTRACT

The nasa yuwe is an official language of Colombia, it is currently in danger of extinction, nowadays advanced strategies are being promoted from different national and indigenous organizatons such as the information technologies to seek to support the visibility of the language and its use through computational tools.

This document describes the development and results in the adaptation of a vector space model for information retrieval of texts written in nasa yuwe by:

Building a closed test collection of texts written in nasa yuwe, which involved:

• Field work with nasa teachers from several nearby community shelters to the city of Popayan

• The establishment of 97 documents written in nasa yuwe

• The definition of 8 queries

• The register of expert judgment about the relevance of the documents for each query.

A prototype of information retrieval system for texts written in nasa yuwe, it was developed taking into account:

• The adaptation of a nasa yuwe tokenizer based on the Lucene .NET standard tokenizer (version 2.9.4)

• The definition of a stopwords removal list to apply on the documents of the nasa yuwe test collection and queries.

• Performance evaluation of the prototype through traditional measures of the research area as the Precision – Recall Curve.

To develop this work it was observed that although the nasa yuwe, is a language in process of description, it was possible to adapt a tokenizer and to define a stopwords removal list for this language, in order to get a prototype of information retrieval systems for texts written in nasa yuwe, and through perfomance evalution of this prototype was possible to see the adaptation of the nasa tokenizer is an important task in the recovery and this Project showed promising results in relation with the baseline, unlike the results obtained with the stopwords removal list, there is not substantial improvements in the performance of the prototype.

Among the keywords are:

• Information retrieval for texts written in nasa yuwe
• Test collection for texts written in nasa yuwe
• Prototype of Information Retrieval System for texts written in nasa yuwe
• Adapting of a tokenizer for nasa yuwe
• Tokenizer nasa
• Stopwords removal list for nasa yuwe

INTRODUCCIÓN

PRESENTACIÓN

El presente trabajo se desarrolla en el marco del Ciclo I de formación Doctoral en Ingeniería Telemática, aborda un tema específico como lo es la recuperación de textos escritos en lengua nasa yuwe, el idioma de la Comunidad Indígena Páez (Colombia).

En esta sección se presenta el planteamiento del problema, los objetivos y los aportes investigativos que dieron origen a este trabajo.

Luego, en el capítulo 1, se presenta un marco de referencia el cual incluye un contexto teórico y los antecedentes más relevantes relacionados con los temas tratados en este trabajo.

En el capítulo 2, se muestra la primera colección cerrada de prueba en nasa yuwe, su proceso de construcción y resultados, los cuales permiten apreciar características propias de la lengua nasa yuwe, las consultas diseñadas, los juicios emitidos por los expertos y los documentos; también se muestra un análisis estadístico de estos datos, incluyendo el análisis a la luz de la Ley de Zip.

En el capítulo 3, se presenta la adaptación del analizador léxico (tokenizer) y la definición de la lista de palabras vacías (stopwords removal list) que son parte integral de la construcción del prototipo de recuperación de información de textos escritos en nasa yuwe. Además, se presentan los resultados de las evaluaciones de desempeño del sistema en relación con la línea base.

En el capítulo 4, se resume la construcción del prototipo de SRI para nasa yuwe desarrollado, presentando los diferentes artefactos de cada una de las fases de la metodología XP. Este prototipo incluye los resultados descritos en los capítulos 2 y 3.

En el capítulo 5, se presentan las conclusiones, recomendaciones a las que se llegaron en el desarrollo de este trabajo y el trabajo futuro que se plantea para dar continuidad al trabajo.

PLANTEAMIENTO DEL PROBLEMA

El estado colombiano concibe la diversidad étnica y cultural dentro de los derechos sociales [15], la cual se ve expresada entre otros aspectos por las múltiples lenguas indígenas que se han mantenido vivas por siglos, sin embargo, los esfuerzos realizados en pro de la conservación y preservación de éstas lenguas no han sido suficientes, para que las minorías indígenas las preserven como parte vital de su identidad cultural, las cuales constituyen un invaluable tesoro de nuestra historia.

Lo anterior, se hace evidente, en la lengua nasa yuwe del pueblo indígena Nasa (Páez[1]), cuya situación sociolingüística, es preocupante, dado que se encuentra en peligro de extinción, como se puede apreciar en el diagnóstico adelantado por el CRIC, la Universidad del Cauca y el Ministerio de Cultura (2007) [9]: el nasa yuwe, es hablado por menos del 40% de las personas consultadas, esta situación varía de una zona a otra dentro de los territorios Nasa, se aprecia que los niveles de conocimiento y de uso del nasa yuwe en los resguardos son bajos [16]. Uno de tantos factores para la debilitación de la lengua obedece a un limitado acceso a ciertos recursos (sociales, educativos y tecnológicos, entre otros) y en algunas oportunidades a la utilización acrítica de modelos desarrollados en y para otras realidades culturales y políticas.

En el contexto del fortalecimiento de la educación propia, existen leyes que buscan materializar lo establecido en la constitución nacional [15] (como la Ley 115 de Febrero 8 de 1994[2] [17], El Decreto 0804 de Mayo 18 de 1995[3] [18], entre otros) y políticas propias (como las establecidas por el Consejo Regional Indígena del Cauca[4] – CRIC [19], el Programa de Educación Bilingüe e Intercultural – PEBI) entre las que se cuenta la creación de los Proyectos Educativos Comunitarios-PEC y el Sistema Educativo Indígena Propio-SEIP [20], los cuales buscan establecer relaciones entre los usos, costumbres, saberes y lenguas de los pueblos indígenas con otras competencias de la educación escolar, capacitando a los jóvenes para ingresar a las instituciones de educación superior y desenvolverse en la sociedad indígena y no indígena. Igualmente, la Organización Nacional Indígena de Colombia - ONIC [21], en la IV Cumbre de Líderes Indígenas de las Américas (11-13 abril/2012), presentó sus propuestas para adoptar medidas sobre el acceso a tecnologías y la utilización de éstas, recomendando facilitar, apoyar y promover el uso y desarrollo adecuados de Tecnologías de la Información y las Comunicaciones - TIC para los pueblos indígenas, hecho relevante, si se tiene en cuenta la pertinencia, para las comunidades de aprovechar las posibilidades proporcionadas por las TIC. Por tanto, si la sociedad nacional utiliza la tecnología como oportunidad estratégica para el progreso (como se puede apreciar a través de programas y convocatorias del Ministerio de Educación, el Ministerio de Tecnologías y las Comunicaciones, el Ministerio de la Cultura, el Departamento Administrativo de Ciencia, Tecnología e Innovación – COLCIENCIAS, Universidades Nacionales e Internacionales), es fácil pensar en la pertinencia de desarrollar propuestas soportadas en las TIC como un elemento incluyente y de valor para motivar,

[1] El pueblo indígena Nasa es el segundo del país en cuanto a su población absoluta y número de hablantes de su lengua el nasa yuwe, son alrededor de 150.000 personas distribuidas en 7 departamentos [4]

[2] Ley General de Educación

[3] desarrolla los principios de la etnoeducación en el país.

[4] Desde sus inicios ha luchado por el mantenimiento de sus tradiciones culturales y sus lenguas durante muchos años.

fortalecer y apoyar su uso en las comunidades indígenas, especialmente, en lo referente a sus lenguas, como es el caso del nasa yuwe, que se encuentra en peligro de extinción.

A este respecto, las instituciones educativas en territorios indígenas en el Cauca, han adquirido una infraestructura informática precaria pero en creciente mejoramiento, que permite impulsar el desarrollo de diferentes procesos de valoración y fortalecimiento de las lenguas y culturas indígenas a través de diversas propuestas y estrategias, como por ejemplo: [22], [23], [24], [25], entre otras.

Es así que, el uso de la tecnología se propone como una oportunidad estratégica para que su adecuación, apropiación y desarrollo al entorno social y cultural del pueblo nasa, incluya un acercamiento metodológico, conceptual y práctico que apoye la preservación del conocimiento ancestral, su uso en el ámbito educativo y cotidiano de las lenguas indígenas como una estrategia para llevarlas a las realidades que plantea el mundo de hoy. Siendo consecuente con esto, es posible pensar en utilizar técnicas computacionales, que permitan el intercambio de información por medio de actividades de búsqueda y recuperación de información [26], que favorezcan diferentes posibilidades para que las personas de esta Comunidad interactúen en nasa yuwe, para lo cual, se requiere la consolidación de documentos escritos (una colección), sobre los cuales se pueda operar una técnica computacional como un modelo de recuperación de información que permita al usuario nasa hacer la consulta en nasa yuwe y que los documentos recuperados estén escritos en esta lengua, con esto se espera contribuir en: 1) el proceso de visualización de la lengua, y 2) en la sensibilización de su uso mediante herramientas computacionales. Es así, que se propone utilizar un modelo de recuperación de información como el descrito anteriormente, que permita realizar la recuperación de textos escritos en nasa yuwe. Contar con este sistema, se espera motive la escritura y consulta en nasa yuwe, lo cual es aplicable para diversas actividades de carácter social, educativo, político, etc., de esta comunidad.

Teniendo en cuenta que el modelo de espacio vectorial[5] es el que mejores resultados obtiene y es el más utilizado en la actualidad, al inicio de este trabajo se planteó la siguiente pregunta de investigación: **¿Cómo adaptar las fases del proceso de recuperación de información a textos escritos en nasa yuwe?**

Para desarrollar esta propuesta, se hizo la adaptación de un analizador léxico y la definición de una lista de palabras vacías, utilizadas en el proceso de recuperación de información (RI) de textos escritos en nasa yuwe, teniendo en cuenta las características de ésta lengua [27] [28]. Finalmente, se realizó la evaluación de este proceso.

A continuación se presentan los objetivos como fueron aprobados por el Consejo de Facultad de la FIET, en el anteproyecto de tesis.

OBJETIVOS

Objetivo general

Proponer un esquema de representación y búsqueda para textos escritos en lengua nasa yuwe basado en el modelo espacio vectorial y la adaptación de las fases del proceso de

[5] Es el modelo de recuperación de información que mejores resultados ofrecen [7]

recuperación información, con el fin de contribuir en el proceso de visualización y sensibilización de uso de esta lengua indígena mediante estrategias computacionales.

Objetivos específicos

- Crear una colección cerrada de prueba de textos escritos en nasa yuwe mediante el juicio de expertos hablantes de esta lengua.

- Adaptar las tareas del proceso de recuperación y búsqueda de información (tokenizer y stopwords removal list) para textos escritos en nasa yuwe, teniendo en cuenta las características gramaticales del nasa yuwe.

- Desarrollar un prototipo de Sistema de Recuperación de Información, soportado en las tareas adaptadas del proceso de recuperación y búsqueda de información para textos escritos en nasa yuwe.

- Definir el nivel de desempeño del prototipo desarrollado para el proceso de recuperación de información de textos escritos en nasa yuwe, usando medidas tradicionales como precisión, recuerdo, y medida F.

APORTES INVESTIGATIVOS

Los principales aportes que se propusieron y lograron con el desarrollo de este trabajo se presentan en la Tabla 1:

Aporte Investigativo	Publicación
La conformación de la primera colección cerrada de prueba para textos escritos en nasa yuwe, desarrollada mediante el juicio de expertos	Artículo publicado en evento Internacional clasificado como A2 por Colciencias. Building a test collection for nasa yuwe language (Ver Anexo 6)
La adaptación de un analizador léxico (tokenizer) para recuperación de textos escritos en nasa yuwe.	Articulo enviado a evaluación en evento internacional titulad “Tokenizer adapted for nasa yuwe language”. En espera de los resultados de evaluación
La definición de algunas características de la lengua nasa yuwe para adaptar el analizador léxico y la definición de una lista de palabras vacías para el proceso de RI.	Artículo enviado a evaluación a revista internacional clasificada como A2 por Colciencias. Titulado; “IRS for Nasa Yuwe: An Efficient and Simple Proposal” En espera de resultados.
El desarrollo de un prototipo de un Sistema de Recuperación de Información para textos escritos en nasa yuwe, en el cual, se implementa el tokenizador adaptado para nasa yuwe, y la lista definida para la remoción de palabras vacías (stopwords removal list)	

Aporte Investigativo	Publicación
basado en la frecuencia de los términos en la colección.	
La definición de la calidad de la recuperación del SRI basado en las medidas precisión, recuerdo, medida F, entre otras.	
Trabajo futuro, propuesta de algoritmos metaheurísticos para utilizar en la construcción de un Identificador de partes del discurso para nasa yuwe.	Artículo publicado en evento internacional clasificado como A2 por Colciencias. Titulo: "Continuos Optimization Based on a Hybridization of Differential Evolution with K-means" (Ver Anexo 7)

Tabla 1. Descripción de Aportes investigativos. Fuente: Elaboración propia

DISEÑO METODOLÓGICO

Como metodología para el desarrollo de este proyecto se utilizó el Patrón Iterativo de Investigación (Iterative Research Pattern) [29], el cual plantea que el diseño de patrones puede ser usado para resolver problemas recurrentes, por tanto, al igual que el diseño de patrones, la investigación iterativa es una metodología que se enfoca en varios ciclos cortos para el desarrollo de un trabajo de investigación (trabajo de campo, desarrollo, pruebas de laboratorio, entre otros). Consta de 4 pasos básicos: observaciones de campo, identificación del problema, desarrollo tecnológico y pruebas de campo, soportado en los pasos del método científico y entendiendo por investigación (research): "*al proceso de generar una pregunta y una hipótesis de investigación, diseñando y desarrollando un experimento y evaluando los resultados para responder la pregunta original y confirmar o negar la hipótesis"* [29].

Dentro de las iteraciones se utilizaron metodologías como:

1) Investigación documental [30, p. 216], para apoyar la revisión y análisis de bibliografía, presentado en el marco de referencia de este documento, el cual ha sido soporte para la construcción de la colección cerrada de prueba y realizar la adaptación de las tareas del proceso de recuperación de información.

2) Proceso de desarrollo de software XP [31], [32], para apoyar el desarrollo del prototipo del SRI para textos escritos en nasa yuwe; en donde, se utilizan la colección cerrada de prueba (construida para nasa yuwe) y los resultados de la adaptación del analizador léxico y la lista de palabras vacías para nasa yuwe.

3) Juicio de expertos [33], [34], para apoyar la conformación de la colección cerrada de prueba y el trabajo requerido con el experto en lingüística y los profesores hablantes de nasa yuwe, quienes emitieron sus juicios para cada par documento – consulta.

4) Adaptación de la metodología propuesta en [35], para la evaluación experimental en Ingeniería de software, la cual se utiliza para definir el nivel de desempeño del prototipo usando las medidas tradicionales (como curva de precisión – recuerdo, medida F, entre otras).

1. MARCO DE REFERENCIA

En esta sección se presenta un breve contexto teórico y algunos antecedentes sobre los temas relacionados con la presente investigación, a saber: colecciones de prueba de textos, sistemas de recuperación de información de textos y medidas de evaluación.

1.1 CONTEXTO TEÓRICO

1.1.1 Recuperación de información (RI)

La recuperación de información se viene trabajando desde los años 50 aproximadamente, sin embargo actualmente, ha cobrado gran relevancia, dada la importancia de contar con información organizada y relevante sobre un tópico deseado. Por tanto, se encuentran varias descripciones de lo que se entiende por recuperación de información, entre las que se tiene: "*Es la búsqueda de material (documentos) de naturaleza no estructurada (usualmente textos) que satisface una necesidad de información dentro de una gran colección*" [36]. Un Sistema de Recuperación de Información -SRI tiene como objetivo encontrar aquellos documentos que satisfacen las necesidades del usuario, las cuales son expresadas en forma de una consulta [37, p. 221]; la respuesta "ideal" de un SRI está formada solamente por documentos relevantes a la consulta [38], de allí que las dos principales metas para un SRI son [2, p. 17]: maximizar la probabilidad de que la consulta encuentre elementos relevantes y producir el mejor ranking como sea posible de acuerdo a los estimados de probabilidad de relevancia. Entre las principales técnicas utilizadas en RI se encuentran [26, p. 446]:

1. Análisis léxico (tokenizer)

La tokenización tiene como base el reconocimiento de unidades léxicas de un texto, de tal forma que se introduce un documento (texto) y se obtiene una lista de unidades léxicas como salida [39]. Es una de las primeras tareas que se ejecuta a la hora de realizar el pre-procesamiento de texto [1]. En la mayoría de los casos los tokens coinciden con las palabras, sin embargo, son llamadas unidades léxicas del texto y se conocen como tokens. De la misma manera, los programas que se encargan de hacer esta descomposición léxica se llaman tokenizadores (tokenizers) [39]. Un tokenizador tiene dos propiedades importantes: 1) Es el único componente que se debe usar en toda aplicación de procesamiento de lenguaje natural que involucre tareas de análisis. 2) Siempre se utiliza al inicio del proceso de análisis [40].

2. Remoción de palabras vacías (stopwords removal list)

Esta tarea se basa en la eliminación de aquellas palabras (que se encuentran en los documentos de la colección) que son más frecuentemente usadas en un determinado lenguaje y que tienen poco poder de selectividad. En diversos lenguajes se ha demostrado que se reduce el tamaño de los índices de consulta y se aumenta la calidad de la

recuperación, determinado por medidas como recall[6] (recuerdo), precisión[7] y medida F[8] (measure F) [2], entre otras.

3. Modelo de espacio vectorial.
La recuperación de información es un proceso complejo que busca entre otras tareas, producir una función de ranking[9] (orden) de los documentos con respecto a una consulta, lo cual implica dos tareas principales [5]: 1- La concepción de un forma lógica de la representación de los documentos y las consultas; 2- la definición de una función de ranking que permite cuantificar la similitud entre los documentos y las consultas.

Existen tres modelos principales de representación para recuperación de información en textos no estructurados [5] [36]: 1- El modelo Booleano, donde las consultas se especifican mediante expresiones booleanas, la similitud entre documentos y consulta se limita a pesos binarios. 2- Modelo probabilístico, trata de estimar la probabilidad de que un documento sea relevante para una consulta del usuario. 3- El modelo de espacio vectorial, que propone un marco en el cual se puede encontrar una coincidencia parcial entre los documentos y la consulta, lo cual se logra mediante la asignación de pesos no binarios a los términos en las consultas y los documentos, los pesos de los términos son usados para computar un grado de similitud entre cada consulta y documento, así los documentos son ordenados por su similitud en relación con la consulta. En 2011 Jamil y otros [41], presenta comparaciones entre los modelos de representación de información (modelo de espacio vectorial, booleano y probabilístico) y lenguajes, resaltando sus ventajas y desventajas.

Entre las ventajas que ofrece el modelo de espacio vectorial se encuentran [5]: los pesos de los términos mejoran la calidad del conjunto de respuesta; la coincidencia parcial permite que la recuperación de documentos se aproxime a las condiciones de la consulta; la similitud de coseno permite ordenar los documentos recuperados de acuerdo al grado de similitud de la consulta. Su desventaja es que asume independencia entre los términos del índice, aunque no hay pruebas de que esto mejore los resultados en colecciones de texto genéricas. Basado en estas ventajas y que es el más usado en los SRI de hoy, este es el modelo de representación que se utilizó en el desarrollo de este proyecto.

1.1.2 Colección de prueba

Las colecciones de prueba permiten evaluar el desempeño de un SRI de forma tradicional. Una colección de prueba, está conformada por [36]: 1- Una colección de documentos; 2- Un conjunto de prueba de necesidades de información, expresadas como consultas de usuario y pueden estar constituidos por una o varias palabras clave; 3- Un conjunto de juicios de relevancia, evaluados generalmente de manera binaria, expresando si el documento es relevante o no para cada consulta.

Los primeros experimentos para realizar la evaluación de los sistemas de Recuperación de información realizados hacia los años 50s, como el caso de Cranfield, utilizaron una pequeña colección de documentos; actualmente se usan colecciones más grandes para simular mejor los requerimientos de búsqueda, pero la conformación y consecución de

[6] Es la fracción de documentos recuperados que son relevantes [36].
[7] Es la fracción de documentos relevantes que son recuperados. [36].
[8] Es una medida sencilla que intercambia precisión y recuerdo (recall), en la cual los pesos se dan por la media armónica de precisión y recuerdo [36].
[9] Una función de ranking ordena los documentos en pro de reflejar la relevancia de cada documento para la consulta específica [7].

colecciones grandes con documentos extensos no es una tarea fácil [42]. Existen varias colecciones de prueba como [3]: 1) The Cranfield collection: colección pionera con la que se definieron medidas para cuantificar la efectividad de la RI; fue recolectada en Reino Unido a final de 1950, contiene 1398 artículos de revistas de aerodinámica, un conjunto de 225 consultas y exhaustivos juicios de relevancia de todos los pares. 2) Text Retrieval Conference (TREC): The U.S. National Institute of Standards and Technology (NIST) construyo esta colección y ha sido de gran importancia para evaluación de RI desde 1992; en total comprende 6 CDs que contiene 1,89 millones de documentos (artículos) y juicios de relevancia para 450 necesidades de información, las colecciones más nuevas, TRECs 6-8 proveen 150 necesidades de información sobre 528000 servicios de noticias y artículos de un servicio de información de difusión extranjera. 3) GOV2: desarrollado por NIST, incluye 25 millones de páginas web, actualmente, es la colección más grande web disponible para propósitos de investigación. 4) NII Test Collections for IR Systems (NTCIR): está enfocada en recuperación de información multilingual para algunas lenguas asiáticas (Japonés, Chino y Coreano), está compuesta por varios cientos de miles de artículos escolares; específicamente la Colección NTCIR-4 PATENT, consta de datos de documentos (Solicitudes de Patentes Japonesas y Resúmenes de patentes de Japón entre 1993-1997), 101 tópicos de búsqueda en Japonés (34 fueron traducidos al inglés, al Chino simplificado y al Koreano tradicional), y juicios de relevancia para cada tópico de búsqueda [43]. 5) Cross Language Evaluation Forum (CLEF): está centrada en recuperación de información multilingual para lenguas europeas. 6) Reuters-21578 y Reuters-RCV1: es una colección de referencia compuesta de los artículos publicados en Reuters. Contiene más de 800 mil documentos organizados en 103 categorías [5]. 7) 20 Newsgroups: Compuesta por miles de mensajes de grupos de noticias, organizados de acuerdo a 20 grupos [5]. A continuación en la Tabla 2 se presenta una comparación entre éstas colecciones.

Colección	Lenguaje	Tópicos	# documentos	Tamaño
Cranfield II	Inglés	Resumen de artículos científicos	1.398	579 KB
TREC-AP	Inglés	Servicios de noticias de AP (1988-1990)	242.918	0.7 GB
GOV2	Inglés	Páginas web gubernamentales	25.205.179	426 GB
NTCIR-4[10] PATENT [43]	Inglés, Chino, Japonés y Koreano	Documentos de patentes	7.000.000	65 GB
CLEF-multi	Holandés, Inglés, finlandés, Francés, Alemán, Italiano, Portugués, Ruso, Español, Suizo	Periódicos y servicios de noticias para un periodo de tiempo (1994 – 1995)	1.869.564	4.7 GB
RCV1[11]	Inglés	Historias de noticias	810.000	2.5 GB
Reuters-21578[12]	Inglés	Historias de servicios de noticias en 5 categorías diferentes [5]	21.578	28 MB

[10] http://research.nii.ac.jp/ntcir/permission/ntcir-4/perm-en-PATENT.html
[11] http://trec.nist.gov/data/reuters/reuters.html
[12] http://www.daviddlewis.com/resources/testcollections/reuters21578/

Colección	Lenguaje	Tópicos	# documentos	Tamaño
20 Newsgroups [5]	Inglés	Grupos de noticias en 20 diferentes grupos	18.846 newsgroup documentos	13.8 MB comprimido

Tabla 2. Descripción de algunas colecciones de prueba. Adaptación de [42]

Entre los métodos más comunes para realizar los juicios de relevancia en una colección de prueba se encuentran: 1) Pooling [44], que crea un subconjunto configurado con la lista de los documentos recuperados con más alto puntaje para una consulta, los cuales han sido obtenidos mediante un sistema de RI. Este método propuesto por Gilbert and Sparck Jones en 1979 [45] ha sido usado entre otros por TREC-6 y es bien conocido por su eficiencia. 2) Move to Front (MTF) Pooling [42], es una mejora al método Pooling standard usando un número variable de documentos recuperados por distintos SRI dependiendo del desempeño de cada sistema. 3) Interactive Searching and Judging (ISJ) [42], busca crear un pool de juicios de documentos con el mínimo esfuerzo, para lograr esto, a un grupo de buscadores se les solicita encontrar y emitir juicios de tantos documentos relevantes como sea posible en un breve período de tiempo. Este método es poco popular, debido a que tiene el riesgo de que los juicios de relevancia puedan resultar sesgados por los sistemas de búsqueda que los ha creado, por tanto, es aplicado como método suplementario para mejorar la calidad de una colección de prueba.

1.1.3 La Comunidad nasa

El pueblo indígena Nasa es el más numeroso del departamento del Cauca (Colombia), con una presencia mayoritaria en los resguardos y cabildos de este departamento. Existe una tendencia a llamarlos etnia Páez ya que a la llegada de los españoles a sus tierras, estos, los españoles, los llamaron pueblo Páez, pero el nombre reconocido por ellos es Nasa que significa “gente”, en adelante se referirá a este pueblo como pueblo Nasa [4].

Los Nasa habitan especialmente ambos costados de la cordillera central y un costado de la cordillera occidental del Departamento del Cauca, además de la planicie entre estas dos cordilleras. Tierradentro es considerado el territorio original de los Nasa, Hoy los Nasa están asentados en el departamento del Valle, en los municipios de Florida, Dagua y Jamundí; en el Cauca en los municipios de Páez – Belalcázar, Inzá, Morales, Jambaló, El Tambo, Caldono, Silvia, Totoró, Toribío, Caloto, Corinto, Santander de Quilichao, Miranda, Buenos Aires, Popayán, Puracé y Cajibío; en el Tolima en los municipio de Planadas y Ortega, en el Huila en los municipios La Plata, Gigante y Palermo. En el Putumayo se encuentran en los municipios de Mocoa, Puerto Asís y Villagarzón; en el Caquetá en Belén de los Andaquíes, Puerto Rico, La Florida, San José del Fragua, San Vicente del Caguán y Solano; y en el Meta, en el municipio de Mesetas [46].

1.1.4 El nasa yuwe

El nasa yuwe es una de las lenguas oficiales de la República de Colombia es hablada por la comunidad nasa (pueblo indígena Páez de Colombia), y es la lengua étnica más importante del territorio colombiano, se habla en varios resguardos y asentamientos nasa ubicados en varios municipios de los departamentos de Caquetá, Putumayo, Meta, Tolima Valle del Cauca, Cauca y Huila [47], se calcula que la población nasa ya bordea los 200.000, de los cuales un 75% son hablantes activos de la lengua nasa [23]. Cabe resaltar que la situación sociolingüística del nasa yuwe la ubica en peligro de extinción, dado que por

diversos factores culturales, sociales, geográficos e incluso históricos se ha ido perdiendo [9].

Es una lengua que se ha transmitido de generación en generación mediante tradición oral. En el siglo XVIII hubo intentos de escribir la lengua pero solo en el siglo XX es cuando se logra el alfabeto y más tarde se da la unificación de este alfabeto, lo cual representó un importante paso para encontrar caminos que coadyuven en su desarrollo y revitalización [10], por tanto, no son muchos los textos que se encuentran escritos en este alfabeto, dado que existen muchas variantes en la representación de los sonidos de la lengua, lo cual varía de una zona a otra. Para efectos del desarrollo de este trabajo se han buscado textos escritos con el alfabeto unificado, el cual está conformado por 32 vocales y 61 consonantes que se presentan en la Figura 1.

Vocales

Oral: i e a u	Nasal: ĩ ẽ ã ũ
Orales Interruptas: i’ e’ a’ u’	Nasal interruptas: ĩ’ ẽ’ ã’ ũ’
Oral aspirada: ih eh ah uh	Nasal aspirada: ĩh ẽh ãh ũh
Oral larga: ii ee aa uu	Nasal larga: ĩĩ ẽẽ ãã ũũ

Consonantes	bilabial	alveolar	palatal	velar
Básicas	p	t	ç	k
Aspiradas	ph	th	çh	kh
Palatalizadas	px	tx	çx	kx
Aspirada Palatalizada	pxh	txh	çxh	kxh
Prenasal	b	d	z	g
Prenasal – palatalizada	bx	dx	zx	gx
Nasal	m	n		
Nasal – palatalizada		nx		
Fricativa		s		j
Fricativa – palatalizada	fx	sx		jx
Lateral		l		
Lateral – palatalizada		lx		
Aproximantes	w		y	
Aproximantes - palatalizadas	vx			

Figura 1. Alfabeto unificado en nasa yuwe. Fuente: [48]

La lengua nasa yuwe[13], había sido incluida en la Familia lingüística Chibcha, pero estudios recientes señalan las incongruencias de esta inclusión, por ello actualmente, se considera como lengua independiente (no clasificada) [28]. La descripción de esta lengua se encuentra en proceso, por tal motivo son pocos los trabajos que se encuentran, entre ellos:

- Gramática del Páez o nasa yuwe [49], que describe la fonología, morfología y sintaxis del Páez o nasa yuwe.
- Diccionario Nasa Yuwe – Castellano [50], contiene vocabulario para editores Nasa y se hace una descripción del nasa yuwe soportado en la gramática del castellano.
- Una mirada al nasa yuwe de Novirao [51], el cual presenta una descripción fonológica y morfológica del nasa yuwe de Novirao y hace una comparación con otras variantes.

[13] Nasa significa: “gente”, “ser humano”, “Páez”; yuwe: “boca”, “palabra”, “lengua” [28] .

A continuación, se presenta una breve descripción de algunas características de la lengua nasa yuwe [8]:

- La lengua nasa tiene lexemas[14] y gramemas[15] cuya estructura es relativamente simple.
- Los morfemas[16] se suman unos a otros en la cadena.
- Se han encontrado afijos (prefijos y sufijos), no se han encontrado infijos.
- La mayor parte de morfemas tienen como significante un fonema (vocálico o consonántico) o una cadena de fonemas.
- La formación de la palabra en nasa yuwe exige la presencia de al menos un radical[17] simple por palabra, el cual deberá aparecer sólo o acompañado de morfemas flexionales[18] o morfemas derivativos[19].
- La flexión[20] se expresa mediante sufijos, existen flexión: modo-personal, declinación y facultativa.

Para la definición de las clases de palabras en nasa yuwe se han propuesto las siguientes [52]:

- Palabra predicativa: formalmente definida por ser la que lleva la base predicativa y la flexión modo – personal.
- Nombre: la construcción resultante de la aplicación de un conjunto de marcas de flexión (declinación e identificadores) a una base.
- Cualificativa: formada por un radical cualificativo que no recibe ningún tipo de flexión ni otra marca, pero en ese caso acompaña otra palabra, que puede ser predicativa o nominal.
- Conectora: no lleva flexión, son empleadas como conectoras de la oración.

1.1.5 Lucene

Es una API de código abierto, que contiene una biblioteca de búsqueda de texto. Adiciona contenido de un documento a un índice, el cual permite ejecutar consultas, retornando resultados ranqueados por relevancia de consulta. Lucene es capaz de obtener mejores respuesta porque crea un índice invertido. En Lucene los documentos son la unidad de búsqueda e indexación, es decir, un índice puede contener uno o más documentos y éste a su vez puede contener uno o varios campos. La indexación en Lucene involucra crear documentos comprimidos de uno o más campos y adicionar estos documentos a un objeto de la clase IndexWriter [53].

Para definir los documentos relevantes en un proceso de recuperación en Lucene se utiliza una clase llamada “Similarity”, en la cual se definen los componentes para que se lleve a

[14] Es la unidad mínima con significado léxico. Ejemplo **deport**-e, **kne**–ne (frente en nasa) [28].
[15] Es un morfema que se suma al lexema para indicar accidentes gramaticales. Ejemplo, perr –**o**- [83].
[16] Es la unidad mínima aislable en al análisis morfológico. Ejemplo Habilidades → tres morfemas: Habil (que aporta el significado léxico), -idad (sufijo derivativo) y –es (sufijo flexivo) [83]
[17] Es la base de una construcción. Ejemplo Yat (casa), yat-we’sx (casas)
[18] Los Morfemas Flexivos tienen como función poner a los lexemas en el modo gramatical adecuado, es decir, con el correspondiente: género o número (para los sustantivos), modo gramatical adecuado (para los verbos [83]
[19] Los Morfemas Derivativos son aquellos que añaden matices al significado de los lexemas son prefijos, sufijos e interfijos [83]
[20] Las variaciones flexivas pueden aportar información relativa al género, al número, a la persona, al tiempo, al aspecto, al modo. [83]

cabo la puntuación de los documentos encontrados. Lucene utiliza una combinación de Modelo Booleano con Modelo espacio vectorial (VSM) [6].

Como se menciona en [3], el modelo espacio vectorial, representa documentos y consultas como vectores ponderados en un espacio multidimensional, donde cada término índice es una dimensión y los pesos son valores TF[21]-IDF[22] . El valor de TF para un término t y un documento (o consulta) X, su valor TF(t,x) varía según la frecuencia de ocurrencia de t en X . El valor de IDF(t) varía con el inverso del número de documentos que contienen el término t. Los valores de TF-IDF son creados para producir resultados de búsquedas de alta calidad, por tanto, Lucene utiliza TF-IDF, la similitud de una consulta en relación con un documento de la colección, se calcula basado en el puntaje (score) dado por la similitud del Coseno [11].Lucene [6] refina la forma de calcular el puntaje (score) de similitud para obtener calidad y lo simplifica llevando a una forma práctica, representada en la Ecuación 1:

$$score(q,d) = \ coord(q,d) \cdot queryNorm(q) \cdot \sum_{t\ in\ q} (tf(t\ in\ d) \cdot idf(t)^2 \cdot tgetBoost(\ \ ) \cdot norm(t,d)$$

Ecuación 1. Función de puntajes práctica de Lucene. Fuente: [6]

Donde,

- tf(t in d), definida como el número de veces que un término aparece en score del documento d actualmente. tf(t in q) se asume como 1, por tanto, no aparece en la ecuación. Su valor se calcula como aparece en la Ecuación 2:

$$\boldsymbol{tf\ (t\ in\ d) = frecuencia^{1/2}}$$

Ecuación 2. Tf (t in d). Fuente: [6]

- Idf(t) representa la frecuencia inversa del documento, este valor correlaciona el número de documento en el cual el termino t aparece, su cálculo está dado por la Ecuación 3:

$$idf(\ t) = 1 + \log(\frac{numDocs}{docFreq + 1})$$

Ecuación 3. Cálculo idf(t). Fuente: [6]

- Coord(q,d) es un factor de puntaje (score) basado en cuantos de los términos de la consulta son encontrados en un documento específico.

- queryNorm(q) es un factor de normalización usado para asignar puntajes entre consultas comparables. Este factor no afecta el ranking de documentos sino que intenta hacer que las puntuaciones de las consultas sean comparables. Se calcula según la Ecuación 4:

[21] TF es la frecuencia absoluta de aparición de un término en un documento [3]
[22] IDF es la frecuencia inversa de documento, que permite potenciar (boost) un término no tan frecuente en un documento de la colección, es decir, un término con menor frecuencia tiene un valor de IDF que uno con más apariciones (TF) [3]

$$\boldsymbol{queryNorm(q) = queryNorm(sumOfsquaredWeights) = 1/(sumOfSquearedWeights^{\frac{1}{2}})}$$

Ecuación 4. Cálculo de queryNorm. Fuente: [6]

Donde, $sumOfSquaredWeights = q\ .\ getBoost()^2\ .\ \sum_{t\ in\ q}(idf(t)\ .t.\ getBoost())^2$

- t.getBoost() es una búsqueda del Boost (valor de preferencia) para el término t en la consulta q, como se especifica en el texto de la consulta.

- Norm(t,d), es calculada cuando un documento se adiciona al índice, incluye varios factores de preferencia o relevancia (boost), que son multiplicados; estos factores se describen a continuación y son calculados o asignados en tiempo de indexación como:

    - document boost (doc.setBoost(): que permite configurar un factor de preferencia (boost factor) para cada acceso por coincidencia en un documento, este valor es multiplicado dentro del puntaje para todos los accesos a ese documento.
    - field bost (field.setBoost): es un factor de preferencia que da pesos a los accesos al documento en donde hay coincidencia. Este valor es multiplicado en el puntaje de todos los accesos por coincidencia de un documento.
    - LengthNorm, se calcula cuando el documento se adiciona al índice según el número de tokens coincidentes en el documento.

En la Figura 2, se presenta el proceso de búsqueda en Lucene, el cual está constituido por la indexación y el procesamiento de la consulta.

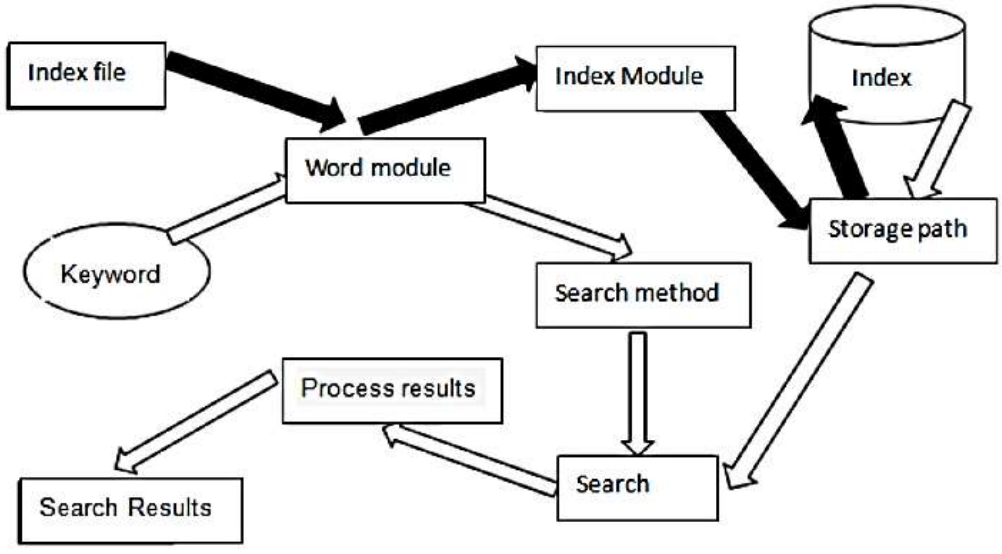


Figura 2. Descripción del flujo de trabajo de Lucene. Fuente: [54]

La arquitectura de Lucene está construida por capas que tienen bajo acoplamiento de características. Lucene consta de un paquete de infraestructura, el indexador e interfaces

externas. El código fuente de Lucene está dividido en 7 módulos y tiene tres clases centrales: análisis, indexación y búsqueda [54].

1.1.6 Medidas de evaluación

Las medidas más ampliamente usadas para la evaluación de un SRI son:

1. Precisión (Precision – P)

Indica la proporción de documentos recuperados que son relevantes, se calcula mediante la Ecuación 5 [3].

$$Precision = {\#(items\ relevantes\ recuperados)}\Big/{\#(items\ recuperados)}.$$

Ecuación 5. Fórmula de Precisión. Fuente: [3]

2. Recuerdo (Recall – R)

Es la porción de documentos relevantes que son recuperados [3], se calcula mediante la Ecuación 6.

$$Recuerdo = {\#(items\ relevantes\ recuperados)}\Big/{\#(items\ relevantes)}.$$

Ecuación 6. Fórmula de Precisión. Fuente: [3]

Por ejemplo, si el sistema recupera 10 documentos de una colección de 100 y éstos previamente han sido juzgados como relevantes o no relevantes. Si se recuperan 8 relevantes de un total de 62 documentos relevantes en la colección, la precisión será igual a 8/10, es decir, 0,8 y el recuerdo será igual a 8/62, es decir, 0,13 [55].

La precisión y el recuerdo son las medidas más frecuentemente usadas en la evaluación del desempeño en recuperación de información, muchas de las otras medidas estándar se basan en estos dos conceptos [14].

3. Medida F

Resume la precisión y el recuerdo en un punto, por tanto, la Medida F es la media armónica entre la precisión y el recuerdo, se calcula según la Ecuación 7.

$$F = \frac{(\beta^2+1)PR}{\beta^2 P+R} \text{ Donde } \beta^2 = \frac{1-\alpha}{\alpha}\text{; cuando } \beta = 1 \rightarrow F_{\beta=1} = \frac{2PR}{P+R}$$

Ecuación 7. Fórmula de Medida F. Fuente: [3]

Si se quiere enfatizar en la precisión se debe escoger un β bajo, y en caso de que se quiera enfatizar en recuerdo β debe ser alto [55].

4. Curva de Precisión Recuerdo (Precision – Recall Curves)

Consiste en dibujar una serie de puntos de corte ranqueados e interpolados de valores de precisión y recuerdo [55], los valores son promediados para todas las consultas del sistema de recuperación en cada nivel de recuerdo, estos puntos que forman la curva, permiten

comparar y analizar el desempeño de un sistema de recuperación de información utilizando diferentes algoritmos, por ejemplo, un algoritmo nuevo propuesto en contraste con un algoritmo tradicional [11]. En la Figura 3, se presenta la curva de precisión promedio versus recuerdo para dos algoritmos de recuperación distintos, en donde, el algoritmo B tiene mayor precisión para valores bajos de recuerdo mientras que el algoritmo A es superior para valores altos de recuerdo [11].

5. *Precisión media en documentos relevantes observados (Average Precision at seen relevant documents - AP) [11]*

Genera un valor resumen de la lista ordenada (ranqueada) de documentos recuperados promediando la precisión que se obtiene después de que un nuevo documento relevante es encontrado; por ejemplo para una consulta q1, se tiene el listado ordenado de documentos recuperados por un SRI, mostrado en la Tabla 3:

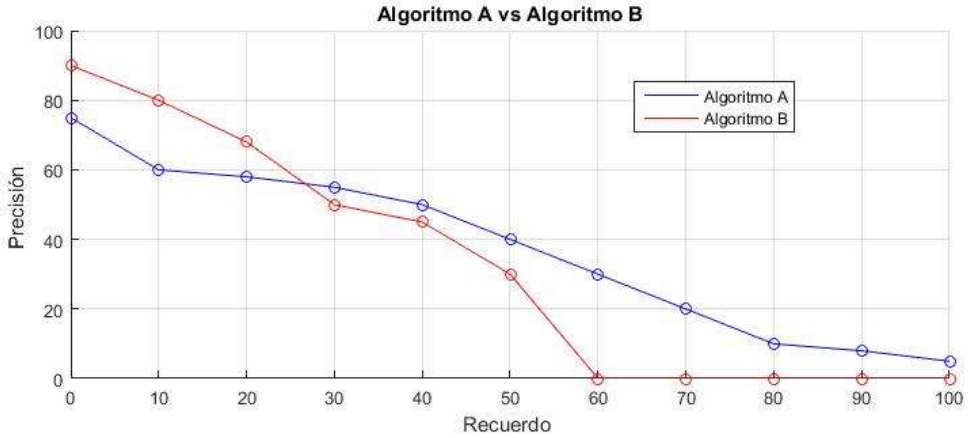


Figura 3. Ejemplo Curva Precisión – Recuerdo. Fuente: [11]

Resultado	Documento Recuperado	Relevante	Resultado	Documento Recuperado	Relevante	Resultado	Documento Recuperado	Relevante
1	Doc_123	Si	6	Doc_9	Si	11	Doc_38	NO
2	Doc_84	NO	7	Doc_511	NO	12	Doc_48	NO
3	Doc_56	Si	8	Doc_129	NO	13	Doc_250	NO
4	Doc_6	NO	9	Doc_187	NO	14	Doc_113	NO
5	Doc_8	NO	10	Doc_25	Si	15	Doc_3	Si

Tabla 3. Ejemplo de documentos recuperados por un SRI. Fuente: [11]

La precisión promedio se calcula con el promedio de la sumatoria de las precisiones después de cada nuevo documento relevante recuperado, como se muestra en la Tabla 4:

Resultado	Documento Relevante Recuperado	Precisión
1	Doc_123	1=1/1
3	Doc_56	0,66=2/3
6	Doc_9	0,5=3/6
10	Doc_25	0,4=4/10
15	Doc_3	0,3=5/15
Sumatoria de precisiones		2,86

Resultado	Documento Relevante Recuperado	Precisión
Numero de precisiones		5
Precisión promedio (AP)		**2,86/5 =0,57**

Tabla 4. Ejemplo de precisión promedio. Fuente: [11]

Esta medida favorece a los sistemas que recuperan rápidamente documentos relevantes.

6. Precisión R (R-Precision)

Genera un valor singular resumen de una lista ordenada mediante el cálculo de la precisión en una posición R de la lista, donde R es el número total de documentos relevantes para la actual consulta [11]. Requiere un conocimiento completo del conjunto de documentos relevantes para una consulta y así se pueden evaluar tanto consultas con muchos documentos relevantes como con pocos. [12], por ejemplo, para el caso mostrado en la Tabla 3 y Tabla 4, para R = 10 hay cuatro documentos relevantes en la lista de 10 documentos recuperados, entonces Precisión R es igual a 0,4.

R-precision puede ser entendida como una aproximación al área bajo la curva de precisión vs recuerdo.

7. Histograma de precisión (Precision Histograms) [11]

La medida de precisión R para varias consultas permite comparar el desempeño histórico de dos algoritmos mediante un histograma de los valores de Precisión R para cada algoritmo en cada consulta, es decir, el valor representado en el histograma se calcula mediante la Ecuación 8, en la cual $RP_{A(i)}$ y $RP_{B(i)}$, son valores de Precisión R para el algoritmo A y B respectivamente para un valor i de la lista ordenada de documentos recuperados.

$$RP_{A/B(i)} = RP_A(i) - RP_B(i)$$

Ecuación 8. Histogramas de Precisión. Fuente: [11]

Haciendo la revisión visual del histograma de precisión se puede concluir que un valor de RPA/B(i) igual a cero indica que el desempeño de ambos algoritmos es similar, un valor positivo indica un mejor desempeño del algoritmo A, y un valor negativo indica un mejor desempeño del algoritmo B, como se puede apreciar en la Figura 4, en donde, para el algoritmo A se aprecia un mejor desempeño para todas las consultas excepto para las consultas 4 y 5, dado que en su mayoría los valores RPA/B(i) son positivos.

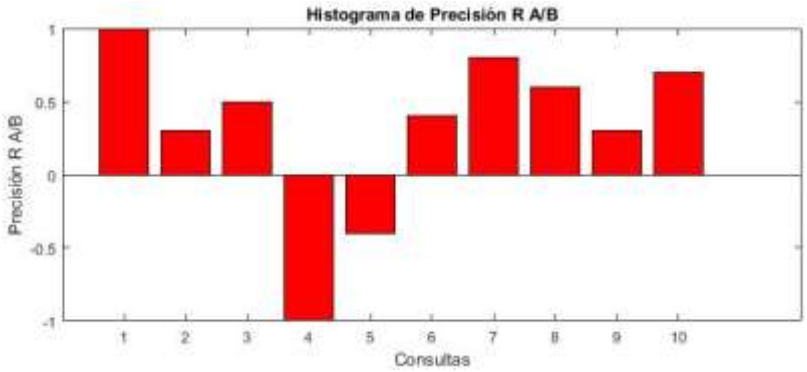


Figura 4. Ejemplo Histograma de Precisión R. Fuente: [11]

1.2 TRABAJOS RELEVANTES

A continuación se presenta una serie de trabajos que han sido seleccionados como relevantes para el desarrollo de esta monografía, porque presentan métodos para la construcción de colecciones de prueba, adaptación y propuesta de analizadores léxicos, y definición de listas de palabras vacías, lo cual es parte fundamental para este desarrollo presentado en este documento.

1.2.1 Colección de prueba

Algunos trabajos relevantes que permiten apreciar la necesidad de construir colecciones de prueba en lenguas específicas para evaluar un SRI: 1) un trabajo sobre el Sorani, una de las principales líneas de la lengua Kurda, su principal contribución es el Pewan, la primera colección de prueba estándar disponible para evaluar sistemas de recuperación de información Sorani [56]; 2) Colección de prueba en Italiano, conformado por noticias completas, consultas hipotéticas, consultas reales de posibles usuarios, una clasificación manual realizada por expertos y un sistema de clasificación automático de documentos [57, p. 281]; 3) una de las primeras colecciones de prueba para Farsi, lengua oficial de Irán, presenta el proceso de construcción de la colección y sus comparaciones con otras colecciones de Farsi [44]; 4) una colección de prueba desarrollada para responder preguntas en lengua macedonio, que puede ser usada para desarrollar y evaluar SRI, la colección consta de 4 documentos y 163 preguntas de múltiple opción tomadas de los cursos de Historia de la Informática y Aplicaciones en computador que son parte del currículo de la Universidad, los resultados preliminares mostraron que a pesar de ser pequeña puede ser efectivamente usada para su propósito [58]; 5) una colección de prueba para Hamshahri, de la lengua Persa (o Farsi) conformada por artículos de periódico de 1996 a 2002, el tamaño de los documentos varía desde noticias cortas (menos de 1KB) a artículos largos en promedio de 1.8 KB [59]. En la Tabla 5 se presenta un resumen general de las colecciones presentadas anteriormente.

Colección	Lenguaje	Tópicos	# documentos	Tamaño
Pewan [56]	Sorani	Artículos de noticias (2003 -2013). Adicionalmente una lista de prefijos y sufijos, stopwords, traducción de inglés de las consultas de Pewan	115.340	97,8 MB
Colección de prueba de Italiano [57, p. 281]	Italiano	Un gran conjunto de documentos completos de noticias de un año de un periódico de italiano	70.000 documentos	No Disponible
Mahak [44]	Farsi	Artículos de noticias de ISNA EN 12 categorías de noticias	3007 documentos 216 consultas	No Disponible
Para responder preguntas [58]	Macedonio	Documentos de cursos de historia y computación de la Institute of Informatics at the Faculty of Natural	4 documentos y 163 preguntas de	No Disponible

Colección	Lenguaje	Tópicos	# documentos	Tamaño
		Sciences and Mathematics en University in Skopje	selección múltiple	
Hamshahri [59]	Farsi (Persa)	Artículos de noticias de periódicos de 1996 a 2002	166.774 documentos y 65 consultas	564 MB

Tabla 5. Descripción de colecciones para lenguas específicas.

1.2.2 Modelo de espacio vectorial

A continuación se presentan algunos trabajos relacionados con la adaptación y utilización del modelo espacio vectorial para la recuperación de textos escritos: Wang [60], presenta un modelo de espacio vectorial para medir la similitud entre la consulta y el documento cuando responden a una consulta con múltiples términos. En el modelo propuesto la consulta se expresa mediante el vector q y el documento mediante el vector D, entonces se calcula el peso de valor S_i que corresponde a cada elemento del vector D así: $S_i = \sum_{j=1}^{N} W_{ij}F(j), donde\ 1 \leq i \leq M$. Un nuevo término W_{ij} es introducido en el modelo relacionado a cada elemento del Vector D con el vector q. Cuando una palabra en posición i en el vector D es una palabra clave (keyword), W_{ij} es el número de grupos tipo J que contiene esta palabra. Cuando la palabra en la posición i no es keyword, W_{ij} es el número de veces que esta palabra aparece en un documento y W_{ij}es cero, cuando j>1. Karshenas y Dimililer [61] presentan el diseño y construcción del Sistema de Recuperación de textos PIRS (Precision Information Retrieval System) basado en una modificación del modelo espacio vectorial que incluye una nueva fórmula con pesos para la consulta y una función de similitud; el sistema incluye 10 módulos entre los que se encuentran: Temporary Collection, ScanDir, ReWrite function, LCollection, Run Function, StemCollection, Indexer, Inverted File, Tree, y Vocabulary, con estas 2 modificaciones se mejoraron los niveles de precisión promedio del sistema. En 1997, Lee y otros [62], revisan varias implementaciones para la representación de documentos y consultas utilizando el modelo espacio vectorial. El método 1, full vector-space model, la representación de los vectores de documento es solo conceptual, rara vez se almacena internamente dado que es grande y disperso. El método 2, utiliza raíz cuadrada de los números de términos en un documento como factor de normalización. El método 3, permite simplificar la computación eliminando el factor de normalización. El método 4, solo hace uso de las frecuencias de los términos e ignora idf. El método 5, ignora la frecuencia de los términos pero conserva los valores de idf para determinar los pesos de los términos. El método 6, ignora los valores de idf y tf y mide el número de términos comunes en los vectores de los documentos y las consultas, obteniendo entre otros resultados: 1) Para consultas conceptuales es mejor utilizar solo frecuencia inversa de documentos e ignorar la frecuencia de términos, pero para consultas en lenguaje natural ambas frecuencias se deben tener en cuenta; 2) El factor de normalización encuentra mejores resultados y computacionalmente es más eficiente; 3) Para desarrollar un sistema de recuperación eficiente y efectivo se deben tener en cuenta varias consideraciones, como por ejemplo, la estructura del index, los métodos de procesamiento de las consultas cuyo índice se base en semánticas y le dan relativa importancia a los términos de las consultas, pueden reducir el tiempo de procesamiento considerablemente.

Estos y otros trabajos permitieron estudiar la aplicabilidad del modelo espacio vectorial en sistemas de recuperación y sus usos en lenguas que tienen poco o ningún antecedente en recuperación de información. También facilitó visualizar los diferentes enfoques que tiene el modelo espacio vectorial.

1.2.3 Análisis Léxico (Tokenizer)

En 2015, Hammo, Yagi, y otros [63], describen como se puede utilizar un proceso de recuperación de información para el Arábigo, que estudia los cambios semánticos de la lengua, en relación con la frecuencia de uso de una palabra a través del tiempo y cómo cambia su significado en el tiempo; el tokenizador implementado para este procesamiento incluye dos salidas: tokens que corresponden a unidades léxicas cuyos caracteres son reconocibles y tokens que necesitan más análisis morfológico y los tokens de un carácter de longitud no se tuvieron en cuenta.

En 2013, Guan y Zhang [64], describen aspectos relevantes sobre el procesamiento de texto Chino (a diferencia del Inglés, en el cual las palabras en las oraciones están separadas por espacios, en el Chino no existe espacios entre las palabras) en la forma como las oraciones están divididas en una secuencia razonable de palabras (Chinese word segmentation); el algoritmo de segmentación de palabras se puede realizar de tres formas: basado en reglas, basado en estadística y basado en comprensión. Este artículo propone un algoritmo que mejora el método de segmentación de texto aplicado al modelo espacio vectorial, que hace una gran mejora en las tasas de similitud sobre textos de noticias en chino. Los resultados del algoritmo mostraron que son casi cercanos a la segmentación humana y puede remplazarla, ahorrando tiempo y dinero y mejorando la eficiencia de la similitud de textos de noticias.

En 2006, Jiang y Zhai [65], presentan la generalización de varias estrategias de tokenización en un conjunto de heurísticas organizadas para ser aplicadas en la recuperación de información en biomedicina, concluyendo que remover caracteres no funcionales es seguro, y que se pueden aplicar diferentes tipos de tokenización (tres métodos de normalización de punto de quiebre y tres conjuntos de punto de quiebre), lo cual presenta mejoras en el desempeño de hasta 8% en la medida de evaluación utilizada (MAP).

En 2004, Klatt y Bohnet [40], presentan un proceso de tokenización para una colección de prueba de Alemán, el cual consta de dos estados, así: 1) El texto de entrada es separado en sus respectivos tokens, mediante una separación de espacios en blanco, luego signos de puntuación del inicio y el final son removidos, para puntuaciones en el medio se obtienen tres tokens y luego se repite el proceso hasta obtener los tokens, con este proceso se obtiene información para tomar decisiones en el siguiente paso; la información se organiza así: la primera columna contiene los tokens separados, la segunda columna señala si los tokens son compuestos por caracteres alfanuméricos o guiones o si tienen algún posible problema léxico; en la tercera columna se codifica el proceso de separación; en la cuarta columna se almacenan los niveles de recursión para obtener el token. 2) Se complementa la información anterior en dos columnas así: en la primera, representa el texto tokenizado, en la segunda, contiene información adicional a la presentada en el anterior estado. Es decir, se incluyen las anotaciones de estructuras lingüísticas tales como oraciones y las decisiones que puedan generar algún inconveniente.

Los antecedentes citados mostraron diferentes enfoques para realizar la adaptación de un analizador léxico para una lengua en recuperación de información, como por ejemplo, definición de reglas heurísticas, lo cual ha sido un aporte significativo para el desarrollo de esta propuesta.

1.2.4 Remoción de palabras vacías (Stopwords Removal List)

En 2015, Hammo, Yagi, y otros [63], también describen que utilizaron una lista de palabras vacías que aportan poco al significado y tienen alta frecuencia, de tal forma, que se mejora la velocidad de búsqueda, adicionalmente, se pudo apreciar que esta lista de palabras difícilmente ha cambiado durante los pasados 16 siglos.

En 2012, Joshi y otros [66], presentan experimentos con la eliminación de palabras vacías en la lengua Gujarati usando la medida MAP (Mean Average Precision). Los resultados muestran que la eliminación de palabras vacías en documentos de textos en Gujarati incrementan significativamente los valores de precisión en las tareas de recuperación de información, también se detectó que ampliando una consulta de Titulo a Titulo-Descripción-Narrativa mejora la efectividad de la recuperación cerca de un 24%.

En 2011, Zaman y otros [67], evalúan el uso de las listas de palabras vacías para inglés en sistemas de recuperación de información basado en Indexación semántica latente (LSI) con grandes datasets de texto. El principal hallazgo es que para sistemas de recuperación de información basados en LSI, el uso arbitrario de una lista de palabras vacías reduce el desempeño de la recuperación y para mejorarla se debe adaptar una lista para cada gran conjunto de datasets.

En 2009, Pandey y otros [68], presentan un experimento incluyendo y excluyendo la tarea de remoción de palabras vacías, en un sistema de RI para Hindi, sobre una colección de prueba elaborada (se tomaron 700 documentos del corpus EMILEE y 70 consultas), obteniendo que el uso de palabras vacías mejora el desempeño del Sistema de RI.

En 2006, Jiang y Zhai [65], también evaluaron empíricamente el impacto de utilizar lista de palabras vacías sobre el desempeño del sistema encontrando que no mejora el desempeño o ligeramente lo mejora, en el experimento se utilizaron 132 palabras vacías de inglés tomadas del PUBMED.

Estos trabajos entre otros, permitieron apreciar los métodos para definir la lista de palabras vacías y los diferentes impactos en el desempeño del sistema de recuperación, es decir, se encontraron trabajos en donde la remoción de las palabras vacías mejoró el desempeño y otros en donde fue lo contrario, lo cual aportó una referente para analizar los resultados obtenidos para el nasa yuwe.

1.2.5 Lucene en la recuperación de información

En 2013, Zhang y Zhan [69], presentan la construcción de un sistema de recuperación basado en Lucene, allí se describe la arquitectura y los componentes de Lucene utilizados para la construcción de Framework Compass el cual está basado en Lucene. En 2012, Li y otros [70], describen los componentes y la arquitectura de Lucene, encontrando que provee una oportunidad para el aprendizaje y diseño de los sistemas de recuperación de información. En 2011, Cui, Chen y Li [54], presentan la construcción de un Sistema de recuperación de información para el chino basado en el modelo de Lucene y es aplicado

para la recuperación de datos de investigación de un sistema de información Forestal, encontrando que el sistema es eficiente y rápido en la recuperación [54].

Estos y otros trabajos sobre Lucene permitieron visualizar la aplicabilidad de la arquitectura de Lucene en la construcción del sistema de recuperación de información para textos escritos en diferentes lenguas.

2. COLECCIÓN DE PRUEBA

En esta sección se introduce el proceso de construcción de los principales componentes de la colección de prueba elaborada para nasa yuwe.

2.1 DESCRIPCIÓN DEL PROCESO DE CONSTRUCCIÓN

Para la construcción de la colección de prueba para textos escritos en nasa yuwe se tuvo en cuenta: 1) trabajo de campo con la comunidad nasa, se utilizó la metodología de Investigación participativa [71], que permitió resolver preguntas sobre la lengua nasa, determinar sus características y así limitar los aspectos de la lengua que se tomaron en cuenta para la construcción de la colección de prueba, involucrando a los miembros de la comunidad como sujetos activos y 2) análisis estadístico de los datos obtenidos tanto de documentos, consultas y juicios. En la Figura 5, se presenta un breve esquema del proceso utilizado para construir la colección, teniendo en cuenta las dinámicas propias de la comunidad nasa y del nasa yuwe.

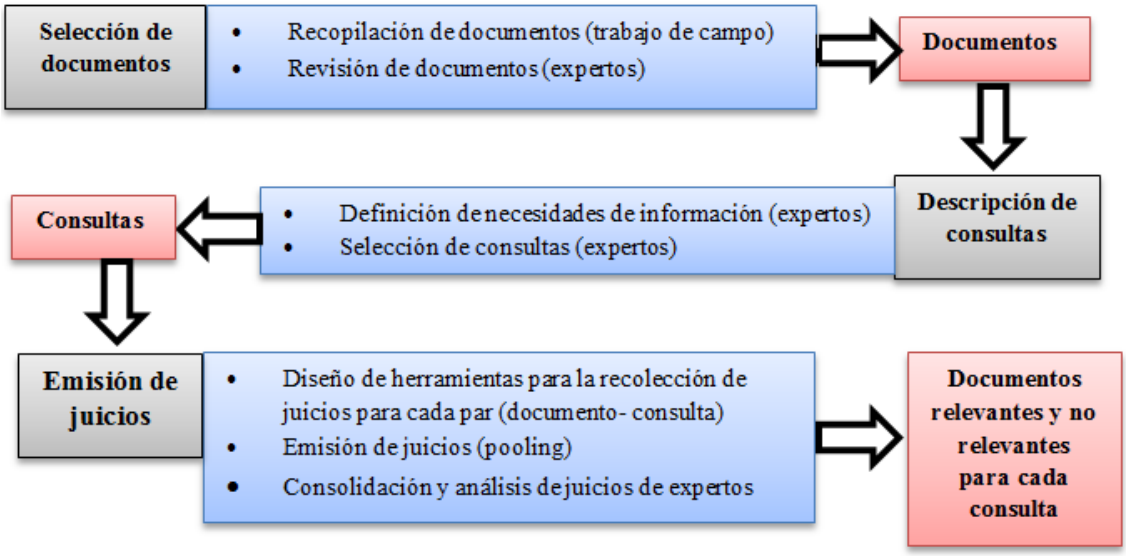


Figura 5. Descripción del proceso de construcción de la colección de prueba. Fuente: Elaboración propia

2.1.1 Selección de documentos

- Para la recopilación de documentos, se realizó trabajo de campo, teniendo en cuenta la existencia de muchas variantes en la lengua nasa yuwe, asociadas a la zona geográfica en la cual se encuentra ubicado el resguardo de cada comunidad, es así, que fue necesario delimitar los documentos de la colección a aquellos escritos en alfabeto unificado [48]. Cabe destacar que no son muchos los documentos escritos dado lo nuevo del alfabeto y la condición sociolingüística de la lengua.

- Para la revisión de los documentos seleccionados se realizó trabajo con expertos hablantes de nasa yuwe, el trabajo consistió en la revisión de estos textos tanto en su escritura como en su gramática.

2.1.2 Descripción de Consultas

- Para la definición de necesidades de información se tuvieron en cuenta los temas tratados en los documentos seleccionados, obteniendo una lista de 20 necesidades de información.

- La selección de consultas, fue realizada mediante priorización de los temas con más necesidad de información en referencia a los textos seleccionados, se trabajó con expertos en el tema para finalmente obtener 8 consultas.

2.1.3 Emisión de juicios

- Para el diseño de herramientas que soportaran el proceso de recolección de los juicios de relevancia para cada par (documento – consulta), se tuvieron en cuenta dos escenarios: 1) Recolección de juicios de expertos de manera virtual, por tal motivo se diseñó una herramienta que permitiese recolectar esta información mediante una aplicación web (ver Figura 8). 2) Recolección de juicios mediante formato impreso en donde cada experto emite su juicio para cada par (ver Figura 9 y Figura 10).

  Los instrumentos desarrollados para la emisión de juicios de cada par (documento - consulta), tuvieron en cuenta la definición de una escala de 4 valores[23]: Muy Relevante (MR), Relevante(R), Poco Relevante (PR) y No Relevante (NR), con los cuales se evita confusiones a los expertos en la recopilación de información, en pro de facilitar el proceso de emisión de juicios para cada par al tener un rango más amplio para emitir el juicio, en lugar de una escala de Relevante y No Relevante.

- Para la recolección de juicios para cada par (documento – consulta), se utilizó el juicio de expertos, dado que no se tenían juicios previos al respecto o preselecciones de algún sistema de recuperación de información. La selección de expertos, fue realizada teniendo en cuenta diversidad de variantes en los hablantes de nasa yuwe, en su mayoría fueron docentes interesados en la visualización de la lengua.

2.2 ELEMENTOS DE LA COLECCIÓN

2.2.1 Documentos de la colección

En primera instancia, se seleccionaron documentos relacionados con cuentos de la cultura y cosmovisión nasa, a continuación se describen los textos seleccionados:

- Area Nasawe'sx Fxinzenxi - Cuentos y Cosmovisión Nasa [72]. En este texto se encuentran historias propias de la vida nasa escritas en nasa yuwe y con traducción contextualizada al castellano.

[23] Se tuvo en cuenta las recomendaciones de la escala Likert, tanto para definir los 4 niveles como para su análisis.

- Eç thegya’ ipi’ki’ tha’w - Te invitamos a leer [73]. Se encuentran textos cortos sobre descripciones de la vida nasa y se proponen algunas actividades para apoyar esas descripciones escritas en nasa yuwe y con traducción al castellano.
- Nasawe’sx Kiwaka Fxi’zenxi Ẽen. [74] – Tiempos de la vida en el territorio nasa. En este texto se encuentra una investigación sobre los conocimientos ancestrales de la cultura nasa escrita tanto en nasa yuwe como en castellano.
- Pees kupx fxi’zenxi - La metamorfosis de la vida. [75]. De este texto se tomaron historias sobre los caciques nasa en tiempos de antaño.

En segunda instancia, se realizó la revisión de los documentos lo cual incluyó: su digitalización y corrección a nivel de caracteres especiales propios de la escritura del nasa yuwe y a nivel de la gramática del nasa yuwe, se contó con el apoyo de dos profesores hablantes de nasa yuwe, con quienes se hizo una evaluación preliminar de los textos en referencia a la emisión de juicios, obteniendo que el texto de [75] a pesar de estar escrito en alfabeto unificado, no era apropiado para integrar la colección dada la dificultad que se presenta en su lectura, por el uso de palabras específicas de la variante de Toribio que no son comúnmente usadas en otras variantes del nasa yuwe, por tanto, se sacaron estas historias de los documentos de la colección. Es así que quedaron los tres primeros textos descritos. Todos los documentos se dejaron en formato texto UTF-8, para soportar los caracteres del alfabeto unificado de nasa yuwe, en su mayoría los documentos cuentan con un título y el texto. En la Tabla 6, se presenta un fragmento de un documento escrito en nasa yuwe.

Texto escrito en nasa yuwe	*Texto escrito en español*
Nasa vxanxi’s pta’sxnxi. Txaniteya’ kiwe wala u’sene’yũ’ aça’ khã'sx ũskiweçxane’yũ’ mẽh kũh jwed ksxa’w ũskiweyũ’ne’sa’, vxite ne'jwe’sxtayu’ aça’ vxitesa’ nuuçxkwẽsane’tayu’. Ne'jwe’sxyũ’ puutx ptamne’tayũ’. Piçthẽ’jsa’ tayne’ yaaseyũ’, uysa’ umane’ yaaseyu’. Naa je’zsa ũusyahtxya’ uweçxa naa kiwete kĩhçxahne’ peejxsa ũsu’ txãatxi’s vxitya’ takhne’nta: Yu’a’s, fxtũu tasxtxi’s, kwettxi’s, kĩnjwã jxukane’nta vxitsa’, nmehte’ naa kiwe’ nasane’ peejxyu’ aça’ nasa’swa vxitya’ yahtxne’nta kxteeçxãh puutx fxi’zekahn jĩçxa.	**El origen del Hombre.** Anteriormente cuando la tierra era joven solo existían espíritus de viento, algunos eran espíritus mayores y otros eran espíritus menores o subalternos, cada uno de estos tenía su propia pareja, uno de los espíritus mayores que se llamaba TAY y la otra que se llamaba UMA, se pusieron a pensar en cómo poblar la tierra y decidieron crear el agua, las plantas, los animales, las piedras y todo lo que hoy en día existe en la tierra; por ultimo decidieron crear al hombre para que este cuidara de ellos y que todos viviéramos en armonía, juntos en un mismo lugar.

Tabla 6. Fragmento de cuento escrito en nasa yuwe: “El origen del pueblo nasa”. Fuente: [72]

Finalmente, se obtuvieron 97 documentos de texto escritos en nasa yuwe para la colección de prueba, con un tamaño promedio de 1.5 KB por documento. En la Figura 6, se presenta la distribución de tamaño de los documentos de la colección; la cantidad de términos en los documentos de la colección oscila entre 15 y 500, y en promedio los documentos tienen 103 términos.

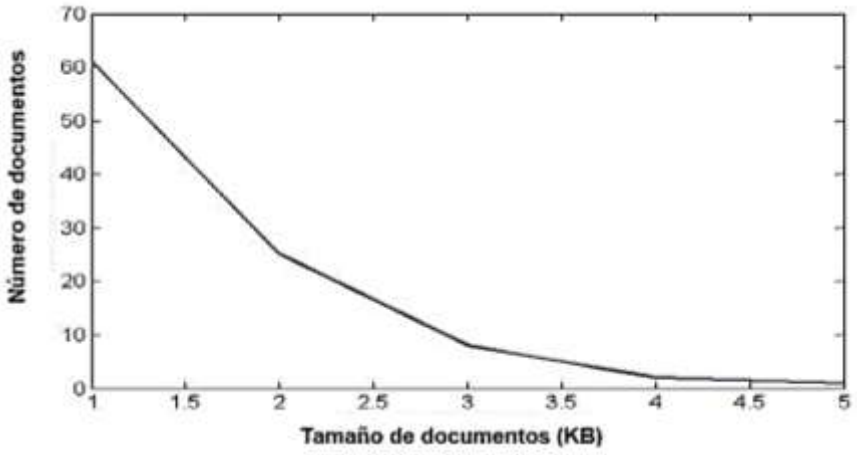


Figura 6. Descripción del tamaño de los documentos de la colección de prueba. Fuente: elaboración propia

2.2.2 Consultas

En Tabla 8, se puede apreciar la lista de consultas seleccionadas para la colección de prueba.

N°	Necesidad de información	Titulo	Descripción Consulta en nasa yuwe	Palabras por consulta	Narrativa
1	Conocer sobre el origen y época (siembra) del maíz	Kutxh	kutxh yuwe's jiyuka ki' uh a'te's txãwẽy	7	Documentos relevantes que podrían informar al usuario nasa del SRI sobre el origen y la época del maíz.
2	Conocer sobre ayuda y trabajo de los Médicos tradicionales	Thẽ' walawe'sx	Thẽ' walawe'sx majii	2	Documentos relevantes que podrían informar al usuario nasa del SRI sobre los médicos tradicionales de la comunidad
3	Conocer sobre épocas de la luna	A'te	A'te dxi'j / a'te yuwe's jiyuna	6	Documentos relevantes que podrían informar al usuario nasa del SIR de información sobre las épocas de la luna.
4	historias sobre origen/nacimiento - aparición en la cultura nasa	Upxhnxi /vanxi	Upxhnxi yuwe / vxanxi yuwe	5	Documentos relevantes que podrían informar al usuario nasa del SRI sobre origen y aparición en la cultura nasa.
5	Sobre la gallina y el pollo en la cultura nasa	Atalx	Atalx wejxa's jiyuna	3	Documentos relevantes que podrían informar al usuario nasa del sistema de recuperación de información sobre el pollo y la gallina en la cultura nasa
6	Historias sobre el Viento en la cultura nasa	Wejxa	Wejxa yuwe's jiyuna	3	Documentos relevantes que podrían informar al usuario nasa SRI sobre las historias del viento en la cultura nasa
7	Historias sobre el sol en la cultura nasa	Sek	Sek yuwe's jiyuna /sek dxi'j	5	Documentos relevantes que podrían informar al usuario nasa del SRI sobre las historias del sol en la cultura nasa

N°	Necesidad de información	Titulo	Descripción Consulta en nasa yuwe	Palabras por consulta	Narrativa
8	Historias sobre caciques en la cultura nasa	Sa'twe'sx	Sa'twe'sx yuwe's jiyuna	3	Documentos relevantes que podrían informar al usuario nasa del SRI sobre historias de los caciques nasa.

Tabla 7. Descripción de las consultas

En la Figura 7, se muestra la distribución de términos por consulta con 4.4 palabras por consulta en promedio, con un mínimo de 3 y un máximo de 7 palabras, lo cual permite apreciar la cantidad de términos requeridos para expresar las necesidades de información en esta lengua. Adicionalmente, también se revisó que la interpretación de la consulta por parte de los expertos depende del contexto y la ubicación de las palabras por tanto, si se adicionan o disminuyen más palabras clave en la consulta el juicio de un par específico podría variar, como se puede apreciar con las consultas 5 (Atalx wejxa's jiyuna) y la consulta 6 (Wejxa yuwe's jiyuna), en la consulta 5 la palabra "wejxa's" habla sobre historias, en la consulta 6 la palabra "Wejxa" hace referencia al viento, es decir, cambiando la vocal oral por una oral interrupta, el significado de una necesidad de información en nasa yuwe puede variar mucho.

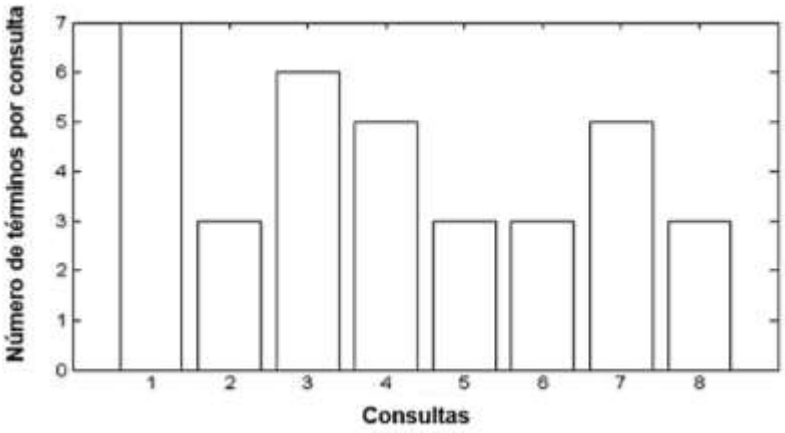


Figura 7. Longitud de las Consultas (# of términos). Fuente: Elaboración propia

2.2.3 Herramientas diseñadas para la recolección de juicios

En la recolección de los juicios por parte de los expertos, en primera instancia, se pensó en hacerlo mediante un prototipo que se desarrolló haciendo uso de un servicio Web que permitiese a los expertos visualizar los documentos e ir emitiendo los juicios, esto no fue posible debido a: las distancias geográficas en las que se encuentra ubicado cada experto, en algunos casos, la carencia de infraestructura tecnológica para soportar el uso de esta aplicación y un bajo nivel de alfabetización digital. En la Figura 8, se presenta la interfaz que fue desarrollada para realizar esta tarea. Aunque esta aplicación software de escritorio no se pudo utilizar, quedo implementada y funcional para ampliar y mejorar posteriormente la colección. A continuación se presenta una breve descripción de cada zona de la interfaz presentada:

- La zona 1, muestra el número y la descripción de la consulta, permite seleccionar cada consulta (1 a 8).
- La zona 2, ofrece la posibilidad de emitir el juicio sobre cada documento en relación con la consulta que este seleccionada, las opciones son Muy Relevante, Relevante, Poco Relevante, No Relevante y Sin Revisar.
- La zona 3, presenta el título de cada documento sobre el cual se va a emitir el juicio de relevancia.
- La zona 4, se visualiza el texto de cada documento.
- La zona 5, enmarcada dentro de las llaves ({}) es la zona de visualización de los documentos, se configuró para mostrar 6 documentos por página.
- La zona 6, permite manipular las páginas con botones anterior y siguiente y la visualización del número de página.

Cabe mencionar que si en algún momento se llega a utilizar esta interfaz deberá ser sometida a evaluaciones de usabilidad en pro de que su uso no afecte la calidad del juicio del experto.

En segunda instancia, se optó por la opción de recolectar los juicios de manera manual, por tanto se entregó a cada experto un paquete conformado por los documentos impresos de la colección y un formato para facilitarle la emisión de los juicios a cada experto, en la Figura 9, se puede apreciar el formato usado para el registro de la evaluación. Cada consulta tiene un color diferente para facilitar su visualización en la emisión del juicio de cada par documento – consulta.

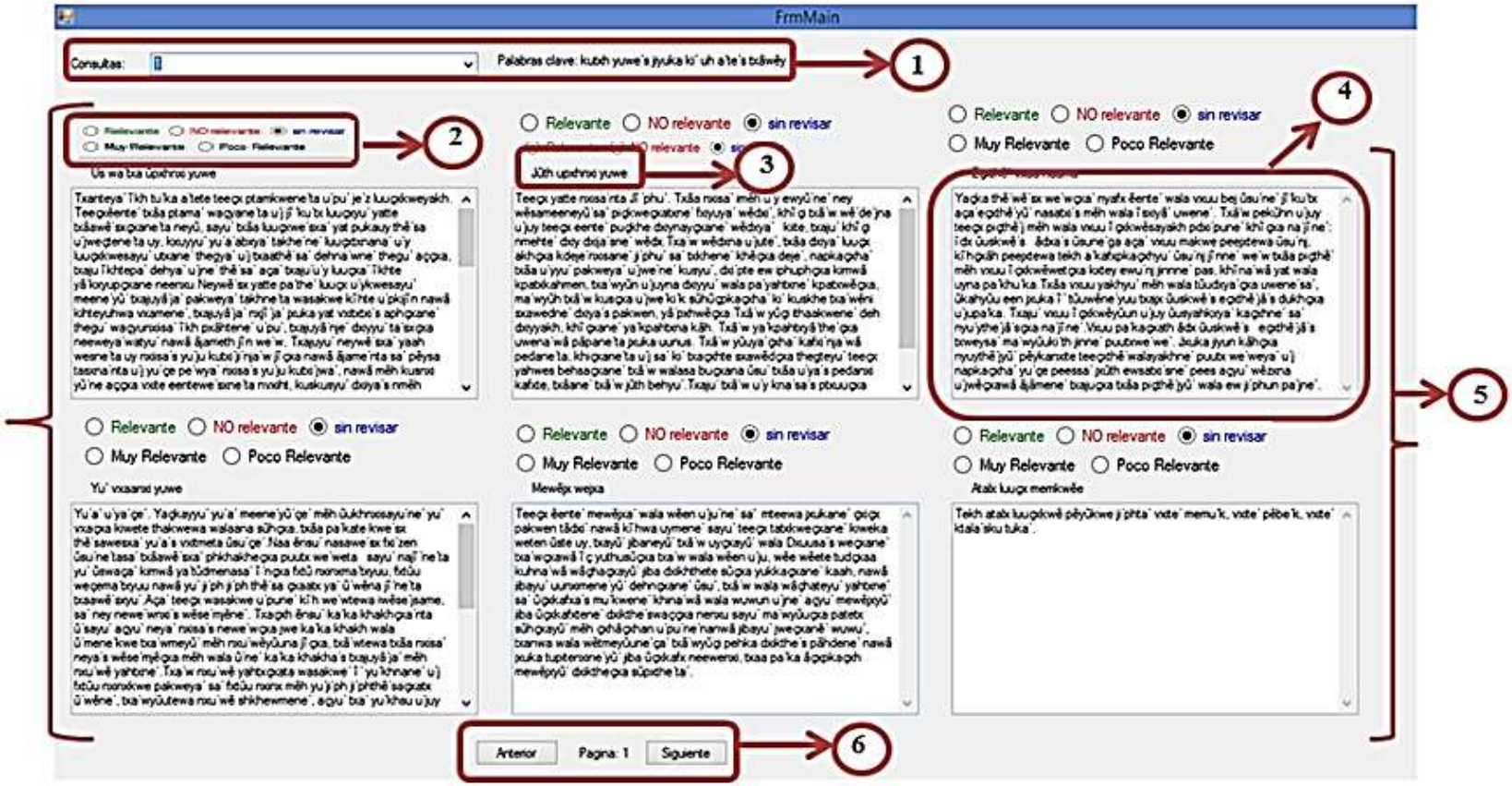


Figura 8. Pantalla interfaz para recolectar juicios. Fuente: elaboración propia.

INSTRUCCIONES: PARA CADA UNA DE LAS 8 CONSULTAS MARQUE CON UNA X SEGÚN SU CRITERIO EL DOCUMENTO SEA MUY RELEVANTE (MR), RELEVANTE (R), POCO RELEVANTE (PR), NO RELEVANTE (NR). SOLO PUEDE MARCAR UNA OPCION POR CADA CONSULTA.

Nombre Experto: _____

	Consulta N°1: kutxh yuwe's jiyuka ki' uh a'te's txãwẽy				Consulta N°2: Thẽ' walawe'sx majii				Consulta N°3: A'te dxi'j / a'te yuwe's jiyuna				Consulta N°4: upxhnxi yuwe / vxanxi yuwe				Consulta N°5: Atalx wejxa's jiyuna				Consulta N°6: Wejxa yuwe's jiyuna				Consulta N°7: Sek yuwe's jiyuna /sek dxi'j				Consulta N°8: Sa'twe'sx yuwe's jiyuna			
	Relevancia de la Consulta				Relevancia de la Consulta				Relevancia de la Consulta				Relevancia de la Consulta				Relevancia de la Consulta				Relevancia de la Consulta				Relevancia de la Consulta				Relevancia de la Consulta			
Libro: Cuentos y Cosmovisión nasa.	MR	R	PR	NR	MR	R	PR	NR	MR	R	PR	NR	MR	R	PR	NR	MR	R	PR	NR	MR	R	PR	NR	MR	R	PR	NR	MR	R	PR	NR
Nasa yxanxi's pta'sxnxi (página 6)																																
Us wa'txa ũpxhnxi yuwe (Página 8)																																
Kutxh wa'txa ũpxhnxi yuwe (Página 10-11)																																
Jũth upxhnxi yuwe (Página 14)																																
Ecxthẽ'yxuu naamu (Página 16)																																
Yu' yxaanxi yuwe (Página 18)																																
Mewẽjx wejxa (Página 20)																																
Atalx luucx memkwẽe (Página 21)																																

Figura 9. Formato para emitir juicios a cada documento de la colección (ejemplo). Fuente: elaboración propia.

Una vez se obtuvieron los juicios de relevancia en formato impreso, se adicionó un formato a un prototipo web que permitiese insertar estos datos de manera rápida, en la Figura 10, se puede apreciar esta interfaz y sus zonas así:

- La zona 1, permite seleccionar el nombre del experto que va a realizar la emisión de juicios.
- La zona 2, visualiza las 8 consultas.
- La zona 3, muestra los títulos de los documentos.
- La zona 4, presenta la lista de chequeo con la escala (MR, R, PR y NR) para emitir el juicio del documento en relación con cada consulta.
- Al final del formulario, se tiene un botón para guardar los datos digitados sobre el formulario.

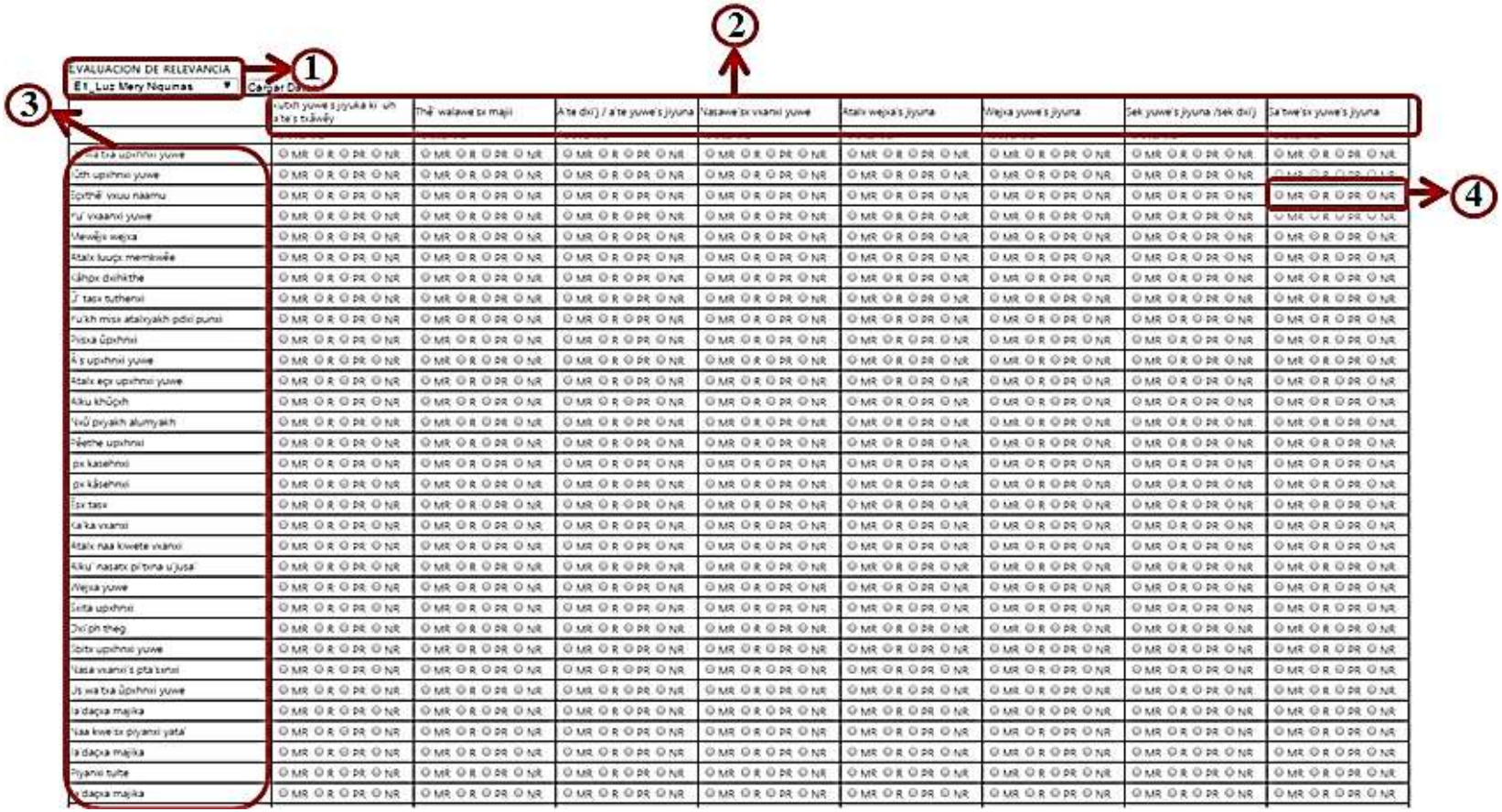


Figura 10. Pantalla interfaz para digitalizar juicios. Fuente: elaboración propia.

Al igual que con la interfaz presentada en la Figura 8, con el prototipo presentado en la Figura 10, cabe mencionar que para masificar su uso esta interfaz deberá ser sometida a evaluaciones de usabilidad en pro de que su uso no afecte la calidad del juicio del experto al momento de su inserción. .

Una vez digitalizados los juicios se realizó una nueva ronda de emisión de juicios con algunos expertos para aquellos pares (documento – consulta) en los cuales los juicios estuvieran con amplia discrepancia en el concepto emitido por ellos, de tal forma, que se pudiese alcanzar un consenso.

2.2.4 Juicios por cada par (documento - consulta)

Los expertos seleccionados para la emisión de juicios fueron profesores hablantes de nasa yuwe y un lingüista experto en nasa yuwe provenientes de diferentes variantes de nasa yuwe. Cabe destacar que el proceso de emisión de juicios por parte de cada experto, fue costoso en tiempo y esfuerzo, dada la magnitud del ejercicio y que no se tenían precedentes al respecto. En Tabla 8, se puede apreciar el listado de expertos.

N°	Experto	Ocupación	Variante
E1	Luz Mery Niquinas	Profesora hablante de nasa yuwe en el programa de etnoeducación de la Universidad del Cauca	Vitoncó (Tierradentro)
E2	Benilda Trochez	Profesora de nasa yuwe en Resguardo López Adentro (Cauca)	Jambaló
E3	Tulio Rojas	Lingüista, hablante y experto en nasa yuwe y profesor de la Universidad del Cauca	Munchique-Tigres
E4	Roucsana Chocue	Profesora de nasa yuwe en Resguardo López Adentro (Cauca)	Caldono
E5	Oscar Guetio	Hablante nasa yuwe en Pueblo Nuevo (Cauca)	Pueblo Nuevo
E6	Oliveiro	Profesor de nasa yuwe en Pueblo Nuevo (Cauca)	Pueblo Nuevo
E7	Sammy Tombe	Profesor de nasa yuwe en pueblo nuevo (Cauca)	Pueblo Nuevo
E8	Luis Alberto Guetoto	Profesor de nasa yuwe en Caldono (Cauca)	Caldono
E9	Jose Libardo Valencia	Profesor de nasa yuwe Resguardo de Munchique	Munchique-Tigres

Tabla 8. Expertos que participaron en la emisión de juicios

Para la consolidación de los juicios se tomaron todos los valores con igual peso y se hicieron dos grupos: En el primero se agruparon los valores obtenidos en la escala de Muy Relevante y Relevante (MR + R), y el segundo grupo se contó con los valores de Poco Relevante y No Relevante (PR + NR).

Finalmente, se tomó como referente para determinar la relevancia o no relevancia de un documento en relación con cada consulta, los valores obtenidos de los grupos mencionados anteriormente así: 1) Se consideran relevantes los documentos que obtuvieran el 60% de

relevancia en el primer grupo (MR + R) del total de respuestas de los expertos en cada consulta. 2) Se consideran no relevantes los documentos que obtuvieron como valor en el segundo grupo (PR + NR) 60%. El determinar el porcentaje de relevancia de los documentos en 60%, permitió diferenciar la escala de relevancia de cada uno con relación a cada consulta. En la colección, el 63% de los documentos son relevantes para uno o más de las 8 consultas. La Figura 11, despliega el número de documentos relevantes para cada consulta.

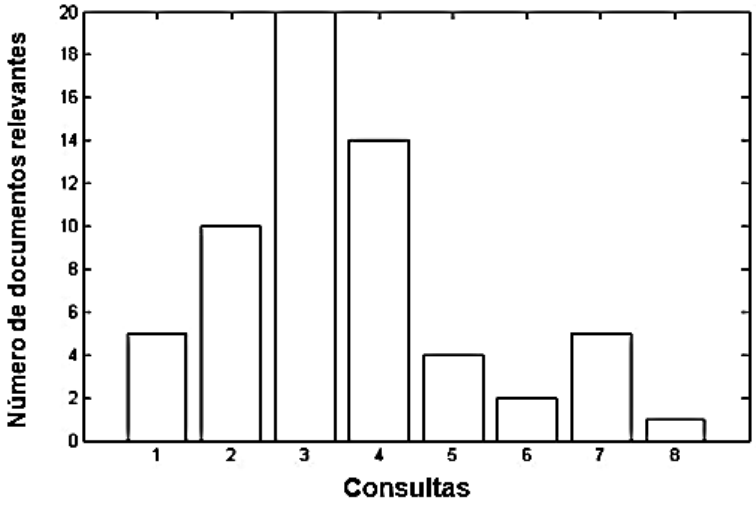


Figura 11. Número de documentos relevantes por cada consulta

A continuación en la Tabla 9 se presentan los documentos relevantes para la consulta 1: kutxh yuwe's jiyuka ki' uh a'te's txãwẽy.

	Documentos de la colección de prueba en nasa yuwe	Consulta 1
1	Kutxh wa'txa ũpxhnxi yuwe (Página 10-11)	X
2	Página 50 1. KHUTX UH A`TE.	X
3	Página 64 9. SXIB VXIS kutxh yuwe's jiyuka ki' uh a'te's txãwẽy A´TE	X
4	Página 72 13. ÇUT A'TE	X
5	Página 80 17. KUTXH SXADE A'TE MEEÇXA Ũ' PKHAKH A' TE	X

Tabla 9. Documentos relevantes para la consulta 1

El listado completo de documentos relevantes por consulta, es decir, el listado de cada par documento-consulta se encuentra en el Anexo 1. Documentos Relevantes por cada consulta, también en este anexo se incluyó el cálculo del nivel de relevancia de los documentos relevantes utilizando la fórmula sugerida por [76], mediante la cual en la mayoría de los casos se encuentra un nivel de relevancia mayor a 50%, de tal forma, que se puede apreciar valores de relevancia muy cercanos a los presentados en el Anexo 2.

En el anexo 2, se encuentra el listado de los resultados obtenidos en la emisión de juicios para cada consulta.

2.2.5 Descripción de la colección de prueba

En la Tabla 10 se resumen algunos atributos de la primera versión de la colección de prueba del nasa yuwe. Dicha colección se encuentra publicada en: http://www.ewa.edu.co/coleccion.

Atributos	Valor
Tamaño de la Colección	113 KB
Formatos de los documentos	Texto (UTF-8)
N° de documentos	97
Promedio de términos por documento	103
Promedio de tamaño de documentos	1.5 KB
N° de consultas	8

Tabla 10. Atributos de la colección de nasa yuwe

2.3 ANÁLISIS ESTADÍSTICO DE LA COLECCIÓN

En la Tabla 11, se presentan los términos con mayor frecuencia en los documentos de la colección, lo cual es de gran importancia para la tarea de remoción de palabras vacías, además presenta la longitud de las palabras teniendo en cuenta el alfabeto nasa unificado [48]. Se encontraron 4801 palabras diferentes incluyendo las conjugaciones. En el anexo 3, se encuentra el listado parcial de las palabras con su frecuencia de términos.

Término	Frecuencia	Longitud	Término	Frecuencia	Longitud
Wala	146	4	Teeçx	77	4
a'te	121	3	txã'w	73	3
Txãa	114	3	kwe'sx	66	3
Naa	108	3	Sa'	57	2
Nasa	92	4	ki'	54	2

Tabla 11. Términos con más frecuencia en los documentos de la colección

Para comprender cómo los términos de la colección de prueba de nasa yuwe se distribuyen en los documentos[24], se presenta la Figura 13, que muestra un gráfico logarítmico con la frecuencia de términos de la colección en función de su rango, permitiendo apreciar el comportamiento de los términos y el cumplimiento de la Ley de Zipf para el nasa yuwe, la cual se aplica para la mayoría de los lenguajes [3].

En la Figura 12, se presenta la aproximación potencial del comportamiento de la frecuencia de los términos de la colección, según la ecuación: $Y = 143.28X^{-0,621}$ y con factor de correlación $R^2 = 0,8958$. Expresando la Ley de Zipf, para la lengua nasa yuwe como [36]:

$$Cf_i = Ci^k \text{ o } Cf_i = \log c + k * \log i\,,\ donde\ K = -1\ y\ C\ es\ una\ constante,$$

De tal forma que, $Y = 143,28X^{-0,621}$ expresada como Logaritmo es:

$$Log\ Y = Log(143,28\ X^{-0,621})$$

[24] Esto permite caracterizar las propiedades de los algoritmos para comprimir las listas de términos (postings lists) [3]

$$Log\ Y = Log(143{,}28) + Log\ (X^{-0{,}621})$$

$$Log\ Y = 2{,}1562 - 0{,}621 * Log(X)$$

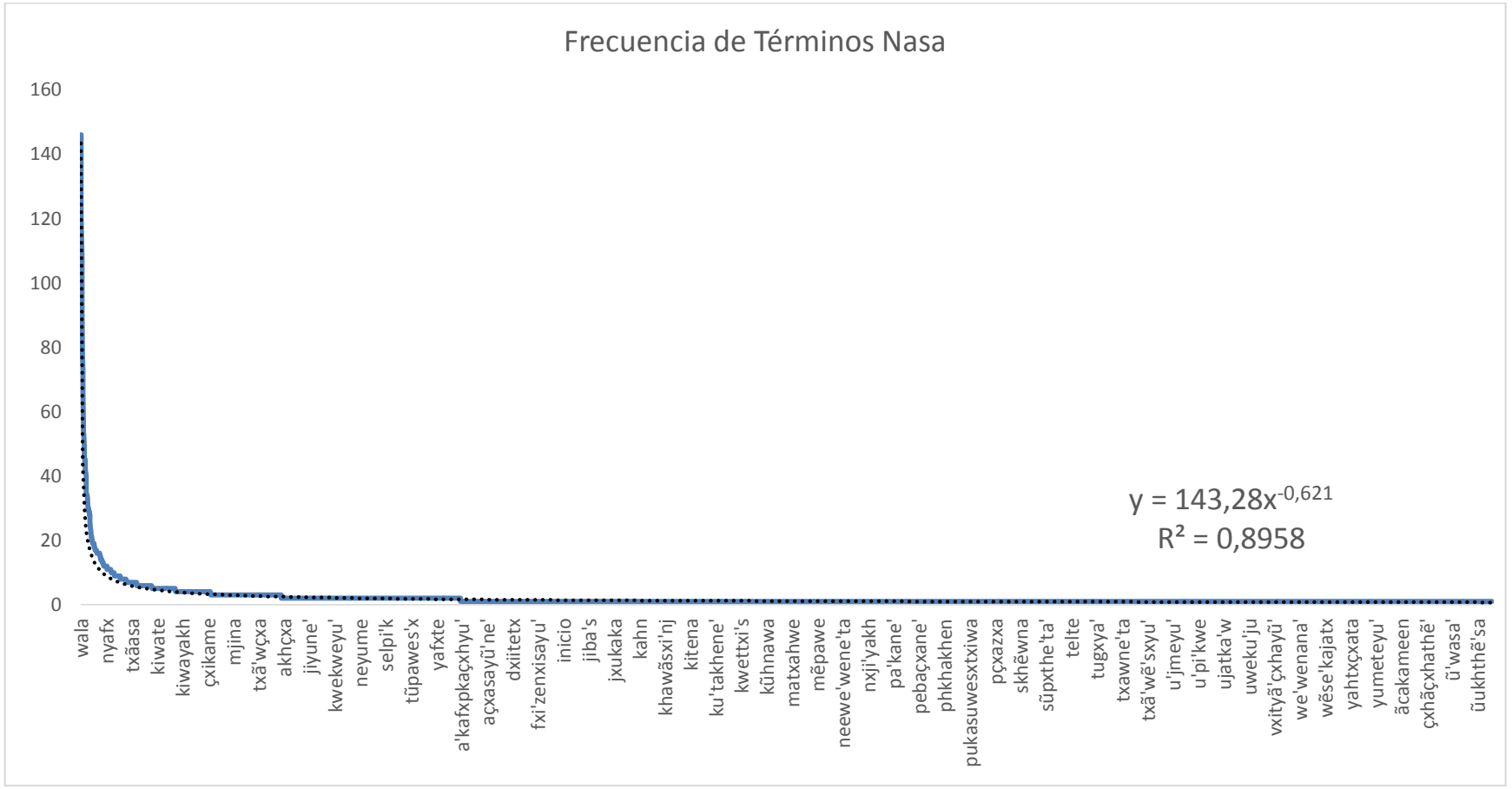
Finalmente, obtenemos la ecuación de siguiente forma:

$$Log\ Cf_i = Log\ (C) + K * \log i$$

$$Log\ Cf_i = Log(143{,}28) - 0{,}621 * \log X$$

Donde, C= 143,28 y K =-0,621

Figura 12. Frecuencia de términos en la colección



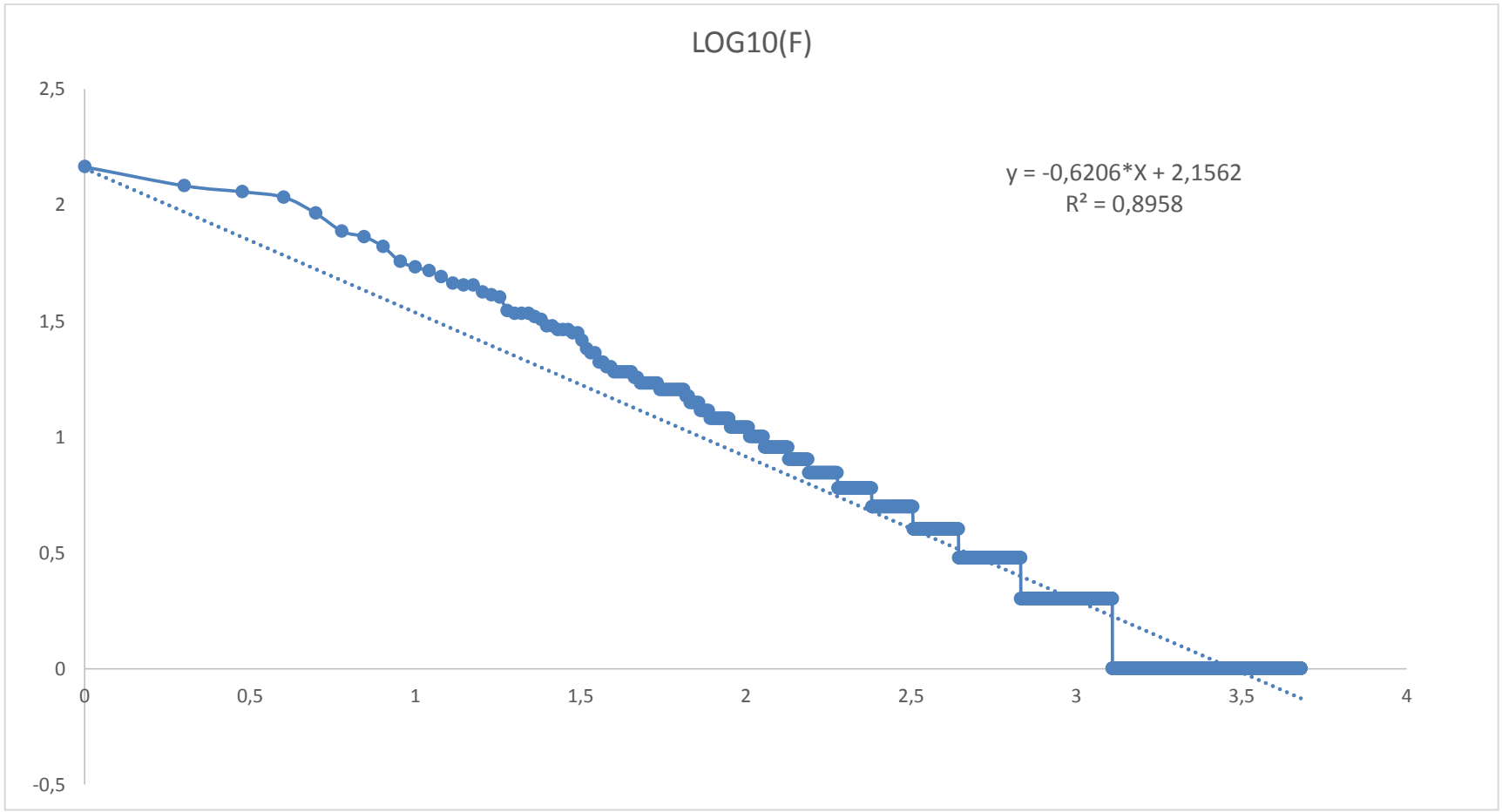


Figura 13. Ley de Zipf para Nasa Yuwe. Elaboración propia

3. ADAPTACIÓN DE UN ESQUEMA DE REPRESENTACIÓN Y BÚSQUEDA PARA TEXTOS ESCRITOS EN NASA YUWE

En esta sección se presenta la propuesta del esquema de representación y búsqueda para textos escritos en lengua nasa yuwe basado en el modelo espacio vectorial y la adaptación de las fases del proceso de recuperación información.

Para el desarrollo de este trabajo se utilizó como metodología de referencia el patrón iterativo de investigación [29]. El trabajo se realizó mediante dos iteraciones, la primera relacionada con la adaptación del tokenizador para nasa yuwe y la segunda con la definición de una lista de palabras vacías para nasa yuwe.

Para definir el modelo de espacio vectorial para el Sistema de Recuperación de textos escritos en nasa yuwe, se utilizó el cálculo de similitud de Lucene [6], que involucra la puntuación de varios componentes, como se presenta en el capítulo 1 (ver sección 1.1.5).

3.1 ADAPTACIÓN DE UN ANALIZADOR LEXICO PARA NASA YUWE

Para iniciar el proceso de adaptación del analizador léxico (tokenizer), se estableció la línea base con el tokenizador estándar de Lucene [53] haciendo uso de las consultas y los documentos de la colección de prueba, descrita en el anterior capitulo. Como medida de evaluación de los resultados se utilizó la Curva Precisión – Recuerdo. Al final de este capítulo se presentan otras medidas de evaluación.

Para cada evaluación se configuró el sistema para que recuperará como máximo 40 resultados al momento de hacer la consulta en el sistema de recuperación de información, se considera que es un número adecuado dado que se le da la posibilidad al sistema de encontrar los documentos relevantes.

La Curva Precisión – Recuerdo obtiene sus valores mediante la acumulación del número de documentos relevantes recuperados y el total de documentos relevantes y posteriormente se calcula la precisión y el recuerdo teniendo en cuenta el resultado y la cantidad de consultas involucradas en la recuperación. Lo anterior, permite apreciar mejor el impacto de cada consulta.

Para la evaluación de resultados de la curva Precisión – Recuerdo, se tienen en cuenta tres casos así:

- Caso 1, todas las consultas de la 1 a la 8 (C 1 a 8) en la colección, lo que permitió evaluar el desempeño del sistema teniendo en cuenta todas las condiciones particulares de la lengua nasa yuwe.

- Caso 2, las consultas de la 1 a 7 (C 1 a 7), de tal forma, que se pueda revisar el impacto de la consulta 8 sobre el desempeño del sistema, dado que esta consulta solo tiene un documento relevante.

- Caso 3, las consultas de la 2 a 4 (C 2 a 4), que permite evaluar el comportamiento del sistema con pocas consultas y con las consultas que más documentos relevantes tienen asociadas.

3.1.1 Tokenizador Estándar de Lucene (ST)

Se toman las medidas utilizando sólo el tokenizador Estándar (ST), el cual separa las palabras según los caracteres de puntuación, removiendo los signos de puntuación, pero los puntos que no están seguidos por un espacio en blanco son considerados parte del token. Separa las palabras unidas por guiones (-) a menos que contenga un número en el token, en cuyo caso no se separa. Identifica direcciones de internet, correos electrónicos y acrónimos [6].

1. *Valores de Precisión Recuerdo utilizando ST*

Precisión Recuerdo	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
ST (Caso 1)	31,25%	17,50%	16,25%	13,97%	13,59%	0,00%	0,00%	0,00%	0,00%	0,00%
ST (Caso 2)	35,71%	20,00%	18,57%	15,08%	13,23%	0,00%	0,00%	0,00%	0,00%	0,00%
ST (Caso 3)	16,67%	21,21%	22,81%	23,19%	20,43%	0,00%	0,00%	0,00%	0,00%	0,00%

Tabla 12. Valores Precisión Recuerdo para ST. Fuente Elaboración Propia.

2. *Curva Precisión Recuerdo comparativa utilizando ST*

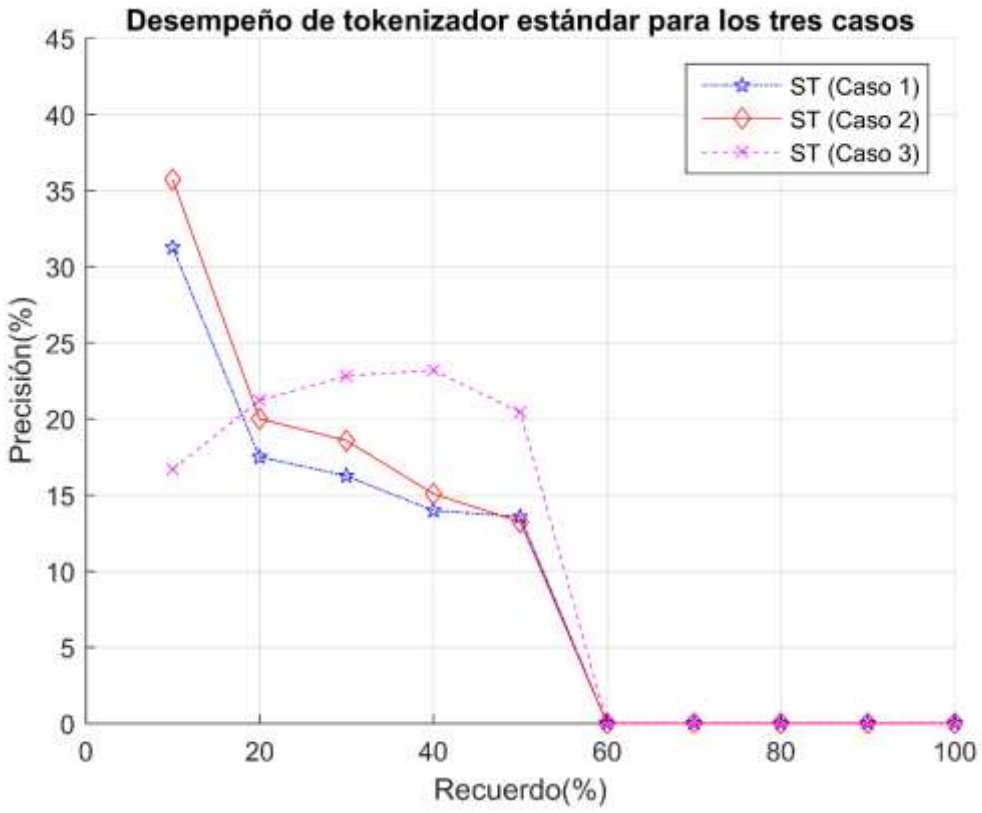


Figura 14. Comparación de casos utilizando ST. Fuente Elaboración Propia.

En la Tabla 12 y en la Figura 14, se puede apreciar que:

- El comportamiento del sistema presenta una mejora en el caso 2 al excluir la consulta 8, lo que indica que la cantidad de documentos relevantes por consulta afectan el desempeño del sistema, por esta razón, el caso 1 no se tendrá en cuenta en los siguientes procesamientos.

- Para el caso 3 se aprecia que para un valor de Recuerdo de 10% el sistema presenta una disminución del desempeño, pero para los otros valores de recuerdo se observan mejores resultados, lo que indica que la precisión en los primeros resultados presentados por el sistemas son bajas, es decir que la cantidad de documentos relevantes recuperados es baja, pero a medida que aumentan los niveles de recuerdo, la precisión mejora obstensiblemente.

3.1.2 Tokenizador Estándar de Lucene + Filtro Estándar

Se adiciona al tokenizador estándar (ST) un filtro estándar (Filter), el cual remueve la terminación **'s** del final de las palabras, y quita los puntos de los acrónimos [6]. Se pudo apreciar en los tres casos que el uso del filtro no genera ningún cambio en el desempeño del sistema, en comparación con los resultados presentados en la Figura 14, por tanto no se incluyó en el siguiente análisis.

3.1.3 Tokenizador Estándar de Lucene + Convertidor de texto a minúsculas

Al tokenizador estándar de Lucene se le adiciona un convertidor de palabras a minúsculas (Lower Case -LC), que transforma todo el texto en minúsculas [6].

1. Valores de Precisión Recuerdo utilizando ST + LC

Precisión Recuerdo	**10%**	**20%**	**30%**	**40%**	**50%**	**60%**	**70%**	**80%**	**90%**	**100%**
ST (Caso 2)	35,71%	20,00%	18,57%	15,08%	13,23%	0,00%	0,00%	0,00%	0,00%	0,00%
ST + LC (Caso 2)	71,43%	50,00%	16,88%	14,29%	14,86%	11,65%	0,00%	0,00%	0,00%	0,00%
ST (Caso 3)	16,67%	21,21%	22,81%	23,19%	20,43%	0,00%	0,00%	0,00%	0,00%	0,00%
ST + LC (Caso 3)	50,00%	22,22%	19,70%	22,67%	19,79%	0,00%	0,00%	0,00%	0,00%	0,00%

Tabla 13. Valores Precisión Recuerdo para ST + LC. Fuente: Elaboración Propia

2. *Curva Precisión – Recuerdo utilizando ST + LC*

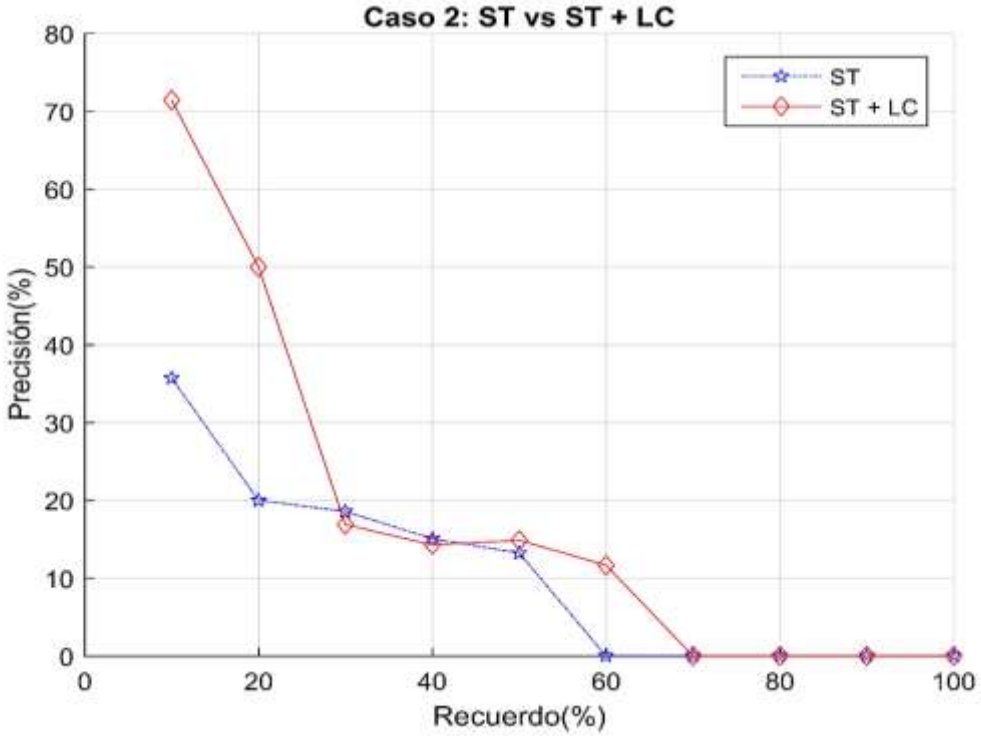


Figura 15. Comparación de ST vs ST + LC Caso 2. Fuente Elaboración Propia.

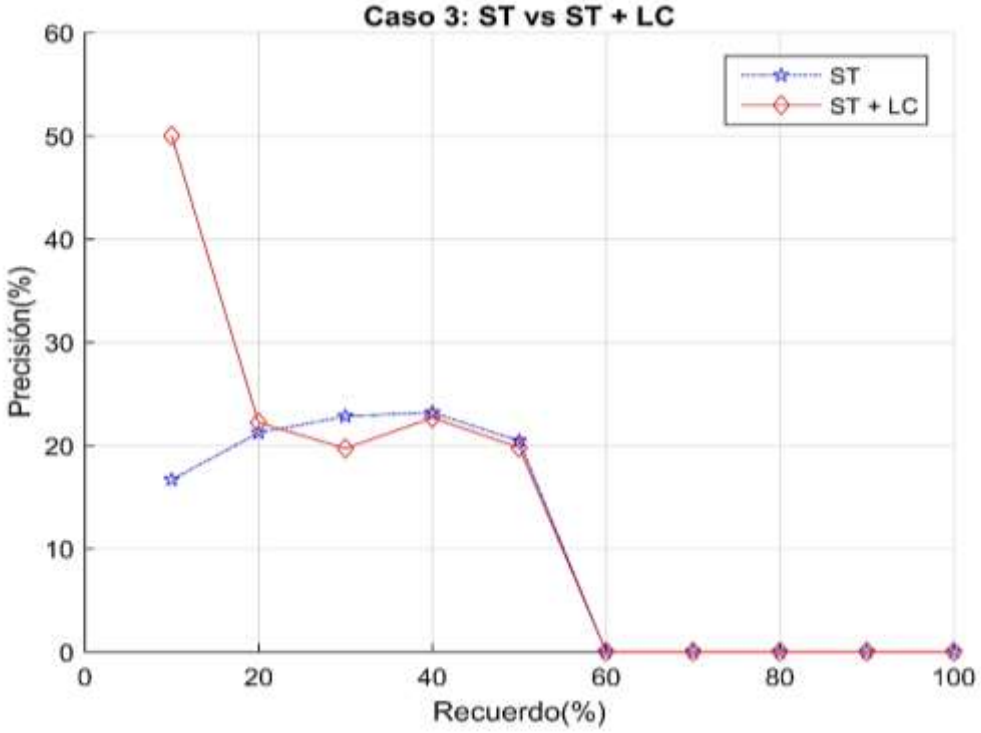


Figura 16. Comparación de ST vs ST + LC Caso 3. Fuente Elaboración Propia.

En la Tabla 13 y en las Figura 15 y 16, se puede apreciar que:

- En todos los casos la aplicación del convertidor de palabras a minúscula, mejora el desempeño del sistema en los primeros niveles de recuerdo, lo cual es entendible dado que aunque los términos no estén correctamente tokenizados, se puede mejorar la comparación de los términos en la consulta con los de los documentos, por ejemplo, sin el convertidor, la palabra **Wala** es diferente a la palabra **wala**, al aplicar el convertidor estas dos palabras se unifican.

- El comportamiento del sistema en los dos casos es similar excepto para el valor de recuerdo de 10% que en el caso 2 presenta una mejora muy alta al aplicar el convertidor a minúsculas, mientras que para los valores de recuerdo de 30% a 50% en los dos casos se observa una pequeña disminución, lo cual se entiende porque la tokenización no es correcta y el desempeño del sistema puede ser impredecible dado que se están separando palabras incorrectamente como por ejemplo en el caso de a'te y a'te' que se convierten en dos tokens unificados /a/ /te/, pero no son los tokens esperados según la lengua nasa yuwe.

A continuación en la Figura 17, se presenta la comparación del desempeño del sistema utilizando ST y ST + LC en cada caso:

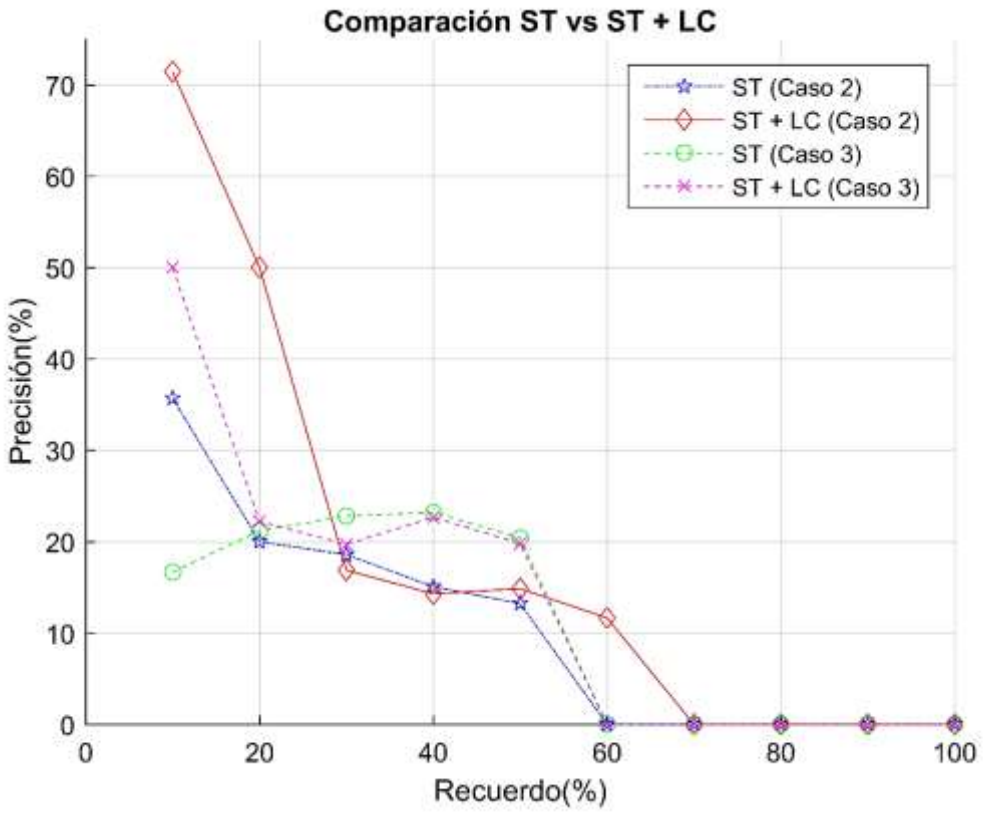


Figura 17. Comparación de desempeño del sistema. Fuente: Elaboración propia

La Figura 15, la Figura 16 y Figura 17 comparan el desempeño del sistema utilizando el tokenizador estándar de Lucene para los casos 2 y 3, las cuales se convierten en la línea base para el proceso de adaptación del tokenizador para nasa yuwe.

En resumen se puede decir que tanto en las tablas anteriores como en las figuras se muestra una mejora en el desempeño del tokenizador estándar de Lucene con el uso del Convertidor de texto en minúsculas para todos los casos.

Otro aspecto que también se revisó en los procesamientos anteriores fue la cantidad de términos (tokens) identificados para la colección de prueba de nasa yuwe para cada uno de los casos en cada procesamiento obteniendo que cuando se utiliza el tokenizador estándar se encuentran 4373 tokens y al adicionarle el convertidor de texto a minúscula son 4149, observándose una diferencia de 183 términos que corresponde a 4,18% menos que al utilizar solo el tokenizador estándar de Lucene.

Al revisar el texto procesado de los documentos de la colección con la línea base se encontraron algunos inconvenientes, como ejemplo se presenta un fragmento del documento C1 Pag-10-11.txt de la colección de prueba nasa yuwe.

Texto nasa yuwe	Texto Procesado con ST + LC
Us wa'txa ũpxhnxi yuwe Txanteya' ĩkh tu'ka a'tete teeçx ptamkwene'ta u'pu' je'z luuçxkweyakh. Teeçx ẽente' txãa ptama' waçyane'ta u'j jĩ'ku'tx luuçxyu' yatte txãawẽ'sxçxane'ta neyũ, sayu' txãa luuçxwe'sxa' yat pukauy thẽ'sa u'jweçtene'ta uy, kxuyyu' yu'a' abxya' takhe'ne' luuçxtxnana' u'y luuçxkwesayu' utxane' thegya' u'j txaa thẽ'sa' dehna'wne' thegu' aççxa,	Us wa'txa ũpxhnxi yuwe Txanteya ĩkh tu ka a tete teeçx ptamkwene ta u pu je z luuçxkweyakh Teeçx ẽente txãa ptama waçyane ta u j jĩ ku tx luuçxyu yatte txãawẽ sxçxane ta neyũ sayu txãa luuçxwe sxa yat pukauy thẽ sa u jweçtene ta uy kxuyyu yu a abxya takhe ne luuçxtxnana u y luuçxkwesayu utxane thegya u j txaa thẽ sa dehna wne thegu aççxa

Tabla 14. Contrastación del Texto Procesado con relación al documento original. Fuente Elaboración Propia.

Entre los errores encontrados en el procesamiento se encuentran:

1. La separación inadecuada de palabras, como se puede apreciar en el texto en color rojo en la Tabla 14 y en lo ejemplos presentados en la Tabla 15.

Palabra nasa	Procesamiento	Número de Tokens	Tokens correctos
tu'ka	tu ka	2	1
a'tete	a tete	2	1
ptamkwene'ta	ptamkwene ta	2	1
u'pu'	u pu	2	1
je'z	je z	2	1

Tabla 15. Ejemplo Separación inadecuada de palabras. Fuente Elaboración Propia.

2. En la Tabla 16, muestra que se cambia una vocal interrupta por una vocal oral cuando está ubicada al finalizar la palabra, como se puede apreciar en el texto en color rojo en la Tabla 14.

Palabra nasa	Procesamiento
Txanteya’	Txanteya
ẽente’	ẽente
sayu’	Sayu

Tabla 16. Ejemplo eliminación del acento en las palabras nasa. Fuente Elaboración Propia.

3. Otro error que fue detectado es que en algunos casos las vocales nasales generaban también división y reemplazo por la correspondiente vocal oral, se muestra un ejemplo en la Tabla 17.

Palabra nasa	Procesamiento
ẽente’	e ente
Mjĩsa	mji sa
kũhku'ka'wã'	ku hku'ka'wa

Tabla 17. Ejemplo separación y reemplazo de vocales nasales. Fuente: Elaboración Propia

A continuación en la Tabla 18, se presenta el fragmento del documento C1 Pag-10-11.txt tokenizado, las celdas resaltadas en color muestran los errores en el procesamiento del texto, como se puede apreciar son bastantes en un fragmento pequeño de texto, por tanto, es necesario enfocar las mejoras en la corrección de estos errores detectados.

Palabra nasa	Token	Palabra nasa	Token	Palabra nasa	Token
Us	us	u’j	u	yu’a’	yu
wa'txa	wa		j		a
	txa	jĩ’ku’tx	jĩ	takhe’ne’	takhe
ũpxhnxi	ũpxhnxi		ku		ne
yuwe	yuwe		tx	luuçxtxnana’	luuçxtxnana
Txanteya’	txanteya	luuçxyu’	luuçxyu	u’y	u
ĩkh	ĩkh	yatte	yatte		y
tu’ka	tu	txãawẽ’sxçxane’ta	txãawẽ	luuçxkwesayu'	luuçxkwesayu
	ka		sxçxane	utxane’	utxane
a’tete	A		ta	thegya’	thegya
	tete	neyũ	neyũ	u’j	u
teeçx	teeçx	sayu	sayu		j
ptamkwene’ta	ptamkwene	txãa	txãa	txaa	txaa
	ta	luuçxwe’sxa	luuçxwe	thẽ’sa’	thẽ
u’pu’	u		sxa		sa
	pu	yat	yat	dehna’wne’	dehna
je’z	je	pukauy	pukauy		wne
	z	thẽ’sa	thẽ	thegu’	thegu
Luuçxkweyakh.	luuçxkweyakh		sa	aççxa	aççxa
Teeçx	teeçx	u’jweçtene’ta	u		
ẽente’	ẽente		jweçtene		
txãa	txãa		ta		
ptama’	ptama	uy	uy		
waçyane’ta	waçyane	kxuyyu’	kxuyyu		
	ta	abxya’	abxya		

Tabla 18. Contrastación de los Tokens con relación al documento original. Fuente Elaboración Propia.

Estos errores también se presentan en buscadores como Google, dadas las características anteriormente descritas del nasa yuwe (Ver Anexo 5).

En virtud de hacer una mejora al tokenizador estándar de Lucene para el procesamiento de textos escritos en nasa yuwe se realizó una versión nasa de este tokenizador que incluye unas mejoras con respecto a los errores señalados y mostrados nuevamente en la tabla anterior.

Con lo anteriormente descrito, se procede a presentar la adaptación de un tokenizador para textos escritos en nasa yuwe, basado en el tokenizador estándar [6].

3.1.4 Proceso de adaptación del tokenizador nasa basado en el tokenizador estándar de Lucene.

La versión .NET de Lucene cuenta con varias utilidades para ejecutar las diferentes tareas de recuperación de información como se presentó en la introducción de esta sección. Entre las tareas que se pueden desarrollar con estas utilidades se encuentran el análisis (tareas como análisis léxico, filtros, convertidores de texto a minúsculas o mayúsculas, entre otros), la indexación, el almacenamiento, la consulta, entre otras.

Para efectos de este trabajo se utiliza el tokenizador estándar de Lucene (en su versión 2.9.4) el cual tiene como principal objetivo la lectura de los datos (como consultas y documentos) y dividirlos en tokens (unidades pequeñas con significado, en su mayoría son palabras) y organizar estos tokens. En la Figura 18, se presenta la implementación de Lucene para .NET, donde se pueden apreciar las diferentes utilidades.

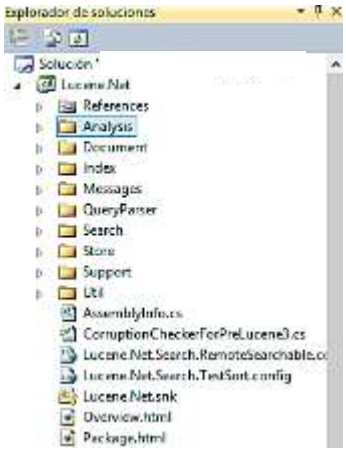


Figura 18. Librería Lucene para .NET. Fuente: Elaboración propia.

La versión nasa del tokenizador estándar de Lucene incluye las siguientes mejoras:

1. Eliminar las tildes (') de las vocales orales y nasales interruptas como elemento separador de palabras, por ejemplo: u'jweçtene'ta, que es separada incorrectamente: /u/ jweçtene /ta/, para lo cual se procedió de la siguiente forma:
   - Se identificaron diferentes tipos de tilde en la forma de escribir documentos en nasa yuwe como son: ´ , ` , ‘ , ’ , es decir, que no solo se podrían encontrar en los documentos de la colección sino en las consultas que haga el usuario, por tanto,

fue necesario unificar estos tipos de tildes por una sola (’) antes de iniciar el procesamiento de textos y consultas en el tokenizador nasa.

- Una vez realizada la unificación de todas las tildes se identificó en el tokenizador de Lucene los caracteres utilizados para separar palabras como espacios en blanco, y signos de puntuación entre otros, y así eliminar la tilde seleccionada como elemento separador de palabras en un texto u oración.

2. Mantener la tilde (’) de las vocales orales y nasales interruptas al final de una palabra, por ejemplo: thegu’, que termina en la vocal oral interrupta u’, pierde el carácter de tilde (’) al ser tokenizada, para lo cual se procedió de la misma manera que en el ítem anterior, es decir, se eliminaron los diferentes tipos de tildes como elemento separador de tokens, haciendo una unificación de estas previamente.

3. Evitar el cambio de las vocales nasales por vocales orales y la división de la palabra, lo cual fue una situación extraña, dado que no con todos los documentos se presentaba esta situación, por tanto, se procedió así.

    - Identificar porqué se presentaba esta situación en unas palabras y en otras no, como el caso de la palabra nasa mjĩsa al tokenizarla quedaba dividida en dos tokens (mji sa), pero en el primer token la i nasal era reemplazada por la i oral, de tal forma, que fue complicado identificar cuál era el error dado que las palabras eran iguales en apariencia. Finalmente, se pudo identificar que las palabras eran diferentes porque la vocal nasal estaba divida en dos, como se muestra en la parte izquierda de la Figura 19, es decir, al ubicar el cursor sobre la letra se aprecia que el carácter está formado por dos partes, pero la vocal de la derecha es un solo carácter, y en apariencia eran iguales. Por tanto, era necesario identificar si era un error en la escritura o una situación que se podía presentar en la escritura normal de la lengua nasa yuwe.

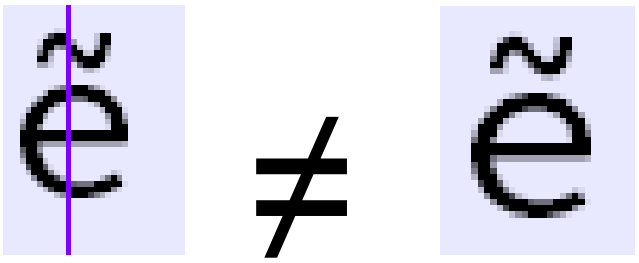


Figura 19. Diferencias en la escritura de la vocal nasal ẽ. Fuente: Archivo nasa revisado en Notepad++

- Teniendo en cuenta lo anterior, fue necesario revisar si era un error en la digitalización de los textos, por ejemplo, con el OCR al escanear los documentos, al aplicar el tipo de fuente (como Charis Sil), al utilizar el software de escritura de nasa yuwe [77], obteniendo que con estos métodos de escritura no se presentaba esta situación con las vocales nasales. Seguidamente, se revisaron las distintas formas en que los profesores nasa pueden escribir textos, una de las formas que generó este error en los documentos y al utilizar una herramienta de inserción como

por ejemplo, la presentada en la página del alfabeto fonético internacional[25], por lo tanto, esta situación fue un problema a considerar en la adaptación del tokenizador para nasa, dado que si bien se podrían corregir los documentos de la colección que presentan esta situación, esta situación se podría presentar al momento de introducir una consulta.

- La situación se corrigió al incluir un paso adicional en el procesamiento de textos nasa que buscó unificar tanto los documentos como las consultas antes de iniciar el procesamiento.

4. Adicionalmente, se identificaron algunas palabras que varían en la escritura por ejemplo: Khutx, Kuthx y Kutxh utilizadas para escribir maíz, a'te y a'te', utilizadas para escribir (época o luna) dependiendo de la variante del idioma, lo cual es un problema dado que si el usuario para escribir maíz lo escribe de manera diferente a como esta en los documentos de la colección, el desempeño del sistema se ve muy afectado, porque no encontraría documentos relevantes relacionados con la consulta. Por esta razón y a pesar, de que esta situación no es parte de los objetivos de este proyecto, se decidió unificar estas palabras en el tokenizador adaptado para nasa yuwe.

A continuación en la Figura 20, se presenta una imagen que esquematiza a grosso modo la forma en que se organizó el tokenizador nasa.

Figura 20. Esquema tokenizador nasa. Fuente: Elaboración propia.

Una vez realizada la adaptación del tokenizador estándar de Lucene para realizar el procesamiento de documentos nasa, se presentan los resultados de lo que en delante se denominará el tokenizador nasa (NT), para revisar su desempeño se utilizan los mismos escenarios y casos descritos al inicio de este capítulo.

[25] http://westonruter.github.io/ipa-chart/keyboard/

3.1.5 Tokenizador Nasa (NT)

En primera instancia, se presenta una tabla con los resultados obtenidos con la aplicación del tokenizador nasa, el cual incluye los cambios mencionados en el ítem anterior, contrastado con la aplicación del tokenizador estándar para cada uno de los casos.

1 Valores de Precisión Recuerdo utilizando NT

En la Tabla 19 se presentan unas mejoras en el desempeño del tokenizador nasa con relación a la aplicación del uso del tokenizador estándar. También para el Caso 1 se observa que el sistema tiene un menor desempeño en comparación con el caso 2, es decir, sin tener en cuenta la Consulta 8, por lo tanto, se puede apreciar el impacto negativo de esta consulta sobre el desempeño del sistema.

Precisión Recuerdo	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
ST (Caso 1)	31,25%	17,50%	16,25%	13,97%	13,59%	0,00%	0,00%	0,00%	0,00%	0,00%
NT (Caso 1)	50,00%	50,00%	20,83%	12,50%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
% de mejora	37,50%	65,00%	21,99%	-11,76%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
ST (Caso 2)	35,71%	20,00%	18,57%	15,08%	13,23%	0,00%	0,00%	0,00%	0,00%	0,00%
NT (Caso 2)	57,14%	57,14%	23,81%	13,67%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
% de mejora	37,51%	65,00%	22,01%	-10,35%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
ST (Caso 3)	16,67%	21,21%	22,81%	23,19%	20,43%	0,00%	0,00%	0,00%	0,00%	0,00%
NT (Caso 3)	66,67%	25,93%	20,00%	20,29%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
% de mejora	75,00%	18,20%	-14,05%	-14,29%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%

Tabla 19. Valores Precisión Recuerdo para NT. Fuente Elaboración Propia.

2 Curva Precisión- Recuerdo utilizando NT

A continuación se presentan las gráficas comparativas de casos utilizando el tokenizador nasa.

En La Figura 21 se presenta la contrastación de la aplicación del tokenizador estándar (ST) vs el tokenizador nasa para el caso 1, se puede apreciar que para los valores del 10% al 40% de recuerdo el tokenizador nasa muestra mejor desempeño del sistema o similar, para valores del 40% al 60% el ST muestra un mejor desempeño y posteriormente el comportamiento es similar.

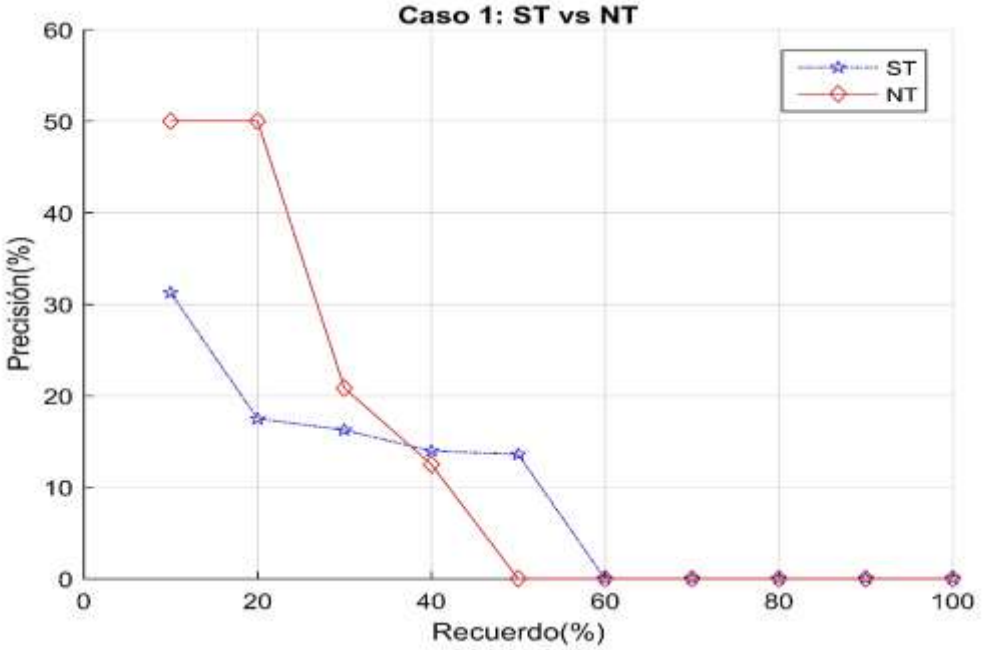


Figura 21. Curva Precisión – Recuerdo NT Vs ST Caso 1. Fuente: Elaboración propia

La Figura 22 presenta la Curva Precisión-Recuerdo para el caso 2, se puede apreciar en primera instancia que el comportamiento es muy similar al de la curva en la Figura 21, de igual forma presenta una mejora en los valores alcanzados en el rango de 10% a 40% de recuerdo. Por tanto, al igual que en las secciones anteriores no se incluirá el caso 1 para los siguientes procesamientos, dado su similitud.

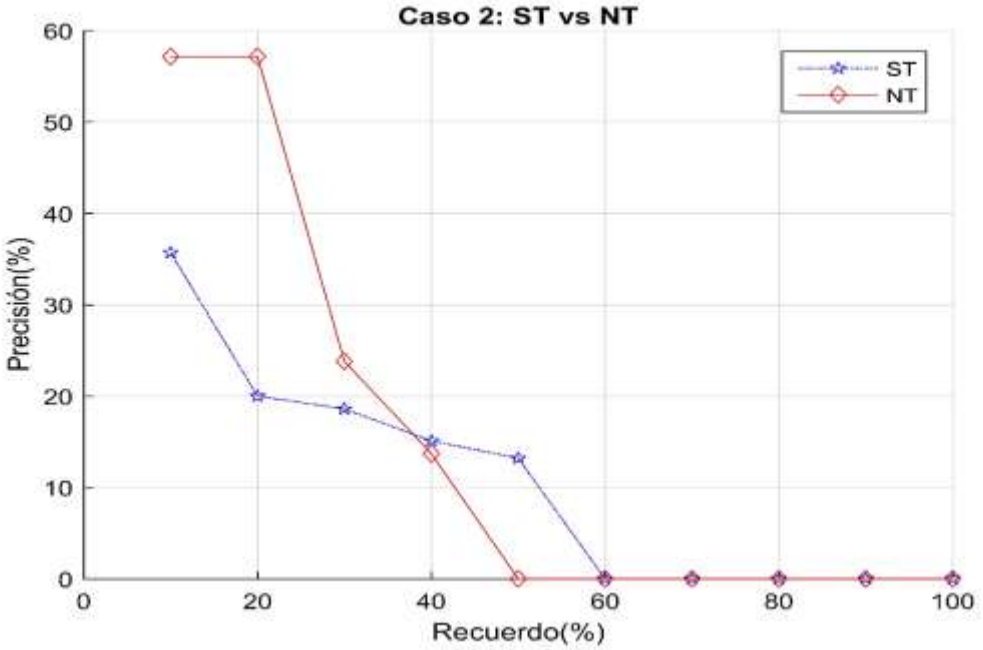


Figura 22. Curva Precisión – Recuerdo NT Vs ST Caso 2. Fuente: Elaboración propia

La Figura 23 presenta la contrastación entre la aplicación del tokenizador estándar (ST) y el tokenizador nasa (NT) para el caso 3, en la cual se puede apreciar:

- Un incremento sustancial en la evaluación para un valor del 10% de Recuerdo, cuando se aplica el tokenizador nasa, lo cual se entiende como consecuencia de que para el caso 3, las consultas tienen una cantidad considerable de documentos relevantes.

- También se puede ver una fuerte caída de la curva en 20% de recuerdo y para los siguientes valores se observa una ligera diferencia entre las curvas de ST y NT.

- Exceptuando el primer valor, el comportamiento de la curva es muy similar al del Caso 2 presentado en la Figura 22.

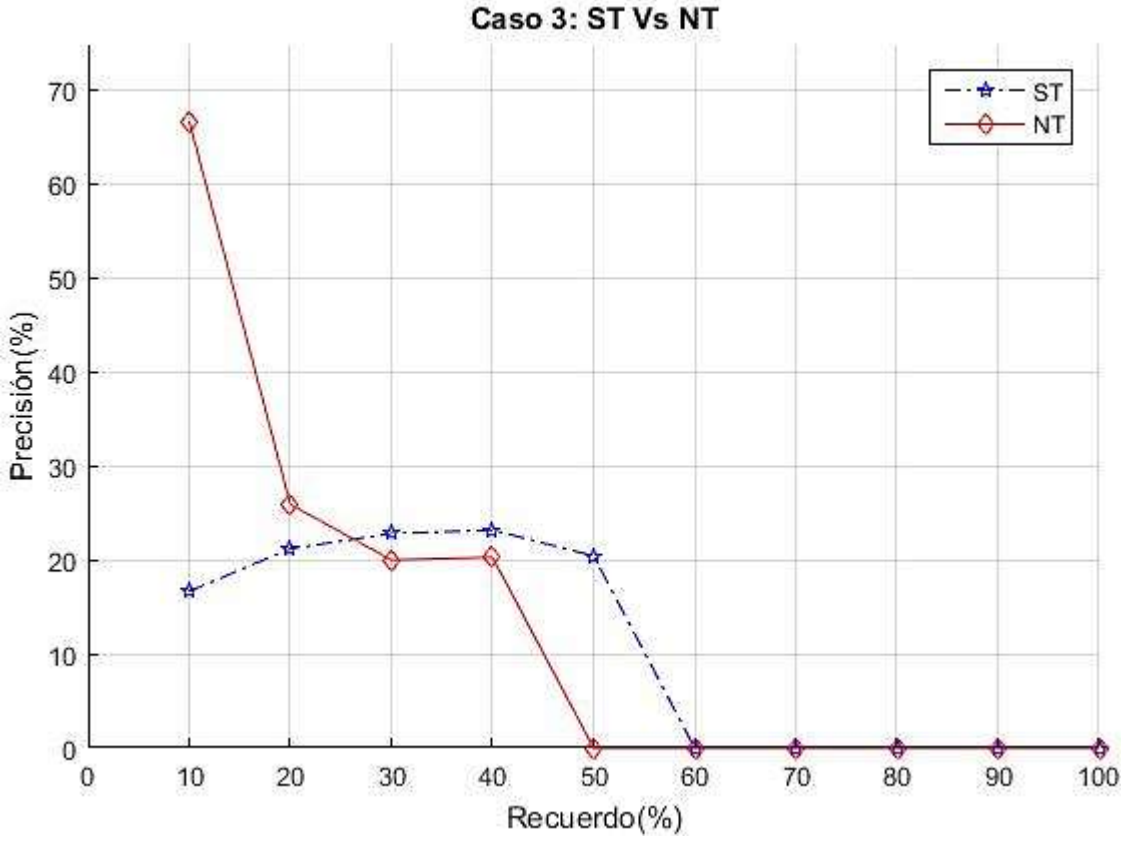


Figura 23. Curva Prec. Rec. Para NT Vs ST Caso 3. Fuente: Elaboración Propia

3.1.6 Tokenizador Nasa (NT) + Filtro

Se presenta la aplicación del tokenizador nasa + la aplicación del mismo filtro descrito en la sección 3.1.2 que elimina signos de puntuación y la terminación ‘s.

En la evaluación realizada se pudo apreciar que la aplicación de este filtro no genera mejoras en el desempeño del tokenizador nasa para ninguno de los casos al igual que cuando se utiliza el ST + el filtro, por tanto, no se tendrá en cuenta para la aplicación del Convertidor de palabras en minúsculas, también teniendo en cuenta que este filtro quita

algunas terminaciones ‘s de las palabras [6], para el caso del nasa yuwe no es aplicable dado que hay algunas palabras en las que se puede presentar esta terminación.

3.1.7 Tokenizador Nasa (NT) + Convertidor de texto en minúsculas (LC)

A continuación se presenta la aplicación del tokenizador nasa + el convertidor de texto a minúsculas (LowerCase - LC).

1 Valores de Precisión Recuerdo utilizando NT

A continuación en la Tabla 20, se presentan los resultados de la evaluación utilizando el tokenizador nasa + el convertidor de texto a minúsculas (LC), contrastado con el tokenizador estándar + el LC. En esta tabla se puede apreciar que:

- Para el valor de recuerdo del 10%, para el caso 2 no se obtiene un mejor desempeño del sistema, lo cual es entendible dado que como no se está tokenizando correctamente el texto nasa con el ST al unificar más valores con el LC la precisión mejora. Mientras que para el Caso 3, se observa una mejora en este valor, lo cual es comprensible dada la cantidad de textos relevantes para estas tres consultas y al unificar tokens la precisión mejore.

- Para valores del 20% al 50% de recuerdo, para el caso 2 se muestran mejoras sustanciales en la evaluación del NT + LC y para el caso 3 se observan mejoras y valores similares en este rango de valor de recuerdo. Lo cual demuestra que el Tokenizador Nasa es mejor que el Tokenizador Estándar comparándolos desde el punto de vista del desempeño.

Precisión – Recuerdo	**10%**	**20%**	**30%**	**40%**	**50%**	**60%**	**70%**	**80%**	**90%**	**100%**
ST+ LC (Caso 2)	71,43%	50,00%	16,88%	14,29%	14,86%	11,65%	0,00%	0,00%	0,00%	0,00%
NT + LC (Caso 2)	57,14%	57,14%	26,53%	16,96%	14,88%	0,00%	0,00%	0,00%	0,00%	0,00%
% de mejora	-25,00%	12,50%	36,37%	15,79%	0,16%	0,00%	0,00%	0,00%	0,00%	0,00%
ST+ LC (Caso 3)	50,00%	22,22%	19,70%	22,67%	19,79%	0,00%	0,00%	0,00%	0,00%	0,00%
NT + LC (Caso 3)	66,67%	25,00%	23,08%	22,73%	19,19%	0,00%	0,00%	0,00%	0,00%	0,00%
% de mejora	25,0%	11,1%	14,6%	0,3%	-3,1%	0,0%	0,0%	0,0%	0,0%	0,0%

Tabla 20. Valores Precisión Recuerdo para NT + LC. Fuente Elaboración Propia.

En la Tabla 21, se presenta una comparación entre el desempeño del sistema utilizando NT y NT + LC, observando que en los dos primeros valores de recuerdo para los dos casos se muestra un desempeño similar, pero para los siguientes valores de recuerdo se observa una mejora sustancial en el desempeño del sistema, de tal forma que la aplicación de LC en términos generales beneficia el desempeño del sistema.

Precisión - Recuerdo	**10%**	**20%**	**30%**	**40%**	**50%**	**60%**	**70%**	**80%**	**90%**	**100%**
NT (Caso 2)	57,14%	57,14%	23,81%	13,67%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
NT + LC (Caso 2)	57,14%	57,14%	26,53%	16,96%	14,88%	0,00%	0,00%	0,00%	0,00%	0,00%
% de mejora	0,01%	0,01%	10,26%	19,42%	100,00%	0,00%	0,00%	0,00%	0,00%	0,00%
NT (Caso 3)	66,67%	25,93%	20,00%	20,29%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
NT + LC (Caso 3)	66,67%	25,00%	23,08%	22,73%	19,19%	0,00%	0,00%	0,00%	0,00%	0,00%

Precisión - Recuerdo	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
% de mejora	0,00%	-3,72%	13,33%	10,72%	100,00%	0,00%	0,00%	0,00%	0,00%	0,00%

Tabla 21. Valores Precisión Recuerdo contrastando NT Vs NT + LC. Fuente Elaboración Propia.

2 Curva Precisión Recuerdo utilizando NT + LC

A continuación en las Figura 24 y Figura 25, se muestra gráficamente el comportamiento descrito en el párrafo anterior.

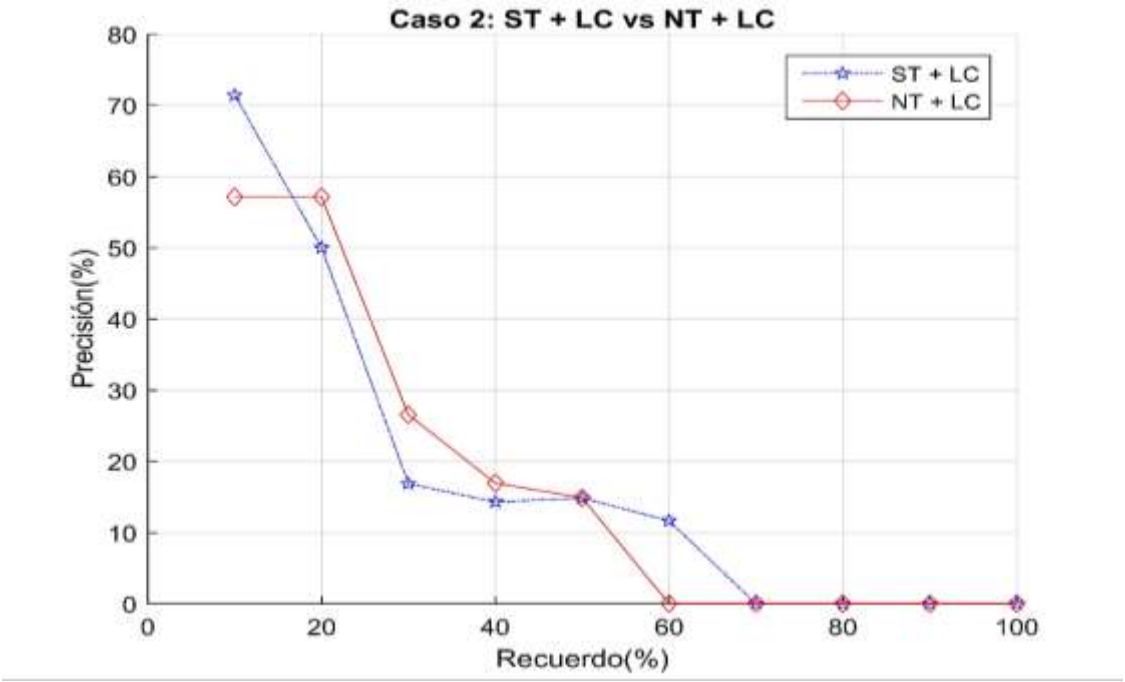


Figura 24. Curva Prec. Rec. Para NT + LC vs ST + LC Caso 2. Fuente: Elaboración Propia

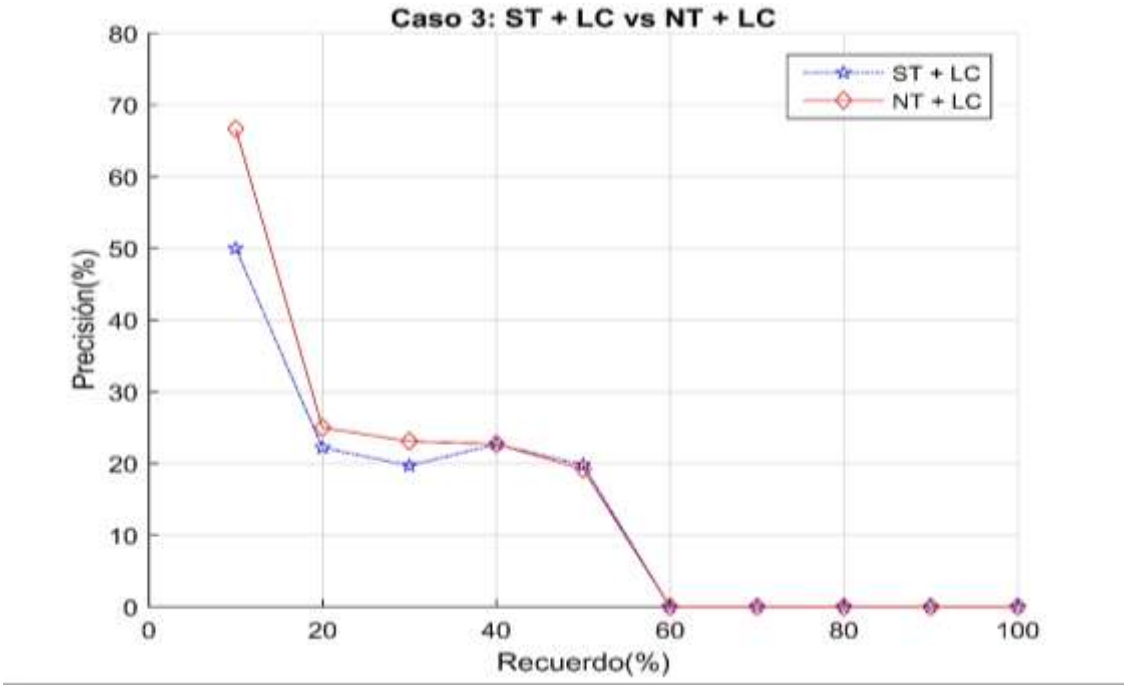


Figura 25. Curva Prec. Rec. Para NT + LC vs ST + LC Caso 3. Fuente: Elaboración Propia

En las Figura 26 y Figura 27se presenta gráficamente la evaluación del sistema utilizando NT vs NT + LC para los dos casos presentados en la Tabla 21.

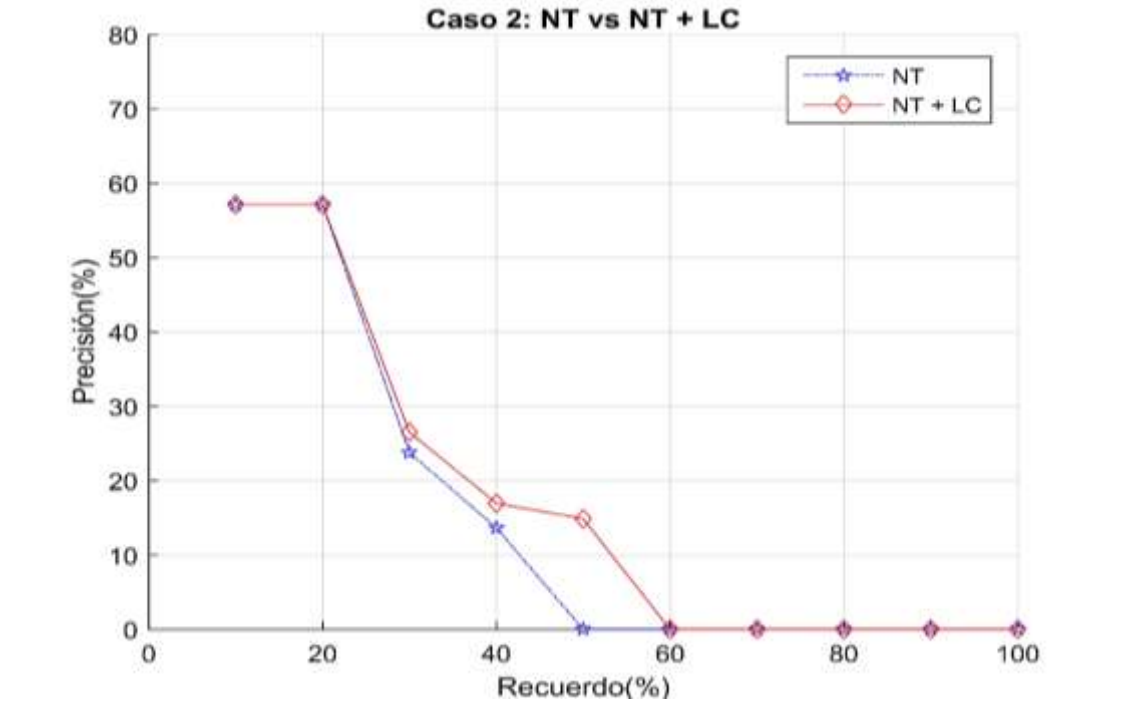


Figura 26. Curva Prec. Rec. NT vs NT + LC Caso 2. Fuente: Elaboración Propia

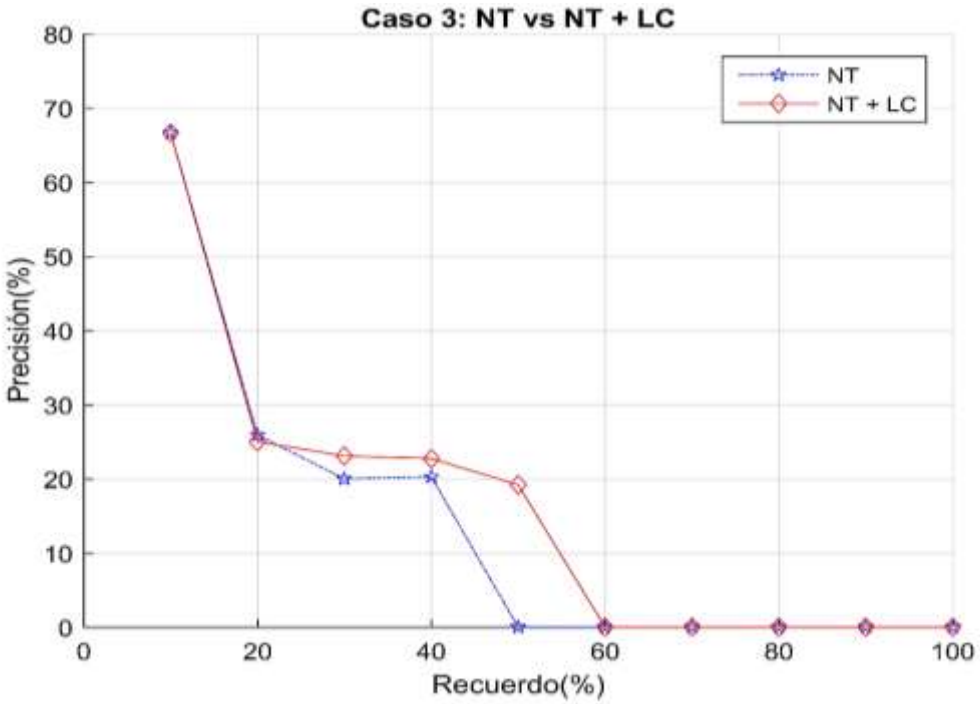


Figura 27. Curva Prec. Rec. NT vs NT + LC Caso 3. Fuente: Elaboración Propia

En pro de revisar porque se presentan las bajas en los valores iniciales de la curva precisión recuerdo, se revisó un fragmento de un texto de la colección (C1 Pag-22.txt), con el fin de revisar cómo se tokeniza.

	Consultas 1 a 8		
Fragmento de texto	**Tokens ST + LC**	**Tokens NT**	**Tokens NT + LC**
Kãhpx	kãhpx	Kãhpx	kãhpx
dxihkthe	dxihkthe	dxihkthe	dxihkthe
Teeçx	teeçx	Teeçx	teeçx
pxãhte’	pxãhte	pxãhte'	pxãhte'
kãhpxne’	kãhpxne	kãhpxne'	kãhpxne'
u’pu’	u	u'pu'	u'pu'
	pu		
nawã	nawã	nawã	nawã
ũ’n	ũ	ũ'n	ũ'n
	n		
ki’	ki	ki'	ki'
dehn,	dehn	dehn	dehn
txa’wçxane’	txa	txa'wçxane'	txa'wçxane'
	wçxane		
yũun	yũun	yũun	yũun
u’psayu’,	u	u'psayu'	u'psayu'
	psayu		
teeçx	teeçx	teeçx	teeçx
ẽente’	ẽente	ẽente'	ẽente'
nasa	nasa	nasa	nasa
tulte	tulte	tulte	tulte
u’kaçxa	u	u'kaçxa	u'kaçxa
	kaçxa		
ã’sx	ã	ã'sx	ã'sx

	Consultas 1 a 8		
Fragmento de texto	**Tokens ST + LC**	**Tokens NT**	**Tokens NT + LC**
uhnxi ũsu'ne' txtene' pesweçxa ũ'yã' u'jwe,	sx uhnxi ũsu ne txtene pesweçxa ũ yã u jwe	uhnxi ũsu'ne' txtene' pesweçxa ũ'yã' u'jwe	uhnxi ũsu'ne' txtene' pesweçxa ũ'yã' u'jwe

Tabla 22. Comparación de Tokens procesados. Fuente Elaboración Propia.

Como se aprecia en la tabla anterior, cuando se utiliza ST + LC, se encuentran errores en la tokenización, ya mencionados anteriormente (Sección 3.1.4), lo que hace pensar que al tokenizar mal las palabras genera una mejor posibilidad de encontrar palabras comunes en más documentos, aunque no sea en realidad la palabra buscada por ejemplo: la palabra U'JWE saca dos tokens de ella U y JWE, en el texto también se encuentra la palabra U'JWE' la cual saca dos tokens U y JWE a pesar de que son palabras de escritura similar son diferentes, pero para ST + LC va a tener mejores valores de precisión porque hay muchos documentos que van a tener estos tokens (U, y JWE) que no necesariamente se refieren a las mismas palabras, esto también hace que al tokenizar una consulta de usuario se tengan más tokens a buscar en los diferentes documentos por tanto, también se aumentará la precisión. Mientras que para el procesamiento con NT y NT + LC las va a tratar como términos diferentes, reduciendo su aparición solo a los tokens correctos en cada documento.

Como otra medida del desempeño del tokenizador nasa + Lower Case se presenta el fragmento del documento C1 Pag-10-11.txt (presentado en la Tabla 18) con los resultados del procesamiento después de haber utilizado el tokenizador nasa + LowerCase, donde se pueden apreciar las mejoras incluidas, dado que los tokens nasa son correctos.

Palabra nasa	Token	Palabra nasa	Token	Palabra nasa	Token
Us	us	ptama'	ptama'	kxuyyu'	kxuyyu'
wa'txa	wa'txa	waçyane'ta	waçyane'ta	abxya'	abxya'
Ũpxhnxi	ũpxhnxi	u'j	u'j	yu'a'	yu'a'
Yuwe	yuwe	jĩ'ku'tx	jĩ'ku'tx	takhe'ne'	takhe'ne'
Txanteya'	txanteya'	luuçxyu'	luuçxyu'	luuçxtxnana'	luuçxtxnana'
Ĩkh	ĩkh	yatte	yatte	u'y	u'y
tu'ka	tu'ka	txãawẽ'sxçxane'ta	txãawẽ'sxçxane'ta	luuçxkwesayu'	luuçxkwesayu'
a'tete	a'tete	neyũ	neyũ	utxane'	utxane'
Teeçx	teeçx	Sayu'	sayu'	thegya'	thegya'
ptamkwene'ta	ptamkwene'ta	txãa	txãa	u'j	u'j
u'pu'	u'pu'	luuçxwe'sxa	luuçxwe'sxa'	txaa	txaa
je'z	je'z	yat	yat	thẽ'sa'	thẽ'sa'
Luuçxkweyakh.	luuçxkweyakh	pukauy	pukauy	dehna'wne'	dehna'wne'
Teeçx	teeçx	thẽ'sa	thẽ'sa	thegu'	thegu'
ẽente'	ẽente'	u'jweçtene'ta	u'jweçtene'ta	aççxa	aççxa
Txãa	txãa	uy	uy		

Tabla 23. Contrastación de los Tokens con relación al documento original. Fuente Elaboración Propia.

Como se puede apreciar las palabras nasa quedaron correctamente tokenizadas, por tanto, la adaptación realizada al tokenizador estándar de Lucene para el procesamiento de textos escritos en nasa yuwe fue exitosa, a pesar de que para algunos niveles de recuerdo, la precisión es inferior cuando se aplica el tokenizador estándar.

La cantidad de tokens para la colección de textos escritos en nasa yuwe en cada caso se dió así:

- Para el procesamiento con ST + LC se tokenizaron 4149 términos
- Para el procesamiento con NT se tokenizaron 5020 términos
- Para el procesamiento con NT + LC se tokenizaron 4801 términos

Adicionalmente, me permito aclarar que las operaciones incluidas en el tokenizador estándar de Lucene para adaptarlo para el procesamiento de textos nasa, tienen un orden n por lo cual el tiempo requerido para procesar documentos de diferentes tamaños conserva el orden de complejidad del tokenizador original.

3.2 DEFINICIÓN DE LISTA DE PALABRAS VACÍAS PARA NASA YUWE

Para la definición de la lista de palabras vacías (STOPWORDS REMOVAL LIST) en primera instancia fue necesario tomar una línea base a partir del tokenizador nasa, y del tokenizador estándar teniendo en cuenta palabras vacías del español, previamente definidas en Lucene.

En segunda instancia, se procedió a definir las palabras vacías para textos escritos en nasa yuwe, utilizando la aproximación de frecuencia de palabras.

3.2.1 Línea base utilizando la lista de palabras vacías en español

Se tomó como base el tokenizador nasa con el convertidor en minúscula (Lower Case) utilizando la lista previamente definida para español en Lucene, ésta lista se pueden apreciar en el Anexo 4.

Precisión Recuerdo	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
NT + LC (Caso 1)	50,00%	50,00%	23,21%	14,84%	14,21%	0,00%	0,00%	0,00%	0,00%	0,00%
NT + LC + STW(ESPAÑOL) (Caso 1)	50,00%	50,00%	23,21%	14,84%	14,21%	0,00%	0,00%	0,00%	0,00%	0,00%
NT + LC (Caso 2)	57,14%	57,14%	26,53%	16,96%	14,88%	0,00%	0,00%	0,00%	0,00%	0,00%
NT + LC + STW(ESPAÑOL) (Caso 2)	57,14%	57,14%	26,53%	16,96%	14,88%	0,00%	0,00%	0,00%	0,00%	0,00%
NT + LC (Caso 3)	66,67%	25,00%	23,08%	22,73%	19,19%	0,00%	0,00%	0,00%	0,00%	0,00%
NT + LC STW(ESPAÑOL) (Caso 3)	66,67%	25,00%	23,08%	22,73%	19,19%	0,00%	0,00%	0,00%	0,00%	0,00%

Tabla 24. Línea base para la remoción de palabras vacías (español). Fuente Elaboración Propia.

Se pueden apreciar que no hay diferencia en el desempeño del sistema al utilizar la lista de palabras vacías del español, para los casos presentados indicando que la lista es poco

adecuada para el tokenizador. Por tanto, en los tres casos se hizo necesario definir una lista de palabras vacías adecuada a los textos escritos en nasa yuwe.

3.2.2 Definición de palabras vacías para nasa yuwe

La definición de la lista de palabras vacías para nasa yuwe tuvo tres momentos así: En un primer momento, se tomaron las primeras 100 palabras con mayor frecuencia (Ver Anexo 3). Luego en un segundo Momento estas palabras se clasificaron según [50] y [8] de tal forma, que se pudo trabajar con el experto en lingüística (profesor Tulio Rojas) quien definió cuáles palabras de las 100 podrían llegar a formar parte de la lista de palabras. Finalmente, en un tercer momento, en donde se evaluaron estas palabras definidas por el experto en relación con el desempeño del sistema.

La definición se hizo teniendo en cuenta la frecuencia de las palabras en la colección, para la clasificación de las palabras se utilizó el diccionario nasa yuwe [50] y la gramática presentada en [8], en el anexo 4, se encuentran las 100 palabras y su clasificación. En la Tabla 25, se presenta la lista de palabras candidatas a formar parte de la lista de palabras vacías para nasa yuwe, es decir, con las que se hicieron las evaluaciones.

N°	**Término en la Colección**	**Freq.**	**Significado**	**Clasificación de la palabra el diccionario nasa yuwe** [50]	**Clasificación de la palabra según el lingüista** [27]
1	txãa	114	el, ella, aquel, aquella, ese, esa	pron.p	deíctico
2	naa	108	este, esta, esto	adjetivo	deíctico
3	teeçx	77	uno	adjetivo	numeral/nombre
4	txã'w	73	así		conector
5	kwe'sx	66	nosotros, nosotras, nuestro	pronombre	deíctico
6	sa'	57	y, entonces	conjunción	conector
7	ki'	54	y		conector
8	jxuka	46	todo	adjetivo	cualificativo
9	nawã	45	así	conjunción	deíctico
10	vxite	41	otro	adjetivo	deíctico
11	seena	40	terrible	adjetivo	cualificativo
12	aça'	35	aproximante a, "y" en castellano/ entonces	conjunción	conector
13	wẽt	34	Agradable, sabroso	adj. adv	calificativo
14	txaju'	33	entonces, luego, después de eso	conjunción	conector
15	mẽh	32	bastante	Adverbio	cualificativo
16	açyu'	29	entonces, así, eso fue	conjunción	conector
17	nawa	28	así	Adverbio, conjunción	deíctico
18	ma'w	24	como	adverbio	pronombre interrogativo
19	sena	20	bastante	Adjetivo, adverbio	cualificativo

N°	Término en la Colección	Freq.	Significado	Clasificación de la palabra el diccionario nasa yuwe [50]	Clasificación de la palabra según el lingüista [27]
20	aççxa	19	solamente entonces	Conjunción, adverbio	conector
21	je'z	17	dos	adjetivo	numeral/nombre
22	maa	16	cual, cualquier, alguno	pronombre	deíctico
23	na'w	16	¿Cómo?	adverbio	
24	pa'ka	16	por, a causa de	preposición	deíctico
25	txa'w	16	Así, asimismo, de la misma manera	adverbio	conector
26	txajuyu'	16	luego	Conjunción	conector
27	tekh	14	tres	adjetivo	Numeral / nombre
28	ũskan	14	Estar, permanecer	verbo	verbo
29	u'jn	13	ida	sustantivo	Verbo
30	majika	12	¿Cómo dice él o ella?		Pronombre interrogativo
31	meeçxa	12	siempre		conector
32	meeçxa'	12	o	conjunción	conector
33	txa's	12	el, ella, aquel, aquella, ese, esa	pron.p	deíctico
34	txajx	12	Así es	pron. pos	

Tabla 25. Lista de palabras candidatas para lista de palabras vacías. Experto en Lingüística prof. Tulio Rojas

3.2.3 Evaluación de palabras candidata a conformar la lista de palabras vacías

Una vez definidas las palabras candidatas se procedió a realizar las evaluaciones requeridas utilizando el tokenizador nasa + LC y se fue revisando el impacto de la palabra vacía candidata sobre su desempeño. Finalmente, los resultados se contrastaron con las líneas base definidas para cada caso.

En primera instancia, se hicieron las evaluaciones utilizando la lista de palabras candidatas, presentada en la Tabla 26, en donde se puede apreciar que en todos los casos hay una disminución en el desempeño del sistema, por tanto, fue necesario ajustar esta lista de 34 palabras candidatas.

Precisión Recuerdo	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
NT + LC + STW(ESPAÑOL) (Caso 1)	50,00%	50,00%	23,21%	14,84%	14,21%	0,00%	0,00%	0,00%	0,00%	0,00%
NT + LC Lista palabras vacías (34 candidatas) (Caso 1)	50,00%	37,50%	23,21%	13,75%	13,59%	0,00%	0,00%	0,00%	0,00%	0,00%
NT + LC + STW(ESPAÑOL) (Caso 2)	57,14%	57,14%	26,53%	16,96%	14,88%	0,00%	0,00%	0,00%	0,00%	0,00%
NT + LC +	57,14%	42,86%	26,53%	15,08%	14,29%	0,00%	0,00%	0,00%	0,00%	0,00%

Precisión Recuerdo	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
Lista palabras vacías (34 candidatas) (Caso 2)										
NT + LC STW(ESPAÑOL) (Caso 3)	66,67%	25,00%	23,08%	22,73%	19,19%	0,00%	0,00%	0,00%	0,00%	0,00%
NT + LC + Lista palabras vacías (34 candidatas) (Caso 3)	66,67%	25,00%	21,43%	21,74%	18,63%	0,00%	0,00%	0,00%	0,00%	0,00%

Tabla 26. Valores de Precisión Recuerdo con Lista de palabras vacías candidatas. Fuente: Elaboración propia

A continuación en las Figura 28 y Figura 29 se puede observar que el comportamiento es similar para los casos 1 y 2 y en ambos casos al utilizar la lista de palabras vacías candidata el desempeño del sistema presenta una ligera disminución, por tanto, fue necesario revisar la lista de palabras candidatas para identificar cuáles son las más acertadas para conformar la lista de palabras vacías para nasa yuwe.

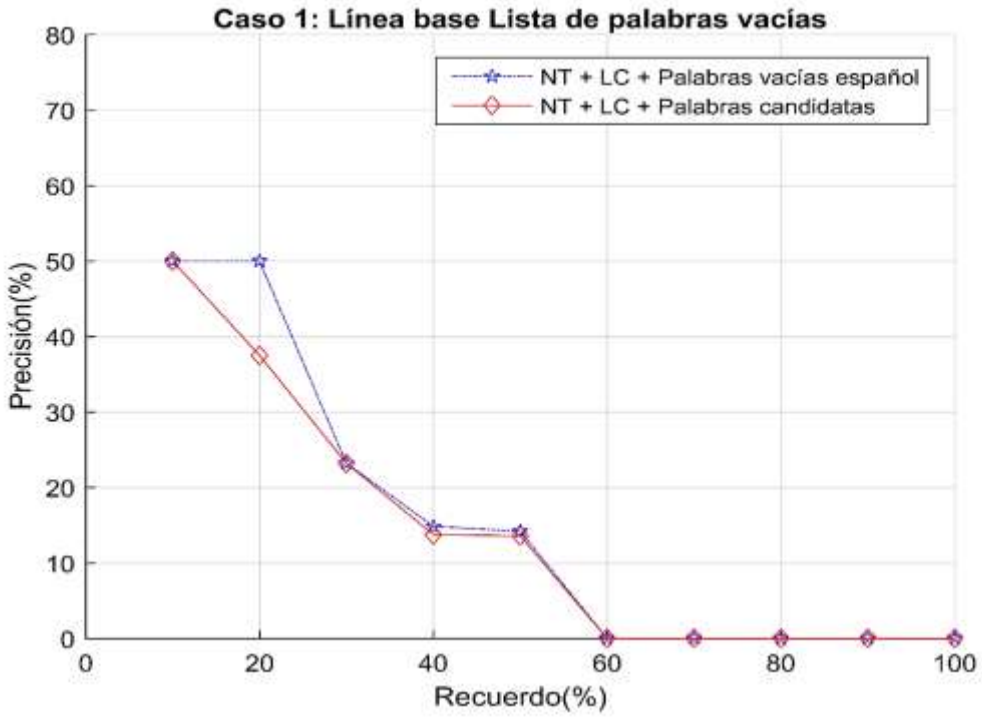


Figura 28. Curva Precisión Recuerdo con línea base Caso 1. Fuente: Elaboración propia

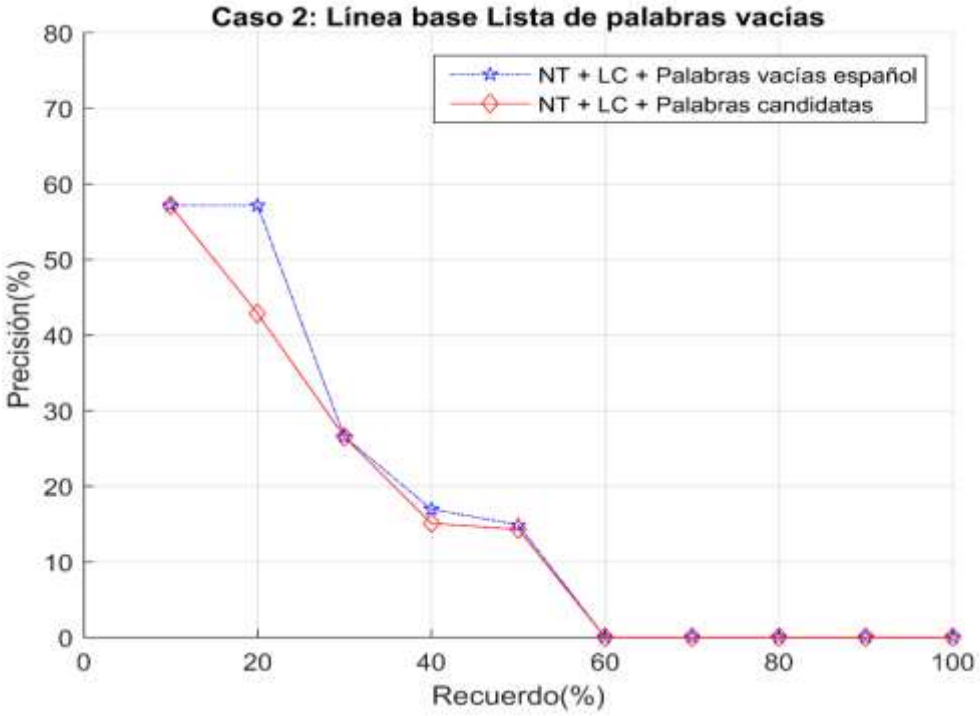


Figura 29. Curva Precisión Recuerdo con línea base Caso 2. Fuente: Elaboración propia

En la Figura 30, se presenta el comportamiento para el caso 3, a diferencia de los dos casos anteriores se puede apreciar que el sistema se comporta de manera similar, y al igual que en los casos anteriores fue necesario revisar la lista de palabras vacías, sin embargo, se puede apreciar que el impacto de las palabras vacías para consultas con muchos resultados relevantes es menor.

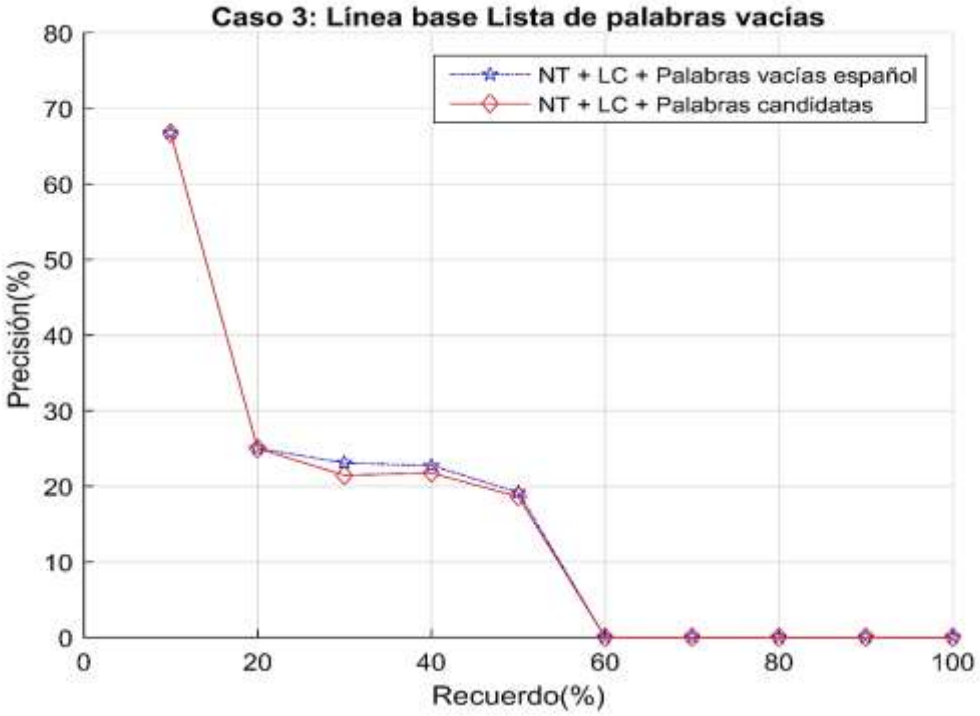


Figura 30. Curva Precisión Recuerdo con línea base Caso 3. Fuente: Elaboración propia

A continuación se presenta la Tabla 27, en donde se muestran los resultados del ajuste de la lista de palabras vacías.

Precisión Recuerdo	**10%**	**20%**	**30%**	**40%**	**50%**	**60%**	**70%**	**80%**	**90%**	**100%**
NT + LC + STW(ESPAÑOL) (Caso 1)	50,00%	50,00%	23,21%	14,84%	14,21%	0,00%	0,00%	0,00%	0,00%	0,00%
NT + LC Lista palabras vacías (34 candidatas) (Caso 1)	50,00%	37,50%	23,21%	13,75%	13,59%	0,00%	0,00%	0,00%	0,00%	0,00%
NT + LC Lista palabras vacías (20 candidatas) (Caso 1)	50,00%	56,25%	23,21%	14,84%	13,59%	0,00%	0,00%	0,00%	0,00%	0,00%

Tabla 27. Valores de Precisión Recuerdo con Lista de palabras vacías Caso 1. Fuente: Elaboración propia

En la Tabla 27, se presentan las siguientes situaciones de palabras vacías, así:

- En la primera fila se presenta los resultados de la evaluación de NT + LC + la lista de palabras vacías en español, que servirá de línea base para compararla con las siguientes evaluaciones y la estructura del índice es de 224 KB.

- En la segunda fila se presenta NT + LC + Lista palabras vacías (34 candidatas), donde se muestra una disminución en el desempeño del sistema y la estructura del índice es de 211 KB.

- En la tercera fila se encuentra NT + LC + Lista palabras vacías (20), el cual muestra una mejora en el desempeño del sistema, la lista de palabras es de 20 ("txãa", "naa", "teeçx", "txã'w","seena", "wẽt", "mẽh", "ma'w", "je'z", "maa", "na'w", "pa'ka", "txa'w", "txajuyu'", "ũskan", “u'jn", "majika", "meeçxa", "meeçxa'", "txa's", "txajx") y el tamaño de la estructura del índice es de 216 KB.

A continuación se presenta la Tabla 28, que presenta los resultados de la evaluación de la lista de palabras vacías con las cuales el sistema se desempeña mejor para el caso 2.

Precisión Recuerdo	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
NT + LC + STW(ESPAÑOL) (Caso 2)	57,14%	57,14%	26,53%	16,96%	14,88%	0,00%	0,00%	0,00%	0,00%	0,00%
NT + LC + Lista palabras vacías (34 candidatas) (Caso 2)	57,14%	42,86%	26,53%	15,08%	14,29%	0,00%	0,00%	0,00%	0,00%	0,00%
NT + LC + Lista palabras vacías (20 candidatas) (Caso 2)	57,14%	64,29%	26,53%	16,96%	14,88%	0,00%	0,00%	0,00%	0,00%	0,00%

Tabla 28. Valores de Precisión Recuerdo con Lista de palabras vacías Caso 2. Fuente: Elaboración propia

En la Tabla 28, se presentan las siguientes situaciones de palabras vacías así:

- En la primera fila se presentan los resultados de las evaluaciones de NT + LC + STW (ESPAÑOL), el cual se toma como línea base y la estructura del índice es de 224 KB.

- En la segunda fila NT + LC + Lista palabras vacías (34 candidatas), donde se muestra una disminución en el desempeño del sistema y la estructura del índice es de 211 KB.

- En la tercera fila se encuentra NT + LC + Lista palabras vacías (20), el cual muestra una mejora en el valor de la curva para el 20% de recuerdo, pero una disminución para otros valores la lista de palabras es "txãa", "naa","teeçx", "txã'w","jxuka", "vxite", "seena", "wẽt", "mẽh", "ma'w", "je'z", "maa", "na'w", "pa'ka", "txa'w", "txajuyu'", "ũskan", "u'jn", "majika","meeçxa'", "txa's", "txajx" y la estructura del índice es de 216 KB.

A continuación se presenta la Tabla 29, con los resultados de la evaluación de la lista de palabras vacías con las cuales el sistema se desempeña mejor para el caso 3. En la cual se puede apreciar que el sistema es menos sensible a la lista de palabras dado que sus resultados son similares a cuando se toman las 34 palabras candidatas.

Precisión Recuerdo	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
NT + LC + STW(ESPAÑOL) (Caso 3)	66,67%	25,00%	23,08%	22,73%	19,19%	0,00%	0,00%	0,00%	0,00%	0,00%
NT + LC + Lista palabras vacías (34 candidatas) (Caso 3)	66,67%	25,00%	21,43%	21,74%	18,63%	0,00%	0,00%	0,00%	0,00%	0,00%
NT + LC + Lista palabras vacías (20 candidatas) (Caso 3)	66,67%	25,00%	23,08%	22,73%	19,19%	0,00%	0,00%	0,00%	0,00%	0,00%

Tabla 29. Valores de Precisión Recuerdo con Lista de palabras vacías Caso 3. Fuente: Elaboración propia

En la Tabla 30, se presenta la lista de palabras que mejor se ajustaron a cada situación previamente descrita.

Término en la Colección		Lista de palabras vacías Caso 1	Lista de palabras vacías Caso 2	Lista de palabras vacías Caso 3
		NT + LC	NT + LC	NT + LC
1	txãa	X	X	X
2	naa	X	X	X
3	teeçx	X	X	X
4	txã'w		X	X
5	kwe'sx			X
6	sa'			X
7	ki'			X
8	jxuka		X	X
9	nawã			X
10	vxite		X	X
11	seena	X	X	X
12	aça'			X
13	wẽt	X	X	X
14	txaju'			X
15	mẽh	X	X	X
16	açyu'			X
17	nawa			X
18	ma'w	X	X	X
19	sena			X
20	aççxa			X
21	je'z	X		X
22	maa	X		X
23	na'w	X		X

Término en la Colección		Lista de palabras vacías Caso 1	Lista de palabras vacías Caso 2	Lista de palabras vacías Caso 3
		NT + LC	NT + LC	NT + LC
24	pa'ka	X	X	X
25	txa'w	X	X	X
26	txajuyu'	X	X	X
27	tekh			X
28	ũskan	X	X	X
29	u'jn	X	X	X
30	majika		X	X
31	meeçxa	X	X	X
32	meeçxa'	X	X	X
33	txa's	X	X	X
34	txajx	X	X	X
Total palabras vacías		**19**	**22**	**34**

Tabla 30. Lista de Palabras vacías para nasa yuwe. Elaboración propia

Como se puede apreciar en la Tabla 30, todas las palabras señaladas por el experto en lingüística se encontraron como palabras vacías, en alguno de los casos presentados, incluso algunas fueron plenamente identificables para todos los casos presentados.

La lista definitiva de palabras vacías para nasa yuwe, se saca de aquellas que son comunes para todos los casos presentados en la Tabla 30 quedando conformada por 16 palabras así: Txãa, naa, teeçx, seena, wẽt, mẽh, ma'w, pa'ka, txa'w, txajuyu', ũskan, u'jn, meeçxa, meeçxa', txa's, txajx.

Otro aspecto a considerar al momento de definir la lista de palabras vacías es su impacto en el tamaño de la estructura del índice, según [11] la eliminación de palabras vacías típicamente reduce 40% o más su tamaño, para el caso de la lista de palabras vacías para nasa yuwe, dado el tamaño de los documentos de la colección y las estadísticas de la colección (Tabla 10), se considera como una colección pequeña, los datos de tamaño a considerar son:

Sistema	Tamaño de la estructura del índice	% de disminución
NT + LC + STW(ESPAÑOL) Caso 1	224 KB	---
NT + LC + Lista palabras vacías (34 candidatas) Caso 1	211 KB	5,80%
NT + LC + Lista palabras vacías (16) Caso 1	218 KB	2,68%
NT + LC + STW(ESPAÑOL) Caso 2	224 KB	-----
NT + LC + Lista palabras vacías (34 candidatas) Caso 2	211 KB	5,80%

Sistema	Tamaño de la estructura del índice	% de disminución
NT + LC + Lista palabras vacías (16) Caso 2	218 KB	2,68%
NT + LC + STW(ESPAÑOL) Caso 3	224 KB	
NT + LC + Lista palabras vacías (34 candidatas) Caso 3	211 KB	5,80%
NT + LC Lista palabras vacías (16) Caso 3	218 KB	2,68%

Tabla 31. Tamaño de la estructura del índice con Lista de palabras vacías. Fuente: Elaboración propia

En el archivo Anexo 4 Evaluación con palabras vacías, se encuentra el detalle de las evaluaciones realizadas para identificar la lista de palabras vacías presentada en la Tabla 30.

3.3 OTRAS MEDIDAS DE EVALUACIÓN

3.3.1 Medida F at K, Precisión at K, Recuerdo at K

A continuación en la Tabla 32 se presenta otras medidas de evaluación calculadas para el sistema utilizando NT + LC + Lista de palabras vacías en contraste con ST + LC, calculadas para cada consulta, se seleccionaron los 8 primeros resultados para hacer el cálculo de F_1 at K, precisión at K, y recuerdo at K. En el archivo Anexo 5. Otras Medidas de Evaluación, se presenta la tabla completa con las medidas para 40 resultados.

En esta tabla se puede apreciar que las medidas son similares para la mayoría de las consultas, sin embargo, para la consulta 2, el desempeño del tokenizador nasa yuwe es superior en comparación con el tokenizador estándar, también se observa que en algunos casos las medidas para el resultado 1 muestran mejor desempeño cuando se utiliza ST + LC + Lista de palabras vacías y para los siguientes resultados el NT + LC + Lista de palabras vacías mejorando su desempeño.

En la Tabla 33, se presentan las medidas de Precisión, Recuerdo y Medida F_1, a manera de resumen por resultado para cada consulta. Allí se puede observar que la evaluación para el NT + LC + Lista de palabras vacías, es mejor en la mayoría de los resultados que las evaluaciones para ST + LC + Lista de palabras vacías. En la tabla se resaltan los valores en donde cada sistema se desempeña mejor que el otro en pro de favorecer la visualización sobre el desempeño.

ST + LC + LISTA DE PALABRAS VACIAS NASA								NT + LC + LISTA DE PALABRAS VACÍAS NASA						
Idconsulta	Resultado	Relev Recup	Recup	Total Relev	Precisión at K	Recall at K	Medida F at K	Resultado	Relev Recup	Recup	Total Relev	P at K	R at K	Medida F at K
1	1	1	1	5	100%	20%	33%	1	0	1	5	0,00%	0,00%	0,00%
1	2	1	2	5	50%	20%	29%	2	1	2	5	50,00%	20,00%	28,57%
1	3	1	3	5	33%	20%	25%	3	1	3	5	33,33%	20,00%	25,00%
1	4	1	4	5	25%	20%	22%	4	1	4	5	25,00%	20,00%	22,22%
1	5	1	5	5	20%	20%	20%	5	1	5	5	20,00%	20,00%	20,00%
1	6	1	6	5	17%	20%	18%	6	1	6	5	16,67%	20,00%	18,18%
1	7	1	7	5	14%	20%	17%	7	1	7	5	14,29%	20,00%	16,67%
1	8	1	8	5	13%	20%	15%	8	1	8	5	12,50%	20,00%	15,38%
2	1	0	1	10	0%	0%	0%	1	1	1	10	100,00%	10,00%	18,18%
2	2	0	2	10	0%	0%	0%	2	1	2	10	50,00%	10,00%	16,67%
2	3	0	3	10	0%	0%	0%	3	1	3	10	33,33%	10,00%	15,38%
2	4	0	4	10	0%	0%	0%	4	1	4	10	25,00%	10,00%	14,29%
2	5	0	5	10	0%	0%	0%	5	1	5	10	20,00%	10,00%	13,33%
2	6	0	6	10	0%	0%	0%	6	1	6	10	16,67%	10,00%	12,50%
2	7	0	7	10	0%	0%	0%	7	2	7	10	28,57%	20,00%	23,53%
2	8	0	8	10	0%	0%	0%	8	2	8	10	25,00%	20,00%	22,22%
3	1	0	1	20	0%	0%	0%	1	0	1	20	0,00%	0,00%	0,00%
3	2	1	2	20	50%	5%	9%	2	0	2	20	0,00%	0,00%	0,00%
3	3	1	3	20	33%	5%	9%	3	1	3	20	33,33%	5,00%	8,70%
3	4	1	4	20	25%	5%	8%	4	1	4	20	25,00%	5,00%	8,33%
3	5	1	5	20	20%	5%	8%	5	1	5	20	20,00%	5,00%	8,00%
3	6	1	6	20	17%	5%	8%	6	1	6	20	16,67%	5,00%	7,69%
3	7	2	7	20	29%	10%	15%	7	1	7	20	14,29%	5,00%	7,41%

ST + LC + LISTA DE PALABRAS VACIAS NASA								NT + LC + LISTA DE PALABRAS VACÍAS NASA						
Idconsulta	**Resultado**	**Relev Recup**	**Recup**	**Total Relev**	**Precisión at K**	**Recall at K**	**Medida F at K**	**Resultado**	**Relev Recup**	**Recup**	**Total Relev**	**P at K**	**R at K**	**Medida F at K**
3	8	2	8	20	25%	10%	14%	8	1	8	20	12,50%	5,00%	7,14%
4	1	1	1	15	100%	7%	13%	1	1	1	15	100,00%	6,67%	12,50%
4	2	2	2	15	100%	13%	24%	2	2	2	15	100,00%	13,33%	23,53%
4	3	2	3	15	67%	13%	22%	3	2	3	15	66,67%	13,33%	22,22%
4	4	2	4	15	50%	13%	21%	4	2	4	15	50,00%	13,33%	21,05%
4	5	2	5	15	40%	13%	20%	5	2	5	15	40,00%	13,33%	20,00%
4	6	2	6	15	33%	13%	19%	6	2	6	15	33,33%	13,33%	19,05%
4	7	2	7	15	29%	13%	18%	7	2	7	15	28,57%	13,33%	18,18%
4	8	2	8	15	25%	13%	17%	8	3	8	15	37,50%	20,00%	26,09%
5	1	1	1	4	100%	25%	40%	1	1	1	4	100,00%	25,00%	40,00%
5	2	1	2	4	50%	25%	33%	2	2	2	4	100,00%	50,00%	66,67%
5	3	1	3	4	33%	25%	29%	3	2	3	4	66,67%	50,00%	57,14%
5	4	1	4	4	25%	25%	25%	4	2	4	4	50,00%	50,00%	50,00%
5	5	1	5	4	20%	25%	22%	5	3	5	4	60,00%	75,00%	66,67%
5	6	2	6	4	33%	50%	40%	6	3	6	4	50,00%	75,00%	60,00%
5	7	2	7	4	29%	50%	36%	7	3	7	4	42,86%	75,00%	54,55%
5	8	3	8	4	38%	75%	50%	8	3	8	4	37,50%	75,00%	50,00%
6	1	1	1	2	100%	50%	67%	1	0	1	2	0,00%	0,00%	0,00%
6	2	1	2	2	50%	50%	50%	2	1	2	2	50,00%	50,00%	50,00%
6	3	1	3	2	33%	50%	40%	3	1	3	2	33,33%	50,00%	40,00%
6	4	1	4	2	25%	50%	33%	4	1	4	2	25,00%	50,00%	33,33%
6	5	1	5	2	20%	50%	29%	5	1	5	2	20,00%	50,00%	28,57%
6	6	1	6	2	17%	50%	25%	6	1	6	2	16,67%	50,00%	25,00%
6	7	1	7	2	14%	50%	22%	7	1	7	2	14,29%	50,00%	22,22%

ST + LC + LISTA DE PALABRAS VACIAS NASA								NT + LC + LISTA DE PALABRAS VACÍAS NASA						
Idconsulta	Resultado	Relev Recup	Recup	Total Relev	Precisión at K	Recall at K	Medida F at K	Resultado	Relev Recup	Recup	Total Relev	P at K	R at K	Medida F at K
6	8	1	8	2	13%	50%	20%	8	1	8	2	12,50%	50,00%	20,00%
7	1	1	1	5	100%	20%	33%	1	1	1	5	100,00%	20,00%	33,33%
7	2	1	2	5	50%	20%	29%	2	2	2	5	100,00%	40,00%	57,14%
7	3	1	3	5	33%	20%	25%	3	3	3	5	100,00%	60,00%	75,00%
7	4	1	4	5	25%	20%	22%	4	3	4	5	75,00%	60,00%	66,67%
7	5	1	5	5	20%	20%	20%	5	3	5	5	60,00%	60,00%	60,00%
7	6	1	6	5	17%	20%	18%	6	3	6	5	50,00%	60,00%	54,55%
7	7	1	7	5	14%	20%	17%	7	3	7	5	42,86%	60,00%	50,00%
7	8	1	8	5	13%	20%	15%	8	3	8	5	37,50%	60,00%	46,15%
8	1	0	1	1	0%	0%	0%	1	0	1	1	0,00%	0,00%	0,00%
8	2	0	2	1	0%	0%	0%	2	0	2	1	0,00%	0,00%	0,00%
8	3	0	3	1	0%	0%	0%	3	0	3	1	0,00%	0,00%	0,00%
8	4	0	4	1	0%	0%	0%	4	0	4	1	0,00%	0,00%	0,00%
8	5	0	5	1	0%	0%	0%	5	0	5	1	0,00%	0,00%	0,00%
8	6	0	6	1	0%	0%	0%	6	0	6	1	0,00%	0,00%	0,00%
8	7	0	7	1	0%	0%	0%	7	0	7	1	0,00%	0,00%	0,00%
8	8	0	8	1	0%	0%	0%	8	0	8	1	0,00%	0,00%	0,00%

Tabla 32. Medida F para ST + LC Vs NT + LC + palabras vacías. Fuente: Elaboración propia

RESULTADO	ST + LC + Lista de palabras vacías nasa			NT + LC + Lista de palabras vacías nasa		
	RECUERDO	PRECISIÓN	MEDIDA F1	RECUERDO	PRECISIÓN	MEDIDA F1
1	8,06%	62,50%	14,28%	6,45%	50,00%	11,43%
2	11,29%	43,75%	17,95%	14,52%	56,25%	23,08%
3	11,29%	29,17%	16,28%	17,74%	45,83%	25,58%
4	11,29%	21,88%	14,89%	17,74%	34,38%	23,40%
5	11,29%	17,50%	13,73%	17,74%	27,50%	21,57%
6	12,90%	16,67%	14,54%	19,35%	25,00%	21,82%
7	14,52%	16,07%	15,26%	20,97%	23,21%	22,03%
8	16,13%	15,63%	15,88%	22,58%	21,88%	22,22%
9	19,35%	16,67%	17,91%	24,19%	20,83%	22,39%
10	19,35%	15,00%	16,90%	25,81%	20,00%	22,54%
11	20,97%	14,77%	17,33%	25,81%	18,18%	21,33%
12	20,97%	13,54%	16,46%	27,42%	17,71%	21,52%
13	22,58%	13,46%	16,87%	29,03%	17,31%	21,69%
14	25,81%	14,29%	18,39%	29,03%	16,07%	20,69%
15	25,81%	13,33%	17,58%	29,03%	15,00%	19,78%
16	27,42%	13,28%	17,89%	30,65%	14,84%	20,00%
17	27,42%	12,50%	17,17%	30,65%	13,97%	19,19%
18	29,03%	12,50%	17,48%	30,65%	13,19%	18,45%
19	30,65%	12,50%	17,76%	30,65%	12,50%	17,76%
20	33,87%	13,13%	18,92%	35,48%	13,75%	19,82%
21	35,48%	13,10%	19,13%	35,48%	13,10%	19,13%
22	37,10%	13,07%	19,33%	38,71%	13,64%	20,17%
23	40,32%	13,59%	20,33%	40,32%	13,59%	20,33%
24	40,32%	13,02%	19,69%	41,94%	13,54%	20,47%
25	43,55%	13,50%	20,61%	41,94%	13,00%	19,85%
26	45,16%	13,46%	20,74%	41,94%	12,50%	19,26%
27	45,16%	12,96%	20,14%	43,55%	12,50%	19,42%
28	45,16%	12,50%	19,58%	43,55%	12,05%	18,88%
29	45,16%	12,07%	19,05%	43,55%	11,64%	18,37%
30	45,16%	11,67%	18,54%	43,55%	11,25%	17,88%
31	45,16%	11,29%	18,06%	43,55%	10,89%	17,42%
32	46,77%	11,33%	18,24%	43,55%	10,55%	16,98%
33	48,39%	11,36%	18,40%	45,16%	10,61%	17,18%
34	48,39%	11,03%	17,96%	46,77%	10,66%	17,37%
35	50,00%	11,07%	18,13%	46,77%	10,36%	16,96%
36	53,23%	11,46%	18,86%	46,77%	10,07%	16,57%
37	53,23%	11,15%	18,44%	48,39%	10,14%	16,76%
38	53,23%	10,86%	18,03%	48,39%	9,87%	16,39%

RESULTADO	ST + LC + Lista de palabras vacías nasa			NT + LC + Lista de palabras vacías nasa		
	RECUERDO	PRECISIÓN	MEDIDA F1	RECUERDO	PRECISIÓN	MEDIDA F1
39	53,23%	10,58%	17,65%	48,39%	9,62%	16,04%
40	53,23%	10,31%	17,28%	48,39%	9,38%	15,71%

Tabla 33. Resumen de medidas por Resultado para todas las consultas. Fuente: Elaboración propia

3.3.2 Precisión Media en documentos relevantes observados (Average Precision at Seen Relevant Documents)

Esta medida permite resumir el comportamiento del sistema, en relación con cada documento nuevo que recupera en un ranking de documentos. A continuación se presenta esta medida para cada consulta con el procesamiento NT + LC + Lista de Palabras Vacías en contraste con ST + LC + Lista de Palabras Vacías, el detalle de los datos se puede apreciar en el archivo Anexo 5. Otras Medidas de Evaluación.

En la Tabla 34, se puede apreciar que el sistema es bastante rápido para la recuperación de documentos relevantes para las consultas 1, 2, 4, 5 y 7. Para las consultas 3 y 8 el desempeño de los dos procesamientos es muy similar.

ASPRD				
Consulta	ST + LC + Lista Nasa		NT + LC + Lista Nasa	
	Ranking de Precisiones	ASPRD	Ranking de Precisiones	ASPRD
1	$Res_1 = 1$ $Res_{35} = 0,57$ $Res_{36} = 0,83$	0,38	$Res_2 = 1$	1
2	$Res_9 = 0,11$ $Res_{25} = 0,77$ $Res_{32} = 0,09$	0,09	$Res_1 = 1$ $Res_7 = 0,29$	0,65
3	$Res_2 = 0,5$ $Res_7 = 0,29$ $Res_{13} = 0,23$ $Res_{14} = 0,29$ $Res_{18} = 0,28$ $Res_{19} = 0,32$ $Res_{20} = 0,35$ $Res_{23} = 0,35$ $Res_{25} = 0,36$ $Res_{33} = 0,30$ $Res_{36} = 0,31$	0,32	$Res_3 = 0,33$ $Res_{10} = 0,20$ $Res_{12} = 0,25$ $Res_{13} = 0,31$ $Res_{16} = 0,31$ $Res_{20} = 0,30$ $Res_{22} = 0,32$ $Res_{26} = 0,31$ $Res_{33} = 0,27$ $Res_{34} = 0,29$ $Res_{37} = 0,30$	0,29

ASPRD				
Consulta	ST + LC + Lista Nasa		NT + LC + Lista Nasa	
	Ranking de Precisiones	ASPRD	Ranking de Precisiones	ASPRD
4	Res_1= 1	0,47	Res_1 = 1	0,48
	Res_2 =1		Res_2 = 1	
	Res_9 = 0,33		Res_8 = 0,38	
	Res_{20} = 0,2		Res_{20} = 0,20	
	Res_{22} =0,23		Res_{21} = 0,24	
	Res_{23} = 0,26		Res_{22} = 0,27	
	Res_{25} = 0,28		Res_{24} = 0,29	
5	Res_1= 1	0,47	Res_1 = 1	0,76
	Res_6 = 0,33		Res_2 = 1	
	Res_8=0,38		Res_5 = 0,6	
	Res_{20} = 0,2		Res_9 = 0,44	
6	Res_1= 1	1	Res_2 = 0,5	0,5
7	Res_1 =1	0,46	Res_1 = 1	1
	Res_{11} = 0,18		Res_2 = 1	
	Res_{14} = 0,21		Res_3 = 1	
8	Res_{16} = 0,06	0,06	Res_{20} = 0,05	0,05

Tabla 34. Valores de APSRD. Fuente: Elaboración propia

3.3.3 Precisión R (R-Precision)

Esta medida es un parámetro útil para observar el comportamiento del sistema (NT + LC + Lista de palabras vacías nasa en contraste con ST + LC + Lista de palabras vacías nasa) para cada consulta. A continuación en la Tabla 35, se presenta la comparación de esta medida y se resaltan las consultas en donde el NT + LC + Lista de palabras vacías tiene una mejor evaluación en su desempeño.

Se puede apreciar en la Tabla 35, que el sistema maneja una buena precisión R para algunas consultas 2 y la 5 y para otras consultas como la 8 y la 6, los valores de Precisión R no son muy buenos.

Consulta	Valor de R	ST + LC Lista de Palabras vacías nasa	NT + LC + Lista de Palabras vacías nasa
		R-Precisión	R-Precisión
1	R=10	10%	10%
	R=20	5%	5%
2	R=10	10%	20%
	R= 20	5%	5%
3	R=10	20%	20%
	R=20	35%	30%
4	R=10	30%	30%

Consulta	Valor de R	ST + LC Lista de Palabras vacías nasa	NT + LC + Lista de Palabras vacías nasa
		R-Precisión	R-Precisión
	R=20	20%	20%
5	R=10	30%	40%
	R=20	20%	20%
6	R=10	10%	10%
	R=20	5%	5%
7	R=10	10%	30%
	R=20	15%	15%
8	R=10	0%	0%
	R=20	5%	5%

Tabla 35. Valores de APSRD para consulta 1 – 8. Fuente: Elaboración propia

3.3.4 Histograma de Precisión

Esta medida permite comparar el desempeño de dos algoritmos para un valor de Precisión R =10 de cada consulta. En la Figura 31, se toma A como NT + LC + LP y a B como ST + LC + LP, observándose un desempeño similar entre las dos ejecuciones para las consultas 1, 3, y 6, un mejor desempeño en las consultas 2, 5 y 7 por parte de A, en la figura se muestra el desempeño de los algoritmos mediante las barras amarillas del histograma.

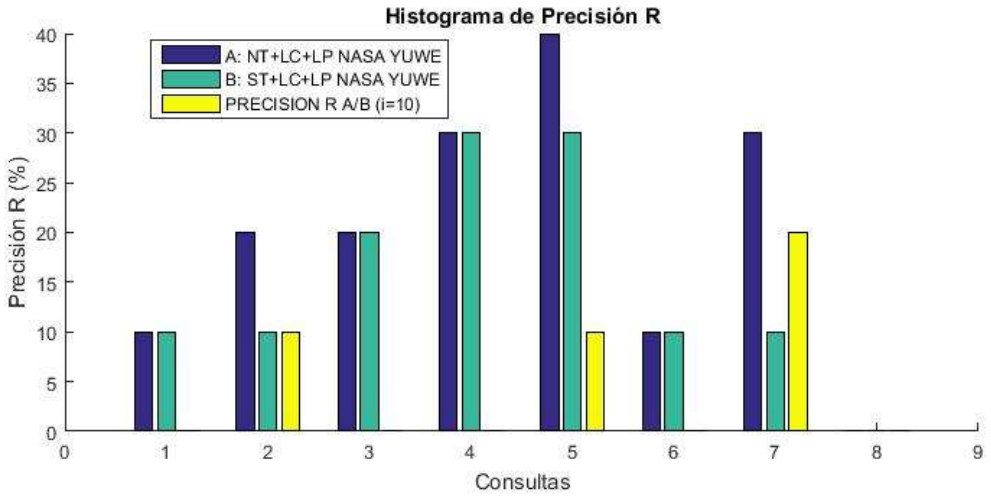


Figura 31. Histograma de Precisión R=10. Fuente: Elaboración propia

Haciendo un resumen de las diferentes evaluaciones mostradas en esta sección, se puede apreciar que:

- El tokenizador nasa tiene un mejor desempeño para valores bajos de recuerdo en contraste con el desempeño del tokenizador estándar, para todos los casos, especialmente, para el caso 3.

- Cuando al tokenizador nasa se le adiciona el convertidor de texto en minúscula, los valores de precisión mejoran sustancialmente para valores del 20% al 50% de recuerdo y se obtiene también un desempeño similar para el resto de los valores de recuerdo.

- Para los valores en los que el tokenizador estándar supera el tokenizador nasa, cabe recordar que como se tokenizan incorrectamente las palabras nasa con el tokenizador estándar, se puede aumentar la posibilidad de coincidencia entre los términos de los documentos de la colección y la consulta, de tal forma, que se alcanza un mejor desempeño pero los resultados no son los adecuados, recuperando documentos no relacionados con la consulta; situación que no se presenta con el tokenizador nasa.

- Al adicionar la lista de palabras vacías el tokenizador nasa mantiene su desempeño, sin embargo, se pudo apreciar que: 1) las palabras incorrectas impactan el desempeño del tokenizador; 2) al seleccionar la lista adecuada de palabras vacías el desempeño se mantiene; 3) el tamaño de la estructura del índice solo se reduce en un 2.68%.

- Con las medidas de Precisión R y Precisión Media en documentos relevantes observado, se puede apreciar que el sistema de recuperación al utilizar el tokenizador nasa muestra buen desempeño en varias consultas, como se aprecia en el histograma de precisión R. Además basado en los resultados de Precisión Media en documentos relevantes observados, se puede concluir que el sistema recupera más rápidamente resultados relevantes con el tokenizador nasa propuesto y el uso de las palabras vacías seleccionadas.

4. SISTEMA DE RECUPERACIÓN DE INFORMACIÓN PARA TEXTOS ESCRITOS EN NASA YUWE

Para la construcción del prototipo de recuperación de textos escritos en nasa yuwe, se utilizó la metodología XP [32]. El objetivo de la construcción del prototipo fue la evaluación del desempeño del sistema de recuperación de información y la recuperación de los textos escritos en nasa yuwe de la colección de prueba construida (capitulo 2).

4.1 FASE DE EXPLORACIÓN

En esta fase se elaboraron las historias de usuario y se definió la arquitectura del prototipo, así como la familiarización con las herramientas de desarrollo y el paquete Lucene .Net (versión 2.9.4).

4.1.1 Historias de usuario

HISTORIA DE USUARIO No.1		Usuario: Profesor Comunidad nasa
Nombre: Indexar archivos		
Prioridad: Alta	Puntos estimados: 5	Riesgo: medio
Responsable: Luz Marina Sierra		
Descripción		
Una de las primeras tareas es indexar los documentos de la colección. Se requiere crear: el índice en el disco, almacenarlo en disco y procesar el texto de los documentos de la colección en tokens.		
Observación: Se utilizan las clases del análisis de Lucene .NET para hacer esta indexación		

Tabla 36. Historia de Usuario No. 1. Fuente: Elaboración propia

HISTORIA DE USUARIO No.2		Usuario: Profesor Comunidad nasa
Nombre: Procesar consulta		
Prioridad: Alta	Puntos estimados: 3	Riesgo: medio
Responsable: Luz Marina Sierra		
Descripción		
Otra tarea del prototipo es procesar las consultas en tokens, de tal forma que se pueda hacer la evaluación de similitud entre tokens de consulta y documentos. Se requiere procesar el texto de las consultas.		
Observación: Se utilizan las clases del análisis de Lucene .NET para hacer este procesamiento.		

Tabla 37. Historia de Usuario No. 2. Fuente: Elaboración propia

HISTORIA DE USUARIO No.3		Usuario: Desarrollador / profesor Comunidad nasa
Nombre: Evaluar desempeño		
Prioridad: Alta	Puntos estimados: 3	Riesgo: medio
Responsable: Luz Marina Sierra		
Descripción		

El prototipo debe utilizar una medida de evaluación para revisar el desempeño en relación al procesamiento de los documentos de la colección nasa yuwe. Para este caso se utiliza la curva precisión recuerdo promediada.
Observación: Los resultados de las medidas para cada consulta deben quedar almacenados en la Base de Datos.

Tabla 38. Historia de Usuario No. 3. Fuente: Elaboración propia

HISTORIA DE USUARIO No.4	Usuario: profesor Comunidad nasa	
Nombre: Almacenar medidas		
Prioridad: Alta	Puntos estimados: 1	Riesgo: medio
Responsable: Luz Marina Sierra		
Descripción		
El prototipo debe permitir al usuario ver el listado de documentos de la colección de prueba recuperados en relación a una consulta ingresada. .		
Observación: Los resultados de las medidas para cada consulta deben quedar almacenados en la Base de Datos.		

Tabla 39. Historia de Usuario No. 4. Fuente: Elaboración propia

4.1.2 Arquitectura

En la Figura 32, se presenta un diseño general de la arquitectura del prototipo, utilizando la arquitectura de 3 capas de Microsoft [78].

Figura 32. Bosquejo de la arquitectura. Fuente: Adaptación Microsoft [78]

Es de tenerse en cuenta que se utiliza una librería adicional que contiene las clases de procesamiento de Lucene .NET (versión 2.9.4).

4.2 FASE DE PLANEACIÓN

En esta fase se priorizaron las historias de usuario. Como producto de esta fase se obtuvo el plan de iteraciones.

4.2.1 Plan de iteraciones

Se realizaron tres iteraciones así:

Iteración	Número de Historia de Usuario	Nombre Historia de Usuario
1	1	Indexar archivos
1	2	Procesar consulta
2	3	Evaluar desempeño
3	4	Almacenar medidas

Tabla 40. Plan de iteraciones. Fuente: Elaboración propia

4.3 FASE DE ITERACIÓN

El prototipo ha sido desarrollado como aplicación de escritorio para Windows, dado que inicialmente se trabaja de manera personal con cada profesor de la comunidad que ha colaborado con este proyecto y de esta manera como trabajo futuro corregir y ajustar la aplicación (especialmente en lo relacionado con interfaz) según lo observado por los profesores para posteriormente pasarla a una versión web y publicarla.

4.3.1 Selección de la tecnología para la construcción del prototipo

Como herramienta de desarrolló se seleccionó Microsoft Visual Studio 2010 utilizando lenguaje C# y Base de datos de SQL Server 2008, debido a:

- Su facilidad para desarrollar aplicaciones complejas y la integración con otros desarrollos.
- La velocidad de procesamiento de las aplicaciones que se construyen en esta plataforma.
- El conocimiento previo por parte de la desarrolladora.
- Desde un principio se pensó en hacer uso de las librerías de Lucene .NET [6].

4.3.2 Iteraciones

En esta sección se presentan agrupados los artefactos de cada iteración, mostrados por las 4 tareas: Análisis, Diseño, Codificación y Despliegue.

1. Análisis.

A continuación se presentan las tareas de ingeniería y las tarjetas Clase Responsabilidad Colaboración – CRC que permitieron obtener los detalles de estas historias.

Tarea de Ingeniería No. 1	Indexación de archivos en disco	Historia de Usuario No. 1
Descripción: Esta tarea está relacionada con: Ubicar los archivos de la colección de prueba, crear el directorio donde se almacenan los índices, y finalmente crear el índice.		

Tabla 41. Tarea de ingeniería No. 1. Fuente: Elaboración propia.

Tarea de Ingeniería No. 2	Indexación documento	Historia de Usuario No. 1
Descripción:		

Esta tarea está relacionada con: Leer cada documento de la colección, procesarlo y adicionarlo al índice.

Tabla 42. Tarea de ingeniería No. 2. Fuente: Elaboración propia.

Tarea de Ingeniería No. 3	Procesar texto	Historia de Usuario No. 1
Descripción: Esta tarea se relaciona con la aplicación de las tareas a los documentos de la colección y al texto de la consulta. Entre las tareas se encuentran: tokenizar (estándar o nasa), utilizar filtros y convertidores de texto a minúsculas y definir la lista de palabras vacías a utilizar para el procesamiento. Se deben tener en cuenta varias opciones independientes para que se puedan realizar las diferentes evaluaciones del sistema.		

Tabla 43. Tarea de ingeniería No. 3. Fuente: Elaboración propia.

Tarea de Ingeniería No. 4	Tokenizar con Lucene .NET	Historia de Usuario No. 1
Descripción: Esta tarea se encarga de organizar los parámetros del texto de los documentos y llamar la clase de Lucene .NET encargada de utilizar el tokenizador estándar de Lucene .NET y como resultado se obtiene los tokens de cada documento de la colección.		

Tabla 44. Tarea de ingeniería No. 4. Fuente: Elaboración propia.

Tarea de Ingeniería No. 5	Aplicar Filtro	Historia de Usuario No. 1
Descripción: Esta tarea se encarga de recibir los tokens de cada documento una vez se aplique el tokenizador (estándar o nasa) y aplicar el filtro de Lucene para quitar la puntuación que aún no se ha retirado y las terminaciones ´s entre otros.		

Tabla 45. Tarea de ingeniería No. 5. Fuente: Elaboración propia.

Tarea de Ingeniería No. 6	Aplicar Convertidor de texto a mayúsculas	Historia de Usuario No. 1
Descripción: Esta tarea se encarga de convertir los tokens de cada documento resultantes en minúsculas.		

Tabla 46. Tarea de ingeniería No. 6. Fuente: Elaboración propia.

Tarea de Ingeniería No. 7	Incluir aspectos del nasa yuwe al tokenizador	Historia de Usuario No. 1
Descripción: Esta tarea se encarga de hacer la adaptación al tokenizador estándar de Lucene con el fin de incluir aquellos aspectos propios del nasa yuwe, descritos en el capítulo anterior (sección 3.1.4)		

Tabla 47. Tarea de ingeniería No. 7. Fuente: Elaboración propia.

Tarea de Ingeniería No. 8	Definir Lista de palabras vacías	Historia de Usuario No. 1
Descripción: Se encarga de definir y hacer uso de las palabras vacías, es decir, los tokens que se deben quitar tanto en nasa yuwe como en español.		

Tabla 48. Tarea de ingeniería No. 8. Fuente: Elaboración propia.

Tarea de Ingeniería No.9	Procesar consulta	Historia de Usuario No. 2
Descripción: Esta tarea se encarga de generar las interfaces para que se puedan procesar las consultas de la misma manera y con las opciones definidas para los documentos de la colección, así como la creación del índice para la consulta o consultas.		

Tabla 49. Tarea de ingeniería No. 9. Fuente: Elaboración propia.

Tarea de Ingeniería No. 10	Evaluar desempeño	Historia de Usuario No. 3
Descripción: Se encarga de ir comparando los documentos de la colección con el texto de la consulta en pro de listar los documentos relevantes e ir calculando los documentos relevantes, los recuperados y el total de documentos relevantes tanto por consulta como por resultado general del sistema. Con estos datos se calculan los valores de precisión para cada valor de recuerdo en cada consulta y se promedia teniendo en cuenta el número de consultas involucradas. Estos valores son los que se muestran al usuario.		

Tabla 50. Tarea de ingeniería No. 10. Fuente: Elaboración propia.

Tarea de Ingeniería No. 11	Almacenamiento de medidas	Historia de Usuario No. 4
Descripción: Se encarga de recibir los datos que se obtienen de la tarea anterior y enviarlo a la Base de Datos.		

Tabla 51. Tarea de ingeniería No. 11. Fuente: Elaboración propia.

A continuación se presentan las principales tarjetas CRC.

Indexar	
Responsabilidades	Colaboradores
Mostrar términos indexados	Tokenizador Nasa
Crear índice	Tokenizador estándar (Lucene .NET)
Indexar documento	Filtro (Lucene .NET)
Procesar Texto de documentos y consultas	Convertir texto en min (Lucene .NET)
	Remoción de palabras vacías

Tabla 52. Tarjeta CRC No. 1. Fuente: Elaboración propia.

Consultar	
Responsabilidades	Colaboradores
Procesar consultas	Indexar Estadísticas
Buscar documentos relevantes	
Llevar y almacenar estadísticas	
Calcular valores de precisión para cada valor recuerdo	
Mostrar valores	

Tabla 53. Tarjeta CRC No. 2. Fuente: Elaboración propia.

Lista para la remoción de palabras vacías	
Responsabilidades	Colaboradores
Definir listas de palabras vacías para nasa yuwe y español	Indexar

Tabla 54. Tarjeta CRC No. 3. Fuente: Elaboración propia

Estadística	
Responsabilidades	Colaboradores
Almacenar datos en base de datos y mostrarlos	Consultar

Tabla 55. Tarjeta CRC No. 4. Fuente: Elaboración propia

2. Diseño

Se definieron las pruebas que permitieran evaluar el desempeño del prototipo.

Caso 1: Tokenizador estándar de Lucene
Determinar que se está tokenizando cada palabra de los documentos de prueba de la colección utilizando el tokenizador estándar de Lucene
Caso 2: Tokenizador nasa
Determinar que se está tokenizando cada palabra de los documentos de prueba de la colección utilizando el tokenizador nasa yuwe
Caso 3: Aplicar Convertidor de texto a minúsculas
Determinar que los tokens se convierten todos en minúsculas
Caso 4: Mostrar los datos de precisión y recuerdo para cada configuración (tokenizador nasa o estándar, filtro (si o no), convertidor de minúsculas (si o no), lista de palabras vacías (si o no)

Tabla 56. Casos configurados para realizar pruebas al prototipo. Fuente: Elaboración propia

Las interfaces del prototipo son muy sencillas, las cuales se han desarrollado para efectos de uso de este trabajo, en el momento en que se vaya a entregar al usuario final se dejarán solo las opciones y funciones requeridas por este. A continuación se presentan:

- Interfaz principal, el sistema debe presentar dos opciones, una para realizar procesamiento para evaluación de desempeño y otra para hacer las consultas y ver resultados (Figura 33).

Figura 33. Ventana principal. Fuente: Elaboración propia

- Interfaz de procesamiento para evaluación de desempeño, se indica la ubicación de los documentos de la colección y en donde se debe ubicar el índice, las opciones de indexar primero y luego evaluar, con esta última se muestra en una tabla o data grid los datos de precisión y recuerdo resultantes del procesamiento, como se muestra en la Figura 34.

Figura 34. Ventana de evaluación de desempeño. Fuente: Elaboración propia

- Interfaz de consultas, debe permitir introducir el texto de la consulta, activar la consulta y mostrar los resultados, como se muestra en la Figura 35.

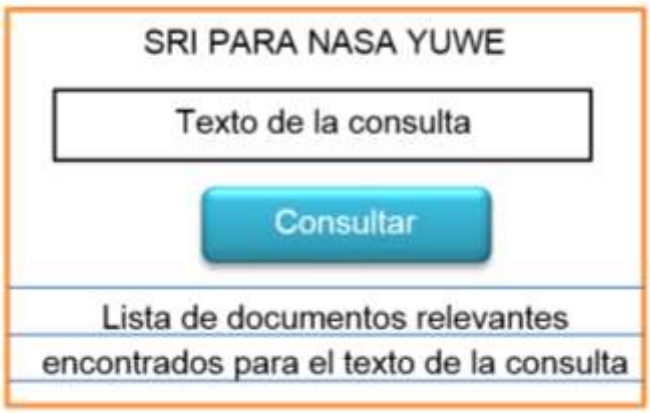


Figura 35. Ventana de consultas. Fuente: Elaboración propia

3. Codificación

Se implementaron las tareas de ingeniería correspondientes a las historias de usuario, utilizando la herramienta de desarrollo seleccionada.

4. Despliegue

A continuación se presentan algunas interfaces de la aplicación. En la Figura 36, se presenta la ventana principal del prototipo de SRI para nasa yuwe desarrollado, allí se puede apreciar en la zona 1, la indexación que se debe realizar de los documentos de la colección de prueba para esto se solicita la ruta en donde se encuentra la colección y la ruta en donde se van a almacenar la estructura de índices. En la zona 2, se presentan dos botones que contienen las dos opciones del SRI para hacer la consulta mediante el botón de recuperación de información y la otra opción es para evaluar el desempeño del sistema mediante el botón Evaluador de rendimiento.

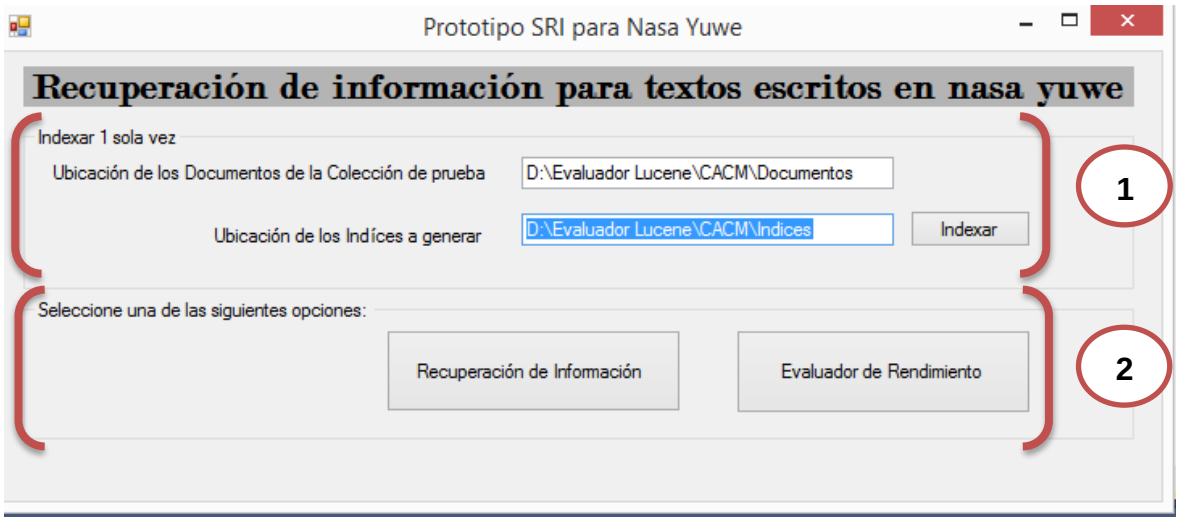


Figura 36. Ventana Principal del prototipo. Fuente: Elaboración propia

En la Figura 37, en la zona 1, se muestra la ruta en donde se encuentran almacenados los índices generados en el formulario principal, en la zona 2, se presenta un campo para indicar la cantidad máxima de registros que se desean que el sistema recupere y también está el botón que realiza la evaluación de la búsqueda y muestra en la tabla de la zona 3, los valores de precisión para cada valor de recuerdo.

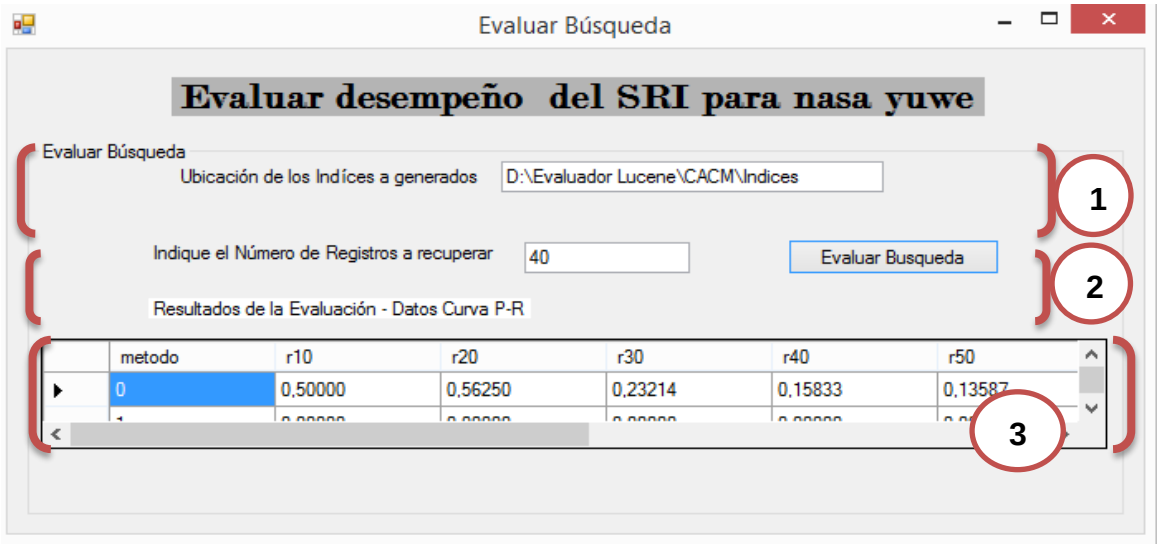


Figura 37. Ventana evaluación del desempeño. Fuente: Elaboración propia

En la Figura 38, se puede apreciar la recuperación de documentos escritos en nasa yuwe que realiza el sistema. En la zona 1 se muestra en donde se encuentran almacenados la estructura de índices generados sobre los cuales se realizan las consultas, seguidamente se encuentra un espacio en donde se digita las palabras de la consulta y el botón para iniciar el proceso. En la zona 2, se presentan los resultados de la búsqueda, se divide en dos partes, en la parte izquierda se encuentra la lista de documentos recuperados mostrada mediante el nombre del archivo y las primeras líneas que contiene el documento

de texto de la colección de prueba recuperado, allí se puede seleccionar el documento que sea de interés y en la parte derecha de la zona 2, se encuentra un botón mediante el cual se puede visualizar el documento seleccionado en la lista de la parte izquierda.

Figura 38. Ventana consultas. Fuente: Elaboración propia

4.4 FASE DE PRODUCCIÓN

Se hicieron algunos ajustes al prototipo en relación con las impresiones que se obtuvieron de algunos usuarios, varias de estas observaciones se tomarán en cuenta para una versión operativa del software.

La experiencia de producción se realizó con las personas de la comunidad nasa que participaron como expertos para la emisión de juicios, dado su conocimiento del contenido de los documentos de la colección de prueba.

A los expertos se les pidió interactuar con el sistema en la opción de recuperación del sistema, obteniendo buenas impresiones en lo relacionado con la facilidad de uso y la eficacia. Adicionalmente, se pudo apreciar que varios aspectos deben ser transparentes para el usuario como la parte de los índices y el botón de evaluación de desempeño, es decir, estos aspectos se deben quitar o limitar el acceso para cierto tipo de usuarios como el administrador o un usuario interesado en estos resultados.

4.5 FASE DE MANTENIMIENTO

Se realizarán más pruebas con los profesores con el fin de formalizar requisitos que se deben incluir en la versión operativa del prototipo. .Por tanto, esta fase no se aplica para efectos del desarrollo de este prototipo.

4.6 FASE DE MUERTE DEL PROYECTO

Para efectos de este proyecto, se dio por concluido el ciclo de desarrollo del prototipo de recuperación de información de textos escritos en nasa yuwe, pero como se mencionó anteriormente, hacen falta algunos ajustes para pasar de prototipo a una versión operacional, entre los que se encuentran: hacer una versión WEB con login para dos tipos de usuario (profesor nasa o hablante nasa y experto), mejorar las interfaces de acuerdo a las sugerencias recibidas por los usuarios de prueba.

5. CONCLUSIONES Y TRABAJO FUTURO

5.1 CONCLUSIONES

La pregunta planteada al inicio de este trabajo: ¿**Cómo adaptar las fases del proceso de recuperación de información a textos escritos en nasa yuwe?,** fue resuelta satisfactoriamente, como se puede apreciar en cada uno de las secciones de este documento, indicando: 1) aspectos metodológicos utilizados y seguidos para su resolución, 2) resultados como son las fases de tokenización y remoción de palabras vacías adaptadas específicamente para la lengua nasa yuwe y 3) evaluaciones de rendimiento que permitieron identificar las diferentes líneas base y su correspondiente contraste con los resultados obtenidos en cada sección.

A continuación se presentan cada una de las conclusiones obtenidas después de la realización de este trabajo.

5.1.1 Sobre la Construcción de la Colección de Prueba

La construcción de una colección de prueba para el nasa yuwe, representa un punto de partida muy importante para el desarrollo de posteriores experimentos en recuperación de información en nasa.

Es la primera colección de prueba desarrollada para la evaluación de sistemas de recuperación de información de textos escritos en nasa yuwe. Se presentó de manera detallada el proceso de construcción, la colección, y algunas estadísticas generales para propósitos de la recuperación de textos y procesamiento de lenguaje, al mismo tiempo que permiten compararla con colecciones de otras lenguas en situación sociolingüística similar a la del nasa yuwe.

Adicionalmente, se presenta un completo referente sobre otras colecciones de prueba existentes que permiten comparar la importancia de este trabajo en términos de construir una colección de prueba adecuada para la lengua nasa yuwe.

El desarrollo de esta experiencia se espera motive la escritura en nasa yuwe y la interacción de las personas de la comunidad nasa que participan en este proceso.

El proceso de construcción de la colección de prueba y los resultados obtenidos fueron publicados en un evento indexado como A2 por Colciencias (Ver Anexo 6 con la información detallada de la publicación y el artículo se puede ver en el archivo Anexo 6. Articulo colección publicado).

La construcción de la colección para textos escritos en nasa yuwe tomó más tiempo del esperado, específicamente en: 1) la selección y revisión de los textos de la colección y 2) en la emisión de juicios por parte de los expertos.

5.1.2 Sobre la Adaptación del tokenizador y la definición de la lista de palabras vacías.

En referencia a la adaptación del tokenizador estándar de Lucene para textos escritos en nasa yuwe, se pudo realizar la tokenización de manera exitosa, realizando un correcto procesamiento de los tokens nasa.

A pesar de que el nasa yuwe cumple la Ley de Zipf, al momento de definir la lista de palabras vacías identificada para esta lengua, no se obtuvo el resultado esperado a nivel de mejorar el desempeño en los valores de precisión de la curva y a nivel de la reducción del tamaño de la estructura del índice, es decir, mientras en la mayoría de las lenguas se pueden obtener reducciones hasta del 40%, en el tamaño en nasa yuwe fue cercana al 3% solamente.

5.1.3 Sobre la Construcción del prototipo

El prototipo permitió la evaluación del desempeño del SRI para nasa yuwe utilizando diferentes parámetros como tokenizador (estándar o para nasa yuwe), aplicación del convertidor de textos a minúsculas (si o no) y utilización de la lista de palabras vacías definida para nasa yuwe (si o no).

5.1.4 Conclusiones generales

Se logró proponer un esquema de representación y búsqueda para textos escritos en nasa yuwe, utilizando el modelo de espacio vectorial y haciendo la adaptación de un tokenizador para nasa y la definición de una lista de palabras vacías para nasa yuwe. Este esquema fue materializado en un prototipo de sistema de recuperación de información para textos escritos en nasa yuwe.

La experiencia obtenida de la realización de este trabajo es muy valiosa, dado que no existen antecedentes similares para nasa yuwe en referencia tanto a la colección de prueba, y de la adaptación de las tareas de análisis léxico (tokenizer) y lista de palabras vacías (stopwords removal list) para el prototipo de SRI, por tanto, se convierte en un referente para futuros trabajos en esta lengua o en otra de similar origen.

5.2 TRABAJO FUTURO

Es necesario ampliar y mejorar varios aspectos que favorezcan la calidad de la colección de prueba construida, en lo relacionado con incrementar la cantidad de documentos de la colección e incluir nuevos juicios de relevancia de más profesores hablantes de la lengua.

El prototipo de SRI debe pasar por más pruebas y verificaciones antes de ser convertido en una herramienta de uso masivo en los profesores participantes de este proyecto, es decir, habrá que revisar diferentes opciones entre las que se encuentran una versión web.

Para mejorar el desempeño del prototipo realizado es necesario adaptar otras tareas de procesamiento como un stemmer para nasa yuwe, dado que esto podría mejorar la similitud entre los términos de los documentos de la colección y la consulta.

Posteriormente, se plantea la construcción de un identificador de partes del discurso para nasa yuwe, utilizando algoritmos heurísticos para su desarrollo, esto facilitará la

construcción de actividades de enseñanza más complejas que apoyarán la visualización de la lengua a través de herramientas computacionales. Como parte de esta exploración para dar continuidad a este trabajo se publicó un artículo en donde se presenta una mejora del algoritmo Evolución, la cual he llamado EDKMeans, algoritmo evolutivo que demuestra buenos resultados y se utilizará como base para la construcción del Identificador de Partes del Discurso. La información detallada de esta publicación A2, se encuentra en el Anexo 7 y el archivo Anexo 7. Articulo Evolucion Diferencial EDKMeans muestra la versión publicada del artículo.

REFERENCIAS

[1] R. Sproat, «Linguistic Processing for Speech Synthesis,» de *Springer Handbook of Speech Processing*, J. Benesty, M. Sondhi y Y. Huang, Edits., USA, Springer Berlin Heidelberg, 2008, pp. 457-470.

[2] C. Peters, M. Braschler y P. Clough, «Within-Language Information Retrieval,» de *Multilingual Information Retrieval*, Springer Berlin Heidelberg, 2012, pp. 17-55.

[3] C. Manning , P. Raghavan y H. Shütze, An Introduction to Information Retrieval, Cambridge University Press, 2009.

[4] X. Pachón C., «Los Nasa o la Gente Páez,» de *Geografía Humana de Colombia. Región Andina Central Tomo IV*, vol. 2, Bogotá, 1996.

[5] B. Ribeiro-Neto y R. Baeza-Yates, Modern Information Retrieval -the concepts and technology behind search, Second ed., Harlow: Addison Wesley, 2011.

[6] Lucene Apache, «Lucene Apache,» 2009. [En línea]. Available: https://lucene.apache.org/core/3_6_2/api/core/org/apache/lucene/search/Similarity.html. [Último acceso: Julio 2009].

[7] Springer Science+Business Media, «F-Measure,» de *Encyclopedia of Machine Learning*, USA, Springer US, 2010, p. 416.

[8] T. E. Rojas, «Esbozo Gramatical de la lengua nasa (lengua Paéz). Atlas Sociolingüístico de pueblos indígenas en América Latina. Tomo I.,» UNICEF, Ed., 2009.

[9] Universidad del Cauca, CRIC-PEBIl-Comisión General de Lenguas, «Estudio Sociolingüistico Fase preliminar. Base de datos - CRIC 01/2007 Lengua Nasa Yuwe y Namtrik. Popayàn, Cauca, Colombia,» CRIC, Popayán - Colombia, 2008.

[10] T. Rojas Curieux, Por los caminos de la recuperación de la lengua Paéz (nasa yuwe), Popayán: Letrarte editores, 2006.

[11] R. Baeza-Yates y B. Ribeiro-Neto, Modern Information Retrieval, New York: Pearson - Addison Wesley, 1999.

[12] N. Craswell, «R-Precision,» de *Encyclopedia of Database Systems*, USA, Springer US, 2009, pp. 2453-2453.

[13] E. Zhang y Y. Zhang, «Average Precision,» de *Encyclopedia of Database Systems*, USA, Springer Us, 2009, pp. 192-193.

[14] E. Zhang y Y. Zhang, «Recall,» de *Encyclopedia of Database Systems*, USA, Springer US, 2009, pp. 2348-2348.

[15] Asamblea Nacional Constituyente, República de Colombia, «Banco de la República,» 1991. [En línea]. Available: http://www.banrep.gov.co/regimen/resoluciones/cp91.pdf. [Último acceso: 18 Octubre 2012].

[16] T. Rojas , L. M. Sierra , E. Meza y J. Villegas , «INTEGRACIÓN METODOLÓGICA PARA EL DESARROLLO DE RECURSOS EDUCATIVOS INFORMÁTICOS PARA APOYAR LA ENSEÑANZA DEL NASA YUWE,» *Gerenc. Tecnol. Inform.,* vol. Vol. 12, nº N° 32 | Ene - Abr |, pp. pp 45 - 60, 2013.

[17] Congreso de Colombia, «Ministerio de Educación,» 8 Febrero 1994. [En línea]. Available: http://www.mineducacion.gov.co/1621/articles-85906_archivo_pdf.pdf. [Último acceso: 18 Octubre 2012].

[18] Presidencia de Colombia, «Ministerio de Educación,» 18 Mayo 1995. [En línea]. Available: http://www.mineducacion.gov.co/1621/articles-86228_archivo_pdf.pdf. [Último acceso: 18 Octubre 2012].

[19] Consejo Regional Indígena del Cauca - CRIC, «Consejo Regional Indígena del Cauca - CRIC,» [En línea]. Available: http://www.cric-colombia.org/portal/estructura-organizativa/plataforma-de-lucha/. [Último acceso: 18 Octubre 2012].

[20] Consejo Regional Indígena del Cauca - CRIC y P. Programa de Educación Bilingüe e Intercultural -, «Sistema Educativo Indígena Propio -SEIP. Primer Documento de Trabajo,» CRIC , 2011.

[21] Organización Nacional Indígena de Colombia - ONIC, «Organización Nacional Indígena de Colombia - ONIC,» Abril 2012. [En línea]. Available: http://cms.onic.org.co/2012/04/es-urgente-aprobar-declaracion-americana-sobre-los-derechos-de-los-pueblos-indigenas/. [Último acceso: 21 Octubre 2012].

[22] L. M. Sierra Martínez, T. Rojas Curieux y R. C. Naranjo, Ewa: Comunidad Virtual de Apoyo a los Procesos de Etnoeducación Nasa, Popayán: Sello Editorial Universidad del Cauca, 2010.

[23] L. M. Sierra, T. Rojas, E. Meza y R. Naranjo, Metodología para Construir Materiales Educativos que Soporten la Enseñanza del Nasa Yuwe, Popayán: Proyecto Financiado por Colciencias y la Universidad del Cauca, 2013.

[24] T. Rojas Curieux, R. Naranjo Cuervo, L. M. Sierra Martínez, L. Besacier y E. Marsico , «Proyecto Investigación: Desarrollo de Herramientas Informáticas para la Revitalización de Lenguas en Peligro del Suroccidente Colombiano,» Unicauca - Colciencias- LIG- ICETEX, Popayán (Colombia) - Grenoble, Lyon (Francia), 2010.

[25] T. Rojas, L. M. Sierra , E. Diaz, G. Gonzalez y R. Naranjo, «Configuración de un corpus en Nasa Yuwe y Nam Trik para el reconocimiento automático de señales,» Universidad del Cauca - VRI- Proyecto de Investigación en Ejecución., Popayán , 2011.

[26] R. Baeza-Yates, «Challenges in the Interaction of Information Retrieval and Natural Language Processing,» de *Computational Linguistics and Intelligent Text Processing*, vol. Volume 2945, Springer Berlin Heidelberg, 2004, pp. 445-456.

[27] T. E. Rojas Curieux, La lengua paéz una visión de su gramática, primera ed., M. d. Cultura, Ed., Bogotá: Panamericana Formas e Impresos S.A., 1998.

[28] T. E. Rojas, «Esbozo Gramatical de la lengua nasa (lengua Paéz).,» de *El Lenguaje en Colombia. Tomo I: Realidad Lingüístidca de Colombia*, UNICEF, Ed., Bogotá, Academia Colombiana de la Lengua e Instituto Caro y Cuervo, 2012.

[29] K. S. Pratt, «Design Patterns for Research Methods: Iterative Field Research,» 2009. [En línea]. Available: http://www.kpratt.net/wp-content/uploads/2009/01/research_methods.pdf. [Último acceso: Octubre 2013].

[30] C. Sabino, El proceso de investigación, Caracas: Panapo, 1992.

[31] D. Wells, «Extreme Programming: A gentle introduction,» 28 Septiembre 2009. [En línea]. Available: http://www.extremeprogramming.org/. [Último acceso: 28 Diciembre 2012].

[32] Tangient LLC., «Entorno virtual de aprendizaje,» 2013. [En línea]. Available: http://programacion-extrema.wikispaces.com/. [Último acceso: 27 Enero 2013].

[33] E. Astigarraga, «Prospectiva,» 21 Octubre 2008. [En línea]. Available: http://www.prospectiva.eu/zaharra/Metodo_delphi.pdf. [Último acceso: 27 Marzo 2013].

[34] Project Management Institute, A guide to the Project Management Body of Knowledge, 5 ed., Project Management Institute, 2013, p. 589.

[35] C. Wohlin, M. C. Ohlsson, P. Runeson, M. Höst, B. Regnell y A. Wesslén, Experimentation in Software Engineering, New York: Springer, 2012.

[36] C. Manning, P. Raghavan y H. Shütze, An Introduction to Information Retrieval, Cambridge University Press, 2009.

[37] L. S. Larkey, L. Ballesteros y M. E. Connell, «Light Stemming for Arabic Information Retrieval,» de *Arabic Computational Morphology Text, Speech and Language Technology*, Springer, 2007, pp. 221-243.

[38] G. H. Tolosa y F. Bordignon, Introducción a la Recuperación de Información, Argentina: Universidad Nacional de Luján, 2005.

[39] V. A. Yatsko, «Methods and algorithms for automatic text analysis,» *Automatic Documentation and Mathematical Linguistics,* vol. 45, nº 5, pp. 224-231, 2011.

[40] S. Klatt y B. Bohnet, «You Don't Have to Think Twice if You Carefully Tokenize,» de *Natural Language Processing – IJCNLP 2004 - Lecture Notes in Computer Science*, vol. 3248, Hainan Island, China: Springer Berlin Heidelberg, 2004, pp. 299-309.

[41] N. Jamil, N. A. Jamaludin, N. Abdul Rahman y N. Sabari, «Implementation of Vector-Space Online Document Retrieval System Using Open Source Technology,» de *IEEE Conference on Open Systems (ICOS)*, Langkawi, 2011.

[42] C. Peters, M. Braschler y P. Clough, «Chapter 5 Evaluation for Multilingual Information Retrieval Systems,» de *Multilingual Information Retrieval*, Springer-Verlag Berlin Heidelberg, 2012, pp. 129-169.

[43] NTCIR Project, «NTCIR Project,» 2007. [En línea]. Available: http://research.nii.ac.jp/ntcir/permission/ntcir-4/perm-en-PATENT.html. [Último acceso: 5 12 2014].

[44] K. Esmaili, H. Abolhassani, M. Neshati, E. Behrangi y other, «Mahak: A Test Collection for Evaluation of Farsi Information Retrieval Systems,» de *IEEE/ACS International Conference on Computer Systems and Applications, 2007. AICCSA '07*, IEEE, 2007, pp. 639 - 644.

[45] K. Kuriyama, N. Kando, T. Nozue y K. Eguchi, «Pooling for a Large-Scale Test Collection: An Analysis of the Search Results from the First NTCIR Workshop,» *Information Retrieval,* vol. 5, nº 1, pp. 41-59, 2002.

[46] A. Fabre, «Diccionario etnolingüístico y guía bibliográfica de los pueblos indígenas sudamericanos. NASA,» 2005. [En línea]. Available: http://butler.cc.tut.fi/~fabre/BookInternetVersio/Dic=Nasa.pdf. [Último acceso: 9 Abril 2012].

[47] Instituto Colombiano de Cultura Hispánica, Geografía Humana de Colombia, Región Andina Central Tomo IV Volumen II, Bogotá: Banco de la República, 2000.

[48] M. Farfán Martínez y T. Rojas Curieux, Zuy Luuçxkwe kwe'kwe'sx ipx kwetuy piyaaka. Cartilla de aprendizaje de nasa yuwe como segunda lengua., Buenos Aires, 2010.

[49] I. Jung, Gramática del Páez o nasa yuwe. Descripción de una Lengua Indígena de Colombia., Published by LINOM GmbH 2008., 1984.

[50] CRIC y el Programa de Dllo Rural en la Región de Tierra Dentro Cxhab Wala - PT/CW, Diccionario Nasa Yuwe - Castellano, Primera ed., Popayán: Litografía San José, 2005.

[51] T. C. Rojas C., A. Perdomo Dizú y M. H. Corrales Carvaja, Una Mirada al nasa yuwe de Novirao, Primera ed., Popayán: Sello Editorial Universidad del Cauca, 2009.

[52] T. E. Rojas C, «Klyum o el polo conceptual que evade todo intento de aprehensión,» de *El léxico del cuerpo humano a través de la gramática y la semántica*, Bogotá, Universidad de los Andes, 1999, p. 59.

[53] K. Tan , Lucene y Solr consultant, «Lucene tutorial.com,» 2015. [En línea]. Available: http://www.lucenetutorial.com/lucene-in-5-minutes.html. [Último acceso: Julio 2015].

[54] Y. Cui, Y. Chen y J. Li, «Research of Information Search Engine in Forestry Based on the Lucene,» de *Advances in Automation and Robotics, Vol. 2. Series Lecture Notes in Electrical Engineering*, Volume 123 of the s pp, Springer Berlin Heidelberg, 2011, pp. 603-609.

[55] B. Carterette y E. M. Voorhees, «Overview of Information Retrieval Evaluation,» de *CURRENT CHALLENGES IN PATENT INFORMATION RETRIEVAL*, Berlin, Springer-Verlag Berlin Hiedelberg, 2011.

[56] K. Sheykh Esmaili, S. Salavati y S. Yosefi, «Building A Test Collection For Sorani Kurdish,» de *ACS International Conference on Computer Systems and Applications (AICCSA)*, Ifrane, 2013.

[57] M. Agosti, M. Bacchin, N. Ferro y M. Melucci, «Improving the Automatic Retrieval of Text Documents,» de *Advances in Cross-Language Information Retrieval*, vol. Volume 2785, C. Peters, M. Braschler y J. Gonzalo, Edits., Springer Berlin Heidelberg, 2003, pp. 279-290.

[58] J. Armenska, A. Tomovski y K. Zd, «Information Retrieval Using a Macedonian Test Collection for Question Answering,» de *ICT Innovations 2010*, Springer Berlin Heidelberg, 2010, pp. 205-214.

[59] A. AleAhmad, H. Amiri, E. Darrudi, M. Rahgozar y F. Oroumchian, «Hamshahri: A standard Persian text collection,» *Knowledge-Based Systems,* vol. 22, nº 5, pp. 382-387, 2009.

[60] L. Wang, «Relevance Weighting of Multi-Term Queries for Vector Space Model,» de *IEEE Symposium on Computational Intelligence and Data Mining*, Nashville, TN, 2009.

[61] A. Karshenas y K. Dimililer, «PIRS: An Information Retrieval System based on the Vector Space Model,» de *23rd International Symposium on Computer and Information Sciences, 2008. ISCIS '08*, Istanbul, 2008.

[62] D. Lee, K. Seamons y . H. Chuang, «Document Ranking and the Vector-Space Model,» *Software, IEEE,* vol. 14, nº 2, pp. 67 - 75, 1997.

[63] B. Hammo , S. Yagi, O. Ismail y M. AbuShariah, «Exploring and exploiting a historical corpus for Arabic,» *Language Resources and Evaluation,* pp. 1-23, 2015.

[64] W. Guan y P. Zhang, «Research and application of news-text similarity algorithm based on Chinese word segmentation,» de *3rd International Conference on Consumer Electronics, Communications and Networks (CECNet)*, Xianning, 2013.

[65] J. Jiang y C. Zhai, «An empirical study of tokenization strategies,» *Information Retrieval,* vol. 10, nº 4, pp. 341-363, October 2007.

[66] H. Joshi, J. Pareek, R. Patel y K. Chauhan, «To Stop Or Not to Stop - Experiments on Stopword Elimination for Information Retrieval of Gujarati Text Documents,» de *Nirma University International Conference on Engineering (NUiCONE)*, Ahmedabad, 2012.

[67] A. N. K. Zaman, P. Matsakis y C. Brown, «Evaluation of Stop Word Lists in Text Retrieval Using latent Semantic Indexing,» de *Sixth International Conference on Digital Information Management (ICDIM)*, Melbourn, QLD, 2011.

[68] A. Kumar Pandey y T. J. Siddiqui, «Evaluating Effect of Stemming and Stop-word Removal on Hindi Text Retrieval,» de *Proceedings of the First International Conference on Intelligent Human Computer Interaction*, Allahabad, India, Springer India, 2009, pp. 316-326.

[69] C. Zhang y S. Zhan, «Research and Implementation of Full-Text Retrieval,» de *Proceedings of the 2012 International Conference on Communication, Electronics and Automation Engineering*, vol. 181, Springer Berlin Heidelberg, 2012, pp. of the series Advances in Intelligent Systems and Computing pp 349-356.

[70] H. Li, W. Li, G. Wang y X. Peng , «Information Retrieval Services Based on Lucene Architecture,» de *Information Computing and Applications Volume 307 of the series Communications in Computer and Information Science*, China, Springer Berlin Heidelberg, 2012, pp. 638-645.

[71] M. Ortiz y B. Borjas, «La Investigación Acción Participativa: aporte de Fals Borda a la educación popular,» *Espacio Abierto Cuaderno Venezolano de Sociología,* vol. 17, nº 4, pp. 615-627, 2008.

[72] Consejo Regional Indígena del Cauca – Programa de Educación Bilingüe e Intercultural (PEBI - CRIC), Cuentos y Cosmovisión Nasa. Area Nasawe'sx Fxinzenxi, Segunda ed., Popayán, 2010.

[73] Consejo Regional Indígena del Cauca – Programa de Educación Bilingüe e Intercultural (PEBI - CRIC), Te invitamos a leer. Eç thegya' ipi'ki' tha'w, Primera ed., Cali: Grafitextos, 2007.

[74] Asociación de Cabildos Ukawe'sx Nasa Çxhab, Consejo Regional Indígena del Cauca – Programa de Educación Bilingüe e Intercultural (PEBI - CRIC), NASAWE'SX KIWAKA FXI'ZENXI ẼEN, Primera ed., Cali: Grafitextos, 2006.

[75] M. Yule Yatacue y C. Vitonas Pavi, Pees kupx fxi'zenxi. La metamorfosis de la vida, Tercera ed., Toribio, Cauca: Grafitextos, 2012.

[76] C. Cobos , H. Ordoñez, L. Krug-Wives y L. Thom, «Collaborative Evaluation to Build Closed Repositories on Business Process Models,» de *16th International Conference on Enterprise Information Systems*, Lisboa, Portugal, 2014.

[77] E. W. Narváez Burbano y M. F. García, Monografía y Anexos: Alternativa para la entrada de caracteres en lenga nasa yuwe aplicada a la producción de materiales tipo texto, Popayán: Tesis Ingeniería de Sistemas Universidad del Cauca, 2014.

[78] Microsoft, «Microsoft,» 2010. [En línea]. Available: https://www.google.com.co/search?q=arquitectura+de+tres+capas+microsoft&oq=a&aqs=chrome.5.69i60l4j69i57j69i59.11903j0j4&sourceid=chrome&espv=210&es_sm=93&ie=UTF-8. [Último acceso: Octubre 2013].

[79] L. M. Sierra Martínez, C. A. Cobos Lozada, J. C. Corrales y T. Rojas Curieux, «Building a nasa yuwe Test Collection,» de *Computational Linguistics and Intelligent Text Processing Volume 9041 of the series Lecture Notes in Computer Science*, El Cairo, Egipto, Springer International Publishing, 2015, pp. 112-123.

[80] L. M. Sierra, C. Cobos, J. C. Corrales y T. Rojas, «: "Continuos Optimization Based on a Hybridization of Differential Evolution with K-means",» de *IBERAMIA 2014*, Santiago de Chile, 2014.

[81] T. E. Rojas, «Esbozo Gramatical de la lengua nasa (lengua Paéz). Atlas Sociolingüístico de pueblos indígenas en América Latina. Tomo I.,» de *El lenguaje en Colombia Tomo I*, UNICEF, Ed., Bogotá, Academía Colombiana de la Lengua e Instituto Caro y Cuervo, 2009, pp. 479-495.

[82] A. P. Ibarra Quiroga, J. C. Mosquera Ramirez y R. F. Zuñiga Muñoz, Proyecto Apoyo Multimedial Indigena "AMI" Propuesta Metodologica para la Construccion de Software Etnoeducativo, U. C. d. Colombia, Ed., Popayán: Universidad Cooperativa de Colombia, 2004.

[83] S. J. Ruano Rincón y Á. C. Checa Hurtado, «Lineamientos para la adecuación de IGUs en el ámbito de la cultura indígena Páez,» *20vo Simposio de Factores Humanos en Telecomunicaciones,* 2006.

[84] J. A. Villegas, E. Solarte y L. M. Sierra, Material Etnoeducativo Informático Tipo Micromundo Para El Apoyo De La Enseñanza Del Nasa-Yuwe. Tesis de Pregrado Monografía y Anexos, Popayán: Universidad del Cauca, 2013.

[85] Real Academia Española, Nueva gramática básica de la lengua española, Primera ed., Bogotá: Editorial Planeta Colombiana S.A., 2011.