GENERACIÓN AUTOMÁTICA DE RESÚMENES DE MÚLTIPLES DOCUMENTOS BASADA EN EL ALGORITMO GHS+LEM

ANEXOS

WILLIAN ANDRES TAMAYO MONJE
MELISSA LYNNETTE VELA CORAL

Director: Dr. (c) MARTHA ELIANA MENDOZA BECERRA

UNIVERSIDAD DEL CAUCA
FACULTAD DE INGENIERÍA ELECTRÓNICA Y TELECOMUNICACIONES
DEPARTAMENTO DE SISTEMAS
GRUPO DE I+D EN TECNOLOGÍAS DE LA INFORMACIÓN
RECUPERACIÓN DE LA INFORMACIÓN
POPAYÁN, Abril 2012

Tabla de Contenido

LISTA DE TABLAS

LISTA DE FIGURAS

ANEXO A - METODOLOGÍA RECOPILACION DE REQUERIMIENTOS Y RESTRICCIONES DEL ALGORITMO

1 DESCRIPCIÓN GENERAL DE LA METODOLOGÍA

La metodología para la parte teórica e investigativa correspondiente a la recopilación de requerimientos y restricciones del algoritmo involucra 3 etapas: *Descripción del problema, Definición de la Arquitectura y Definición del entorno de evaluación.*

1.1 DESCRIPCIÓN DEL PROBLEMA

En esta etapa se recopilan los datos referentes a los requerimientos del problema planteado, se hacen los análisis de entradas y de salidas del algoritmo, se realiza una investigación para determinar si el algoritmo (HS) ya se encuentra implementado para generación automática de resúmenes, con el objetivo de evitar su implementación y se define el entorno[1] a utilizar. Como resultado de esta fase se obtendrá un documento con las especificaciones de los algoritmos a implementar.

1.1.1 Requerimientos del problema planteado

Para la recolección de requerimientos del problema planteado se realizó una amplia investigación sobre los trabajos que se han realizado alrededor del tema de generación automática de resúmenes y algoritmos de búsqueda armónica.
A continuación se presentan las fichas bibliográficas correspondientes a los núcleos temáticos definidos para este proyecto, donde los núcleos temáticos fundamentales son: generación automática de resúmenes mediante algoritmos evolutivos, generación automática de resúmenes de múltiples documentos, algoritmos de búsqueda armónica.

Ficha Bibliográfica	
Aspectos formales sobre el documento	
Autor: Ehsan Shareghi, Leila Sharif Hassanabadi Titulo: Text Summarization with Harmony Search Algorithm-Based Sentence Extraction (Generación de resúmenes con Extracción de Oraciones basada en Búsqueda Armónica) Tipo de material: Artículo Informativo	
Enfoque	
Disciplina	Recuperación y Almacenamiento de Información
Paradigma conceptual	Ingeniería del Conocimiento
Referentes teóricos	• Qazvinian, V., Sharif, L. and Halavati R., Summarization Text with a Genetic Algorithm-Based Sentence Extraction, 2008. • Lee K. S. and Geem Z. W., A new meta-heuristic algorithm for continuous engineering optimization: harmony search theory and practice, 2005. • Luhn, H., The automatic creation of literature abstracts. IBM Journal of Research and Development, 1958.

[1] Entorno: Conjunto de condiciones extrínsecas que necesita un sistema informático para funcionar, como el tipo de programación, de proceso, las características de las máquinas que lo componen.

Conceptos principales	• Procesamiento del lenguaje natural • Generación automática de resúmenes. • Algoritmos evolutivos. • Algoritmo de búsqueda armónica
Hipótesis	Un buen resumen debe contener 3 factores importantes: Relación con el tema, cohesión y legibilidad.
Tesis	A continuación se muestran las principales ideas expuestas en el artículo: En el artículo se proponen tres factores que son usados para la generación de un buen resumen: El factor de relación con el tema mide la similitud de las oraciones del resumen con el título. Con el factor de cohesión se trata de identificar si las oraciones del resumen tratan de la misma información. Y mediante el factor de legibilidad intentan determinar que tan comprensible es el resumen a través del cálculo de similitud de cada oración del resumen con la siguiente. La representación de los documentos para poder usar el algoritmo de búsqueda armónica se realiza a través de un mapeo uno a uno entre las oraciones del resumen candidato y el documento original. Esto para poder tener una representación vectorial de los documentos que permitiera hacer uso del algoritmo de búsqueda armónica. En el trabajo presentado hicieron uso del algoritmo de búsqueda armónica para la selección de las oraciones que conforman el resumen. La evaluación de la calidad del resumen generado por el sistema propuesto se realizó mediante el cálculo de precisión y recuerdo y con los documentos proporcionados por DUC2 2002 haciendo una comparación entre el resumen generado y los resúmenes ideales. Características El sistema permite realizar resúmenes de un documento, extractivo, es mono-lenguaje y tiene un nivel de procesamiento superficial. Ventajas Hacen uso de un algoritmo no supervisado, lo que le da al sistema una mayor flexibilidad. Desventajas Omiten el parámetro PAR del algoritmo original de búsqueda armónica. Perdiendo así la posible explotación que podría

[2] Document Understanding Conferences (http://duc.nist.gov/)

	ocurrir en la búsqueda sobre una posible área prometedora. Tomando en cuenta la anterior figura, podría describirse en términos generales, el funcionamiento de un algoritmo de generación automática de resúmenes basado en un algoritmo evolutivo.
Tipo de investigación	Exploratoria
Metodología	
Tipo de metodología	Tipo de Metodología: Mixta (Cualitativa y Cuantitativa)
Técnicas	Similitud cosenoidal Representación vectorial de los documentos ROUGE para la evaluación de la calidad de los resúmenes
Resultados	
Conclusiones	La generación automática de resúmenes a través de algoritmos evolutivos, específicamente con búsqueda armónica es un campo que requiere de mucha más exploración no solo en la forma de usar el algoritmo sino también en las características adicionales que permitan mejorar la calidad de los resúmenes basados en éste tipo de algoritmo como las etapas de pre-procesamiento, en la forma de representar los documentos, en el cálculo de similitud, en la creación de la función objetivo, etc. Las comparaciones realizadas con previos trabajos mediante el uso de los conjuntos de datos de DUC 2002 muestran que el método propuesto para obtener resúmenes en la mayoría de los casos es mucho mejor que los trabajos previos.
Recomendaciones	Podría lograrse un mejor desempeño del resumen cuando los coeficientes de la función objetivo sean establecidos de acuerdo al conocimiento de un experto y explorando otras técnicas en las etapas independientes del algoritmo de búsqueda armónica.
Análisis del documento	
Resumen del documento	Este artículo muestra un nuevo método de generación de resúmenes de un documento a través del algoritmo de búsqueda armónica cuya función objetivo se basa en tres factores (relación con el tema, cohesión y legibilidad) que determinan el resumen.
Palabras claves	Generación automática de resúmenes, Algoritmo evolutivo, Algoritmo de Búsqueda Armónica, Extracción de oraciones.

Comentarios	En base a la revisión de este artículo, cabe resaltar que éste es la base de nuestro trabajo dado que hacemos uso del algoritmo de búsqueda armónica más dos variaciones del mismo para la generación de nuestro resumen multi-documento.

Tabla 1. Ficha Bibliográfica - Generación de resúmenes con Extracción de Oraciones basada en Búsqueda Armónica

Ficha Bibliográfica	
Aspectos formales sobre el documento	
Autor: Rasim M. Alguliev, Ramiz M. Aliguliyev, Makrufa S. Hajirahimova, Chingiz A. Mehdiyev Título: MCMR: Maximum coverage and minimum redundant text summarization model. (Modelo de generación de resúmenes con máxima cobertura y mínima redundancia) Tipo de material: Artículo Informativo	
Enfoque	
Disciplina	Recuperación y Almacenamiento de Información
Paradigma conceptual	Ingeniería del Conocimiento
Referentes teóricos	• Carbonell, J., & Goldstein, J., The use of MMR, diversity-based reranking for reordering documents and producing summaries, 1998. • Binwahlan, M. S., Salim, N., Suanmali, L., Swarm based text summarization, 2009. • Radev, D. R., Blair-Goldensohn, S., & Zhang, Z., Experiments in single and multidocument summarization using MEAD. 2001. • Aliguliyev, R. M., Clustering techniques and discrete particle swarm optimization algorithm for multi-document summarization. 2010.
Conceptos principales	• Procesamiento del lenguaje natural • Generación automática de resúmenes de múltiples documentos. • Algoritmos evolutivos. • Algoritmo de optimización de enjambre de partículas. • Algoritmo de Ramificación y poda.
Hipótesis	A través del modelo propuesto, el sistema es capaz de descubrir las oraciones clave en un documento dado y cubrir el contenido principal del documento original, también garantiza que el resumen no contenga oraciones que tengan la misma información.
Tesis	A continuación se muestran las principales ideas expuestas en el artículo: En el artículo se propone un modelo de generación de resúmenes no supervisado que intenta optimizar tres propiedades: relevancia, redundancia y longitud.

	Realizan la representación de los documentos para poder aplicar los algoritmos utilizados para la selección de las oraciones relevantes y la posterior generación del resumen. Prueban dos técnicas diferentes para el cálculo de similitud: la medida cosenoidal y similitud basada en la distancia de google normalizada. Utilizan dos algoritmos de optimización para determinar cuál de los dos tiene un mejor comportamiento para realizar la selección de las oraciones del resumen. La evaluación de la calidad del resumen generado por el sistema propuesto se realizó mediante el cálculo de precisión y recuerdo con n-gramas haciendo uso de la herramienta ROUGE y con los documentos proporcionados por DUC 2005 y 2007 haciendo una comparación entre el resumen generado y los resúmenes ideales. Características 2. El sistema permite realizar resúmenes multi-documento, de tipo extractivo, es mono-lenguaje y tiene un nivel de procesamiento superficial. Ventajas 3. El enfoque utilizado permite que el modelo se adapte a generación de resúmenes tanto de un documento como de múltiples documentos. 4. Evaluaron dos técnicas de similitud y dos algoritmos distintos logrando así una mayor exploración en el comportamiento de los mismos lo que permitió obtener más posibilidades de encontrar una mejor solución en relación con trabajos previos. Desventajas 5. Requieren del cálculo de dos medidas de similitud distintas para poder obtener mejores resultados, esto hace que el algoritmo probablemente se tarde demasiado tiempo en la generación del resumen.
Tipo de investigación	Exploratoria
Metodología	
Tipo de metodología	Tipo de Metodología: Mixta (Cualitativa y Cuantitativa)
Técnicas	Similitud cosenoidal Similitud basada en Distancia de Google Normalizada Representación vectorial de los documentos ROUGE para la evaluación de la calidad de los resúmenes
Resultados	
Conclusiones	En el artículo se presenta un enfoque modelado como un problema de programación lineal entera cuyo objetivo es optimizar tres propiedades: relevancia, redundancia y longitud para la generación automática de resúmenes a través de

	algoritmos de optimización como partículas de enjambre y ramificación y poda. Las comparaciones realizadas con previos trabajos mediante el uso de los conjuntos de datos de DUC 2005 y 2007 y con las medidas de ROUGE-2 y ROUGE-SU4 muestran que el método propuesto mejora significativamente. También se demuestra que los resultados dependen de la medida de similitud, concluyendo que la combinación de la medida cosenoidal y basada en la distancia de google normalizada conduce a mejores resultados que usandolas de manera individual.
Recomendaciones	
Análisis del documento	
Resumen del documento	En este trabajo se propone un modelo de generación de resúmenes no supervisado el cual genera un resumen mediante la extracción de las oraciones sobresalientes en uno o varios documentos dados. El modelo puede directamente descubrir oraciones clave en un documento dado, cubrir el contenido principal del documento(s) original y garantiza que el resumen no contenga oraciones redundantes. Se basa en los algoritmos de: optimización de partículas de enjambre y ramificación y poda.
Palabras claves	Generación automática de resúmenes, máxima cobertura, menor redundancia, programación lineal entera, optimización de enjambre de partículas, ramificación y poda.
Anexos	
Comentarios	En base a la revisión de este artículo, cabe resaltar que nos basamos en él para la representación de múltiples documentos y también tomamos muy en cuenta la función objetivo definida dado que tuvieron muy buenos resultados.

Tabla 2. Ficha Bibliográfica - MCMR: Modelo de generación de resúmenes con máxima cobertura y mínima redundancia

Ficha Bibliográfica	
Aspectos formales sobre el documento	
Autor: Carbonell, J., Goldstein, J. Titulo: The use of MMR, diversity-based reranking for reordering documents and producing summaries (El uso de MMR, reclasificación basada en diversidad para reorganizar los documentos y producir resúmenes) Tipo de material: Artículo Informativo	
Enfoque	
Disciplina	Recuperación y Almacenamiento de Información
Paradigma conceptual	Ingeniería del Conocimiento
Referentes teóricos	• Julian, Kupiec, Jan, Pedersen, Francine, Chen, A trainable document summarizer, 1995. • Carbonell, J., Automated query-relevant summarization

	and diversity-based reranking, 1997.
Conceptos principales	• Generación de resúmenes de un documento basado en consultas. • Relevancia marginal máxima para reducir la redundancia y mantener la relevancia con la consulta.
Hipótesis	Un criterio potencial es la "novedad relevante" que se puede lograr a través de la aplicación de Relevancia Marginal Máxima que permite reducir la redundancia y conservar la relevancia con una consulta.
Tesis	A continuación se muestran las principales ideas expuestas en el artículo: Relevancia marginal máxima evita la redundancia y a la vez mantiene la relevancia hacia la consulta. La relevancia marginal realiza una combinación lineal en la que un documento con alta relevancia marginal es relevante para la consulta y contiene una mínima similitud con previos documentos seleccionados anteriormente. Para determinar la relevancia se calcula la similitud entre la consulta y el documento y para evitar la redundancia se compara la oración seleccionada con las oraciones previamente seleccionadas. Realizan la evaluación manualmente con documentos proporcionados en las conferencias de SUMMAC3 y a través del cálculo de precisión, recuerdo y medida-F. Características El sistema permite realizar resúmenes mono-documento, de tipo extractivo, es mono-lenguaje y tiene un nivel de procesamiento superficial. Ventajas La función de MMR propuesta tiene un coeficiente que permite darle más peso a la relevancia o a la redundancia lo que le da al sistema propuesto una mayor flexibilidad para que el usuario pueda darle mayor o menor peso a la característica deseada. Desventajas El método MMR proporciona un gran beneficio al usuario permitiendo minimizar la redundancia, sin embargo ésta característica aplica especialmente en la generación de resúmenes multi-documento dado que es menos probable encontrar redundancia de información entre las oraciones de un mismo documento.
Tipo de investigación	Exploratoria
Metodología	
Tipo de metodología	Tipo de Metodología: Mixta (Cualitativa y Cuantitativa)
Técnicas	

[3] http://www-nlpir.nist.gov/related_projects/tipster_summac/

Resultados	Los resultados muestran que el método de selección de oraciones para generación de resúmenes trabaja mejor para documentos largos, los cuales típicamente contienen oraciones redundantes a través de las distintas secciones del documento como el resumen, la introducción y la conclusión.
Conclusiones	MMR proporciona una manera útil de proporcionar información al usuario minimizando la redundancia, esto es verdadero especialmente en el caso de encontrar la relevancia de con una consulta para múltiples documentos.
Recomendaciones	Realización de estudios sobre cómo extender el método a colecciones de documentos para generación de resúmenes.
Análisis del documento	
Resumen del documento	El artículo propone un método para combinar la relevancia de una consulta con novedad de la información en el contexto de recuperación de texto y generación de resúmenes. El método es basado en el criterio de Relevancia Marginal Máxima cuyo objetivo es reducir la redundancia mientras se mantiene la relevancia con respecto a una consulta y seleccionar oraciones apropiadas para generación de resúmenes. Para la parte de generación de resúmenes, la oración con la mayor puntuación es agregada al resumen, antes de agregar cualquier otra oración relevante primero chequean si esta oración podría introducir redundancia.
Palabras claves	Generación automática de resúmenes, relevancia marginal máxima, extracción de oraciones.
Comentarios	

Tabla 3. Ficha Bibliográfica - El uso de MMR, reclasificación basada en diversidad para reorganizar los documentos y producir resúmenes

Ficha Bibliográfica	
Aspectos formales sobre el documento	
Autor: Carlos Cobos, Dario Estupiñán, José Pérez. Titulo: GHS + LEM: Global-best Harmony Search using learnable evolution models. (Mejor búsqueda armónica global usando modelos evolucionarios que aprenden) Tipo de material: Artículo Informativo	
Enfoque	
Disciplina	Gestión de la Información
Paradigma conceptual	Ingeniería del Conocimiento
Referentes teóricos	Z. Geem, J. Kim, G.V. Loganathan, A new heuristic optimization algorithm: harmony search, 2001. R.S. Michalski, LEARNABLE EVOLUTION MODEL: evolutionary processes guided by machine learning
Conceptos principales	El algoritmo de búsqueda armónica (HS) se basa en el proceso de improvisación musical, ha sido satisfactoriamente aplicado a muchos problemas de optimización y ha

	experimentado muchos cambios en combinación con otras técnicas de optimización de lo que se han obtenido variaciones del algoritmo como Mejor Búsqueda Armónica Global (GHS) y Búsqueda Armónica Mejorada (IHS). Un modelo evolutivo que aprende (LEM) es un modo de aprendizaje de máquina que puede determinar cuáles individuos en una población son superiores a otros en la realización de ciertas tareas.
Hipótesis	La adaptación de LEM a GHS logra mejorar la convergencia y la exactitud del algoritmo.
Tesis	A continuación se muestran las principales ideas expuestas en el artículo: Inspirados en el concepto de modelo evolutivo que aprende en este trabajo proponen una nueva variación del algoritmo GHS. El proceso de LEM lo resumen en los siguientes pasos: - Generar una población - Ejecutar el modo de aprendizaje de maquina - Ejecutar el modo de aprendizaje Darwiniano - Alternar entre los dos modos hasta que se llegue al criterio de parada. Para realizar el modo de aprendizaje de máquina usan una variación del algoritmo PRISM el cual toma como entrada un conjunto de entrenamiento con valores de atributos determinados por una clasificación a partir de los cuales se generan reglas que permiten la inferencia de nuevos candidatos en la población que emergen no solo de una exploración aleatoria. Características El algoritmo está realizado para trabajar con variables continuas.
Tipo de investigación	Exploratoria
Metodología	
Tipo de metodología	Tipo de Metodología: Mixta (Cualitativa y Cuantitativa)
Técnicas	
Resultados	Realizaron una afinación de parámetros a partir de la cual obtuvieron que los mejores resultados se obtienen con los siguientes valores: - HMCR >=0.9 favorece la convergencia. - HMS entre 5 y 10 - PAR dinámico - RCR entre 0.7 y 1, el mejor es 0.9 Los resultados obtenidos por GHS+LEM exceden la exactitud obtenida por HS, IHS y GHS en todas las funciones de optimización evaluadas. También se realizaron pruebas de escalabilidad. GHS+LEM

	mejora la exactitud en cada una de las funciones usadas, probando ser mejores que HS, IHS y GHS en condiciones de alta dimensionalidad.
Conclusiones	Después de investigar los efectos de los parámetros HMCR, HMS, PAR y RCR en el desempeño del algoritmo propuesto el resultado fue que el valor para HMCR mayor o igual que 0.9 el algoritmo generalmente mejora la eficiencia. Con respecto al tamaño de la memoria armónica, los resultados muestran que el algoritmo propuesto tiene un mejor desempeño cuando el tamaño es entre 5 y 10 y cuando el PAR es dinamico. Con respecto a la variación del parámetro RCR se logra un mejor desempeño cuando se usa una probabilidad entre 0.7 y 1. También se demuestra que el algoritmo tiene una mayor exactitud que otros cuando el número de iteraciones es 10 veces menor que el usado por otros algoritmos de búsqueda armónica. Siguiendo una prueba de escalabilidad, también concluyen que el algoritmo mantiene su exactitud aún cuando se tienen muchas dimensiones (igual o más de 30).
Recomendaciones	
Análisis del documento	
Resumen del documento	El trabajo presenta un nuevo algoritmo de optimización denominado GHS+LEM el cual se basa en el algoritmo de la mejor Búsqueda Armónica Global y técnicas de modelos evolutivos que aprenden para mejorar la convergencia y exactitud del algoritmo. El funcionamiento del algoritmo se evalúa con 15 funciones de optimización comúnmente usadas y los resultados obtenidos se comparan contra HS, IHS y GHS. La evaluación demuestra que el algoritmo propuesto mejora la exactitud de los resultados obtenidos por los otros algoritmos produciendo mejores resultados específicamente en problemas con alta dimensionalidad donde ofrece una convergencia con menor cantidad de iteraciones.
Palabras claves	Búsqueda Armónica, Meta-heurísticas, Algoritmos Evolucionarios, Optimización, modelos evolutivos que aprenden, Aprendizaje de Máquina, Prisma.
Comentarios	

Tabla 4. Ficha Bibliográfica - Mejor búsqueda armónica global usando modelos evolucionarios que aprenden

1.1.2 Especificaciones de los algoritmos a implementar

Se debe desarrollar una aplicación de escritorio la cual permita seleccionar los documentos a resumir, definir la longitud del resumen deseada, definir el algoritmo a ejecutar (HS, GHS, GHS+LEM), finalmente la aplicación debe entregar al usuario un resumen y una hoja de Excel donde estén los resultados de la evaluación del resumen.

1.1.2.1 Búsqueda Armónica (HS)

Entradas:

- **Documentos:** Los documentos para la ejecución de las pruebas no tienen extensión y el usuario puede ingresar un documento, un conjunto de documentos o un grupo de conjuntos de documentos que serán procesados para la generación del resumen. El contenido de un documento está sujeto a la estructura que se muestra en la Figura 1, la cual es la dispuesta por DUC.

- **Parámetros**: El usuario debe ajustar los parámetros del algoritmo que haya seleccionado, si así lo desea, de lo contrario se ejecutara con los parámetros que se afinaron en las pruebas de laboratorio. Los parámetros requeridos para el algoritmo son: Tamaño de la memoria armónica (HMS), Tasa de consideración de la memoria armónica (HMCR), Número de improvisaciones (NI).

- **Longitud del resumen deseado:** La longitud del resumen debe ser aproximadamente de 250 palabras para la realización de las pruebas, debido a que los resúmenes modelo proporcionados por DUC tienen estas longitudes, sin embargo el usuario tiene la posibilidad de realizar un resumen con una cantidad mayor o menor de palabras.

Salidas: La salida del sistema es el resumen generado por el algoritmo de Búsqueda Armónica. Adicional al resumen, se puede realizar una evaluación, en la cual se compara el resumen generado con los resúmenes ideales proporcionados por DUC. Para esto, el sistema después de haber realizado el resumen, debe leer los resúmenes modelo y realizar la evaluación por medio de ROUGE 1.5.5, hay que tener en cuenta que ROUGE nos puede arrojar el promedio de todos los resúmenes modelo o el mejor resultado. Finalmente se obtiene una hoja de Excel donde están todos los valores de ROUGE para cada uno de los conjuntos de documentos.

Condiciones:
Ejecución del algoritmo HS cuya función objetivo consta de los factores Cobertura (FCb) y Eliminación de Redundancia (FER) es:

$$f(x) = \beta * FCb + (1 - \beta) * FER$$

Donde, β, es un valor entre 0 y 1 .

Para que el algoritmo se ejecute correctamente, se deben haber introducido todos los parámetros requeridos: HMS, HMCR y NI.

1.1.2.2 Búsqueda Armónica Global (GHS)

Entradas: Documentos, parámetros, longitud del resumen deseado.

- **Documentos:** Los documentos para la ejecución de las pruebas no tienen extensión y el usuario puede ingresar un documento, un conjunto de

documentos o un grupo de conjuntos de documentos que serán procesados para la generación del resumen. El contenido de un documento está sujeto a la estructura que se muestra en la Figura 1, la cual es la dispuesta por DUC.

- **Parámetros**: El usuario debe ajustar los parámetros del algoritmo que haya seleccionado, si así lo desea, de lo contrario se ejecutara con los parámetros que se afinaron en las pruebas de laboratorio. Los parámetros requeridos para el algoritmo son: Tamaño de la memoria armónica (HMS), Tasa de consideración de la memoria armónica (HMCR), Número de improvisaciones (NI), PAR (Tasa de ajuste del tono).

- **Longitud del resumen deseado:** La longitud del resumen debe ser aproximadamente de 250 palabras para la realización de las pruebas, debido a que los resúmenes modelo proporcionados por DUC tienen estas longitudes, sin embargo el usuario tiene la posibilidad de realizar un resumen con una cantidad mayor o menor de palabras.

Salidas: La salida del sistema es el resumen generado por el algoritmo de Búsqueda Armónica Global. Adicional al resumen, se puede realizar una evaluación, en la cual se compara el resumen generado con los resúmenes ideales proporcionados por DUC. Para esto, el sistema después de haber realizado el resumen, debe leer los resúmenes modelo y realizar la evaluación por medio de ROUGE 1.5.5, hay que tener en cuenta que ROUGE nos puede arrojar el promedio de todos los resúmenes modelo o el mejor resultado. Finalmente se obtiene una hoja de Excel donde están todos los valores de ROUGE para cada uno de los conjuntos de documentos.

Condiciones: Ejecución del algoritmo GHS cuya función objetivo consta de los factores Cobertura (FCb) y Redundancia(FR) es:

$$f(x) = \beta * FCb + (1 - \beta) * FR$$

Donde, β, es un valor entre 0 y 1 .

Para que el algoritmo se ejecute correctamente, se deben haber introducido todos los parámetros requeridos: HMS, HMCR, PAR y NI.

1.1.2.3 Búsqueda Armónica Global Con Modelos Evolutivos Que Aprenden (GHS+LEM)

Entradas: Documentos, parámetros, longitud del resumen deseado.

- **Documentos:** Los documentos para la ejecución de las pruebas no tienen extensión y el usuario puede ingresar un documento, un conjunto de documentos o un grupo de conjuntos de documentos que serán procesados

para la generación del resumen. El contenido de un documento está sujeto a la estructura que se muestra en la Figura 1, la cual es la dispuesta por DUC.

- **Parámetros**: El usuario debe ajustar los parámetros del algoritmo que haya seleccionado, si así lo desea, de lo contrario se ejecutara con los parámetros que se afinaron en las pruebas de laboratorio. Los parámetros requeridos para el algoritmo son: Tamaño de la memoria armónica (HMS), Tasa de consideración de la memoria armónica (HMCR), Número de improvisaciones (NI), PAR (Tasa de ajuste del tono), RCR(Tasa de consideración de reglas), RRU (Tasa de actualización de reglas), HLGS (el tamaño de los grupos de alto y bajo rendimiento).

- **Longitud del resumen deseado:** La longitud del resumen debe ser aproximadamente de 250 palabras para la realización de las pruebas, debido a que los resúmenes modelo proporcionados por DUC tienen estas longitudes, sin embargo el usuario tiene la posibilidad de realizar un resumen con una cantidad mayor o menor de palabras.

Salidas: La salida del sistema es el resumen generado por el algoritmo de Búsqueda Armónica Global. Adicional al resumen, se puede realizar una evaluación, en la cual se compara el resumen generado con los resúmenes ideales proporcionados por DUC. Para esto, el sistema después de haber realizado el resumen, debe leer los resúmenes modelo y realizar la evaluación por medio de ROUGE 1.5.5, hay que tener en cuenta que ROUGE nos puede arrojar el promedio de todos los resúmenes modelo o el mejor resultado. Finalmente se obtiene una hoja de Excel donde están todos los valores de ROUGE para cada uno de los conjuntos de documentos.

Condiciones:

- Ejecución del algoritmo GHS+LEM cuya función objetivo consta de los factores Cobertura (FCb) y Eliminación de Redundancia(FR) es:

$$f(x) = \beta * FCb + (1 - \beta) * FER$$

Donde, β, es un valor entre 0 y 1 .

Para que el algoritmo se ejecute correctamente, se deben haber introducido todos los parámetros requeridos: HMS, HMCR, PAR, RCR, RRU, HLGS y NI.

- Lista de reglas para almacenar la probabilidad de ceros y unos:

$$\mathrm{P}_{\mathrm{s_i}}(1) = \begin{cases} 50\% + (\mathrm{F1A}_{\mathrm{s_i}} - \mathrm{F1B}_{\mathrm{s_i}})/\mathrm{TTG}_{\mathrm{s_i}} & \mathrm{F1A}_{\mathrm{s_i}} - \mathrm{F0A}_{\mathrm{s_i}} \geq 0 \\ 50\% - (\mathrm{F0A}_{\mathrm{s_i}} - \mathrm{F0B}_{\mathrm{s_i}})/\mathrm{TTG} & \mathrm{d.o.m.} \end{cases}$$

Donde *F1A* es la frecuencia de unos en altos, *F0A* frecuencia de ceros en altos, *F1B* frecuencia de unos en bajos, *F0B* frecuencia de ceros en bajos.

```
<DOC>
        <HEADLINE>
        Titulo documento
        </HEADLINE>
        <TEXT>
                <P>
                Texto del documento
                </P>
        </TEXT>
</DOC>
```

Figura 1. Estructura de los documentos de evaluación

1.1.2.4 Algoritmo De Afinación de Parámetros (GHSO, Global-best Harmony Search Optimization)

Entradas: los valores establecidos para ejecución del algoritmo GHSO son los siguientes.

HMS = 10, HMCR = 0,9 PAR = 0.3, NI = 1000

Salidas: Vector de armonía con los mejores valores encontrados para los parámetros después de cumplida las iteraciones.

Condiciones: Ejecución del algoritmo GHSO cuya función objetivo está dada por los valores de ROUGE, es decir, partiendo de los rangos de valores de la entrada del algoritmo se llena la memoria armónica. Con esos valores se ejecuta HS, GHS o GHS+LEM para la generación de los resúmenes y posteriormente los resúmenes obtenidos se evalúan por medio de ROUGE.

1.2 DEFINICIÓN DE LA ARQUITECTURA

Esto se refiere a patrones que brindan un esquema de referencia útil para guiarse en el desarrollo de software dentro de un sistema informático. Los objetivos que persigue esta etapa son: que el software pueda ser sostenible, esto es, fácilmente analizable, modificable, corregible; también se tienen en cuenta otros factores tales como la interacción con otros sistemas informáticos y escalabilidad.
La parte fundamental del sistema propuesto es realizar la ejecución del algoritmo GHS+LEM a partir de 3 características fundamentales: cobertura, cohesión y redundancia. Lo primero que se requiere en el sistema es la etapa de pre-procesamiento en la cual se realiza la segmentación de oraciones, filtro de palabras vacías y lematización. Con ésta etapa se logra una considerable reducción de palabras poco significativas que podrían generar ruido para el proceso de selección de oraciones relevantes. Adicionalmente se logra un tiempo de procesamiento mucho menor.
Posteriormente se realiza la etapa de representación de documentos de origen cuyo objetivo es facilitar la interpretación de los documentos para que puedan ser procesados por el algoritmo GHS+LEM.
En el proceso de selección de oraciones relevantes se ejecuta el algoritmo GHS+LEM cuyo resultado es el conjunto de oraciones que conforman el resumen.

De acuerdo a la exploración realizada se definió un diagrama de procesos general del sistema el cual se muestra en la Figura 2.

Figura 2. Diagrama de procesos del sistema de generación de resúmenes basado en GHS+LEM.

1.3 DEFINICIÓN DEL ENTORNO DE EVALUACIÓN

En este punto se realiza la selección de los documentos proporcionados por DUC, con los cuales se hace la evaluación del algoritmo y también se define el entorno con el cual se realiza la evaluación, como resultado de esta fase se obtiene el conjunto de documentos a utilizar y las características del entorno.

1.3.1 Conjunto de documentos a utilizar

Para la selección de los documentos proporcionados por DUC se realizó una exploración de los documentos proporcionados en cada año y las respectivas tareas propuestas para las conferencias. A continuación se muestra el resultado obtenido a partir de la exploración realizada:

Conferencias de Comprensión de Documentos (Document Understanding Conference, DUC)

Hay muchas actividades enfocadas en la construcción de poderosos sistemas de información. Para fomentar el progreso en el área de generación de resúmenes y permitir a los investigadores participar en proyectos a gran escala, NIST[4](Instituto nacional de

[4] http://www.nist.gov/index.html

estándares y tecnología) crea DUC como una forma de estandarizar la evaluación de los diferentes sistemas de generación de resúmenes propuestos alrededor del mundo.

En DUC se han evaluado sistemas de generación de resúmenes de diversos tipos como: mono-documento, multi-documento, genéricos, basados en consultas, de diferentes tamaños, etc. Los documentos son de noticias y provienen de AQUAINT[5], TIPSTER[6] y TREC[7].

Los artículos publicados en DUC sirven como medio para la difusión de los trabajos realizados por los participantes en las conferencias. Los documentos de DUC contienen información desde el año 2001 hasta el 2007, posteriormente DUC se convirtió en TAC[8](Conferencia de Análisis de Texto).

Los documentos proporcionados por DUC incluyen:

- Documentos
- Resúmenes, resultados, tablas con resultados de evaluaciones, soporte adicional y software.

DUC usa códigos para referirse a trabajos en las comparaciones para conservar la anonimidad de los sistemas que trabajaron en él.

De la Tabla 5 a la Tabla 12 se muestra un breve resumen de las características de DUC desde el año 2001 hasta 2007 y TAC.

DUC 2001	
Año	**2001**
Datos	60 conjuntos de 10 documentos (30 entrenamiento, 30 prueba) Donde cada conjunto contiene resúmenes por documento y resúmenes de múltiples documentos Se crean líneas base automáticas que constan de las primeras n palabras del documento. Los documentos están estructurados con DTD's.(Definición del Tipo de Documento)
Tareas	- Resumen mono-documento: Se crea un resumen de las primeras 100 palabras. - Resúmenes Multi-documento: 4 resúmenes genéricos del conjunto completo (400, 200, 200 y 50 primeras palabras)
Evaluación	Los resúmenes de un documento son evaluados en 2 categorías: Humanos y Automáticos. Para la generación de resumenes multi-documento la evaluación sólo fue humana.
Observaciones	

Tabla 5. Características documentos de DUC 2001

[5] http://www-nlpir.nist.gov/projects/aquaint/
[6] http://www-nlpir.nist.gov/related_projects/tipster/
[7] http://trec.nist.gov/
[8] http://www.nist.gov/tac/

DUC 2002	
Año	**2002**
Datos	60 conjuntos de documentos donde cada conjunto contiene: resúmenes abstractos de un documento y para múltiples documentos resúmenes abstractos/extractos. Los documentos están estructurados con DTD's.
Tareas	- Resúmenes mono-documento: 60 conjuntos de aproximadamente 10 documentos. Se creó un abstract genérico del documento con una longitud de aproximadamente 100 palabras o menos. - Resúmenes multi-documento: 60 conjuntos de aproximadamente 10 documentos. 4 abstracts del conjunto completo con longitudes de aproximadamente 200, 100, 50 y 10 palabras. Dado un conjunto de documentos, se crearon 2 extractos genéricos de 400 y 200 palabras.
Evaluación	Los abstractos son manualmente evaluados por NIST Extractos fueron evaluados automáticamente empleando técnicas de evaluación usadas por NIST como el cálculo de recuerdo.
Observaciones	

Tabla 6. Características documentos de DUC 2002

DUC 2003	
Año	**2003**
Datos	Los documentos son estructurados con DTDs 30 conjuntos de documentos de TREC *Resúmenes manuales:* - Resúmenes muy cortos de un documento (aprox. 10 palabras) - Resúmenes cortos de 100 palabras de cada grupo diseñados para reflejar el punto de vista del evaluador. 30 conjuntos de documentos TDT(Detección y Seguimiento de Tópicos) orientados a tópicos, eventos y a periodos de tiempo.
Tareas	**Tarea1:** Resúmenes muy cortos (dado un documento, crea un resumen de 10 palabras) **Tarea2:** Resúmenes cortos enfocados a eventos (dado un grupo asociado a un tema TDT, se crea un resumen corto de 100 palabras) **Tarea3:** Resúmenes enfocados en puntos de vista (crea un resumen corto de 100 palabras desde un punto de vista específico) **Tarea4:** Resúmenes cortos en respuesta a una consulta.
Evaluación	
Observaciones	

Tabla 7. Características documentos de DUC 2003

DUC 2004	
Año	**2004**
Datos	Documentos de noticias provenientes de las colecciones de TDT y TREC (idiomas: inglés y árabe). - 50 conjuntos de documentos TDT en inglés (cada subconjunto tiene aproximadamente 10 documentos) - 25 conjuntos de documentos árabes (cada subconjunto tiene aproximadamente 10 documentos) - 50 conjuntos de documentos TREC en inglés (cada subconjunto tiene aproximadamente 10 documentos)
Tareas	**Tareas 1 y 2:** Son las mismas que se definieron para DUC 2003. **Tarea3:** Resúmenes de un documento en lenguaje cruzado.
Evaluación	Resúmenes evaluados con ROUGE.
Observaciones	

Tabla 8. Características documentos de DUC 2004

DUC 2005	
Año	**2005**
Datos	Los evaluadores de NIST seleccionan temas de interés, cada tema tiene al menos 35 documentos relevantes asociados, los evaluadores leen los documentos para cada tema y seleccionan un subconjunto de 25 a 50 documentos relevantes.
Tareas	- Resumen de múltiples documentos enfocado en consultas complejas - Generación de un resumen breve, bien organizado, fluido y con un nivel de granularidad. El resumen no debe tener más de 250 palabras.
Evaluación	Nist calcula dos puntuaciones de ROUGE oficiales: recuerdo con ROUGE-2 y ROUGE-SU4.
Observaciones	

Tabla 9. Características documentos de DUC 2005

DUC 2006	
Año	**2006**
Datos	Los evaluadores de NIST desarrollan temas de interés. Crean un tema y seleccionan un conjunto de 25 docs relevantes. Los documentos provienen del corpus AQUAINT, comprenden artículos de noticias de Associated Press and New York Times.
Tareas	

Evaluación	
Observaciones	

Tabla 10. Características documentos de DUC 2006

DUC 2007	
Año	**2007**
Datos	Los datos provienen del corpus de AQUAINT que comprenden artículos de noticias, los evaluadores crean un tema y seleccionan un subconjunto de 25 documentos relevantes.
Tareas	
Evaluación	
Observaciones	

Tabla 11. Características documentos de DUC 2007

Conferencia de análisis de texto TAC	
Año	**2008 a 2011**
Observaciones	TAC es una serie de trabajos para la evaluación cuyo objetivo mejorar la investigación en el procesamiento de lenguaje natural y aplicaciones relacionadas. Proporciona una gran colección de prueba, procedimientos de evaluación común, y un foro de organizaciones que comparten sus resultados. Las tareas propuestas en estos años están enfocadas en procesamiento lingüístico profundo, multilenguaje, evaluación.

Tabla 12. Características documentos de DUC 2008 a 2011

A partir de la exploración realizada se puede concluir que basta con utilizar los documentos de DUC2002 para la evaluación de un sistema de generación de resúmenes multi-documento, sin embargo, también se encuentran trabajos actuales que usaron corpus de años posteriores a DUC2002, esto es, porque aunque las tareas definidas para generación de resúmenes multi-documento hayan sido propuestas en las primeras conferencias de DUC, en las conferencias de años posteriores se pueden encontrar más documentos y con una mejor calidad además de ser usados por sistemas más recientes. Esto da la posibilidad de compararse con trabajos que están enfocados en la tarea de resúmenes multi-documento genéricos y han utilizado documentos de DUC de otros años lo que permite una flexibilidad mayor en la selección de los conjuntos de documentos.

1.3.2 Características del entorno

Para la realización de las pruebas de laboratorio y para el desarrollo del software se utilizó un equipo con las siguientes características:
Sistema operativo: Windows 7 Professional, 32 bits.
Procesador: Intel(R) Pentium(R) 4 CPU 3.00 GHz, 2992 Mhz.
Ram: 1GB.
Hard disk: 74.50 GB
Fabricante del sistema: Dell Computer Corporation

Modelo del sistema: Dimension 8300

El entorno de desarrollo fue Microsoft Visual Studio 2010 con el lenguaje C#.
Para la evaluación de la calidad de los resúmenes generados por los algoritmos se hizo uso de la herramienta ROUGE la cual estaba realizada en el lenguaje Perl y fue ejecutada en el entorno de Windows.

ANEXO B - METODOLOGÍA DESARROLLO DE LOS ALGORITMOS

2 DESCRIPCIÓN GENERAL DE LA METODOLOGÍA RUP

La metodología para la elaboración de los algoritmos y la evaluación es una instanciación del Proceso Unificado, la cual tuvo en cuenta las fases de Iniciación, Elaboración, Construcción y Transición, se incluyo en la fase de transición la realización de pruebas experimentales y análisis de resultados, a continuación se describen cada una de las fases.

2.1 INICIACIÓN

En esta etapa se realiza un diseño preliminar del sistema y de la arquitectura general teniendo en cuenta los requisitos planteados en la primera metodología. Como resultado de esta fase se obtuvo el diagrama general de casos de uso del sistema con el objetivo de establecer los requisitos mínimos para el desarrollo del sistema.

2.1.1 Diagrama general de casos de uso del sistema

El diagrama de casos de uso muestra el posible comportamiento del sistema. Por tal motivo, es útil para determinar los requisitos funcionales del sistema, es decir, representan las funciones que el algoritmo puede ejecutar o realizar. En la Figura 3 se muestra el diagrama de casos de uso.

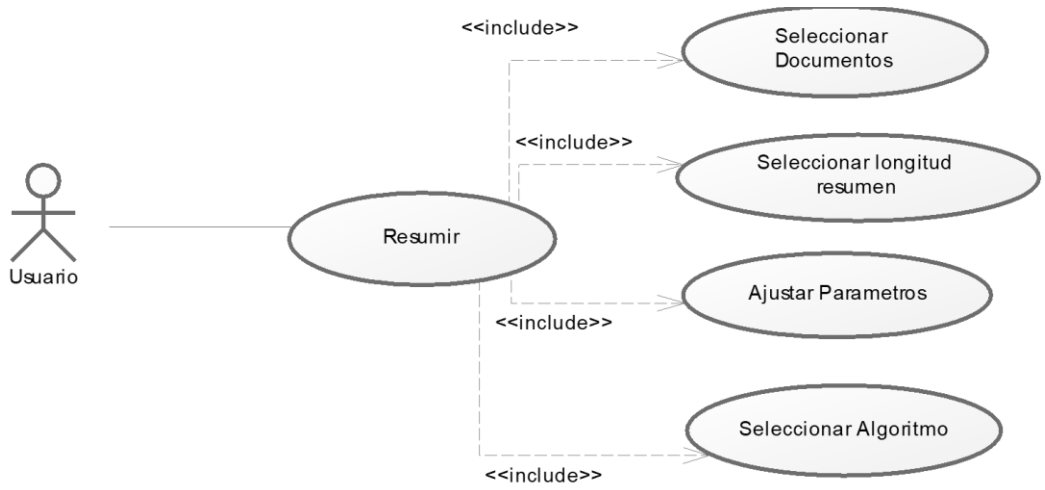

Figura 3. Diagrama general de casos de uso del sistema

2.2 ELABORACIÓN

En esta etapa se comprendió con mayor detalle los requerimientos para el modelado y diseño del algoritmo y el afinamiento de la arquitectura general del sistema. Como

resultado de esta fase se obtuvo: Casos de Uso de alto nivel, Diagrama de Clases y Arquitectura base.

2.2.1 Caso de uso de alto nivel

En la Figura 4 se muestra las operaciones que el usuario del sistema puede realizar, a saber: seleccionar los documentos que va a resumir, seleccionar la longitud deseada del resumen, ajustar los parámetros del algoritmo seleccionado y realizar el resumen de los documentos.

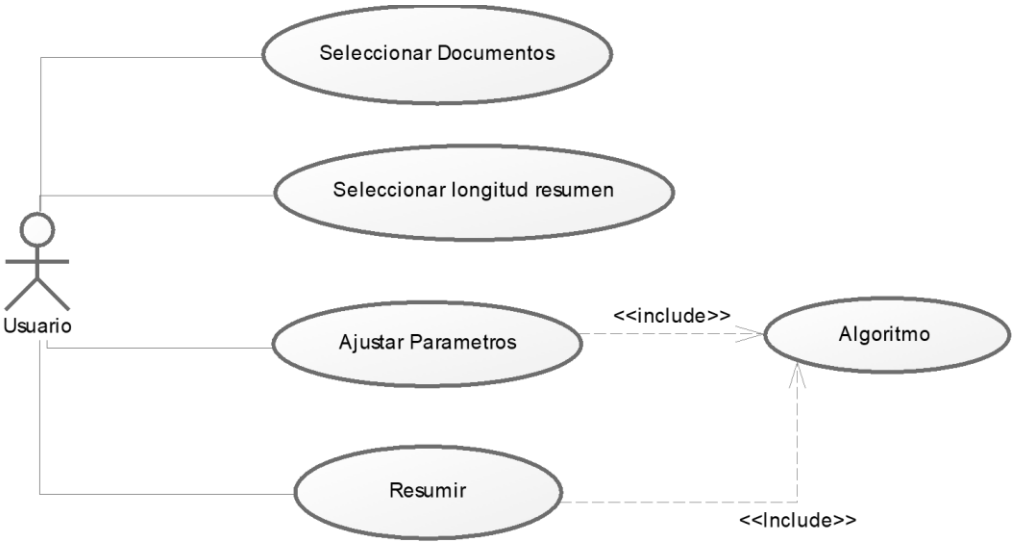

Figura 4. Casos de uso de alto nivel

2.2.2 Diagrama de clases.

El sistema de generación de resúmenes de múltiples documentos, se encuentra dividido en dos módulos.

- Pre-procesamiento
- Algoritmos de búsqueda armónica y evaluación

En la Figura 5 y Figura 6 se muestran los diagramas de clase para cada uno de los módulos y en la Tabla 13. Descripción de cada una de las clases se describe la funcionalidad de cada una de las clases.

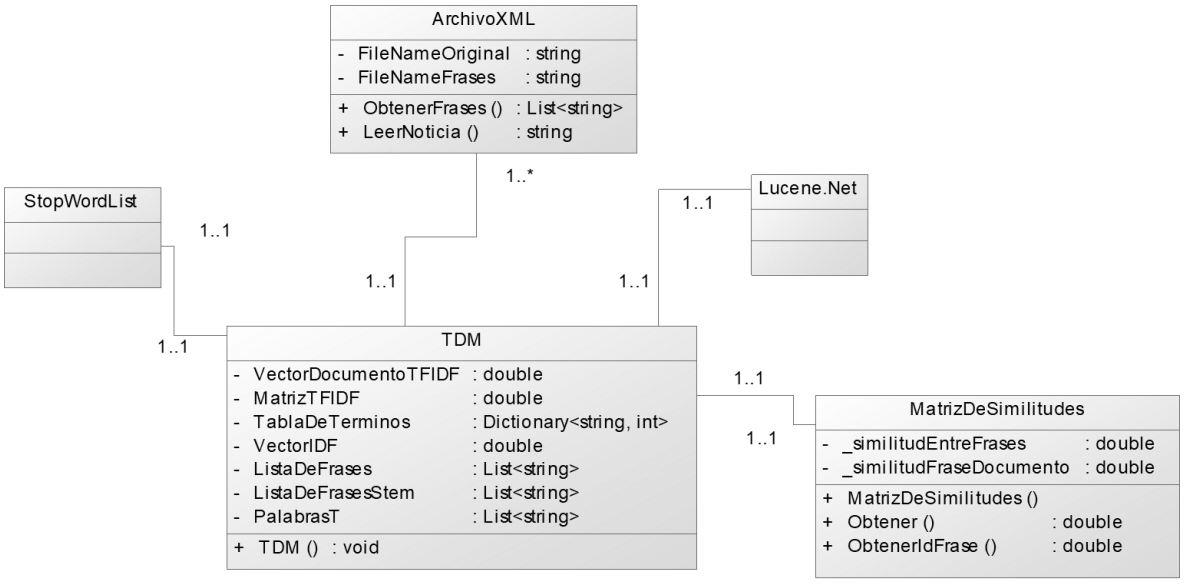

Figura 5. Diagrama clases pre-procesamiento

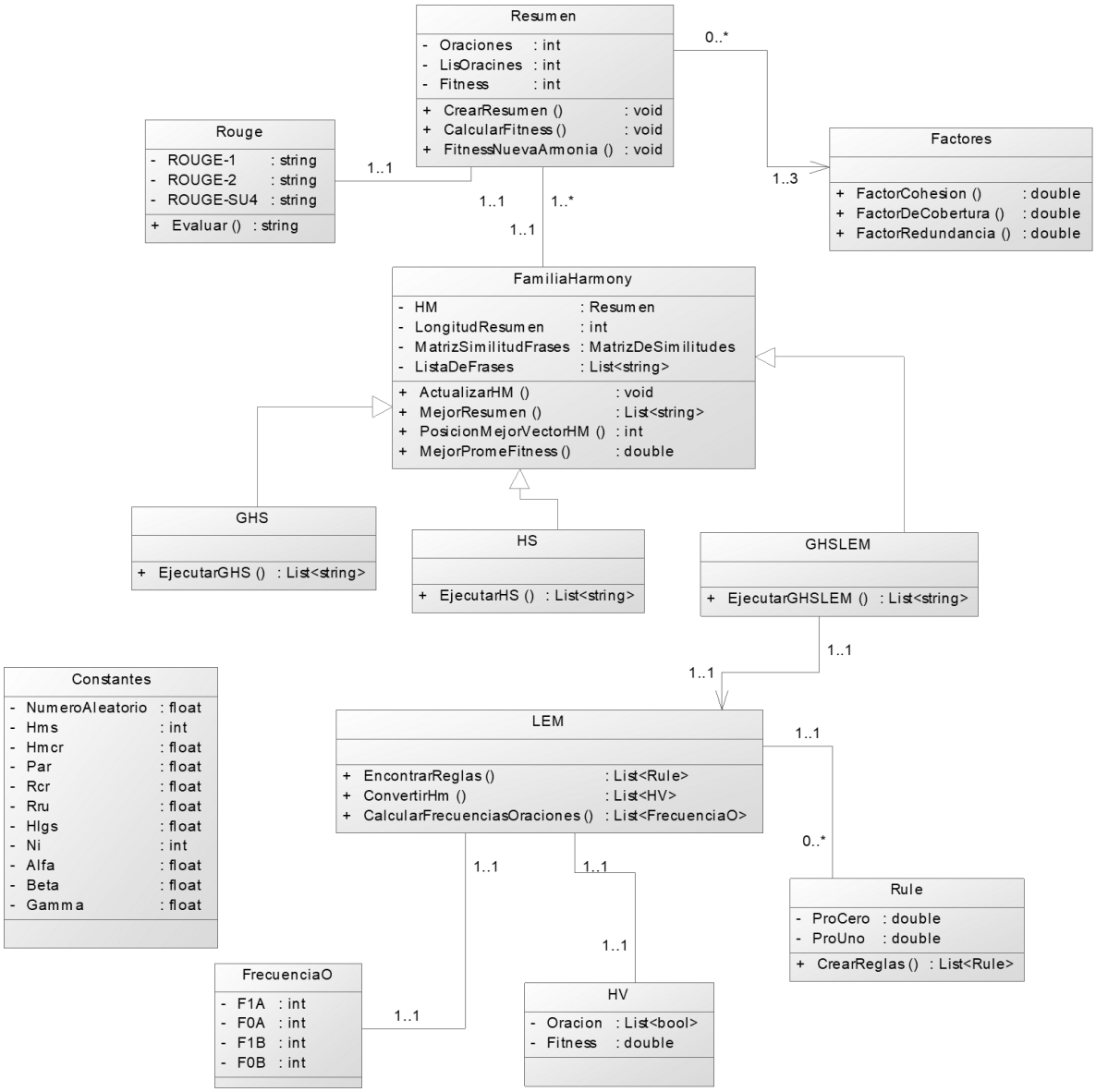

Figura 6. Diagrama clases algoritmos de búsqueda armónica

CLASE	FUNCIÓN
ArchivoXML	En esta clase se implementan funciones que permiten leer los documentos de DUC y realizar la segmentación de las oraciones
StopWordList	Es una clase estática que contiene todas las palabras vacías (stopword) que se eliminarán de los documentos a resumir.
Lucene	Es la clase de lucene.net que permite hacer eliminación de palabras vacías (stopwords), aplicar lematización (stemming), segmentación de oraciones en palabras.
TDM	Clase donde se implementan funciones que permiten realizar la matriz TF-ISF y la matriz de pesos.
MatrizDeSimilitudes	Implementa funciones que permiten el cálculo de la matriz triangular superior de la similitud entre las oraciones y el documento.
Resumen	Representa la funcionalidad para la creación de un resumen.
Factores	Representa los principales factores que se necesitan para un buen resumen.
FamiliaHarmony	Clase abstracta que implementa las funciones comunes para los algoritmos HS, GHS, GHS+LEM
HS	Representa la funcionalidad (pasos) del algoritmo
GHS	Representa la funcionalidad (pasos) del algoritmo GHS
GHS+LEM	Representa la funcionalidad (pasos) del algoritmo GHSLEM
LEM	Contiene la funcionalidad necesaria para la ejecución de LEM (Modelo Evolutivo que Aprende).
FrecuenciaO	Clase utilizada para almacenar las frecuencias de cada oración en la matriz E
HV	Clase usada para hallar las probabilidades cuando se ejecuta el proceso de inferencia de reglas.
Reglas	Especifica las características de una regla
Rouge	Realiza el llamado a ROUGE-1.5.5.pl para realizar la evaluación del resumen generado.
Constantes	Clase utilizada para almacenar todas las variables necesarias para la correcta ejecución de los algoritmos.

Tabla 13. Descripción de cada una de las clases

2.2.3 Arquitectura base:

Para el sistema se definió una arquitectura multinivel que consta de 3 niveles, lógica de presentación, lógica de negocio y persistencia. Entre las ventajas que se obtienen mediante el uso de este tipo de arquitectura están la flexibilidad, la escalabilidad y el mantenimiento eficiente del sistema. En la Figura 7. Arquitectura del Sistema. se muestra la arquitectura del sistema y sus componentes.

A continuación se hace una breve descripción de las funciones que se realizan en cada uno de los niveles de la arquitectura.

- *Presentación:* En este nivel se incluyen los componentes de la interfaz del usuario que permiten seleccionar las opciones necesarias para la ejecución de las pruebas. Este nivel se comunica únicamente con la capa de negocio por medio de las interfaces.

- *Lógica de Negocio*: Este nivel se divide en tres módulos:

    - Módulo de Pre-procesamiento: Contiene la lógica necesaria para seleccionar los documentos y hacer los cálculos requeridos para el proceso de generación de resúmenes de múltiples documentos (segmentación de oraciones, cálculo de matriz TF-ISF).

    - Módulo HarmonySearch: Contiene los algoritmos armónicos que realizarán el proceso de generación de resúmenes de los documentos.

    - Módulo LEM: Contiene el proceso de inferencia de reglas necesario para la ejecución del algoritmo armónico propuesto en el proyecto.

- *Persistencia:* En este nivel es donde residen los datos. Tiene como objetivo almacenar los resultados de la ejecución de las pruebas en una hoja de Excel este nivel contiene el siguiente subnivel:

    - Lógica de servicios: Implementa la persistencia de la información obtenida por el programa ocultando los detalles de los repositorios de datos a los niveles superiores.

2.3 CONSTRUCCIÓN

Una vez realizadas las fases de Iniciación y Elaboración, se obtuvo un prototipo funcional del algoritmo a través de las siguientes actividades:

- **Análisis**: Se hace una profundización sobre los artefactos generados durante la fase de elaboración para la construcción del sistema.

- **Diseño**: Se realizaron los casos de uso reales que sirvieron como guía para la construcción de las diferentes funcionalidades.

- **Implementación**: Se implementó el sistema (en los primeros ciclos el algoritmo) con los artefactos obtenidos en las anteriores actividades (Análisis y Diseño), posteriormente se realizaron pruebas alfa para garantizar su funcionalidad.

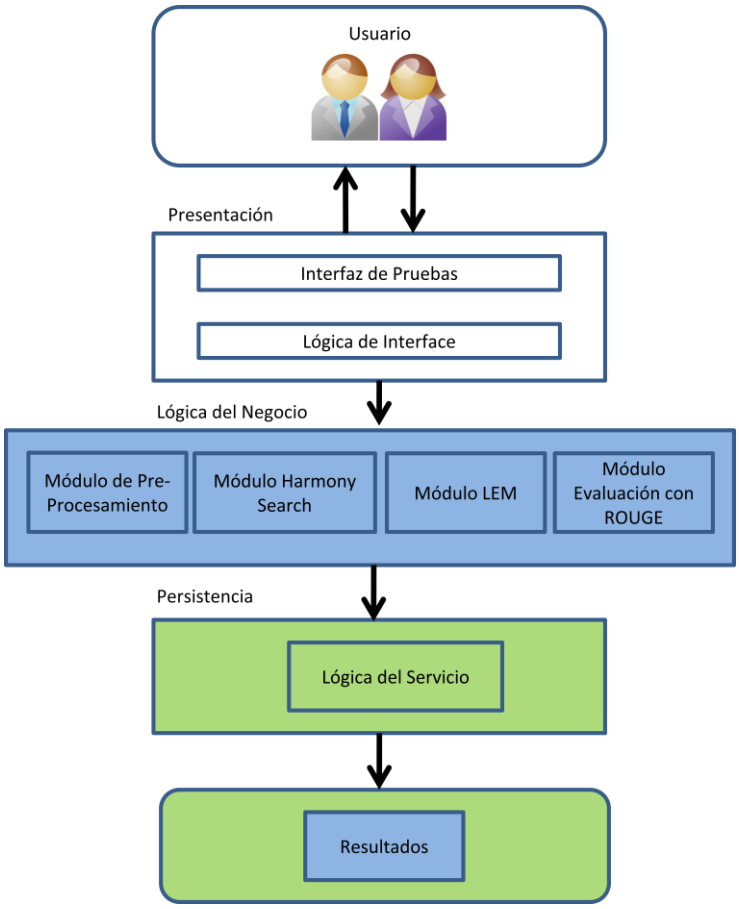

Figura 7. Arquitectura del Sistema.

- **Pruebas**: Al finalizar la implementación del algoritmo se define un conjunto de pruebas de caja negra, que serán aplicadas a las siguientes funcionalidades: 1. Algoritmo de generación automática de resúmenes para uno y múltiples documentos 2. Algoritmo de evaluación de resultados 3. Algoritmo para Tabulación y reporte de resultados de la evaluación.

2.3.1 Ciclos de Desarrollo

Los ciclos de desarrollo permitieron dividir la funcionalidad del sistema en funciones más pequeñas que facilitaron la labor de construcción del sistema cumpliendo con cada una de las fases mencionadas anteriormente. Los ciclos desarrollados fueron:

- Ciclo 1. Algoritmo de pre-procesamiento de documentos: en esta parte se adapto el programa de la segmentación de oraciones, se realizó la construcción de la matriz TF-ISF y matriz de similitud entre oraciones y el documento, se utilizó como lenguaje de implementación Microsoft Visual Studio 2010 y C#.
- Ciclo 2. Adaptación del algoritmo HS: En este ciclo, tomando como base el algoritmo [1] se adaptó el algoritmo HS para generación de resúmenes de múltiples documentos.
- Ciclo 3. Algoritmos GHS y GHS+LEM para generación de resúmenes: Se implementó el algoritmo de generación de reglas y posteriormente se adapto a

GHS, también se implementó GHS para generación de resúmenes de múltiples documentos.

- Ciclo 4. Adaptación de ROUGE: Se incluyó la herramienta de ROUGE dentro de los algoritmos implementados para poder realizar las pruebas.
- Ciclo 5. Tabulación y reporte de resultados: Se implementó una función que permitió guardar los resultados en una hoja de Excel para después realizar el análisis de las pruebas.

2.3.2 Casos de uso reales

A continuación se muestra los casos de uso reales del sistema

<table>
<tr><th colspan="2">CASO DE USO REAL: REALIZAR RESUMEN DE COLECCIÓN DE DOCUMENTOS</th></tr>
<tr><td colspan="2">Actores: Usuario.</td></tr>
<tr><td colspan="2">Propósito: Realizar evaluación de los documentos seleccionados</td></tr>
<tr><td colspan="2">Resumen: El usuario selecciona la colección de documentos que desea evaluar, escoge el algoritmo, ejecuta todo el proceso y finalmente se entrega una hoja de Excel con los resultados de la evaluación.</td></tr>
<tr><td colspan="2">Tipo: Primario.</td></tr>
<tr><th colspan="2">CURSO NORMAL DE LOS EVENTOS</th></tr>
<tr><th>Acción del actor</th><th>Respuesta del sistema</th></tr>
<tr><td>1. El usuario ejecuta la opción que da inicio a este caso de uso.</td><td></td></tr>
<tr><td></td><td>2. El sistema presenta al usuario una interfaz con las opciones: Directorio documentos, Algoritmo de búsqueda armónica (HS, GHS, GHS+LEM)</td></tr>
<tr><td>3. El usuario da clic en el botón Examinar y selecciona los documentos</td><td>4. El sistema toma los documentos y los guarda en memoria temporal.</td></tr>
<tr><td>5. El usuario selecciona el algoritmo a utilizar para realizar el resumen (HS, GHS, GHS+LEM)</td><td>6. El sistema guarda en memoria temporal el algoritmo a utilizar.</td></tr>
<tr><td>7. El usuario da clic en Evaluar</td><td>8. El sistema procesa la opción:
a. El sistema realiza el pre-procesamiento de los documentos, lo que incluye, segmentación de oraciones, construcción de matriz de frecuencias, matriz de pesos y construcción de matriz de similitudes entre oraciones y colección de documentos.
b. Realización del resumen aplicando el algoritmo seleccionado con todos los parámetros que se afinaron.
c. Evaluación del resumen por medio de la herramienta ROUGE.
d. Generar de la hoja de Excel con</td></tr>
</table>

	los resultados de la evaluación.
	9. Una vez terminada la ejecución, el sistema finaliza y en el directorio evaluación se crea una hoja de Excel con los resultados de la evaluación.
CURSO ALTERNO	
Acción del actor	**Respuesta del sistema**
10. El usuario no selecciona los documentos a evaluar.	11. El sistema le informa que debe seleccionar los documentos.

Tabla 14. Caso de uso real: realizar evaluación de colección de documentos

2.3.3 Pruebas de caja negra

Las pruebas de caja negra fueron aplicadas a los algoritmos de pruebas y están divididas en cuatro módulos los cuales describiremos a continuación:

2.3.3.1 Modulo generación automática de resúmenes

En este modulo se seleccionan desde un conjunto de documentos a resumir los cuales deben cumplir con el formato que se mencionó en la **Figura 1.** Estructura de los documentos de evaluación, también se debe seleccionar la longitud del resumen la cual debe ser mayor a 200 y menor a 250 palabras, como resultado final este modulo entrega una lista de oraciones que contienen una cantidad aproximada de palabras igual a la especificada.

En este modulo se debe validar que los documentos que se leen del archivo cumplan con el formato especificado y que la longitud de la colección de los documentos sea mayor a la longitud del resumen, en caso contrario se muestra un error, también se debe validar que la longitud del resumen esté en los rangos especificados.

- Tabla de clases equivalentes
  Lr = Longitud del resumen
  Dr = Directorio conjunto de documentos
  Lcd = Longitud Colección Documentos

Asume	Id	Condición	Clases Correctas	Clases Erróneas
	A	N° de parámetros	{ n = 2 } **1**	{ n < 2 } **2.1** { no selecciono ningún documento } **2.2**
	B	Tipo de parámetros	{ Lr ϵ N } **3**	{ Lr no ϵ N } **4**
A,B	C	Datos correctos	{ Lr > 199 , Lr < 251 } **5.1** { Lcd > Lr } **5.2** { Dr debe ser un directorio que contenga documentos validos} **5.3**	{ Lr < 200 , Lr >250 } **6.1** { Lcd < Lr } **6.2** { Dr no contiene directorios validos } **6.3**

Tabla 15. Clases equivalentes modulo 1

- Batería de pruebas

	Entradas	**Salidas**	**Clases Cubiertas**	**Valores Limite**	**Salidas**
Clases correctas	Conjuntos de documentos validos. Lr = 220	Resumen realizado con éxito	1, 3, 5.1, 5.2, 5.3	Lr = 200, Lr = 250	Resumen realizado con éxito
Clases erróneas	Sin parámetros	Debe seleccionar por lo menos un documento para resumir y una longitud de resumen valida	2.1, 2.2		
	Dr valida y Lr nula	Debe seleccionar una longitud de resumen	2.1		
	Lr = 200.5	Longitud del resumen no valida, debe ser un valor entero entre 200 y 250	4		
	Lr = “cien”	Longitud del resumen no valida, debe ser un valor entero entre 200 y 250	4		
	Lr = 100	Longitud del resumen no valida, debe ser un valor entero entre 200 y 250	6.1	Lr = 200,	Resumen realizado con éxito
	Lr = 300	Longitud del resumen no valida, debe ser un valor entero entre 200 y 250	6.1	Lr = 250	Resumen realizado con éxito
	Dr con documentos no validos	No se pueden leer los documentos a resumir	6.3		
	Lr > Lcd	Longitud del resumen mayor	6.2	Lr < Lcd	Resumen realizado

		a la colección de documentos.			con éxito

Tabla 16. Batería de pruebas modulo 1

2.3.4 Modulo evaluación de resultados

Este modulo recibe como entrada un archivo con el resumen realizado por el algoritmo, los resúmenes modelo que se utilizan para la evaluación son seleccionados de manera automática por la herramienta ROUGE 1.5.5. La salida de este modelo es un archivo con los resultados de la evaluación.

El archivo resumen no tiene ningún tipo de etiqueta, solo contiene las oraciones seleccionadas y no tiene extensión.

- Tabla de clases equivalentes
  Ar = Archivo resumen

Asume	Id	Condición	Clases Correctas	Clases Erróneas
	A	N° de parámetros	{ n = 1 } **1**	{ n < 1 } **2**
	B	Datos correctos	Ar es archivo resumen **3**	Ar no contiene ningún archivo resumen **4**

Tabla 17. Clases equivalentes modulo 2

- Batería de pruebas

	Entradas	Salidas	Clases Cubiertas	Valores Limite	Salidas
Clases correctas	Ar es un archivo resumen valido	Archivo con resultados	1,3		
Clases erróneas	Ar no contiene ningún archivo resumen valido	Error! no se puede leer el resumen	4		
	Ar es nulo	Debes seleccionar el archivo resumen	2		

Tabla 18. Batería de pruebas modulo 2

2.3.5 Modulo tabulación de resultados

Este modulo toma el archivo de evaluación de resultados, lee los valores y genera una hoja de Excel donde se puede graficar y analizar los resultados.

El archivo de evaluación de resultado es un archivo .txt sin ningún tipo de etiqueta.

- Tabla de clases equivalentes

AER = Archivo de evaluación de resultados

Asume	Id	Condición	Clases Correctas	Clases Erróneas
	A	N° de parámetros	{ n = 1 } **1**	{ n < 1 } **2**
	B	Datos correctos	AER es archivo valido **3**	AER no contiene ningún archivo valido **4**

Tabla 19. Clases equivalentes modulo 3

- Batería de pruebas

	Entradas	Salidas	Clases Cubiertas	Valores Limite	Salidas
Clases correctas	AER es un archivo valido	Hoja de Excel creada	1,3		
Clases erróneas	AER no contiene ningún archivo valido	Error! no se puede crear la hoja de excel	4		
	AER es nulo	Debes seleccionar el archivo resumen	2		

Tabla 20. Batería de pruebas modulo 3

2.4 TRANSICIÓN

En esta fase se verifica la funcionalidad del sistema, se realiza el afinamiento de parámetros, ajuste de coeficientes y la evaluación por medio de la utilización de documentos de DUC. Finalmente se realiza el análisis de los resultados obtenidos en la evaluación.

ANEXO C – LEM para Generación de Resúmenes

3 EJEMPLO DE LEM

El siguiente ejemplo muestra en detalle la operación del algoritmo básico de PRISM cuando es aplicado a una memoria armónica que contiene 10 armonías. El conjunto de datos se presenta en la Tabla 21, La cual se encuentra ordenada según el fitness de mayor a menor.

S1	**S2**	**S3**	**S4**	**S5**	**S6**	**F(x)**
1	1	0	0	1	1	5,2342
0	0	0	1	1	1	5,2342
0	1	0	1	0	0	4,5000
0	1	0	1	1	0	4,2342
0	1	0	0	0	1	3,4540
0	0	0	1	0	0	3,3453
1	0	0	0	1	1	2,4345
1	0	0	0	1	0	2,2342
0	1	0	1	1	1	1,2342
1	0	0	1	0	0	-2,3423

Tabla 21. Memoria Armónica

A. Dividimos la memoria amónica en dos grupos (altos y bajos) cada grupo con una cantidad de armonías según el valor de HLGS para este caso 4, en la Tabla 22 se muestran los grupos.

S1	S2	S3	S4	S5	S6	F(x)
1	1	0	0	1	1	5,2342
0	0	0	1	1	1	5,2342
0	1	0	1	0	0	4,5000
0	1	0	1	1	0	4,2342
1	0	0	0	1	1	2,4345
1	0	0	0	1	0	2,2342
0	1	0	1	1	1	1,2342
1	0	0	1	0	0	-2,3423

Tabla 22. Grupos de alto y bajo rendimiento

B. Teniendo los grupos de bajo y alto rendimiento, se calcula la frecuencia de ceros y unos, de la siguiente manera:
   - Grupo de altos: Se calcula frecuencia de unos en altos (*F1A*) y frecuencia de ceros en altos (*F0A*).
   - Grupo de bajos: Se calcula frecuencia de unos en bajos (*F1B*) y frecuencia de ceros en bajos (*F0B*).

En la Tabla 23 se muestran las frecuencias de cada una de las oraciones y el tamaño total de los grupos (TTG).

	F1A	F0A	F1B	F0B	TTG
S1	1	3	3	1	8
S2	3	1	1	0	8
S3	0	4	0	4	8
S4	1	0	2	2	8
S5	3	1	1	0	8
S6	2	2	2	2	8

Tabla 23. Frecuencia de oraciones

C. Se crea una lista de reglas donde se almacena la probabilidad de ocurrencia de ceros y unos para cada dimensión de acuerdo a las siguiente fórmulas:

$$\mathrm{P}_{\mathrm{s_i}}(1) = \begin{cases} 50\% + (\mathrm{F1A}_{\mathrm{s_i}} - \mathrm{F1B}_{\mathrm{s_i}})/\mathrm{TTG}_{\mathrm{s_i}} & \mathrm{F1A}_{\mathrm{s_i}} - \mathrm{F0A}_{\mathrm{s_i}} \geq 0 \\ 50\% - (\mathrm{F0A}_{\mathrm{s_i}} - \mathrm{F0B}_{\mathrm{s_i}})/\ \mathrm{TTG}_{\mathrm{s_i}} & \mathrm{d.o.m.} \end{cases}$$

$$P_{s_i}(0) = 1 - P_{s_i}(1)$$

La lista de reglas para cada una de las oraciones se muestra en la Tabla 24

	P(1)	P(0)
S1	0,25	0,75
S2	0,75	0,25
S3	0,5	0,5
S4	0,375	0,625
S5	0,75	0,25
S6	0,5	0,5

Tabla 24. Probabilidad de ocurrencias

Como podemos observar en la Tabla 24 para el caso de S1 la probabilidad de que se de un uno es de tan solo 25% esto quiere decir que no es conveniente seleccionar esta oración para el resumen.

ANEXO D – SELECCIÓN DE FUNCIÓN OBJETIVO

4 SELECCIÓN DE FUNCIÓN OBJETIVO

La función objetivo es muy importante en los algoritmos evolutivos, dado que es la función que se busca optimizar. En generación automática de resúmenes de múltiples documentos se plantean dos factores muy importantes que hacen parte de la función objetivo, estos son cobertura y eliminación de redundancia. A continuación se muestran las dos funciones objetivos que se probaron en esta investigación.

4.1 Función objetivo de MCMR

Alguliev R. [2] propone una función objetivo en la cual plantea dos factores, el primero es cobertura, con el cual trata de garantizar que las oraciones que están en el resumen sean las más relevantes puesto que abarcan la mayor cantidad del contenido posible, el segundo factor es la redundancia con la cual se pretende reducir la información repetida. Para lograr su objetivo plantean la siguiente ecuación:

$$F = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} [sim(D, Si) + sim(D, Sj)] - Sim(Si, Sj) \quad (1)$$

Donde *Si* y *Sj* son oraciones del resumen, *D* es la colección de oraciones de todos los documentos, $sim(D, Sj)$ es la similitud de la oración *i* del resumen con todas las oraciones del conjunto de documentos, $sim(D, Sj)$ es la similitud de la oración *j* con todas las oraciones del conjunto de documentos, *n* es la cantidad de oraciones que hay en el resumen y $Sim(Si, Sj)$, es la similitud entre la oración *i* y *j.*

4.2 Función objetivo propuesta

Después de haber analizado la función propuesta por Alguliev R. y la función de HS para un solo documento [1] propuesta por Sharegui. E. se llegó a la hipótesis de que el factor de cohesión planteado por Shareghi E. puede ayudar en la eliminación de la redundancia, por consiguiente para el presente trabajo se decide modificar la ecuación 1 de la siguiente manera.

$$F = \left[\frac{\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} [\mathrm{sim(D,Si)} + \mathrm{sim(D,Sj)}]}{m * 2} \right] - \frac{\log(C * 9 + 1)}{\log(M * 9 + 1)} \quad (2)$$

Donde *Si* y *Sj* son oraciones del resumen, *D* es la colección de oraciones de todos los documentos, *sim(D,Si)* es la similitud de la oración *i* del resumen con todas las oraciones del conjunto de documentos, *sim(D,Sj)* es la similitud de la oración *j* con todas las oraciones del conjunto de documentos, *n* es la cantidad de oraciones que hay en el resumen, *m* es la cantidad de combinaciones entre pares de oraciones del resumen y el *2* es una constante que se utiliza para sumar todas las similitudes calculadas entre las oraciones, debido a cada combinación de las sumatorias (*m*) involucra dos oraciones.

4.3 Comparación funciones objetivo

Para poder encontrar la mejor función objetivo para la generación automática de resúmenes se decidió hacer una prueba donde se ejecutó el algoritmo GHS+LEM para las

dos funciones. Los parámetros utilizados fueron los que se muestran en la Tabla 25. Parámetros evaluación funciones

Ni	HMS	HMCR	PAR	RCR	RRU	Beta
15000	200	0,95	0,3	0,9	0,3	0,5

Tabla 25. Parámetros evaluación funciones

Los resultados de la evaluación se muestran en la Tabla 26. Resultados comparación funciones objetivo

	Rouge-1	Rouge-2	Rouge-su4
F. GHS+LEM	0,4553343	0,1169584	0,1733990
F. MCMR	0,4416327	0,1023909	0,1599842

Tabla 26. Resultados comparación funciones objetivo

Como podemos observar en la Tabla 26. Resultados comparación funciones objetivo, los mejores resultados se presentan con la función objetivo propuesta, por lo tanto la función seleccionada para la presente investigación es la presentada en la Ecuación 2.

ANEXO D – PRUEBAS DE AFINACIÓN

5 AFINACIÓN PRELIMINAR DE HS, GHS, GHS+LEM

Antes de realizar la afinación con el algoritmo GHSO se realizaron pruebas preliminares con el objetivo de encontrar rangos apropiados para los parámetros para el espacio de búsqueda. A continuación se muestran las pruebas preliminares más significativas que se hicieron para los algoritmos HS, GHS y GHS+LEM en generación de resúmenes de múltiples documentos, todas la pruebas se evalúan con ROUGE-2 por ser una de las medidas más utilizadas.

5.1 PRUEBAS ALGORITMO HS

ID Prueba	1		
Nombre prueba	Evaluación para algoritmo HS para múltiples documentos – parámetros variables, similitud hacia el título		
Descripción	En esta prueba se calcula la similitud de cada documento con su respectivo título, se tomaron 30 conjuntos de DUC 2005 y se aplicó el algoritmo HS adaptado a cada uno de los conjuntos, se evaluó cada resumen con ROUGE específicamente ROUGE-2 y finalmente se promedió todos los resultados. También se varió los parámetros de HMS y HMCR, para encontrar la mejor combinación.		
Medida de ROUGE	ROUGE-2		
Parámetros	**N° Iteraciones**	**HMCR**	**HMS**
	5000	Desde 0,5 hasta 1 con incrementos de 0,05	70, 80, 90, 100
Beta	0.5		
Mejores resultados por HMS	HMS	HMCR	ROUGE-2
	70	0,35	0,03401
	80	0,4	0,0345
	90	**0,3**	**0,0364**
	100	0,5	0,03509
Análisis	Los mejores resultados se presentan con la combinación		

	Hms de 90 y Hmcr de 0,3 obteniendo un promedio de ROUGE-2 de 0,0364, como podemos observar estos valores son muy bajos comparados con los planteados por Aliquliyev R. [2].

Tabla 27. Prueba 2 HS, similitud con el título variación, HMCR y HMS

ID Prueba	2		
Nombre prueba	Evaluación para algoritmo HS para múltiples documentos – parámetros variables, umbral de similitud con todos los documentos		
Descripción	En esta prueba se tomaron 30 conjuntos de DUC 2005 y se aplicó el algoritmo HS adaptado a cada uno de los conjuntos, se evaluó cada resumen con ROUGE específicamente ROUGE-2 y finalmente se promedió todos los resultados. Se tiene en cuenta cobertura y redundancia. También se hace una variación al algoritmo, la cual consiste en eliminar todas aquellas oraciones que están por debajo de un umbral establecido (0.1), para encontrar este valor se calculo la similitud de todas las oraciones contra el documento y se tomo valor medio de similitud de oraciones con el documento.		
Medida de ROUGE	ROUGE-2		
Parámetros	**N° Iteraciones**	**HMCR**	**HMS**
	5000	Desde 0,5 hasta 1 con incrementos de 0,05	70, 100
Beta	0.5		
Mejores	HMS	HMCR	ROUGE-2

resultados por HMS	70	0,95	0,05926
	100	**0,95**	**0,059558**
Análisis	Los mejores resultados se presentan con la combinación HMS de 100 y HMCR de 0.95 los cuales dan un promedio de 0.059558, como podemos observar esta prueba arroja mejores resultados que la prueba 1 por consiguiente se hace necesario afinar el umbral para la selección de las oraciones.		

Tabla 28. Prueba 2 HS utilizando umbral de similitud

5.2 Afinación umbral de similitud

ID Prueba	3		
Nombre prueba	Afinación del umbral para selección de oraciones para el resumen.		
Descripción	En esta prueba se tomaron 30 conjuntos de DUC 2005 y se aplicó el algoritmo HS adaptado a cada uno de los conjuntos, se evaluó cada resumen con ROUGE específicamente ROUGE-2. Se establece un umbral variable desde 0.03 a 0.2 con incrementos de 0.01, cada conjunto se evaluó con cada uno de estos umbrales y finalmente se calculó el promedio de todos los conjuntos de documentos, para así encontrar el umbral que arroje mejores resultados		
Medida de ROUGE	ROUGE-2		
Parámetros	**N° Iteraciones**	**HMCR**	**HMS**
	5000	0.95	100
Beta	0.5		

Mejores Umbrales	Umbral	ROUGE-2
	0,11	0,0558
	0,13	**0,05683**
	0,18	0,05573
Análisis	Como podemos observar en la gráfica, los mejores promedios de ROUGE-2 se presentaron entre 0.11 y 0.18, específicamente el mejor se presenta con un umbral de 0.13	
Gráfica		

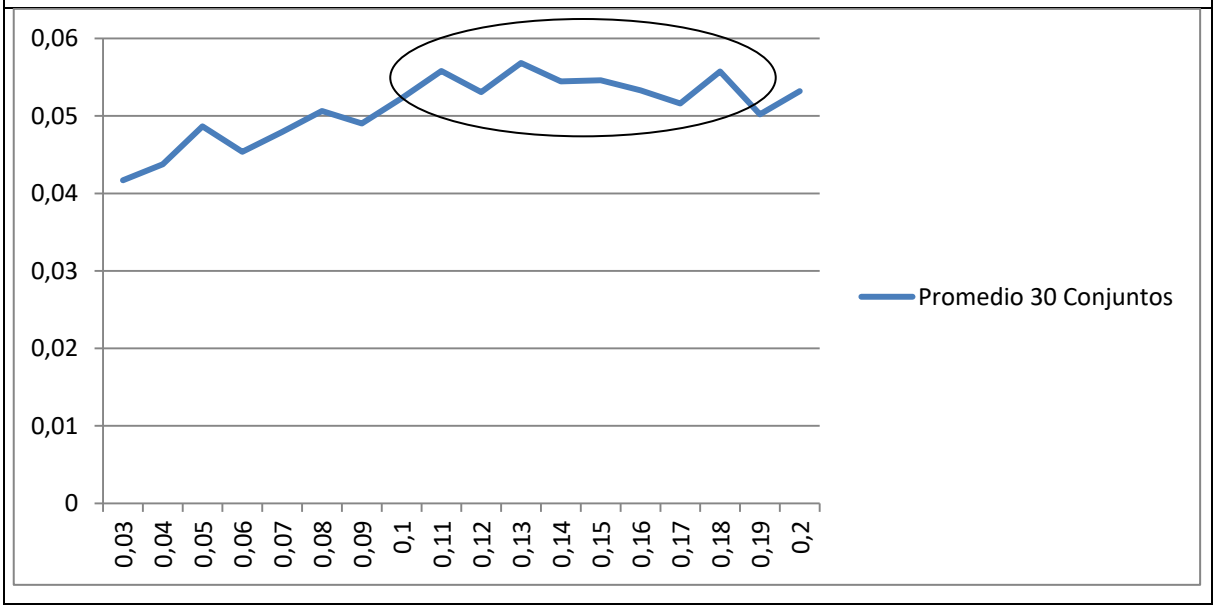

Tabla 29. Afinación umbral de similitud

5.3 Prueba Algoritmo GHS

ID Prueba	4
Nombre prueba	Evaluación para algoritmo GHS para múltiples documentos – parámetros variables.
Descripción	En esta prueba se tomaron 30 conjuntos de DUC 2005 y se aplicó el algoritmo GHS a cada uno de los conjuntos, se evaluó cada resumen con ROUGE específicamente ROUGE-2 y finalmente se promedió todos los resultados. También se varió los parámetros de HMS y HMCR, para encontrar la mejor combinación y se aplicó el umbral de

<table>
<tr><td></td><td colspan="4">selección de oraciones de 0.13 encontrado en la prueba 3</td></tr>
<tr><td>Medida de ROUGE</td><td colspan="4">ROUGE-2</td></tr>
<tr><td rowspan="2">Parámetros</td><td>N° Iteraciones</td><td>HMCR</td><td>HMS</td><td>PAR</td></tr>
<tr><td>5000</td><td>Desde 0,5 hasta 1 con incrementos de 0,05</td><td>70, 80, 90, 100</td><td>0.35</td></tr>
<tr><td>Beta</td><td colspan="4">0.5</td></tr>
<tr><td rowspan="5">Mejores resultados por HMS</td><td>HMS</td><td>HMCR</td><td colspan="2">ROUGE-2</td></tr>
<tr><td>70</td><td>0,4</td><td colspan="2">0,057851</td></tr>
<tr><td>80</td><td>1</td><td colspan="2">0,057973</td></tr>
<tr><td>90</td><td>1</td><td colspan="2">0,059749</td></tr>
<tr><td>100</td><td>1</td><td colspan="2">0,063029</td></tr>
<tr><td>Análisis</td><td colspan="4">Como podemos observar la mejor combinación es HMS de 100 y HMCR de 1 con un promedio de 0.063029, si comparamos este valor con el de la prueba 2, estos son mejores, por lo tanto podemos concluir que la aplicación del algoritmo GHS mejora los resultados, también podemos notar que el valor de HMS y HMCR deben ser altos, lo que nos lleva a concluir que sería conveniente utilizar un valor de HMS mayor.</td></tr>
</table>

Tabla 30. Prueba 4 GHS variación de HMCR y HMS

5.4 Prueba Algoritmo GHS+LEM

ID Prueba	5
Nombre prueba	Evaluación para algoritmo GHS+LEM para múltiples
Descripción	En esta prueba se tomaron 30 conjuntos de DUC 2005 y se aplicó el algoritmo GHS+LEM a cada uno de los conjuntos, se

<table>
<tr><td></td><td colspan="7">evaluó cada resumen con ROUGE específicamente ROUGE-2 y finalmente se promedió todos los resultados.
También se variaron los parámetros de HMS y HMCR, para encontrar la mejor combinación y se aplicó el umbral de selección de oraciones de 0.13 , los valores Par, Rcr, Rru y Hlgs fueron iguales a los utilizados por Calor C. [3]</td></tr>
<tr><td>Medida de ROUGE</td><td colspan="7">ROUGE-2</td></tr>
<tr><td rowspan="2">Parámetros</td><td>Ni</td><td>HMS</td><td>HMCR</td><td>PAR</td><td>RCR</td><td>RRU</td><td>%HLGS</td></tr>
<tr><td>5000</td><td>70, 80, 90, 100</td><td>0,5-1</td><td>0,35</td><td>0,9</td><td>0,3</td><td>40</td></tr>
<tr><td>Beta</td><td colspan="7">0.5</td></tr>
<tr><td rowspan="5">Mejores resultados por HMS</td><td colspan="2">HMS</td><td colspan="2">HMCR</td><td colspan="3">ROUGE-2</td></tr>
<tr><td colspan="2">70</td><td colspan="2">0,95</td><td colspan="3">0,066756333</td></tr>
<tr><td colspan="2">80</td><td colspan="2">0,95</td><td colspan="3">0,068803</td></tr>
<tr><td colspan="2">90</td><td colspan="2">0,95</td><td colspan="3">0,068809</td></tr>
<tr><td colspan="2">100</td><td colspan="2">0,95</td><td colspan="3">0,068047</td></tr>
<tr><td>Análisis</td><td colspan="7">Como podemos observar la mejor combinación es HMS de 90 y HMCR de 0.95 con un promedio de 0,068809, si comparamos este valor con el de la prueba 4, este es mejor, por lo tanto podemos concluir que la aplicación del algoritmo GHS+LEM mejora los resultados</td></tr>
</table>

Tabla 31. Prueba 5 GHS+LEM variación de HMCR y HMS

5.5 Afinación de Parámetros

Para encontrar la mejor combinación de parámetros para los algoritmos HS y GHS+LEM, se tienen dos opciones, la primera es tomar todos los parámetros y hacer todas las posibles combinaciones, pero esta operación tardaría mucho tiempo de ejecución teniendo en cuenta que el espacio de búsqueda es bastante amplio, la segunda opción es aplicar el algoritmo de optimización GHS y encontrar un optimo global. Para la afinación de los parámetros de los algoritmos HS y GHS+LEM se decidió tomar la segunda opción la cual denominamos GHSO (mejor búsqueda armónica global para optimización), los parámetros usados para este algoritmo fueron los más comúnmente usados [4].
La afinación consistió en tomar cada parámetro de HS y GHS+LEM como variables del vector armónico, es decir: $X_i = ((NI)_i, (HMS)_i, (HMCR)_i, \dots)$. Donde X_i es un vector armónico.
Las variables de un vector armónico para la afinación de los parámetros de HS son las que se muestran en la Tabla 32.

Ni	*Hms*	*Hmcr*	*Alfa*	*Beta*	*Gamma*	*F(HS)*

Tabla 32. Vector armónico afinación de parámetros HS

Donde *F(HS)* es el valor de ROUGE-2 obtenido al ejecutar el algoritmo HS para generación de resúmenes de múltiples documentos con los parámetros establecidos en el vector armónico de afinación de parámetros.

Las variables de un vector armónico para afinación de parámetros de GHS+LEM son los que se muestran en la Tabla 33.

Ni	*Hms*	*Hmcr*	*Par*	*Rcr*	*Rru*	*Hlgs*	*Alfa*	*Beta*	*Gamma*	*F(GHS+LEM)*

Tabla 33. Vector armónico afinación parámetros GHS+LEM

Donde *F(GHS+LEM)* es el valor de ROUGE-2 obtenido al ejecutar el algoritmo GHS+LEM para generación de resúmenes de múltiples documentos con los parámetros establecidos en el vector armónico de afinación de parámetros.

Para la afinación se tomaron de forma aleatoria 30 conjuntos de DUC2007, el cálculo del fitness de una armonía consistió en ejecutar cinco veces el algoritmo GHS+LEM con lo que se obtuvieron cinco medidas de ROUGE-2 por cada conjunto de documentos y se promediaron los resultados. A cada conjunto se le realizó el mismo proceso y al final se promediaron los resultados de los 30 conjuntos, el cual representa el fitness de la armonía actual, en la

Figura **8** se muestra un ejemplo para dos conjuntos de documentos.

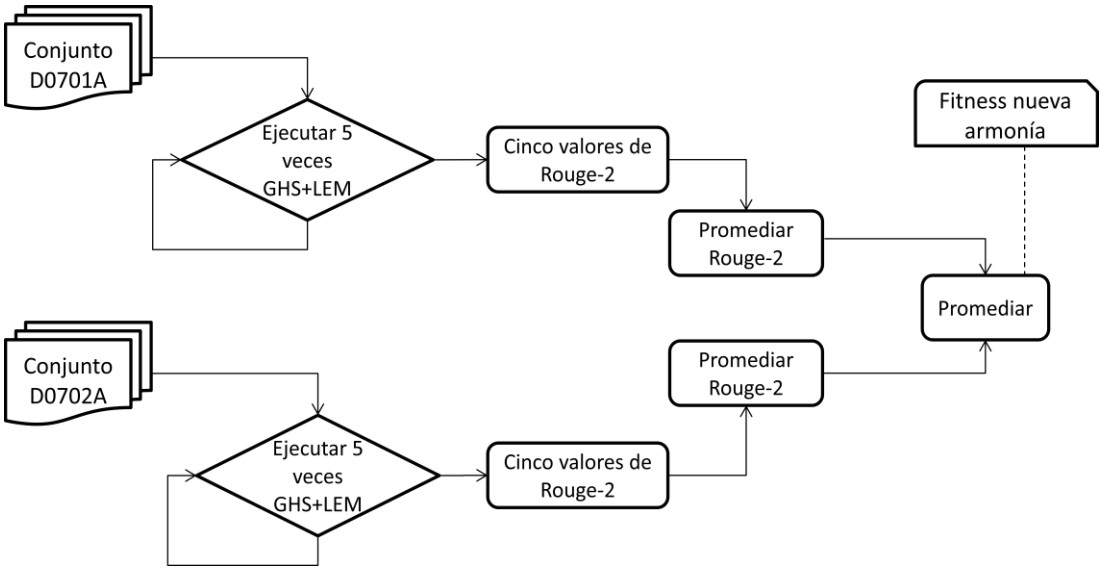

Figura 8. Proceso calculo de fitness de una armonía

5.5.1 Afinación Parámetros HS

Los parámetros que se afinaron para este algoritmo fueron:

- Número de iteraciones (NI): Basados en [1] se partió de 1500 iteraciones hasta 10000, con incrementos de 500 iteraciones, es decir, {1500, 2000, …, 10000}.
- Tamaño de la memoria armónica (HM): Para este valor se partió de un tamaño de 100 con incrementos de 100 hasta llegar a 400.
- Tasa de consideración de la memoria armónica (HMCR): Partiendo de los resultados obtenidos en los trabajos más relevantes para ésta investigación [1, 3, 5], este parámetro puede tomar los valores de {0.9, 0.93, 0,95, 0.98}
- Coeficiente beta: Se varía de cero a uno. Esta variación se realiza de manera similar a la realizada en el trabajo de Sharegui. E. [1].

5.5.1.1 Resultados de la afinación

La memoria armónica inicial se muestra en la Tabla 34 después de 500 iteraciones del algoritmo llegamos a una memoria armónica como la que se muestra en la Tabla 35 donde se puede ver que la mejor combinación de parámetros para HS es la siguiente:

NI	HMS	HMCR	Beta
6000	300	0.98	0.6

5.5.2 Afinación Parámetros GHS+LEM

Los parámetros afinados para GHS+LEM fueron:

- Número de iteraciones (NI): basados en [1] se partió de 1500 iteraciones hasta 10000 con incrementos de 500 iteraciones es decir {1500, 2000, …, 10000}.
- Tamaño de la memoria armónica (HM): Para este valor se partió de un tamaño de 100 con incrementos de 100 hasta llegar a 400.
- Tasa de consideración de la memoria armónica (HMCR): Partiendo de [1, 3, 5], este parámetro podía tomar los valores de {0.9, 0.93, 0,95, 0.98}

- Tasa de ajuste del tono (PAR): Partiendo de [3, 4] los valores que podía tomar esta parámetro fueron {0.1, 0.3, 0.5, 0.7, 0.9}
- Tasa de consideración de reglas (RCR): basados en [3] este parámetro podía variar de la siguiente manera { 0.5, 0.6, 0.7, 0.8, 0.9}
- Tasa de actualización de reglas (RRU): partiendo de [3] este parámetro podía tomar los siguientes valores {0.1, 0.2, 0.3, 0.4, 0.5}.
- Tamaño de los grupos de alto y bajo rendimiento: este es un nuevo parámetro el cual se decidió afinar considerando que es importante saber cuál es el tamaño más adecuado de los grupos de alto y bajo rendimiento, para afinar este valor se decidió variar la distancia que puede existir entre los dos grupos de 0% y 33%, el criterio que se escogió para decidir que el porcentaje máximo puede ser 33% es porque es la tercera parte de la memoria armónica y un valor más grande dejaría pocos datos para realizar un buen análisis para la creación de las reglas.
- El coeficiente Beta se varió de cero a uno igual que en [1].

5.5.2.1 Resultados de la afinación

La memoria armónica inicial se muestra en la Tabla 36 y después de 500 iteraciones del algoritmo GHSO se llega a una memoria como la que se muestra en la Tabla 37, como se puede observar la mejor combinación de parámetros es la siguiente:

NI	HMS	HMCR	PAR	RCR	RRU	%HLGS	Beta
6000	300	0,98	0,7	0,7	0,3	47	0,3

Al algoritmo GHS no se le realizó afinación de parámetros, debido a que no estaba dentro del alcance de este proyecto, por consiguiente los parámetros que se utilizaron para la evaluación fueron los encontrados para GHS+LEM.

NI	HMS	HMCR	Beta	Rouge-2
8000	200	0,96	0,2	0,1006095
6000	400	0,93	0,6	0,10268725
4000	300	0,96	0,3	0,0921564
6000	200	0,96	0,6	0,10180275
6000	300	0,96	0,2	0,099348
8000	100	0,98	0,6	0,0989435
6000	300	0,98	0,6	0,10068725
4000	200	0,96	0,6	0,09986075
6000	200	0,96	0,6	0,10180275
5000	200	0,98	0,3	0,10012375

Tabla 34. Memoria armónica Inicial para HS

NI	HMS	HMCR	Beta	Rouge-2
6000	300	0,98	0,6	0,104687
6000	300	0,98	0,6	0,104687
6000	300	0,98	0,6	0,104687
6000	300	0,98	0,6	0,104687
6000	300	0,98	0,6	0,104687
6000	300	0,98	0,6	0,104687
6000	300	0,98	0,6	0,104687
6000	300	0,98	0,6	0,104687
6000	300	0,98	0,6	0,104687
6000	300	0,98	0,6	0,104687

Tabla 35. Memoria armónica Final para HS

NI	HMS	HMCR	PAR	RCR	RRU	%HLGS	Beta	Rouge-2
4000	100	0,98	0,7	0,8	0,2	49,5	0,5	0,1148250
8000	100	0,98	0,3	0,7	0,2	36,5	0,7	0,1143578
6000	400	0,98	0,7	0,7	0,3	47	0,2	0,1159273
3000	100	0,9	0,7	0,8	0,3	34	0,1	0,1058535
5000	100	0,98	0,5	0,7	0,2	42,5	0,1	0,1093860
5000	200	0,93	0,7	0,8	0,3	50	0,4	0,1079890
4000	100	0,93	0,7	0,7	0,4	47	0,7	0,1063280
3000	300	0,9	0,7	0,8	0,2	49,5	0,7	0,1017650
8000	100	0,98	0,5	0,7	0,4	46,5	0,9	0,1034583
4000	200	0,93	0,5	0,7	0,4	48	0,9	0,1002400

Tabla 36 Memoria armónica inicial para GHS+LEM

NI	HMS	HMCR	PAR	RCR	RRU	%HLGS	Beta	Rouge-2

6000	300	0,98	0,7	0,7	0,3	47	0,3	0,121968
5000	300	0,98	0,6	0,8	0,3	43	0,3	0,120232
5000	300	0,98	0,5	0,7	0,3	47	0,3	0,120232
5000	300	0,98	0,5	0,7	0,3	47	0,3	0,120232
5000	300	0,98	0,7	0,8	0,3	47	0,3	0,118968
6000	300	0,98	0,5	0,8	0,2	47	0,3	0,117802
5000	300	0,98	0,7	0,7	0,3	49,5	0,3	0,117793
5000	400	0,98	0,5	0,8	0,2	47	0,3	0,117564
6000	100	0,98	0,5	0,7	0,3	47	0,3	0,116789
5000	300	0,98	0,5	0,7	0,2	45	0,3	0,116384

Tabla 37 Memoria armónica final para GHS+LEM

Bibliografía

1. Ehsan, S. and H. Leila Sharif, *Text summarization with harmony search algorithm-based sentence extraction*, in *Proceedings of the 5th international conference on Soft computing as transdisciplinary science and technology*. 2008. ACM: Cergy-Pontoise, France.
2. Alguliev, R.M., et al., *MCMR: Maximum coverage and minimum redundant text summarization model.* Expert Systems with Applications, 2011. **In Press, Corrected Proof**.
3. Cobos, C., D. Estupiñan, and J. Pérez, *GHS+LEM: Global-best Harmony Search using learnable evolution models.* Applied Mathematics and Computation. **218**(6): p. 2558-2578.
4. Omran, M.G.H. and M. Mahdavi, *Global-best harmony search.* Applied Mathematics and Computation, 2008. **198**(2): p. 643-656.
5. Mahdavi, M., et al., *Novel meta-heuristic algorithms for clustering web documents.* Applied Mathematics and Computation, 2008. **201**(1-2): p. 441-451.