UNIVERSIDAD AUTÓNOMA DEL ESTADO DE MÉXICO

UNIDAD ACADÉMICA PROFESIONAL TIANGUISTENCO

CARACTERIZACIÓN DE ORACIONES CLAVE DE RESÚMENES MEDIANTE MEDIDAS DE CALIDAD DE AGRUPACIÓN INTERNA

TESIS

PARA OBTENER EL TÍTULO DE

INGENIERO EN SOFTWARE

Q U E P R E S E N T A

NÉSTOR HERNÁNDEZ CASTAÑEDA

ASESOR:

DR. RENÉ ARNULFO GARCÍA HERNÁNDEZ

TIANGUISTENCO, MÉX. JULIO 2017

Resumen

El gran aumento de información digital compartida a través de internet y de otros medios ha hecho necesaria la creación de sistemas que permitan la generación de resúmenes automáticos con el objetivo de presentar a los usuarios la información más relevante del texto o el documento, lo que permite reducir los tiempos de búsqueda y obtención de la información.

Los resúmenes se pueden generar por diversos métodos, pero de forma general se clasifican en dos métodos. Los métodos abstractivos y los métodos extractivos. Estos últimos son los que vamos a utilizar para el propósito de este trabajo.

Existen técnicas de generación de resúmenes extractivos que difieren en la forma de generar el resumen. Algunas de estas técnicas se basan en la selección de frases similares al título del documento, otras por la posición de frases u oraciones en el texto o asignando pesos a las oraciones. Generalmente, estas técnicas de generación de resúmenes son dependientes del idioma o del dominio. Por esta razón se han desarrollado técnicas de generación de resúmenes independientes del idioma y del dominio, estas técnicas también difieren en la forma de generar el resumen. En este trabajo se va estudiar la generación de resúmenes extractivos por agrupamiento ya que existe gran incertidumbre sobre la relación que existe entre la calidad de las agrupaciones generadas y la calidad del resumen obtenido. Debido a que estos resúmenes son generados por agrupamiento obtienen características propias de los grupos, como pueden ser: compactación, separación, distribución y densidad. Por lo que algunos algoritmos de agrupación son incapaces de evaluar características propias de los grupos. Por esta razón en este trabajo se utilizan medidas de calidad interna de agrupación, las cuales mantienen independencia del algoritmo empleado. A través de estas medidas se evalúa la relación que existe entre la calidad de los grupos y la calidad de los resúmenes obtenidos. Además, en

este trabajo se hace un estudio para saber cómo afectan las características de los grupos en la calidad de la agrupación. A través de los experimentos realizados se determina que dos medidas de calidad interna de agrupación pueden evaluar correctamente la relación entre la calidad de los grupos generados con la calidad de los resúmenes utilizados, así como las características de los grupos que son: separación, compactación, ruido, densidad y distribución. Estas medidas son el índice Silhouette y el índice Davies Bouldin.

Índice General

Índice de fórmulas

Índice de tablas

Índice de gráficas y figuras

Capítulo 1:
Introducción

Día con día son generados grandes volúmenes de información digital que es compartida a través de internet. Un usuario promedio busca información específica a través de motores de búsqueda como google para encontrar la información deseada, pero antes de decidir si esta información es la que necesita, el usuario tiene que leer todo el texto ya sea de una página web o de un documento para decidir si efectivamente, esa es la información que él estaba buscando. En caso contario repetirá este proceso hasta encontrar la información deseada.

Una forma de resolver este problema es mediante un sistema que permita obtener las ideas principales de un documento, con el objetivo de obtener la información más relevante del documento para tener una idea general del mismo; es decir un resumen [Soto, 2009].

La generación automática de resúmenes consiste en extraer las ideas principales de un texto con el objetivo de presentar al usuario un texto más breve, pero sin perder su significado original [Ledeneva, 2014] [Soto, 2009].

De acuerdo con [Maña, 2003], en función del nivel lingüístico, las técnicas de generación automática de resúmenes se clasifican en dos: Extractivas y Abstractivas.

Los resúmenes abstractivos se basan en el análisis semántico de documentos haciendo enfoque a las frases que hay en él. Este tipo de análisis tiene mayor profundidad del análisis extractivo, y se conjuga con técnicas de revisión del lenguaje natural, para generar el resumen [Maña, 2003],

Por otra parte, los resúmenes extractivos son el producto de un análisis superficial del texto; llegando solamente al nivel sintáctico. Como resultado se tiene un

resumen generado a partir de los elementos más importantes del texto, por ejemplo: palabras, oraciones o párrafos; por mencionar algunos [Maña, 2003]. Un resumen extractivo de un documento debe presentar las ideas principales del texto utilizando solo fragmentos originales del autor [Soto, 2009].

El principal problema para generar un resumen extractivo es detectar la información más relevante en el documento de origen [García, 2008].

Existen diversos trabajos de investigación que se enfocan a este problema. La mayoría de estos trabajos resuelven el problema basándose en frases clave y en la posición de las frases dentro del texto [Acero, 2001]. Otros trabajos miden la frecuencia de las palabras e índices estadísticos como el trabajo de Iria Cunha [Cunha, 2007], y finalmente, en el trabajo de Fu Lee Wang [Lee, 2006] donde asigna *n* grados de importancia a las oraciones.

Generalmente, todos los trabajos antes mencionados son resúmenes extractivos dependientes del idioma o del dominio, por lo que no pueden ser adaptados a otros idiomas o dominios.

Otras técnicas de generación de resúmenes extractivos que son independientes del dominio y el lenguaje se presentan en los siguientes trabajos, por ejemplo, Villatoro [Villatoro, 2006] donde se extraen todas las secuencias de *n* palabras (n-gramas) del documento las cuales son representadas como características de un modelo. Ledeneva [Ledeneva, 2014] extrae todas las secuencias de gramas utilizando un modelo de secuencias frecuentes maximales. Rosales [Rosales, 2007] extrae todos los n-gramas del documento y construye párrafos con los términos que se consideran más representativos del documento.

Recientemente se han probado métodos mediante aprendizaje no supervisado que no son dependientes del lenguaje ni del dominio, como el trabajo de García [García, 2008], donde a través de un algoritmo de agrupamiento se generan *n* grupos de oraciones basados en la estructura y frecuencia de las palabras. A partir de los grupos formados por oraciones se toma la oración más representativa del grupo para formar parte del resumen.

A pesar de todos los avances presentados sigue habiendo una gran brecha que investigar. Por un lado todos los resúmenes generados en estos trabajos son evaluados con medidas de calidad externas como F-measure, donde se ha alcanzado cerca del 50%. Por otro lado, debido a que los trabajos de resúmenes basados en agrupamiento sugieren que el agrupamiento puede recuperar las ideas importantes de un documento, no se ha estudiado la correspondencia que existe entre los diferentes agrupamientos que se pueden formar y la calidad de los resúmenes generados.

Para comprobar la relevancia que presentan los trabajos de generación automática de resúmenes se utilizan como referencia tres líneas base. Estas líneas base son: línea tope (*top line*), línea base de primeras oraciones (*first*) y línea base de oraciones aleatorias (*random*).

Cabe mencionar que la línea tope, es obtenida a través de la muestra de oro, es decir, a partir de las oraciones que el humano selecciona para generar sus resúmenes. La línea base de primeras oraciones consiste en extraer de los documentos las primeras *n* oraciones, con el cual se obtienen buenos resúmenes. Por último, la línea de base oraciones aleatorias consiste en seleccionar oraciones aleatorias del documento para obtener un resumen [Villatoro, 2006], [Ledeneva, 2014], [Sidorov, 2013].

Para evaluar la calidad de los grupos de oraciones reconstruidos a partir de oraciones clave de resúmenes usando cada línea base es posible usar medidas de calidad interna de agrupación. Cabe recordar que las medidas de calidad interna de agrupación no necesitan tener conocimiento previo de los datos a evaluar; se basan en la medición de la cohesión interna y la separación entre otros grupos. Es decir, las medidas de calidad interna de agrupación evalúan qué tanto se parecen los elementos de un mismo grupo y qué tan diferentes son los elementos de un grupo con los elementos de otros grupos [Rendón, 2011].

Las medidas de calidad internas de agrupación no necesitan de clases previamente definidas para poder evaluar agrupaciones por lo que se consideran independientes del idioma y del dominio.

Ejemplos de estas medidas internas de calidad son: Silhouette, Dunn y Davies Bouldin [Xiong, 2013] [Bolshakova et al, 2002], [Rendón, 2011], solo por mencionar algunas.

En esta investigación se requiere saber si las evaluaciones de las medidas de calidad interna de agrupación mantienen una relación entre la calidad de los grupos generados con base en oraciones clave y la calidad de los resúmenes usados; lo que va a permitir saber cuáles de las medidas pueden servir para las tareas de generación de resúmenes extractivos por agrupación.

1.1 Planteamiento del problema

¿Es posible evaluar la calidad de los resúmenes generados mediante las características del agrupamiento al ser evaluados mediante medidas internas de calidad de agrupación?, incluyendo características como:

- Ruido
- Homogeneidad
- Separación
- Compactación
- Distribución
- Densidad

1.2 Justificación

La necesidad de evaluar la calidad de las oraciones que conforman un resumen independientemente del idioma y del dominio con el que se esté trabajando ha culminado para este trabajo en el uso de medidas internas de calidad de agrupación. Sin embargo, es necesario saber si estas medidas evalúan de forma correcta estos grupos de oraciones candidatas a resumen. Ya que por un lado, las ideas de un documento son redactadas en oraciones; el agrupamiento de oraciones permite detectar las ideas principales de oraciones que conforman un resumen. Y por otro

lado, las medidas internas de calidad de agrupación permiten evaluar la calidad de los grupos de oraciones, por lo tanto las medidas internas de calidad de agrupación deben guardar una relación coherente con la calidad de los resúmenes.

1.3 Hipótesis

Si las medidas de calidad interna de agrupación permiten evaluar cuando un agrupamiento está mejor formado que otro, entonces debería haber una relación con la calidad de los resúmenes que se obtienen a partir de dichos agrupamientos, lo cual se puede comprobar con las líneas base de las colecciones de documentos DUC-2001 y DUC-2002 utilizadas para la generación de resúmenes.

1.4 Alcances y limitaciones

El presente trabajo no busca generar resúmenes extractivos. Se basa en el estudio del comportamiento que tienen oraciones clave de resúmenes al ser evaluadas mediante medidas internas de calidad de agrupación. Basándonos en la hipótesis que una oración clave de un resumen como centro de grupo puede agrupar ideas similares a ella y por consecuencia de acuerdo a la línea base usada debe haber una relación en la calidad de los grupos de oraciones y las evaluaciones generadas por cada una de las medidas de calidad internas.

Algunas limitantes a este trabajo son:

- Necesidad de colecciones de documentos que contengan muestra de oro o de colecciones de documentos que contengan evaluadores humanos.
- Hacer las medidas de calidad internas de agrupación dependientes del idioma y del dominio.

1.5 Objetivos de la tesis

Objetivo general

Caracterizar grupos basados en oraciones clave de resúmenes pertenecientes a tres líneas base obtenidas del corpus DUC-2002 y DUC-2001 al ser evaluados mediante medidas internas de calidad de agrupación.

Objetivos específicos

- Definir qué medida de calidad interna de agrupación es la que puede evaluar correctamente grupos de oraciones generados mediante un algoritmo de aprendizaje no supervisado.
- Conocer las características de los grupos de oraciones que evalúan las medidas internas de calidad, con el fin de posteriormente, caracterizar, adaptar y modelar estas oraciones de acuerdo a los requerimientos de cada medida interna de calidad.
- Obtener un modelo de referencia que permita evaluar conjuntos de oraciones candidatas a resumen generados de forma automática con conjuntos de oraciones seleccionadas por el humano.

1.6 Estructura de la tesis

La estructura de esta tesis es la siguiente:
En el presente capítulo se hace una breve introducción al problema que aborda esta tesis, se incluyen conceptos esenciales para comprender la descripción del problema descrito en este capítulo así como los métodos que se han usado para la tarea de agrupamiento y las métricas para evaluar la calidad de las agrupaciones.

A continuación se describe el resto de capítulos que conforman esta tesis.
En el capítulo 2 se dan a conocer los conceptos necesarios para abordar el problema presentado en esta tesis. Se manejan conceptos generales de minería de

texto como son el preprocesamiento, eliminación de *stop words* y lematización. También se definen modelos de espacio vectorial, modelos de representación de texto y pesado de términos.

Finalmente se presentan conceptos de agrupamiento y medidas de calidad de agrupaciones.

En el capítulo 3 es el estado del arte y se describen los trabajos relacionados con el problema que se desea resolver en esta tesis. De cada trabajo se presentan las principales características que se usaron para resolver el problema propuesto.

En el capítulo 4 se explica el método propuesto para la solución de nuestro problema planteado, incluyendo los distintos índices de validación internos de agrupaciones que vamos a utilizar, el corpus con el cual vamos a trabajar, los distintos modelos de representación de texto y el procedimiento para ocupar cada uno de los elementos mencionados.

En el capítulo 5 se presentan los experimentos correspondientes al proceso del método propuesto. Los experimentos están divididos en experimentos sin preprocesamiento y experimentos con preprocesamiento, se usan los modelos de texto descritos en el marco teórico, así como tres pesados de términos, también descritos en el capítulo ese capítulo.

El capítulo 6 muestra las conclusiones de este trabajo de tesis, se incluyen las aportaciones realizadas con este trabajo de investigación y el trabajo futuro.

Capítulo 2:
Marco teórico

En el presente capítulo se presentan conceptos fundamentales que permiten comprender el entorno sobre el cual se desarrolla este trabajo de investigación. Inicialmente, se describen las etapas de la minería de texto sobre las cuales se basa este trabajo, así como también el preprocesamiento de los datos (*stemming,* lematización).

Posteriormente se describen los diferentes modelos de texto usados para la representación de texto, pesado de términos, así como también los conceptos de agrupamiento*,* líneas base y finalmente, se describen las medidas de calidad internas usadas para evaluar la calidad de las agrupaciones generadas

2.1 Minería de texto

La minería de texto busca encontrar patrones en texto de lenguaje natural, y se puede definir como el proceso de análisis de texto para extraer información útil para propósitos particulares [Waikato, 2002].

2.2 Etapas de la minería de texto

De acuerdo con la tesis de doctorado de Manuel Montes y Gómez [Montes y Gómez, 2002] las etapas de la minería de texto son las siguientes:

- Textos

- Preprocesamiento
- Representación intermedia
- Descubrimiento
- Conocimiento

2.2.1 Etapa de preprocesamiento

Consiste en transformar el texto en una forma estructurada o semiestructurada, en este proceso se obtienen representaciones intermedias del texto. Estas deben ser sencillas para facilitar el análisis del texto, pero también, completas, completas para permitir el descubrimiento de patrones interesantes o en algunos casos nuevos patrones de aprendizaje.

2.2.2 Eliminación de palabras vacías (stopwords)

Son palabras demasiado frecuentes, aparecen en el 80% de un texto y son poco relevantes, por lo general son preposiciones, conjunciones, artículos, etc. Se consideran palabras vacías y son quitadas del texto para evitar que sean importantes [Moreira, 2002].

2.2.3 Lematización

Busca minimizar una palabra a su lema raíz, la palabra raíz puede carecer de significado y el objetivo es reducir la información lingüística. Al hacer esto, las diferentes formas que puede adoptar una palabra quedan reducidas a una forma común, lo cual es llamada *stem o lema*.

El *stem o lema* es la palabra resultante después de liminar sus afijos. Por ejemplo la palabra **perro, perros, perritos, perrotes** quedaría acotada como **perr** [kryscia, 2007].

A diferencia de la *lematización, el stemming* es un proceso de acotamiento que gestiona de manera automática las formas de una palabra. Un algoritmo de *stemming* busca pseudo sufijos de una palabra de acuerdo a la terminación de la palabra y crea una psuedo raíz. Por ejemplo la palabra **pago o pagar,** quedaría reducida a **pag**, el problema de este acotamiento es que la pseudo raíz, también funciona para las palabras **pagano, página o pagoda,** a esto se le conoce como ruido [Valderrábanos, 2004].

El algoritmo Porter es el más utilizado para lematización en el idioma inglés [Peinado, 2003].

Existen algoritmos de stemming para otros idiomas como el español, el griego, el holandés, entre otros idiomas; pero en general todos se basan en reglas sencillas que buscan acotar una palabra para obtener su raíz común [Deco, 2007].

En la actualidad existen varios algoritmos de lematización con los que se puede trabajar. Los algoritmos de stemming son un intento de lematización, la única diferencia es que un algoritmo de stemming solo trunca la palabra, mientras que el algoritmo de lematización, busca la palabra raíz exacta de cada palabra. A continuación se describen algunos algoritmos de lematización de acuerdo con el autor Jesús Peinado [Peinado, 2003]:

A. Lematizador S: Es un lematizador simple en el que las terminaciones “as”, “os” y “s”, son removidas sin excepción. Es muy conservativo, además, es difícil saber dónde se realiza el corte.
B. Lematizador de Lovins: Fue desarrollado por Jule Beth Lovins en 1968 y utiliza un algoritmo único el cual maneja varias listas de excepciones lo que permite remover más de 260 terminaciones.
C. Lematizador de Porter: Es un lematizador línea secuencial considerado como uno de los mejores. Este lematizador remueve más de 60 terminaciones incluyendo terminaciones cortas en 5 pasos. Cada paso da como resultado la remoción de un término o la transformación de la palabra raíz.

2.2.4 Etapa de descubrimiento

En la minería de texto los descubrimientos se clasifican de dos tipos: descriptivos y predictivos, y de ellos se derivan tres enfoques: descubrimientos a nivel representación, descubrimientos a nivel texto, y descubrimientos a nivel mundo. Montes y Gómez [Montes y Gómez, 2002] los explican de la siguiente manera:

a) Descubrimientos a nivel representación.

En este enfoque los métodos buscan construir o descubrir una representación estructurada de los textos.

b) Descubrimientos a nivel texto.

Hay dos tipos de métodos en este enfoque: métodos que descubren patrones a partir de una colección de textos, y métodos que descubren una organización oculta de una colección de textos.

c) En los descubrimientos a nivel texto.

Existen dos tipos de métodos: Identificación de patrones de texto y clasificación de textos.

- Identificación de patrones de lenguaje:

  Los métodos de esta categoría tienen las siguientes características en común:

  1. Consideran todas las palabras del texto y su orden relativo.

  2. Intentan aplicar la máxima cantidad de técnicas provenientes de la minería de datos.

  Estos métodos detectan secuencias frecuentes de palabras, de las cuales en algunas ocasiones se pueden generar reglas asociativas para generar combinaciones de palabras de uso común [Montes y Gómez, 2002].

- Agrupación de textos

Las siguientes características, son propias del agrupamiento de texto:

1. Las medidas de similitud son diversas. Estas van desde la distancia euclidiana entre dos textos, hasta medidas sofisticadas basadas en redes neuronales de tipos mapas auto -organizados.

2. En los resultados se enfatiza la interpretación y visualización de los resultados, por ejemplo, los métodos que implementan interfaces gráficas para poder analizar los resultados o los métodos que asignan una etiqueta para poder clasificar los resultados.

d) Descubrimiento a nivel mundo

Este enfoque considera tareas tales como: análisis de tendencias y descubrimiento de asociaciones. Los métodos correspondientes a este enfoque, tienen las siguientes características en común:

1. Utilizan dos tipos de representaciones de los textos, a nivel documento y a nivel concepto.

2. Utilizan conocimiento de dominio, los cuales son expresados en jerarquías de conceptos.

3. Permite que el usuario guie el proceso, eligiendo las áreas de búsqueda y los conceptos de mayor interés.

2.3 Modelos de representación de textos

Los modelos de texto se basan en técnicas de extracción de términos. La diferencia entre un modelo de texto y otro es la técnica usada para extraer los términos del

documento. Los términos extraídos del documento son convertidos en patrones. Los patrones resultantes pueden ser utilizados para diferentes propósitos que van desde ser analizados, hasta utilizarlos con técnicas de pesados de términos. Estos modelos permiten evaluar la similitud entre documentos mediante técnicas de álgebra lineal como la suma y la multiplicación [Salton et al., 1975].

2.3.1 Selección de términos

La selección de un subconjunto de palabras para representar documentos es conocida como selección de características o selección de términos. Existen diferentes técnicas para realizar la selección de términos, la más sencilla descarta términos poco discriminantes. Un término poco discriminante es aquel que proporciona poca o ninguna información acerca del contenido del documento, por ejemplo las preposiciones, conjunciones y artículos. Estas palabras, conocidas como palabras vacías (en inglés “Stop Words”), se pueden descartar del conjunto de términos y de esta manera se reduce considerablemente la dimensión del vector [Van Rijsbergen, 1979; Hotho & Stumme, 2002; Wang & Hodges, 2006].

2.3.2 Modelo espacio vectorial usando bolsa de palabras

Este modelo de representación de texto consiste en extraer todas las palabras que son diferentes en un documento. Estas palabras comienzan a formar una “bolsa de palabras” en donde las palabras que contiene no guardan el orden ni la posición que tenían anteriormente en el documento de donde fueron extraídas. Todas las palabras que fueron indexadas a la bolsa de palabras son representadas mediante un vector de términos.

2.3.3 Modelo espacio vectorial usando n-gramas

Los n-gramas son una subsecuencias de *n* elementos de una secuencia dada, los elementos de cada n-grama pueden ser palabras o caracteres. Sin embargo, el tamaño de *n* puede ser mayor para problemas particulares.

En este modelo el documento también es representado como un vector de términos en el cual cada término corresponde a un n-grama [Hernández, 2006].

2.4 Pesado de términos

Dentro de los modelos de espacio vectorial existen diferentes formas de pesado de términos, de acuerdo con [Steinbach et al, 2000], existen diversas formas de pesar un término, entre ellas destacan cuatro, las cuales son: ausencia o presencia del término (pesado booleano), frecuencia del término (pesado tf), frecuencia inversa de términos (pesado itf), y pesado tf-idf. Los cuales se describen a continuación.

2.4.1 Pesado booleano

Es la forma más sencilla de pesar un término, ya que se asigna un uno si aparece y un cero de lo contrario, ver fórmula 2.1.

$$P_{i(t_j)} = \begin{cases} \text{1 si aparece en el} \\ \\ \text{0 si no aparece en el} \end{cases}$$

Fórmula 2.1 Pesado de término booleano

Donde $P_{i(t_j)}$: Es el peso del término j en el Documento i

2.4.2 Frecuencia de término (tf)

Fue propuesto por [Luhn, 1957], en este pesado se toma en cuenta la frecuencia con la que aparece un término en el texto, ya que esto puede reflejar mejor la idea del documento, a un término que aparece pocas veces. En este pesado de términos, se asigna un mayor peso a los términos que aparecen con más frecuencia en el documento, los cuales también son evaluados, ver fórmula 2.2.

$$P_{i(t_j)=} F_{ij}$$

Fórmula 2.2 Pesado de término por frecuencia del término

Dónde:F_{ij}, es la frecuencia en que aparece el termino *j* en el documento *i*.

2.4.3 Pesado frecuencia inversa del documento (idf)

Este pesado de término toma en cuenta que un término muy frecuente en varios documentos es menos útil que un término que no es tan frecuente, ya que evalúa la distribución de los términos en el documento. El pesado de frecuencia inversa del documento se define en la siguiente formula, ver fórmula 2.3.

$$P_{i(t_j)=} \boldsymbol{log}\left(\frac{N}{nj}\right)$$

Fórmula 2.3 Pesado de término frecuencia del término menos frecuencia inversa del documento

Dónde:

(t_j) = peso del término j en el documento i

N = número de documentos de la colección

jn = número de documentos en los que el término j aparece.

2.5 Medidas de similitud en patrones

El objetivo de construir un modelo de texto, es permitir usar los patrones generados para medir su similitud con otros patrones. Por este motivo, fue necesario el desarrollo de medidas que permitan medir la similitud entre los patrones que representan a estas oraciones.

Antes del agrupamiento se debe determinar una medida de similitud-distancia. La medida refleja el grado de cercanía o separación de los objetos y debe corresponder a las características que se cree que distinguen a los elementos de cada grupo. En muchos casos, estas características dependen de los datos o el contexto del problema en cuestión, y no hay ninguna medida que sea universalmente mejor para todo tipo de problemas de agrupación [Huang, 2008].

2.5.1 Similitud coseno

Una de las medidas de similitud comúnmente usadas para medir la similitud entre dos patrones representados como vectores de términos es la medida coseno [Manning & Schûtze, 1999], con la cual podemos medir el ángulo que genera dicha comparación. Cuando los dos patrones contienen los mismos términos el valor coseno del ángulo da 1 como resultado, mientras que entre menos términos en común tengan estos patrones, el resultado del coseno del ángulo se acerca a 0 y si los patrones no tienen nada en común el resultado del coseno del ángulo es igual a 0. Cabe aclarar que, aunque el coseno del ángulo da como resultado 1, no quiere decir que los patrones son idénticos, sino que contienen los mismos términos aunque tal vez no en el mismo orden. Esta medida de similitud se representa mediante la fórmula 2.4:

$$Cos(di, dj) \quad = 1 + \frac{d_i^1 . d_j}{\left|\left|di\right|\right| \; ||dj||},$$

Fórmula 2.4 Medida Coseno utilizada para medir la similitud entre patrones.

Además, los vectores de las oraciones pertenecientes a un documento son una unidad de longitud, la fórmula anterior se simplifica a Cos(d_i y d_j)= $d_i^1 . d_j$,

Otro método para calcular la similitud entre dos documentos es el uso de la distancia Euclidiana descrito a continuación:

2.5.2 Distancia euclidiana

De acuerdo con [Huang, 2008] la distancia euclidiana es una métrica estándar para problemas geométricos. Es la distancia ordinaria entre dos puntos y se puede medir fácilmente con una regla en el espacio de dos o tres dimensiones. La distancia euclidiana es ampliamente utilizada en los problemas de agrupamiento, incluyendo el texto agrupado. Para calcular la distancia euclidiana se utiliza la fórmula 2.5.

$$\mathrm{dis}(\mathrm{di}, \mathrm{dj}) = \sqrt{(di - dj)^t}(di - dj) = ||di - dj||$$

Fórmula 2.5 Distancia euclidiana usada para medir la similitud entre patrones.

Si la distancia es cero, entonces los documentos son idénticos, y si no hay nada en común en la distancia de los documentos la distancia es √2 [Zhao, 2005].

2.6 Aprendizaje no supervisado

Dentro del área computacional existe un método llamado como aprendizaje no supervisado, cuya principal característica es que las técnicas basadas en este tipo de aprendizaje no necesitan un conocimiento previo de los datos para poder ser analizados y procesados.

Una de las técnicas de aprendizaje no supervisado es el agrupamiento. El cual puede ser usado para agrupar oraciones [García, 2008].

En el trabajo de García [García, 2008] se encuentra la siguiente definición:

Un algoritmo de aprendizaje no supervisado forma grupos de objetos para lograr, por un lado, la mayor similitud posible entre los objetos de un grupo, por otro lado la mayor disimilitud posible entre objetos de diferentes grupos.

De acuerdo con Lezcano [Lezcano, 2006], el agrupamiento se define como: "Proceso de dividir un conjunto de datos en grupos mutuamente excluyentes de tal manera que cada miembro de un grupo esté lo más cercano posible a otro, y grupos diferentes estén lo más lejos posible uno del otro, donde la distancia está medida con respecto a todas las variables disponibles".

Estos algoritmos tratarán de maximizar las semejanzas entre los objetos de un mismo grupo y maximizar las diferencias entre los objetos en grupos diferentes. Son sumamente utilizados en muchos campos de la ciencia como recuperación de información (categorización de documentos), minería de datos, genética (agrupamiento de genes y proteínas similares), visión por ordenador (segmentación de imágenes), biología, etc. [Jiménez, 2005].

Principalmente se usaban tres tipos de algoritmos para la tarea de agrupamiento: algoritmos jerárquicos, de partición y de optimización. A continuación se describen cada uno de ellos:

2.6.1 Algoritmos jerárquicos

De acuerdo con [Dash & Liu, 2001], estos fueron los primeros algoritmos usados para la tarea de agrupación y pueden ser identificados porque generan una estructura jerárquica o de árbol también llamada dendograma a la hora de generar grupos.

Los algoritmos jerárquicos se caracterizan por agrupar un conjunto de objetos generando una estructura jerárquica, o de árbol, denominada dendograma. Cada nodo hoja del dendograma corresponde a un objeto y la raíz representa el

conjunto completo de objetos. Un corte en algún nivel del dendograma indica un agrupamiento específico [Steinbach et al, 2000].

Los algoritmos jerárquicos se dividen en dos tipos: aglomerativos y divisivos.

Los algoritmos de agrupamiento jerárquico proceden ya sea por una serie de fusiones sucesivas o una serie de divisiones sucesivas. El resultado es la construcción de un árbol estructurado o agrupaciones jerárquicas las cuales pueden ser mostradas como un diagrama llamado dendograma [Han et al, 2001].

Los métodos jerárquicos aglomerativos comienzan con la inserción de cada elemento en un grupo diferente, después cada grupo se fusiona con otro según su similitud (los grupos más similares son fusionados en cada etapa), hasta que solo un grupo se obtiene.

Los métodos jerárquicos divisivos trabajan en forma contraria. Un grupo inicial que contiene todos los objetos es dividido en sub-grupos (basados en disimilitud) hasta que cada objeto queda en un grupo propio.

2.6.2 Algoritmos de partición

A diferencia de los algoritmos jerárquicos, los algoritmos de partición trabajan en un solo nivel. En éste tipo de algoritmos de agrupamiento se especifica previamente el número de grupos a generar y se crea una partición única del conjunto de objetos [Steinbach et al, 2000].

2.6.3 Algoritmos de optimización

Los algoritmos de optimización producen una sola agrupación la cual es optimizada con un criterio predefinido o una función objetivo.

Los algoritmos de optimización comienzan con una partición inicial de objetos dentro de un número de grupos específico. Los objetos son reasignados a los grupos de acuerdo con la función objetivo hasta que se llega a un criterio de finalización. Estos métodos difieren respecto a las particiones de inicio, la función objetivo, el proceso de reasignación, y el criterio de terminación o finalización.

Los métodos de optimización pueden usar particiones iniciales aleatorias o particiones generadas a partir de una semilla [Cole, 1998].

Actualmente también se implementan algoritmos genéticos para la tarea de agrupamiento, los cuales se describen a continuación:

2.6.4 Algoritmos Genéticos

Los Algoritmos Genéticos son métodos adaptativos, generalmente usados en problemas de búsqueda y optimización de parámetros, basados en la reproducción sexual y en el principio de supervivencia del más apto [Rivero, 2010].

En forma simple un algoritmo genético es un proceso iterativo que aplica una serie de operadores genéticos, tales como, selección, cruza y mutación a una población de elementos llamados individuos o cromosomas, representan posibles soluciones al problema, los cuales son seleccionados aleatoriamente del espacio de búsqueda. Los operadores genéticos combinan la información genética de los individuos para formar nuevas generaciones, este proceso es conocido como reproducción. Cada individuo tiene a un valor de aptitud asociado, el cual es evaluado por la función de aptitud para el problema [Valencia, 1997].

2.7 Agrupamiento mediante aprendizaje no supervisado

En el trabajo de García [García, 2009] se propone un enfoque de generación de resúmenes automáticos independiente del lenguaje y dominio usando un algoritmo

de aprendizaje no supervisado, la hipótesis es que si se agrupan ideas similares (oraciones) mediante este algoritmo y posteriormente se selecciona la oraciones más representativa de cada grupo para formar parte de un resumen se obtienen mejores resultados que otros enfoques dependientes del dominio y el lenguaje.

2.8 Línea Base

Es un concepto aplicado al diseño de experimentos. Comúnmente corresponde a un método del estado del arte que resuelve el mismo problema y ha sido aceptado, este método debe ser superado por la tesis propuesta. Normalmente la línea base es un método no muy sofisticado [Sidorov, 2013].

2.8.1 Línea base en resúmenes extractivos: heurística de primeras oraciones (first)

Esta heurística consiste en extraer las primeras oraciones de un documento para formar su resumen. De acuerdo con Ledeneva [Ledeneva, 2014], este método solo trabaja con textos de géneros específicos, por lo que no se podrá trabajar con documentos oficiales como mensajes de correo electrónico, páginas Web o novelas literarias. Sin embargo, aunque es muy sencillo ha resultado tener un alto parecido con el humano, por lo que pocos métodos han logrado superarlo.

2.8.2 Línea base en resúmenes extractivos: oraciones aleatorias (random)

Esta heurística consiste en extraer oraciones aleatorias de un documento para conformar su resumen. La cual se considera una línea base más propia de documentos que no son noticias.

Las líneas base presentadas nos sirven para el propósito de comparar métodos extracción de oraciones automáticos contra la selección de oraciones realizada por humanos. Las líneas base donde se toman en cuenta la selección de oraciones hechas por el humano son las siguientes.

2.8.3 Línea base: Top line

También conocida como línea tope es el valor máximo que puede ser alcanzado por un programa dado la falta de concordancia entre varios evaluadores. Las oraciones pertenecientes al *top line* son obtenidas a partir de la muestra de oro o *gold standard* que donde se encuentran los resúmenes generados por los evaluadores de la colección de documentos. Esta línea base también sirve para modelar las oraciones que el humano elige para crear sus resúmenes.

Las oraciones obtenidas de la línea base *top line* son consideradas para esta tesis oraciones clave de resúmenes, ya que nos permiten modelar el conjunto de oraciones seleccionadas por el humano para generar resúmenes.

Debido a que *top line* contiene el conjunto de oraciones que mejor representan a cada documento de la colección, se espera tener mejores evaluaciones por parte de las medidas de calidad de agrupación usadas en esta línea base, mientras que en las líneas base: primeras oraciones y oraciones aleatorias, se espera obtener evaluaciones de acuerdo a la colección de datos que se ocupe, pero la evaluación de la calidad de los grupos oraciones no debe ser mejor que *top line*.

2.9 Agrupamiento con base en oraciones clave

Para este trabajo se utilizaron algoritmos de agrupación para generar nuestros grupos de oraciones, a estos algoritmos se les hicieron modificaciones para no

hacer recalculo de centroides debido a que cada centroide representa a una oración clave.

2.10 Medidas de calidad de agrupación

Debido a que la tarea de agrupamiento es una tarea de aprendizaje no supervisado, es necesario contar con métricas o técnicas que nos permitan evaluar la calidad de dichas agrupaciones. Las medidas de calidad permiten evaluar la calidad de las agrupaciones generadas por un algoritmo de aprendizaje no supervisado.

Existen dos tipos de medidas para evaluar la calidad del agrupamiento: medidas internas y medidas externas de calidad.

2.10.1 Medidas externas de calidad de agrupación

Estas medidas evalúan la calidad del agrupamiento comparando los grupos obtenidos automáticamente contra clases previamente definidas, usualmente de manera manual. Las medias externas utilizadas en este trabajo son entropía total y F-measure global [Steinbach et al, 2000].

2.10.2 Medidas internas de calidad de agrupación

Las medidas internas de calidad permiten evaluar la calidad del agrupamiento sin tener algún conocimiento externo, se basan en la medición de la cohesión interna, es decir, qué tanto se parecen los documentos de un mismo grupo, y la separación externa, es decir qué tan diferentes son de los documentos de otros grupos. Ejemplo de este tipo de medidas son la similitud global [Steinbach et al, 2000]; y silhouette global [Bolshakova et al, 2002].

El presente trabajo de investigación se enfoca en el uso y estudio de medidas internas de calidad, a continuación se detallan las medidas usadas para este trabajo:

2.10.2.1 Índice Silhoutte

El índice de Silhoutte (S) [Rousseeuw, 1987] valida el rendimiento de agrupación en función de la diferencia distancia por pares de objetos entre y dentro de los grupos. Además, el número de clúster óptimo se determina maximizando el valor de este índice.

Esta métrica define el grado de confianza de que la oración *i* pertenezca al grupo *j* y se define de la siguiente manera, véase fórmula 2.6:

$$S(i) = \frac{AVGD\_BETWEEN(i,k) - AVGD\_WITHIN(i)}{MAX(AVGD\_WITHIN(i),\ AVGD\_BETWEEN(i,k))}$$

Fórmula 2.6 Índice Silhouette utilizado para evaluar la calidad de agrupación de forma intrínseca.

Donde AVGD_BETWEEN (i, k) es la distancia promedio del documento *i* a todos los documentos de los otros grupos, y AVGD_WITHIN (i) es la distancia promedio del documento *i* a los documentos de su mismo grupo. *S (i)* puede tomar valores entre -1 y 1, cuando el valor de s(i) es cercano a 1, se puede inferir que el documento *i* ha sido asignado a un grupo adecuado. Cuando el valor de S(i) es cercano a cero sugiere que el documento pudo haber sido asignado al grupo vecino más cercano o que el documento se encuentra a la misma distancia de ambos grupos. Por otro lado, cuando S(i) es cercano a -1 quiere decir que el documento fue asignado a un grupo incorrecto.

También es posible calcular el valor silhoutte *Sj* de un grupo *Cj* para calcular sus propiedades de heterogeneidad y aislamiento. El cálculo se realiza sumando los valores silhouette s(i) de los documentos del grupo *Cj* como lo indica la siguiente expresión, véase fórmula 2.7:

$$S_j = \sum_{i=1}^{|c_j|} s(i)$$

Fórmula 2.7 Fórmula del índice Silhouette para calcular propiedades de heterogeneidad y aislamiento de un grupo en específico.

Dónde: $|c_j|$= número de documentos en el grupo c_j

Dónde: k= número de grupos.

Nota: Entre mayor sea el valor silhouette, mejor es la calidad del agrupamiento.

2.10.2.2 Índice Dunn

El índice de Dunn [Dunn, 1974] mide la relación entre la distancia más pequeña del grupo y la mayor distancia entre grupos en una partición, véase fórmula 2.8.

$$Dunn = min_{1<i<c}\left\{min\left\{\frac{d\left(c_{i,}\, c_j\right)}{max_{1<k<c}\left(d(x_k)\right)}\right\}\right\}$$

Fórmula 2.8 Índice de Dunn utilizado para evaluar la calidad de agrupación de forma intrínseca.

- D(ci, cj): define la distancia entre grupos ci y cj
- d (Xk) representa la distancia dentro del grupo (Xk)
- c es el número de grupos del conjunto de datos.

Nota: Valores grandes en el índice de Dunn indican buenas agrupaciones

2.10.2.3 Índice Davies Bouldin

El índice de Davies-Bouldin [Davies, 1979] se calcula de la siguiente manera. Para cada grupo C, las similitudes entre C y todos los otros grupos se calculan, y el valor más alto se asigna a C como su similitud de grupo. A continuación, el índice de DB se puede obtener promediando todas las similitudes de clúster. Cuanto menor sea el índice, mejor será el resultado de la agrupación. Al minimizar este índice, los clústeres son los más distintos entre sí y, por lo tanto, lograr la mejor partición. Para evaluar la calidad de una agrupación mediante el índice Davies Bouldin se utiliza la fórmula 2.8.

$$DB = \frac{1}{k} \sum_{i=1, i!=j}^{k} \max\left(\frac{\sigma i + \sigma j}{d(Ci, Cj)}\right)$$

Fórmula 2.9 Índice de Davies Bouldin utilizado para evaluar la calidad de agrupación de forma intrínseca.

Dónde:

- k es el número de grupos.
- σi es la distancia promedio entre cada punto del grupo i y el centroide del grupo.
- σj es la distancia promedio entre cada punto del grupo j y el centroide del grupo.
- Ci, cj es la distancia entre los centroides de los grupos.

Nota: Valores pequeños para el índice DB indica clústeres compactos, y cuyos centros están bien separados los unos de los otros.

2.11 Resumen

En el presente capítulo se dieron a conocer los conceptos necesarios para abordar el problema en cuestión, los conceptos se van presentando de acuerdo a los pasos de la metodología usada para este trabajo la cual es definida al comienzo de este capítulo. Primero se muestran conceptos sobre el procesamiento y modelado de la información, pesados de términos y selección de términos. Posteriormente se presentan términos referentes al agrupamiento y medidas de similitud entre patrones para finalmente concluir con las medidas internas de calidad de agrupación usadas para este trabajo.

Capítulo 3:
Estado del arte

En este capítulo se presentan los trabajos de investigación relacionados con esta tesis. A continuación se explica brevemente en qué consiste cada trabajo de investigación.

3.1 Índices de validación de agrupación internos vs externos

Eréndira Rendón, Itzel Abundez, Alejandra Arizmendi, Elvia M. Quiroz [Rendón, 2011].

En este trabajo se comparan medidas de calidad internas con medidas de calidad externas de validación de agrupaciones para saber cuál de los dos enfoques es capaz de evaluar el número de grupos óptimos de un conjunto de datos sin ninguna información adicional sobre las clases del conjunto de datos. Para generar los grupos se usa el algoritmo k-means, posteriormente se utilizan cuatro medidas de calidad externas de entre ellas las más destacadas son: F-measure y NMI-Measure, así mismo se utilizan cinco medidas de calidad internas de entre ellas las más destacadas son índice Silhouette, índice Dunn, índice Davies bouldin.

La conclusión de este trabajo es que las medidas de calidad internas de validación de agrupación pueden evaluar con mayor precisión el número de grupos óptimos de un conjunto de datos mejor que las medidas de calidad externas.

3.2 Evaluación de las medidas de calidad de agrupación interna

Yanchi Liu, Zhongmou Li, Hui Xiong, Xuedong Gao, Junjie Wu
[Xiong, 2013].

Hui Xong et. al, realizan una investigación sobre características en los datos que pueden afectar el desempeño en la evaluación que realizan medidas de calidad internas de agrupación sobre agrupaciones. En este trabajo se analizan 12 medidas de calidad internas de agrupación, cada una de ellas es sometida a pruebas para comprobar qué tanto afecta su desempeño las siguientes características: ruido, densidad, distribuciones desiguales en los grupos y agrupaciones con formas arbitrarias.

3.3 Comparación de técnicas de agrupación de documentos

[Steinbach et al, 2000]

En este trabajo realizan un estudio experimental sobre algunas de las técnicas de agrupamiento más comunes comparando algoritmos jerárquicos aglomerativos contra algoritmos de partición como k-means y bisecting k-means. La representación que utilizan es el modelo de espacio vectorial usando palabras y en todos los documentos se eliminan las palabras vacías y se realiza truncamiento. Para evaluar la calidad de los agrupamientos obtenidos, con los diferentes algoritmos de agrupamiento, se utilizaron medidas de calidad internas y externas. Sus resultados experimentales demostraron que los algoritmos de partición como k-means y bisecting k-means tienen un mejor desempeño que los algoritmos de agrupamiento aglomerativos, sin embargo k-means es el algoritmo que obtuvo los mejores resultados.

3.4 Agrupamiento de documentos basado en Secuencias Frecuentes Maximales

[Hernández, 2006]

Este trabajo de investigación se agrupan documentos con un modelo de representación de texto que se basa en las secuencias frecuentes maximales de palabras que aparecen en un documento, una secuencia se considera que es frecuente si aparece al menos β veces en el documento o en la colección de documentos, donde β es el umbral de frecuencia dado [Ahonen, 99].

Las principales ventajas de este modelo de texto es que mantiene el orden en el que aparecen las palabas en el texto, además de su representación es compacta.

Se hace una comparación con los modelos de texto n-gramas y bolsa de palabras evaluando los resultados con medidas de calidad externas e internas de agrupación, la representación de secuencias frecuentes maximales obtiene resultados similares al vector de palabras frecuentes y bigramas, superándolos en algunos casos. Sin embargo, el número de términos utilizados con el vector de SFM's es mucho menor que él número de términos utilizados por las otras dos representaciones.

Para realizar los experimentos de este trabajo se hace uso de medidas de calidad internas y externas, las medidas usadas son: silhouette global, similitud global y F-measure.

3.5 Generación automática de resúmenes mediante aprendizaje no supervisado

[Soto et al., 2008]

Este trabajo de investigación consiste en generar grupos de oraciones previamente procesados, los cuales, después del proceso de agrupación y a través de un proceso de extracción de oraciones, pasan a formar parte de un resumen.

Se utilizan tres modelos de texto para el pre procesamiento de las oraciones los cuales son: bolsa de palabras, n-gramas y secuencias frecuentes maximales, estos modelos de texto son combinados con cuatro pesado de términos, que son: pesado

booleano, frecuencia del término, frecuencia inversa del documento y frecuencia inversa del documento menos frecuencia del término.

Concluida la etapa de pre procesamiento se procede a agrupar las oraciones con un algoritmo *K-means.* Los resultados de este proceso de agrupación son evaluados con la medida F (*F-measure*), dando los mejores resultados las siguientes combinaciones: bolsa de palabras - frecuencia de término, n-gramas – frecuencia inversa del documento, secuencias frecuentes maximales – frecuencia inversa del documento.

3.6 Resumen

En este capítulo se presentaron los trabajos relacionados con esta tesis. Primero presentamos las medidas de calidad internas de calidad de agrupación las cuales van a ser utilizadas en esta tesis, posteriormente se hace un estudio de factores que afectan el desempeño de evaluación por parte de estas medidas, después se muestran trabajos donde se han aplicado estas medidas para la evaluación de grupos de documentos y finalmente se aborda el tema de evaluación de oraciones como posibles candidatas a resúmenes por medio de aprendizaje no supervisado.

Capítulo 4:
Método propuesto

En el presente capítulo se describe el método propuesto para la caracterización de oraciones clave de resúmenes mediante medidas internas de calidad de agrupación.

4.1 Descripción del método propuesto

El agrupamiento de oraciones permite detectar las ideas principales de un documento y las medidas de calidad internas de agrupación permiten evaluar la calidad de estos agrupamientos, pero debido a que las medidas internas de calidad de agrupación no están estandarizadas no se puede saber con certeza si la evaluación de estas medidas es correcta. Así mismo se desconoce si a través de estas medidas es posible caracterizar las oraciones que un humano elige para generar un resumen. Debido a esto se utilizan tres líneas base que nos sirven para comparar la relación que existe entre la calidad de los resúmenes generados a partir del agrupamiento y las evaluaciones por parte de las medidas de calidad internas.

En base a esto se construyen grupos de oraciones donde se toma a cada oración clave como representante de un grupo, posteriormente se mide la similitud de cada oración respecto a las oraciones clave del documento, siendo asignada la oración al grupo donde sea más similar con la oración clave. Finalmente, estas agrupaciones son evaluadas con tres medidas internas de calidad de agrupación, las cuales son: índice Davies Bouldin, índice Silhouette, índice Dunn. Las evaluaciones hechas por estos tres índices nos permiten caracterizar los resúmenes generados por medio del agrupamiento a través de tres líneas base las cuales

pertenecen a dos colecciones de documentos que son DUC-2001 y DUC-2002. Estas tres líneas base explicadas anteriormente se forman a partir de heurísticas de selección de oraciones que sirven para conformar el resumen de un documento y de las cuales se sabe la calidad de los resúmenes que generan. Por lo que se espera obtener las mejores evaluaciones en la línea tope (*top line*) mientras que en las otras líneas base (primeras oraciones y oraciones aleatorias) se espera obtener evaluaciones variadas pero no mejores que *top line.*

4.2 Etapas del método propuesto

La figura 4.1 presenta el método propuesto, se usa un diagrama de flujo para mostrar las etapas que conforman este método y se describe cada una de ellas.

4.3 Preprocesamiento

En esta etapa se removieron caracteres que no aportan ningún significado al texto, por ejemplo, llaves, apostrofes, arrobas, conjunciones y artículos.

4.4 Representación del documento

Se usan dos modelos de representación de texto, los cuales son bolsa de palabras y n-gramas, esto se hace con el objetivo de obtener mayor cantidad de resultados y en consecuencia, tener más datos para un análisis posterior de los experimentos.

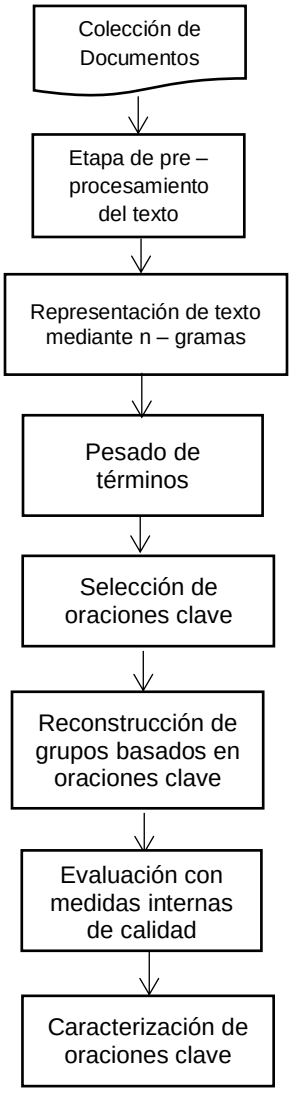


Figura 4.1 Diagrama de etapas del método propuesto

4.5 Pesado de términos

En esta etapa se hace uso de tres pesados de términos que son: frecuencia de término y frecuencia inversa del documento y pesado booleano en conjunto con dos modelos de representación de texto que son: bolsa de palabras y n-gramas; para el modelo n-gramas se van a utilizar bigramas (n=2) y trigramas (n=3) esto con el fin de tener una mejor caracterización de las oraciones clave de resúmenes.

4.6 Reconstrucción de grupos con base en oraciones clave

Una vez que hemos representado los documentos mediante un modelo de representación de texto, el paso siguiente consiste en generar grupos de oraciones. Estos grupos de oraciones son generados mediante un proceso el cual consiste en considerar a cada oración clave como el centro de cada grupo, a partir de los cuales cada oración del texto original se asigna al centro más cercano. Como medida de similitud se usa la distancia euclidiana. Las oraciones más similares con la oración clave se juntan para comenzar a formar grupos. La generación de los grupos se realiza una sola vez, por lo que los centros de cada grupo no cambian.

4.7 Evaluación mediante medidas de calidad internas

Una vez teniendo la reconstrucción de los grupos para cada documento, es necesario saber la calidad que tiene cada agrupación. Para esto se pueden utilizar medidas de calidad internas, que sirven para evaluar la calidad de las agrupaciones sin necesidad de tener un conocimiento de los datos a evaluar. A través de la evaluación de estas medidas se pueden conocer las características que contiene cada agrupamiento en relación con las oraciones que un humano elige para generar un resumen.

Así mismo se espera conocer, cúal o cúales son las medidas capaces de evaluar de forma correcta este tipo de agrupaciones.

4.8 Caracterización de oraciones clave de resúmenes

Después de evaluar las agrupaciones, el siguiente paso consiste en analizar si estos grupos están teniendo evaluaciones coherentes por parte de cada medida de calidad interna, o si por el contrario las agrupaciones están teniendo resultados

incoherentes. También se puede saber con la evaluación de las medidas de calidad internas en qué nivel de elección, respecto a estas medidas, se encuentran las oraciones que el humano elige para formar sus resúmenes.

4.9 Ejemplo del método propuesto

Con el fin de poner en práctica los conceptos mencionados en este capítulo, se procede a realizar un ejemplo siguiendo los pasos descritos en el método propuesto. Los pasos descritos en el método propuesto se van a aplicar a un documento de ejemplo el cual se va ir modificando de acuerdo a la etapa en que se encuentre.

No.	ORACIONES
1	**The government of Egypt protects the pyramids**
2	**The pyramids of Egypt are cultural heritage**
3	**The pyramids were built by the Pharaohs**
4	**A good government protects its cultural heritage**
5	**The pyramids of Egypt were tombs for pharaohs**

Tabla 4.1 Documento ejemplo con 5 oraciones

4.10 Eliminación de stopwords

Siguiendo las etapas de la minería de texto mencionadas en el marco teórico, se procesan los documentos quitando *stopwords* que como se menciona en el marco teórico, son palabras que aparecen en el 80% del texto y no aportan ningún significado al texto.

No.	ORACIONES
1	**Government, Egypt protects, pyramids**
2	**Pyramids, Egypt, cultural, heritage**
3	**Pyramids, built, Pharaohs**
4	**Good, government, protects, cultural, heritage**
5	**Pyramids, Egypt, tombs, pharaohs**

Tabla 4.2 Documento después de eliminar palabras vacías (*stopwords*)

4.11 Lematización (Stemming)

Esta etapa consiste en reducir las variantes de una misma palabra a su raíz. Este procesamiento se hace con el objetivo de reducir la dimensión de los términos, lo cual tiene múltiples beneficios que van desde la reducción de dimensión de los patrones, hasta una mejor detección de frases dentro del texto. Para esta tarea se usa el algoritmo Porter, el cual puede ser encontrado en el sitio oficial de Martin Porter [Porter, 2006] en diferentes lenguajes de programación.

No.	ORACIONES
1	**Govern, Egypt protect, pyramid**
2	**Pyramid, Egypt, cultur, heritag**
3	**Pyramid, built, Pharaoh**
4	**Good, govern, protect, cultur, heritag**
5	**Pyramid, Egypt, tomb, pharaoh**

Tabla 4.3 Documento con aplicación de lematización (*stemming*)

4.12 Representación del documento mediante modelos de texto

Con esta etapa se busca transformar el texto en un modelo de espacio vectorial para que posteriormente pueda ser convertido en un patrón con el uso de algún pesado de término. Para este trabajo se manejan dos modelos de representación

de textos los cuales son: bolsa de palabras y n-gramas. Los n-gramas pueden ser de diferente longitud, para esta tarea son de longitud 2 y 3. Como pesado de término se van a utilizar los pesados de frecuencia del término (tf), frecuencia inversa del documento (idf) y pesado booleano (bool).

La tabla 4.4 muestra las oraciones del documento de la tabla 4.3 representado mediante el modelo de bolsa de palabras y frecuencia de término

Palabras	O1	O2	03	O4	O5
GOVERN	1	0	0	1	0
EGYPT	1	1	0	0	1
PROTECT	1	0	0	1	0
PYRAMID	1	1	1	0	1
CULTUR	0	1	0	1	0
HERITAG	0	1	0	1	0
BUILT	0	0	1	0	0
PHARAOH	0	0	1	0	1
GOOD	0	0	0	1	0
TOMB	0	0	0	0	1

Tabla 4.4 Modelo de bolsa de palabras con pesado booleano.

La tabla 4.5 muestra las oraciones del documento de la tabla 4.2 representado mediante el modelo de n-gramas, donde el valor de n es igual a 2 y el pesado de término ocupado es frecuencia del término (tf)

Palabras	O1	O2	03	O4	O5
GOVERN, EGYPT	1	0	0	0	0
EGYPT, PROTECT	1	0	0	0	0
PROTECT, PYRAMID	1	0	0	0	0
PYRAMID, EGYPT	0	0	0	0	1
EGYPT, CULTUR	0	1	0	0	0
CULTUR, HERITAG	0	1	0	0	0

PYRAMID, BUILT	0	1	0	1	0
BUILT, PHARAOH	0	0	1	0	0
PHARAOH, GOOD	0	0	1	0	0
GOOD, GOVERN	0	0	0	1	0
GOVERN, PROTEC	0	0	0	1	0
PROTEC, CULTURE	0	0	0	1	0
EGYPT, TOMB	0	0	0	0	1
TOMB, PHARAOH	0	0	0	0	1

Tabla 4.5 Modelo de n-gramas con pesado de frecuencia de término.

4.13 Reconstrucción de grupos con base en oraciones clave

En esta etapa se desarrolló un algoritmo de agrupación mediante aprendizaje no supervisado, el cual consiste en generar grupos de oraciones con base en la similitud con las oraciones clave de ese documento. La generación de grupos se hace siguiendo los pasos del algoritmo *K-means*, pero sin recalcular centroides, ya que para este trabajo en específico se necesita saber qué características tiene el agrupamiento cuando las oraciones clave de resúmenes funcionan como centroides de cada grupo.

Para comprobar el funcionamiento de este algoritmo vamos a generar grupos del documento presentado en la tabla 4.3 al cual se le aplicó preprocesamiento, este documento cuenta con un total de 20 palabras, distribuidas en 5 oraciones.

Las oraciones clave de este documento elegidas aleatoriamente para este ejemplo, son las oraciones 2 y 5.

Los pasos para la generación de los grupos son los siguientes.

1) Representación de oraciones y oraciones clave. En este paso se busca representar las oraciones del texto mediante un modelo de texto y pesado de término.
2) Cálculo de distancia. Este paso consiste en calcular la similitud de los patrones resultantes del paso 1 para asignarlos a un grupo. La similitud es

calculada entre las oraciones del texto original y la oración clave medida con distancia euclidiana.

Puesto que las oraciones clave toman el papel de centroides y se eligieron las oraciones 2 y 5, se procede a calcular la distancia euclidiana entre cada oración del documento y la oración clave como a continuación:

Centroide: Oración 2⟶ 0101110000
Oración: Oración 1⟶ 1111000000

Para medir la similitud entre estas dos oraciones, se utiliza la distancia euclidiana

$$d_{c2,o1} = \sqrt{(0-1)^2} + (1-1)^2 + (0-1)^2 + (1-1)^2 + (1-0)^2 + (1-0)^2 + (0-0)^2 + (0-0)^2 + (0-0)^2 + (0-0)^2$$

La distancia resultante entre estas dos oraciones es:

$$d_{c2,o1} = 2$$

Este cálculo se hace para todas las oraciones del texto original con cada una de las oraciones clave.
Ahora procedemos a calcular la distancia entre la oración clave respecto a las demás oraciones:

Centroide: Oración 2⟶ 0101110000
Oración: Oración 2⟶ 0101110000
Oración: Oración 3⟶ 0010011000
Oración: Oración 4⟶ 1010110010
Oración: Oración 5⟶ 0101000101

$$d_{c1,o2} = \sqrt{(0-0)^2} + (1-1)^2 + (0-0)^2 + (1-1)^2 + (1-1)^2 + (1-1)^2 + (0-0)^2 + (0-0)^2 + (0-0)^2 + (0-0)^2 =$$

$$d_{c2,o1} = 0$$

$$d_{c1,o3} = \sqrt{(0-0)^2} + (1-0)^2 + (0-1)^2 + (1-0)^2 + (1-0)^2 + (1-1)^2 + (0-1)^2 + (0-0)^2 + (0-0)^2 + (0-0)^2 =$$

$$d_{c2,o1} = 2.23$$

$$d_{c1,o4} = \sqrt{(0-1)^2} + (1-0)^2 + (0-1)^2 + (1-0)^2 + (1-1)^2 + (1-1)^2 + (0-0)^2 + (0-0)^2 + (0-1)^2 + (0-0)^2 =$$

$$d_{c2,o1} = 2.23$$

$$d_{c1,o5} = \sqrt{(0-0)^2} + (1-0)^2 + (0-1)^2 + (1-0)^2 + (1-0)^2 + (1-0)^2 + (0-1)^2 + (0-1)^2 + (0-0)^2 + (0-1)^2 =$$

$$d_{c2,o1} = 2.82$$

El mismo procedimiento se realiza para calcular la distancia de las oraciones del documento respecto a la segunda oración clave.

Centroide:	Oración 5 ⟶ 0101000101
Oración:	Oración 1 ⟶ 1111000000
Oración:	Oración 2 ⟶ 0101110000
Oración:	Oración 3 ⟶ 0010011000
Oración:	Oración 4 ⟶ 1010110010
Oración:	Oración 5 ⟶ 0101000101

$$d_{c2,o1} = \sqrt{(0-1)^2} + (1-1)^2 + (0-1)^2 + (1-1)^2 + (0-0)^2 + (0-0)^2 + (0-0)^2 + (1-0)^2 + (0-0)^2 + (1-0)^2 =$$

$$d_{c2,o1} = 1.73$$

$$d_{c2,o2} = \sqrt{(0-0)^2} + (1-1)^2 + (0-0)^2 + (1-1)^2 + (0-1)^2 + (0-1)^2 + (0-0)^2 + (1-0)^2 + (0-0)^2 + (1-0)^2 =$$

$$d_{c2,o1} = 2$$

$$d_{c2,o3} = \sqrt{(0-0)^2} + (1-0)^2 + (0-1)^2 + (1-0)^2 + (0-0)^2 + (0-1)^2 + (0-1)^2 + (1-0)^2 + (0-0)^2 + (1-0)^2 =$$

$$d_{c2,o1} = 2.64$$

$$d_{c2,o4} = \sqrt{(0-1)^2} + (1-0)^2 + (0-1)^2 + (1-0)^2 + (0-1)^2 + (0-1)^2 + (0-0)^2 + (1-0)^2 + (0-1)^2 + (1-0)^2 =$$

$$d_{c2,o1} = 3$$

$$d_{c2,o5} = \sqrt{(0-0)^2} + (1-1)^2 + (0-0)^2 + (1-1)^2 + (0-0)^2 + (0-0)^2 + (0-0)^2 + (1-1)^2 + (0-0)^2 + (1-1)^2 =$$

$$d_{c2,o1} = 2$$

3) Agrupación. En este paso es donde se comienzan a formar los grupos. De acuerdo a la distancia euclidiana obtenida entre cada oración y la oración clave, se asigna la oración al centroide con el cual la distancia sea menor.
En la siguiente tabla podemos observar como quedaron organizados los grupos.

No. Oración	Distancia a la oración clave 1	Distancia a la oración clave 2	Grupo asignado
1	2	1.73	2
2	0	2	1
3	2.23	2.64	1
4	2.23	3	1
5	2.82	0	2

Tabla 4.6 Agrupación basada en oraciones clave como centroides de grupo

El algoritmo propuesto para esta etapa termina aquí, ya que no se necesita recalcular los centroides, debido a que perderíamos la relación de los grupos respecto a las oraciones clave.

4.14 Evaluación de grupos mediante medidas de calidad internas

En esta etapa se usan tres medidas de calidad internas de agrupación para evaluar la calidad de los grupos generados anteriormente. Estas medidas de calidad interna se basan en calcular dos características: separación y compactación.

Estos índices buscan que, a través de iteraciones, los grupos sean lo más compactos, pero al mismo tiempo exista la mayor separación entre un grupo y otro.

4.14.1 Evaluación con la medida de calidad interna: índice Dunn

En primer lugar tenemos el índice Dunn, cuya fórmula es la siguiente:

$$Dunn = min_{1<i<c}\left\{ min\left\{ \frac{d\left(c_{i,}\, c_j\right)}{max_{1<k<c}\left(d(x_k)\right)} \right\} \right\}$$

El índice Dunn es una medida de calidad interna de maximización, esto quiere decir que entre más altos sean los valores de esta medida, mejor es la calidad de los grupos que está evaluando.

A continuación se describen los pasos para calcular la calidad de la agrupación con Dunn.

1.- El primer paso consiste en calcular el la distancia entre centroides que son nuestras oraciones clave de cada grupo por medio de distancia Euclidiana.

Oración clave: Oración 2 ⟶ 0101110000
Oración clave: Oración 5 ⟶ 0101000101

$$d_{c2,o1} = \sqrt{(0-0)^2} + (1-1)^2 + (0-0)^2 + (1-1)^2 + (1-0)^2 + (1-0)^2 + (0-0)^2 + (0-1)^2 + (0-0)^2 + (0-1)^2 =$$

$$d_{c1,c2} = 2$$

2.- El paso número dos consiste en calcular la distancia máxima que existe entre los elementos de cada grupo. Por lo tanto, se calcula la distancia entre las oraciones del grupo 1 con las oraciones del grupo 2.

Oraciones Grupo 1

Oración: Oración 2 ⟶ 0101110000
Oración: Oración 3 ⟶ 0010011000
Oración: Oración 4 ⟶ 1010110010

Oraciones grupo 2

Oración: Oración 1 ⟶ 1111000000
Oración: Oración 5 ⟶ 0101000101

Calculamos distancia entre las oraciones con distancia euclidiana.

$$d_{o2,o1} = \sqrt{(0-1)^2} + (0-1)^2 + (1-1)^2 + (0-1)^2 + (1-0)^2 + (1-0)^2 + (0-0)^2 + (0-0)^2 + (0-0)^2 + (0-0)^2 =$$

$$d_{c2,o1} = 2.23$$

$$d_{o2,o5} = \sqrt{(0-0)^2} + (1-1)^2 + (0-0)^2 + (1-1)^2 + (1-0)^2 + (1-0)^2 + (0-0)^2 + (0-1)^2 + (0-0)^2 + (0-1)^2 =$$

$$d_{c2,o1} = 2$$

$$d_{o3,o1} = \sqrt{(0-1)^2} + (0-1)^2 + (1-1)^2 + (0-1)^2 + (0-0)^2 + (1-0)^2 + (1-0)^2 + (0-0)^2 + (0-0)^2 + (0-0)^2 =$$

$$d_{c2,o1} = 2.23$$

$$d_{o3,o5} = \sqrt{(0-0)^2} + (0-1)^2 + (1-0)^2 + (0-1)^2 + (0-0)^2 + (1-0)^2 + (1-0)^2 + (0-1)^2 + (0-0)^2 + (0-1)^2 =$$

$$d_{c3,o1} = 2.64$$

$$d_{o1,o4} = \sqrt{(0-1)^2} + (1-0)^2 + (0-1)^2 + (1-0)^2 + (1-1)^2 + (1-1)^2 + (0-0)^2 + (0-0)^2 + (0-1)^2 + (0-0)^2 =$$

$$d_{c2,o1} = 2.23$$

$$d_{o1,o5} = \sqrt{(1-0)^2} + (0-1)^2 + (1-0)^2 + (0-1)^2 + (1-0)^2 + (1-0)^2 + (0-0)^2 + (0-1)^2 + (1-0)^2 + (0-1)^2 =$$

$$d_{c2,o1} = 3$$

La distancia máxima entre las oraciones de cada grupo es 3.

3.- EL siguiente paso consiste en obtener el valor mínimo al dividir la distancia entre los centroides de cada grupo obtenida en el paso 2 y la distancia máxima obtenida en el paso 3.

$$Dunn = min_{1<i<c}\left\{min\left\{\frac{2}{5}\right\}\right\}$$

Debido a que solo existen dos grupos el valor mínimo no tiene utilidad, pero cuando existen más de 2 grupos que van a ser evaluados con esta medida, el valor mínimo puede ir cambiando.

Por tanto, la calidad de la agrupación de acuerdo a esta medida es

$$Dunn = .4$$

4.14.2 Evaluación con la medida de calidad interna: Davies Bouldin

Otra de las medidas de calidad que usamos para medir la calidad de los grupos de oraciones es el índice Davies Bouldin cuya fórmula es la siguiente:

$$DB = \frac{1}{k} \sum_{i=1, i!=j}^{k} \max\left(\frac{\sigma i + \sigma j}{d(Ci, Cj)}\right)$$

El índice Davies Bouldin es una medida de minimización donde la mejor evaluación que puede dar sobre una agrupación es 0.

A continuación se describen los pasos para calcular la calidad de la agrupación con el índice Davies Bouldin.

Debido a que el índice Davies Bouldin es una medida que se basa en centroides para calcular la cohesión y la separación de los grupos, es necesario hacer el cálculo de centroides antes de comenzar con los pasos a seguir para la evaluación de esta medida sobre los grupos de oraciones.

Definición del “centroide” de un grupo

Dado un grupo de elementos *S,* que contiene *h* elementos s_i, se define a su centroide como el promedio de los vectores que componen el grupo [Zhao, 2005]:

$$C = \frac{\sum_{i=1}^{h} s_i}{h}$$

Centroide grupo 1:

Oración: Oración 2 ⟶ 0101110000 = 4/3 = 1.3
Oración: Oración 3 ⟶ 0010011000 = 3/3 = 1
Oración: Oración 4 ⟶ 1010110010 = 5/3 = 1.6

Centroide del grupo 1= 3.9

Centroide grupo 2:

Oración: Oración 1 ⟶ 1111000000 = 4/3 = 1.3
Oración: Oración 5 ⟶ 0101000101 = 4/3 = 1.3

Centroide del grupo 2= 2.6.

Ya que realizamos el cálculo del centroide de cada grupo, ahora procedemos a evaluar la calidad de los grupos conforme a la fórmula.

1.- El primer paso consiste en calcular σi, que es la distancia promedio entre cada oración del grupo i y el centroide del grupo. Tomando a i como nuestro grupo 1 la operación es la siguiente:

Oración: Oración 2 ⟶ 0101110000 = 4 / 3.9 = 1.02
Oración: Oración 3 ⟶ 0010011000 = 3 / 3.9 = 0.76
Oración: Oración 4 ⟶ 1010110010 = 5 / 3.9 = 1.28

σi = 3.06

2.- El segundo paso consiste en calcular σj, que es la distancia promedio entre cada oración del grupo j y el centroide del grupo. Tomando a j como nuestro grupo 2 la operación es la siguiente:

Oración: Oración 1 ⟶ 1111000000 = 4 / 2.6 = 1.53
Oración: Oración 5 ⟶ 0101000101 = 4 / 2.6 = 1.53

σi= sumatoria grupo i / centroide.

σj = 3.06

3.- El tercer paso consiste en sumar σi y σj.

$$\sigma\text{i} + \sigma\text{j} = 6.12$$

4.- El cuarto paso consiste en calcular la distancia que existe entre los dos centroides de cada grupo $d = (Ci, Cj)$.

$$d = 3.9/2.6$$

$$d = 1.5$$

5.-Por último se divide σi + σj entre la distancia de los centroides $d = (Ci, Cj)$.

$$DB = \frac{1}{k} \sum_{i=1, i!=j}^{k} \max\left(\frac{6.12}{1.5}\right)$$

$$DB = \frac{1}{2}\left(\frac{6.12}{1.5}\right)$$

$$DB = 0.816$$

4.14.3 Evaluación con la medida de calidad interna: Silhouette

Por último se hace uso del índice Silhouette, una medida evalúa grupos con base en un rango que va de 1 a -1. Valores cercanos a 1 indican que las oraciones fueron asignadas al grupo adecuado, valores cercanos a 0 indican que las oraciones pudieron ser asignadas al grupo vecino más cercano y por último valores cercanos a -1 indican una mala agrupación.

La fórmula para evaluar la calidad de las agrupaciones es la siguiente:

$$S(i) = \frac{AVGD\_BETWEEN(i,k) - AVGD\_WITHIN(i)}{MAX(AVGD\_WITHIN(i),\ AVGD\_BETWEEN(i,k))}$$

A continuación se describen los pasos para calcular la calidad de la agrupación de la tabla 3.5 con el índice Silhouette

1.- El primer paso consiste en obtener $AVGD\_BETWEEN(i,k)$, que es la distancia promedio de la oración *i* a todas las oraciones de los demás grupos.

La oración *i* es la oración clave del grupo 1, la cual se muestra a continuación.

Oración clave: Oración 2 ⟶ 0101110000

Procedemos a medir la distancia entre la oración clave del grupo 1 y las oraciones del grupo 2:

Oraciones grupo 2

Oración: Oración 1 ⟶ 1111000000
Oración: Oración 5 ⟶ 0101000101

$$d_{o2,o1} = \sqrt{(0-1)^2} + (0-1)^2 + (1-1)^2 + (0-1)^2 + (1-0)^2 + (1-0)^2 + (0-0)^2 + (0-0)^2 + (0-0)^2 + (0-0)^2 =$$

$$d_{o2,o1} = 2.23$$

$$d_{o2,o5} = \sqrt{(0-0)^2} + (1-1)^2 + (0-0)^2 + (1-1)^2 + (1-0)^2 + (1-0)^2 + (0-0)^2 + (0-1)^2 + (0-0)^2 + (0-1)^2 =$$

$$d_{o2,o5} = 2$$

$$AVGD\_BETWEEN(i,k) = 2.11$$

2.- El segundo paso consiste en calcular AVGD_WITHIN (i) es la distancia promedio de la oración *i* a las oraciones de su mismo grupo.

Oración clave: Oración 2 ⟶ 0101110000

Oraciones Grupo 1

Oración: Oración 3 ⟶ 0010011000
Oración: Oración 4 ⟶ 1010110010

$$d_{o2,o3} = \sqrt{(0-0)^2} + (1-0)^2 + (0-1)^2 + (1-0)^2 + (1-0)^2 + (1-1)^2 + (0-1)^2 + (0-0)^2 + (0-0)^2 + (0-0)^2 =$$

$$d_{o2,o3} = 2.23$$

$$d_{o2,o4} = \sqrt{(0-1)^2} + (1-0)^2 + (0-1)^2 + (1-0)^2 + (1-1)^2 + (1-1)^2 + (0-0)^2 + (0-0)^2 + (1-0)^2 + (0-0)^2 =$$

$$d_{o2,o4} = 2.23$$

AVGD_WITHIN (i) = 2.23

3.- El siguiente paso es obtener el valor máximo de la multiplicación entre $AVGD\_BETWEEN(i,k)$ y *$AVGD\_WITHIN\ (i)$.*

Procedemos a multiplicar los valores de la distancia promedio de la oración *i* a todas las oraciones de los demás grupos:

$$d_{o1,o2} \rightarrow 2.23 * d_{o2,o5} \rightarrow 2 = 4.46$$

$$d_{o2,o4} \rightarrow 2.11 * d_{o2,o4} \rightarrow 2.23 = 4.70$$

$$d_{o1,o2} \rightarrow 2.23 * d_{o2,o4} \rightarrow 2.23 = 4.97 \rightarrow \text{MAX}$$

4.- Calculamos el valor $S(i)$, sustituyendo los valores en la fórmula original.

$$S(i) = \frac{2.11 - 2.23}{4.97}$$

$$S(i) = -0.02$$

Como podemos ver, la evaluación asignada a la agrupación formada en la tabla 3.5, tiene un valor cercano a 0, lo que quiere decir que las oraciones pudieron ser asignadas al grupo vecino más cercano.

4.15 Resumen del método propuesto

En el presente capítulo se realiza un ejemplo para demostrar que las medidas de calidad internas de agrupación usadas en este trabajo son capaces de evaluar grupos de oraciones. Sin embargo, si se evalúan colecciones de documentos el resultado puede variar. Así mismo es necesaria la evaluación de todos los documentos de la línea base *top line* para caracterizar las oraciones clave de resúmenes a través de medidas internas de calidad.

Capítulo 5:
Experimentación

En el presente capítulo se muestran los experimentos realizados para demostrar el comportamiento que tienen los grupos de oraciones seleccionadas por humanos para generar sus resúmenes al ser evaluados mediante medidas de calidad internas. Los experimentos se dividen en dos secciones principales, por una parte se muestran los experimentos en donde los documentos no cuentan con preprocesamiento y por otra parte se muestran los experimentos en donde a los documentos se les ha aplicado preprocesamiento. Los experimentos fueron desarrollados siguiendo el método propuesto en este trabajo.

Antes de presentar los experimentos primero vamos a describir la colección de documentos usada en este trabajo. Después el método de agrupación usado se describe, y por último se describe el método de evaluación para los grupos generados.

5.1 Colecciones de documentos

Para realizar los experimentos se utiliza el corpus DUC-2002. Este corpus consta de 567 documentos de noticias en inglés sobre desastres naturales.

También se realizan experimentos con el corpus DUC-2001, el cual contiene 308 documentos los cuales abarcan temáticas como acontecimientos de desastres naturales, información biográfica sobre un individuo, entre otros.

5.2 Preprocesamiento

Consiste en eliminar palabras vacías en el texto, es decir, aquellas palabras que proporcionan poca o ninguna información acerca del contenido del documento, por ejemplo las preposiciones, conjunciones y artículos. Estas palabras se pueden descartar del conjunto de términos lo que reduce considerablemente la dimensión del vector

5.3 Modelos de representación de texto

Para todos los experimentos se utilizan dos modelos de representación de texto, los cuales son bolsa de palabras y n-gramas.

5.4 Pesado de términos

Cada modelo de representación de texto es conjugado con tres pesados de términos los cuales son: pesado booleano, frecuencia del término (tf) y frecuencia inversa del documento (idf).

5.5 Evaluación de los experimentos

Para evaluar los experimentos desarrollados durante este trabajo se usan medidas de calidad internas de agrupación. Se usan tres medidas presentadas en nuestro marco teórico que son índice Silhouette, índice Dunn, índice Davies Bouldin.

5.6 Experimentos sin preprocesamiento en DUC-2002

Los experimentos presentados en esta sección son documentos sin preprocesamiento, pero han sido representados mediante dos modelos de texto que son bolsa de palabras y n-gramas con tres pesados de términos que son frecuencia del término (TF), pesado booleano frecuencia inversa del documento (IDF).

5.6.1 Experimentos con bolsa de palabras

Los documentos que se utilizaron para los experimentos de esta sección están representados con el modelo de representación de texto de bolsa de palabras y tres pesados de términos.

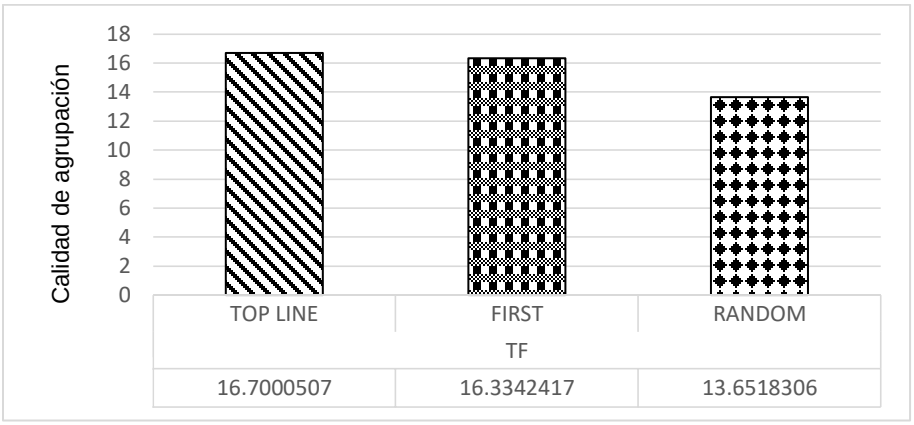
Los siguientes experimentos muestran el desempeño que tuvieron las medidas de calidad internas ocupadas para este trabajo respecto a las líneas base: *top line*, línea base de primeras oraciones y línea base de oraciones aleatorias.

5.6.1.1 Índice de Dunn

Dunn es un índice de maximización, entre más altos sean los valores mejor es la calidad de los grupos que evalúa. Las siguientes gráficas muestran el promedio de evaluación de los 567 documentos pertenecientes a la colección de documentos DUC-2002.

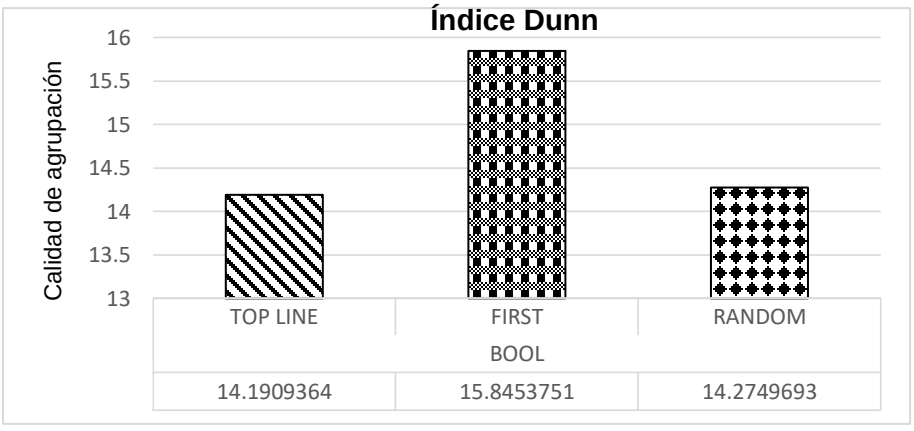
De acuerdo con la hipótesis se esperan obtener las mejores evaluaciones en la calidad de los grupos generados a partir de la línea base tope (*top line*), seguido de la línea base de primeras oraciones (*first*) y las agrupaciones con la calidad más baja en la línea base de oraciones aleatorias (*random*).

En la gráfica 5.1 podemos observar como Dunn es capaz de evaluar correctamente la relación entre la calidad de los grupos de oraciones generados y la calidad de los resúmenes utilizados cuando se usa el pesado de término por frecuencia de término (*tf*), pero no existe una clara diferencia, por lo que se comienza a sospechar que este índice no puede evaluar correctamente la calidad de una agrupación de oraciones.



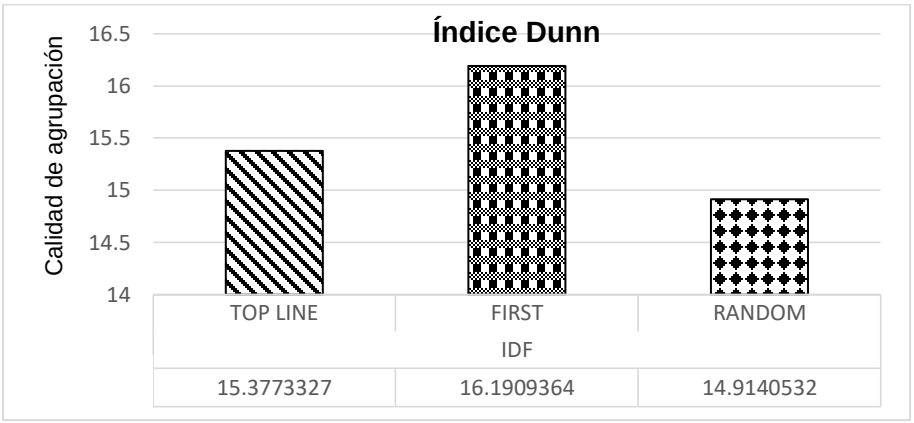
Gráfica 5.1 Evaluación de calidad de agrupación con el índice Dunn usando el modelo de bolsa de palabras con el pesado por frecuencia de término.

En la gráfica 5.2 podemos observar que Dunn es un índice incapaz de evaluar la relación entre la calidad de los grupos de oraciones generados y la calidad de los resúmenes cuando se utiliza el pesado de término booleano, ya que las mejores evaluaciones se dan en la línea base primeras oraciones (*first*).



Gráfica 5.2 Evaluación de calidad de agrupación con el índice Dunn usando el modelo de bolsa de palabras con el pesado booleano en el corpus DUC-2002.

En la gráfica 5.3 se observa un patrón similar a la gráfica anterior; obteniendo mejores evaluaciones de calidad de agrupación en la línea base de primeras oraciones (*first*), debiendo haber obtenido mejor evaluación en la calidad de las agrupaciones generadas a partir de la línea base tope (*top line*).



Gráfica 5.3 Evaluación de calidad de agrupación con el índice Dunn usando el modelo de bolsa de palabras con el pesado idf en el corpus DUC-2002.

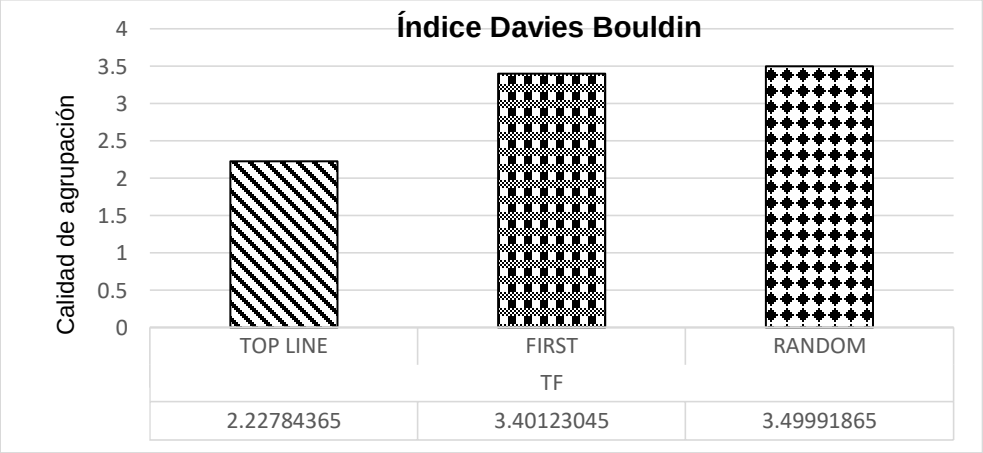
En experimentos sin preprocesamiento utilizando tres pesados de términos: frecuencia del término (*tf*), pesado booleano (*bool*) y frecuencia inversa del documento (*idf*) con el modelo de bolsa de palabras, se concluye que Duun no puede evaluar la calidad de grupos de oraciones generados a partir de tres líneas base correctamente debido a que no mantiene una relación entre la calidad de los grupos generados y la calidad de los resúmenes utilizados.

5.6.1.2 Índice Davies Bouldin

Davies Bouldin es un índice de minimización, esto quiere decir que entre más cercana a cero sea la evaluación, mejor es la calidad de la agrupación. Las siguientes gráficas muestran el promedio de evaluación de los 567 documentos pertenecientes a la colección de documentos DUC-2002.

De acuerdo con la hipótesis se esperan obtener las mejores evaluaciones en la calidad de los grupos generados a partir de la línea base tope (*top line*), seguido de la línea base: primeras oraciones (*first*) y las agrupaciones con la calidad más baja en la línea base oraciones aleatorias (*random*).

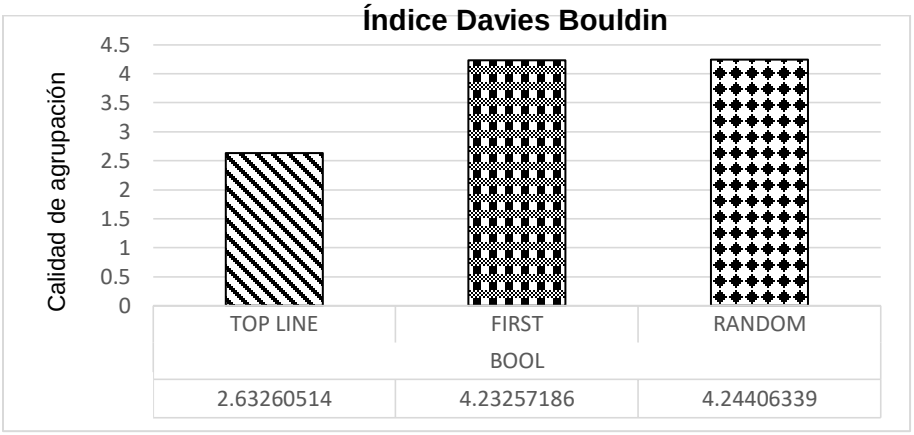
En la gráfica 5.4 podemos ver que el índice Davies Bouldin puede evaluar correctamente la relación entre la calidad de los grupos de oraciones generados y la calidad de los resúmenes utilizados. Las evaluaciones muestran mejor calidad en la línea base tope (*top line*), seguido de la línea base (*first)* de primeras oraciones y (random) de oraciones aleatorias.



Gráfica 5.4 Evaluación de calidad de agrupación con el índice Davies Bouldin usando el modelo de bolsa de palabras con el pesado por frecuencia de término en el corpus DUC-2002.

En la gráfica 5.5 podemos un patrón muy similar a la gráfica 5.4, a pesar de usar otro pesado de términos (pesado booleano) seguimos teniendo las mejores

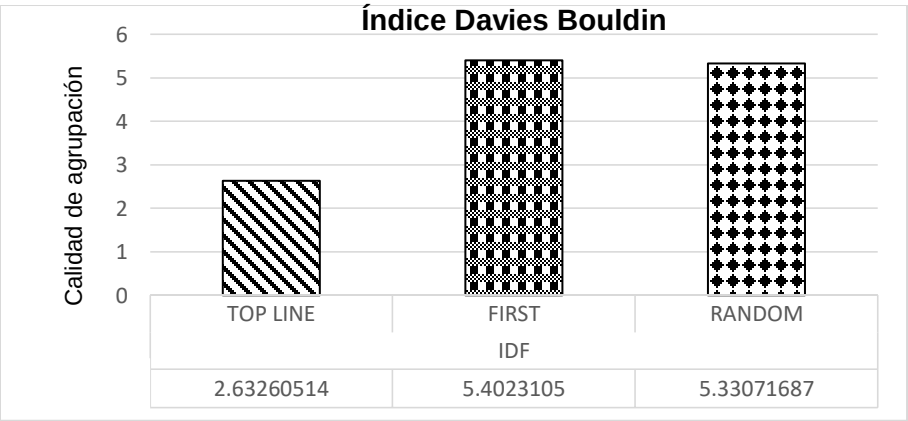
evaluaciones en los grupos de oraciones generados a partir de la línea base tope (*top line*), las líneas base primeras oraciones (*first*) y oraciones aleatorias (*random*) guardan gran similitud en la calidad de sus evaluaciones, suponemos que se debe a que las oraciones pasan por un proceso de agrupación, debido a esto los grupos obtenidos no pueden tener un pésima calidad en la evaluación de los grupos.



Gráfica 5.5 Evaluación de calidad de agrupación con el índice Davies Bouldin usando el modelo de bolsa de palabras con el pesado booleano en el corpus DUC-2002.

La gráfica 5.6 muestra que el índice Davies Bouldin puede evaluar correctamente la relación entre la calidad de los grupos de oraciones generados y la calidad de los resúmenes utilizados. Las evaluaciones muestran mejor calidad en la línea base tope (*top line*), seguido de la línea base (*first)* de primeras oraciones y *(random)* de oraciones aleatorias.

Este experimento se realiza con el pesado de término frecuencia inversa del documento (idf), a pesar de que este pesado de término ha sido ideado para buscar palabras únicas de cada documento y debido a que en este trabajo se usan oraciones, ha demostrado dar buenos resultados.



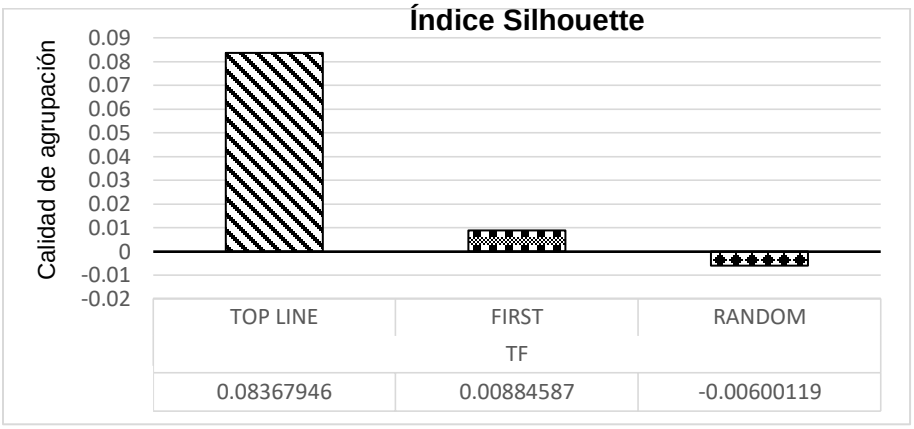
Gráfica 5.6 Evaluación de calidad de agrupación con el índice Davies Bouldin usando el modelo de bolsa de palabras con el pesado idf en el corpus DUC-2002.

5.6.1.3 Índice Silhouette

Silhouette es un índice que puede asignar valores entre 1 y -1 al evaluar la calidad de las agrupaciones. Evaluaciones cercanas a 1 indican que las oraciones están en los grupos correctos, evaluaciones cercanas a 0 indican que las oraciones pudieron ser asignadas al grupo vecino más cercano y evaluaciones cercanas a -1 indican que las oraciones fueron asignadas al grupo incorrecto.

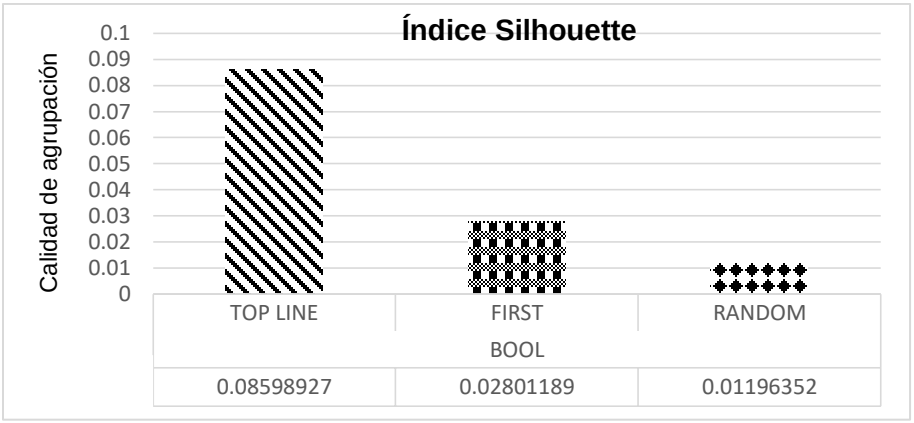
Como podemos ver en la gráfica 5.7 este índice es capaz de evaluar la relación en la calidad de los grupos de oraciones generados y la calidad de los resúmenes

obtenidos. Se usa el pesado por frecuencia de término el cual demuestra buen acoplamiento con este índice. Las evaluaciones de este índice permiten distinguir claramente la calidad de una línea base de otra. Las evaluaciones muestran mejor calidad en la línea base tope (*top line*), seguido de la línea base (*first)* de primeras oraciones y *(random)* de oraciones aleatorias.



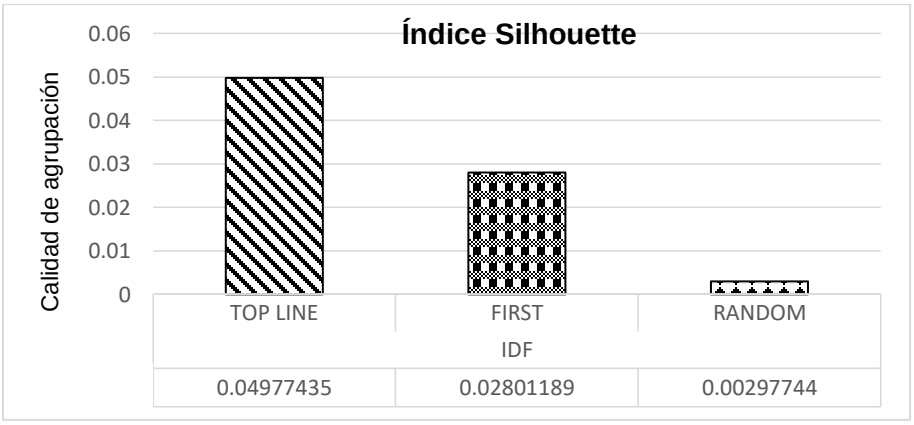
Gráfica 5.7 Evaluación de calidad de agrupación con el índice Silhouette usando el modelo de bolsa de palabras con el pesado por frecuencia de término en el corpus DUC-2002.

La gráfica 5.8 muestra que este índice es capaz de evaluar la relación en la calidad de los grupos de oraciones generados y la calidad de los resúmenes obtenidos. Se usa el pesado por booleano el cual demuestra buen acoplamiento con este índice. Las evaluaciones de este índice permiten distinguir claramente la calidad de una línea base de otra. Las evaluaciones muestran mejor calidad en la línea base tope (*top line*), seguido de la línea base (*first) primeras oraciones y (random) oraciones aleatorias.*



Gráfica 5.8 Evaluación de calidad de agrupación con el índice Silhouette usando el modelo de bolsa de palabras con el pesado booleano en el corpus DUC-2002.

La gráfica 5.9 muestra un patrón similar a la gráfica 5.7 y 5.8, con la diferencia que los grupos de oraciones generados a partir de la línea base de primeras oraciones (*first*) tienen mejor calidad que en las gráficas antes mencionadas. Esto puede ser debido a que este pesado de términos ha sido ideado para buscar palabras únicas de cada documento. Las evaluaciones de este índice permiten distinguir claramente la calidad de una línea base de otra. Las evaluaciones muestran mejor calidad en la línea base tope (*top line*), seguido de la línea base (*first*) de primeras oraciones y *(random)* de oraciones aleatorias.



Gráfica 5.9 Evaluación de calidad de agrupación con el índice Silhouette usando el modelo de bolsa de palabras con el pesado idf en el corpus DUC-2002.

5.7 Experimentos con preprocesamiento en DUC-2002

Después de aplicar el preprocesamiento a todos los documentos de la colección, procedemos a utilizar dos modelos de texto para su representación en conjunto con tres pesados de términos.

Al aplicar preprocesamiento a los documentos notamos que hubo una mejora general en la evaluación de la calidad de los grupos.

Los siguientes experimentos muestran el desempeño que tuvieron las medidas de calidad internas ocupadas para este trabajo respecto a las líneas base: top line, primeras oraciones y oraciones aleatorias.

Cabe recordar que los resultados mostrados por cada índice son el promedio de las evaluaciones realizadas a los 567 documentos de la colección de documentos.

5.7.1 Experimentos con bolsa de palabras

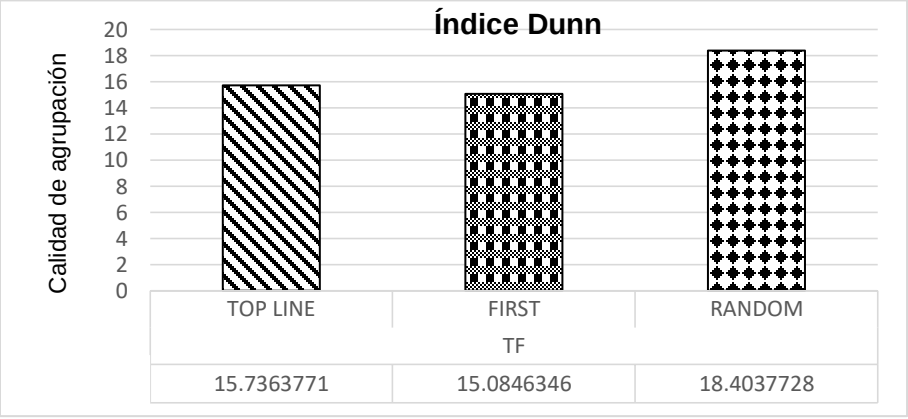
Los documentos utilizados para estos experimentos tienen preprocesamiento y están representados con el modelo de representación de bolsa de palabras con tres pesados de términos.

Los siguientes experimentos muestran el desempeño que tuvieron las medidas de calidad internas ocupadas para este trabajo respecto a las líneas base: top line, primeras oraciones y oraciones aleatorias.

5.7.1.1 Índice de Dunn

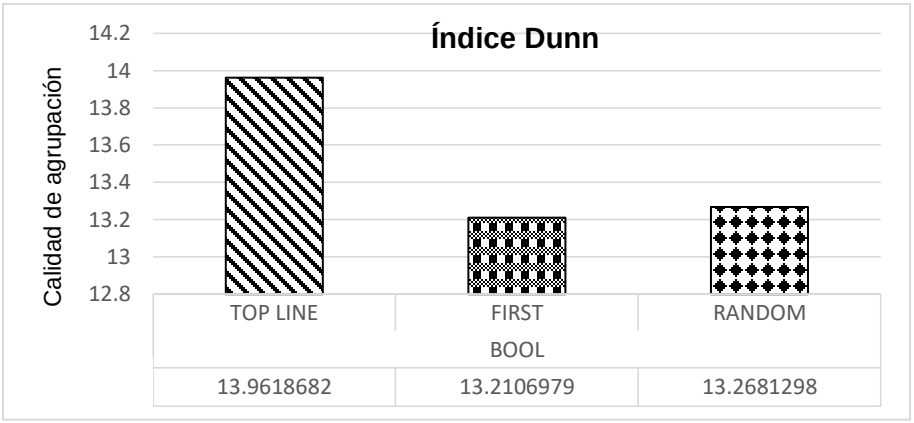
De acuerdo con la hipótesis se esperan obtener las mejores evaluaciones en la calidad de los grupos generados a partir de la línea base tope (*top line*), seguido de la línea base: primeras oraciones (*first*) y las agrupaciones con la calidad más baja en la línea base oraciones aleatorias (*random*).

En la gráfica 5.10 podemos observar como Dunn no es capaz de evaluar correctamente la relación entre la calidad de los grupos de oraciones generados y la calidad de los resúmenes utilizados cuando se usa el pesado de término por frecuencia de término (*tf*). A lo largo de los experimentos mostrados, este índice no puede evaluar de forma correcta la calidad de las agrupaciones generadas a partir de las tres líneas base usadas en este trabajo.



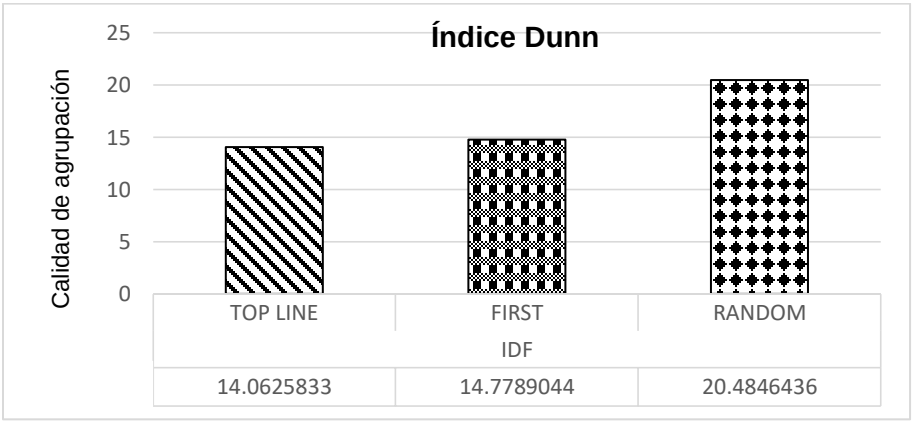
Gráfica 5.10 Evaluación de calidad de agrupación con el índice Dunn usando el modelo de bolsa de palabras con el pesado por frecuencia de término en el corpus DUC-2002 aplicando preprocesamiento.

En la gráfica 5.11 se puede observar que no existe una coherencia en las evaluaciones realizadas por el índice de Dunn. A pesar de aplicar preprocesamiento a los documentos, las evaluaciones de calidad de los grupos siguen siendo erróneas. El pesado de término booleano (*bool*) tampoco muestra mejoría.



Gráfica 5.11 Evaluación de calidad de agrupación con el índice Dunn usando el modelo de bolsa de palabras con el pesado booleano en el corpus DUC-2002 aplicando preprocesamiento.

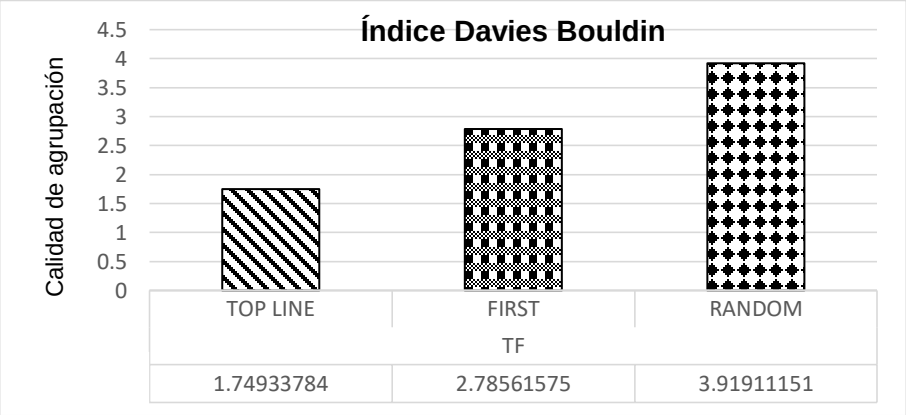
En la gráfica 5.12 se puede observar que no existe una coherencia en las evaluaciones realizadas por el índice de Dunn. A pesar de aplicar preprocesamiento a los documentos, las evaluaciones de calidad de los grupos siguen siendo erróneas. El pesado de término frecuencia inversa del documento (*idf*) tampoco muestra mejoría.



Gráfica 5.12 Evaluación de calidad de agrupación con el índice Dunn usando el modelo de bolsa de palabras con el pesado idf en el corpus DUC-2002 aplicando preprocesamiento.

5.7.1.2 Índice Davies Bouldin

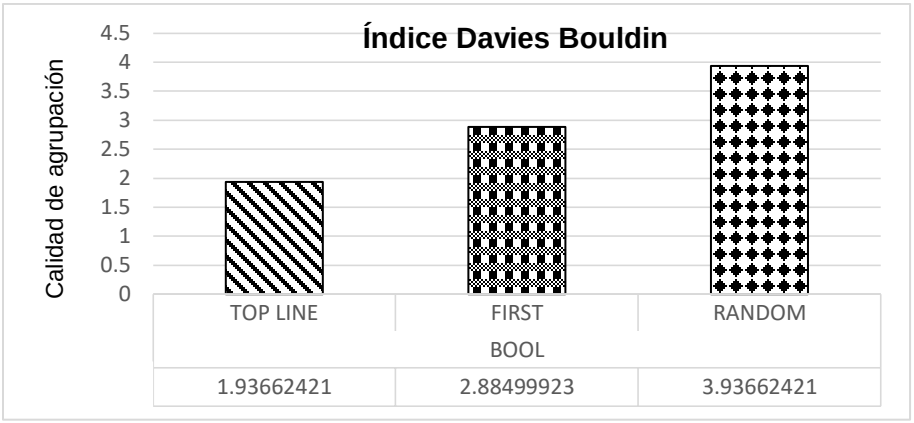
En la gráfica 5.13 podemos ver que el índice Davies Bouldin puede evaluar correctamente la relación entre la calidad de los grupos de oraciones generados y la calidad de los resúmenes utilizados al usar el pesado por frecuencia de término. Las evaluaciones muestran mejor calidad en la línea base tope (*top line*), seguido de la línea base (*first) primeras oraciones y (random) oraciones aleatorias.* Al aplicar preprocesamiento a los documentos se obtiene una mejoría en la calidad de los grupos, es decir, la evaluación de la calidad de las agrupaciones es mejor.



Gráfica 5.13 Evaluación de calidad de agrupación con el índice Davies Bouldin usando el modelo de bolsa de palabras con el pesado por frecuencia del término en el corpus DUC-2002 aplicando preprocesamiento.

En la gráfica 5.14 tenemos un patrón muy similar a la gráfica 5.13 A pesar de usar otro pesado de términos (pesado booleano) seguimos teniendo las mejores evaluaciones en los grupos de oraciones generados a partir de la línea base tope (*top line*), las líneas base de primeras oraciones (*first*) y de oraciones aleatorias (*random*) guardan gran similitud en la calidad de sus evaluaciones, creemos que se debe a que las oraciones pasan por un proceso de agrupación, debido a esto los grupos obtenidos no pueden tener un pésima calidad en la evaluación de los grupos.

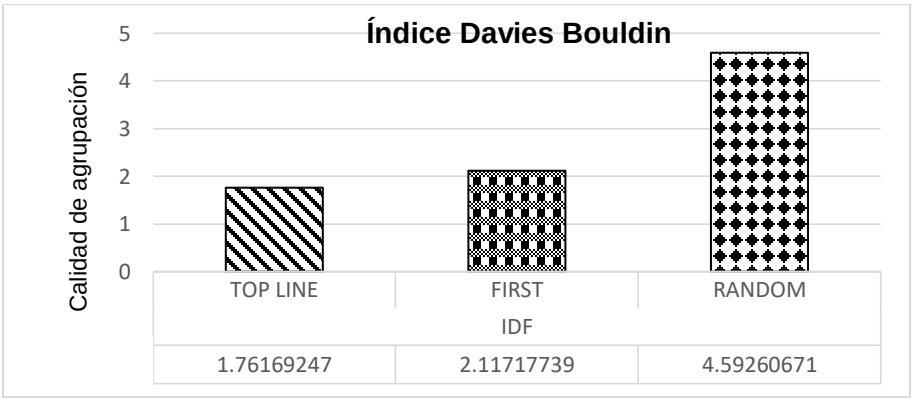
Al aplicar preprocesamiento a los documentos se obtiene una mejoría en la calidad de los grupos, es decir, la evaluación de la calidad de las agrupaciones es mejor.



Gráfica 5.14 Evaluación de calidad de agrupación con el índice Davies Bouldin usando el modelo de bolsa de palabras con el pesado booleano en el corpus DUC-2002 aplicando preprocesamiento.

La gráfica 5.15 que el índice Davies Bouldin puede evaluar correctamente la relación entre la calidad de los grupos de oraciones generados y la calidad de los resúmenes utilizados. Las evaluaciones muestran mejor calidad en la línea base tope (*top line*), seguido de la línea base (*first) primeras oraciones y (random) oraciones aleatorias.*

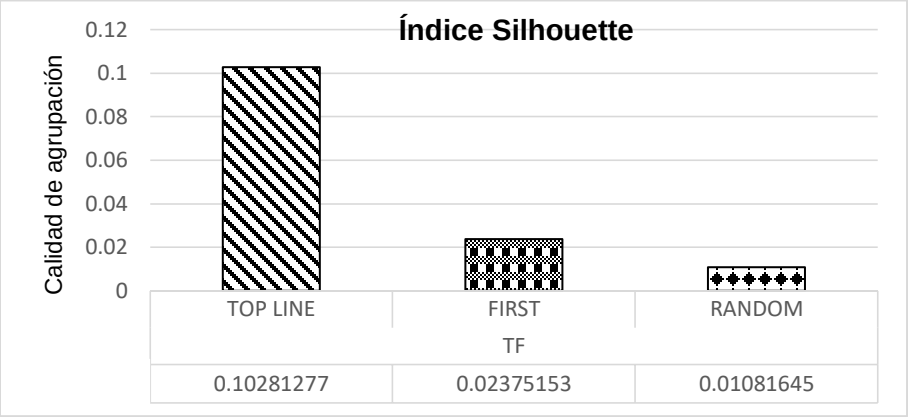
Este experimento se realizó con el pesado de término frecuencia inversa del documento (idf), a pesar de que este pesado de término ha sido ideado para buscar palabras únicas de cada documento y debido a que en este trabajo se usan oraciones, ha demostrado dar buenos resultados. Al aplicar preprocesamiento a los documentos se obtiene una mejoría en la calidad de los grupos, es decir, la evaluación de la calidad de las agrupaciones es mejor.



Gráfica 5.15 Evaluación de calidad de agrupación con el índice Davies Bouldin usando el modelo de bolsa de palabras con el pesado idf en el corpus DUC-2002 aplicando preprocesamiento.

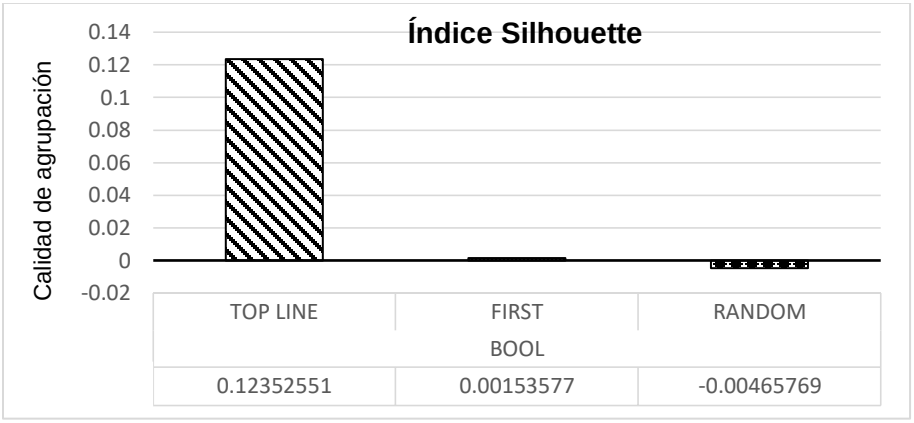
5.7.1.3 Índice Silhouette

Las evaluaciones por parte de este índice también fueron mejores al aplicar preprocesamiento a los documentos. La gráfica 5.16 muestra cómo el índice Silhouette sigue manteniendo el comportamiento esperado respecto a las líneas base. Las evaluaciones muestran mejor calidad en la línea base tope (*top line*), seguido de la línea base (*first)* de primeras oraciones *y (random)* de oraciones aleatorias*.* Al aplicar preprocesamiento a los documentos se obtiene una mejoría en la calidad de los grupos, es decir, la evaluación de la calidad de las agrupaciones es mejor.



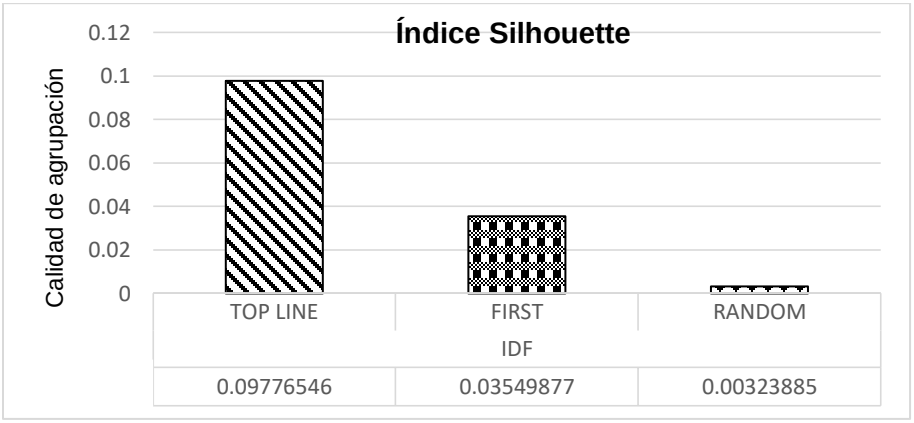
Gráfica 5.16 Evaluación de calidad de agrupación con el índice Silhouette usando el modelo de bolsa de palabras con el pesado por frecuencia de término en el corpus DUC-2002.

La gráfica 5.17 muestra que el índice Silhouette es capaz de evaluar la relación en la calidad de los grupos de oraciones generados y la calidad de los resúmenes obtenidos. Se usa el pesado booleano el cual demuestra buen acoplamiento con este índice. También las evaluaciones de este índice permiten distinguir claramente la calidad de una línea base de otra. Las evaluaciones muestran mejor calidad en la línea base tope (*top line*), seguido de la línea base (*first)* de primeras oraciones y *(random)* de oraciones aleatorias*.* Al aplicar preprocesamiento a los documentos se obtiene una mejoría en la calidad de los grupos, es decir, la evaluación de la calidad de las agrupaciones es mejor.



Gráfica 5.17 Evaluación de calidad de agrupación con el índice Silhouette usando el modelo de bolsa de palabras con el pesado booleano en el corpus DUC-2002.

La gráfica 5.18 muestra un patrón similar a la gráfica 5.16 y 5.17, con la diferencia que los grupos de oraciones generados a partir de la línea base primeras oraciones (*first*) tienen mejor calidad que en las gráficas antes mencionadas. Esto puede ser debido a que el pesado de término frecuencia inversa del documento (*idf*) ha sido ideado para buscar palabras únicas de cada documento. Las evaluaciones de este índice permiten distinguir claramente la calidad de una línea base de otra. Las evaluaciones muestran mejor calidad en la línea base tope (*top line*), seguido de la línea base (*first) primeras oraciones y (random) oraciones aleatorias.*



Gráfica 5.18 Evaluación de calidad de agrupación con el índice Silhouette usando el modelo de bolsa de palabras con el pesado idf en el corpus DUC-2002.

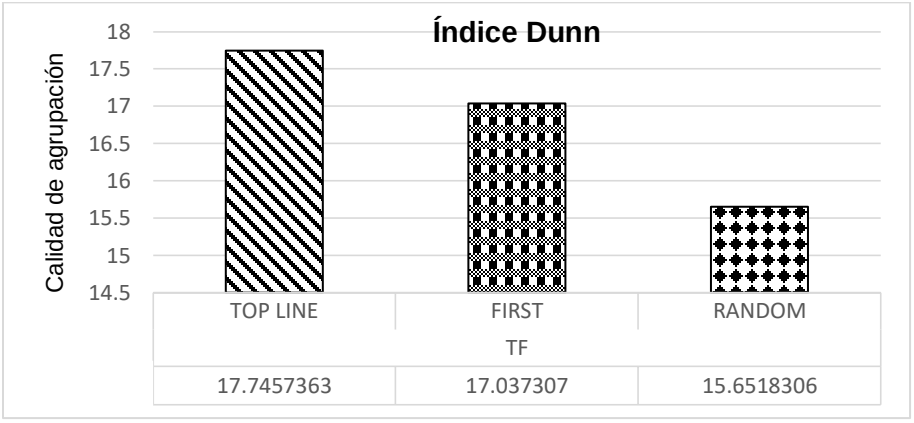
5.7.2 Experimentos con n-gramas

Los experimentos llevados a cabo con n-gramas, es donde se obtuvieron las mejores evaluaciones de calidad en las agrupaciones, ya que se puede distinguir claramente que la línea base de oraciones aleatorias se tienen las peores evaluaciones, mientras que en *top line* las mejores, con excepción del índice de Dunn.

A continuación se muestran los experimentos con bigramas, es decir, con n=2.

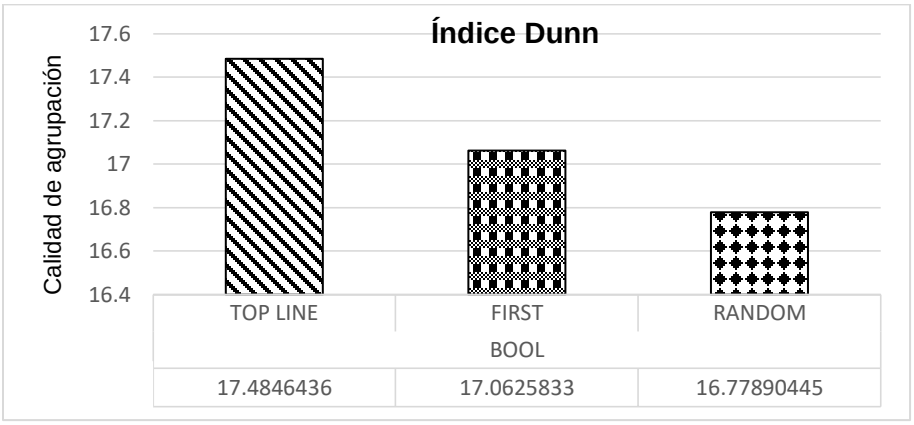
5.7.2.1 Índice de Dunn

En la gráfica 5.19 podemos observar que el índice de Dunn es capaz de evaluar la calidad de las agrupaciones generadas con base en oraciones clave de cada línea base. Esto es debido al modelo de representación de texto usado para estos experimentos (es bigramas). El pesado es por frecuencia de término.



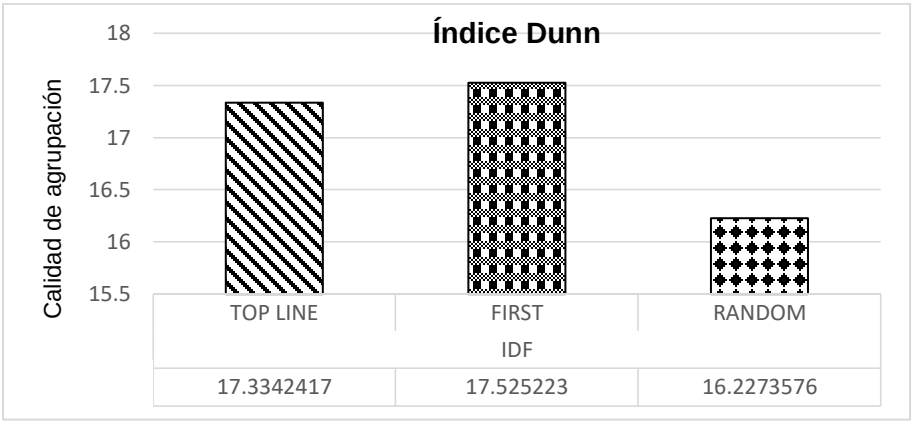
Gráfica 5.19 Evaluación de calidad de agrupación con el índice Dunn usando bigramas con el pesado por frecuencia de término en el corpus DUC-2002.

En la gráfica 5.20 podemos observar que el índice de Dunn es capaz de evaluar la calidad de las agrupaciones generadas con base en oraciones clave de cada línea base. Esto es debido a que el modelo de representación de texto usado para estos experimentos (bigramas). El pesado de término es pesado booleano.



Gráfica 5.20 Evaluación de calidad de agrupación con el índice Dunn usando bigramas con el pesado booleano en el corpus DUC-2002.

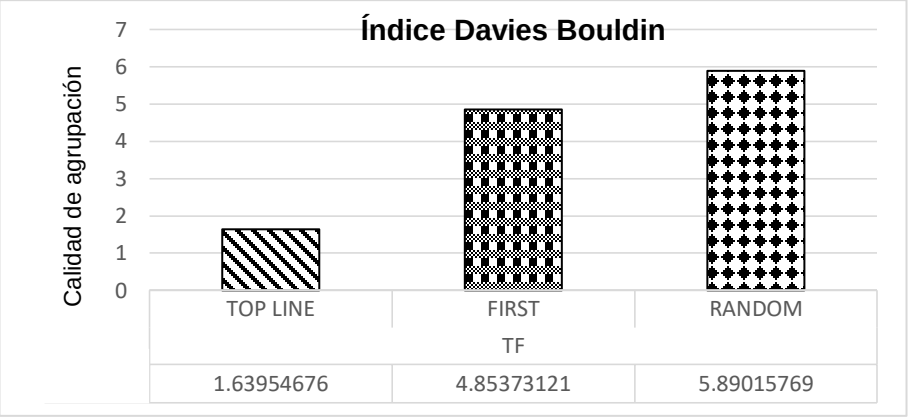
Como podemos observar en la gráfica 5.21, el índice de Dunn no puede mantener la coherencia en las evaluaciones de calidad de los grupos generados a partir de las tres líneas base con el pesado por frecuencia inversa del documento (*idf*). Esto es debido a que la fórmula de este índice no se acopla con este pesado de término.



Gráfica 5.21 Evaluación de calidad de agrupación con el índice Dunn usando bigramas con el pesado por frecuencia inversa del documento en el corpus DUC-2002.

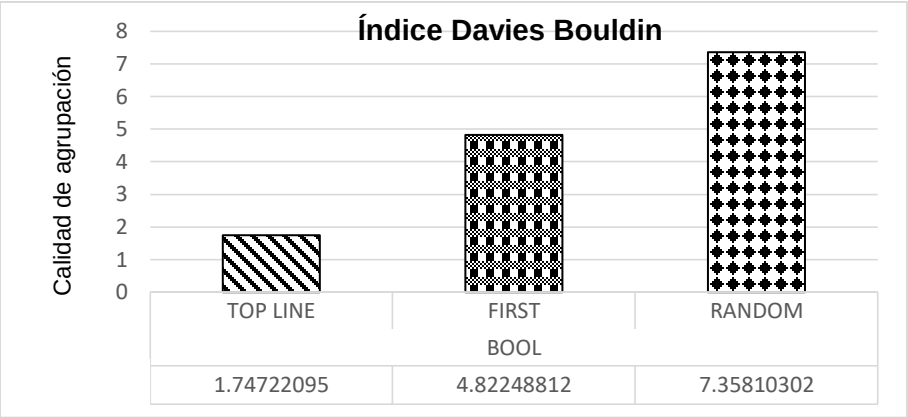
5.7.2.2 Índice Davies Bouldin

De acuerdo a la gráfica 5.22 podemos ver que el índice Davies Bouldin puede evaluar correctamente la relación entre la calidad de los grupos de oraciones generados y la calidad de los resúmenes utilizados usando el pesado por frecuencia de término (*tf*). Las evaluaciones muestran mejor calidad en la línea base tope (top line), seguido de la línea base (*first*) de primeras oraciones y (*random*) de oraciones aleatorias. Al aplicar preprocesamiento y usar bigramas, las calidad de las agrupaciones aumenta considerablemente.



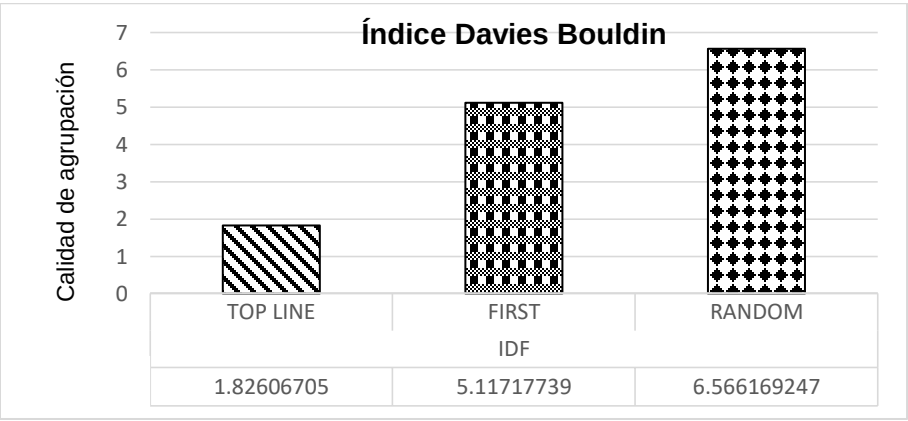
Gráfica 5.22 Evaluación de calidad de agrupación con el índice Davies Bouldin usando bigramas con el pesado por frecuencia del término en el corpus DUC-2002.

De acuerdo a la gráfica 5.23 podemos ver que el índice Davies Bouldin puede evaluar correctamente la relación entre la calidad de los grupos de oraciones generados y la calidad de los resúmenes utilizados usando el pesado booleano (*bool*). Las evaluaciones muestran mejor calidad en la línea base tope (*top line*), seguido de la línea base (*first*) primeras oraciones y (*random*) oraciones aleatorias. Al aplicar preprocesamiento y usar bigramas, las calidad de las agrupaciones aumenta considerablemente.



Gráfica 5.23 Evaluación de calidad de agrupación con el índice Davies Bouldin usando bigramas con el pesado booleano en el corpus DUC-2002.

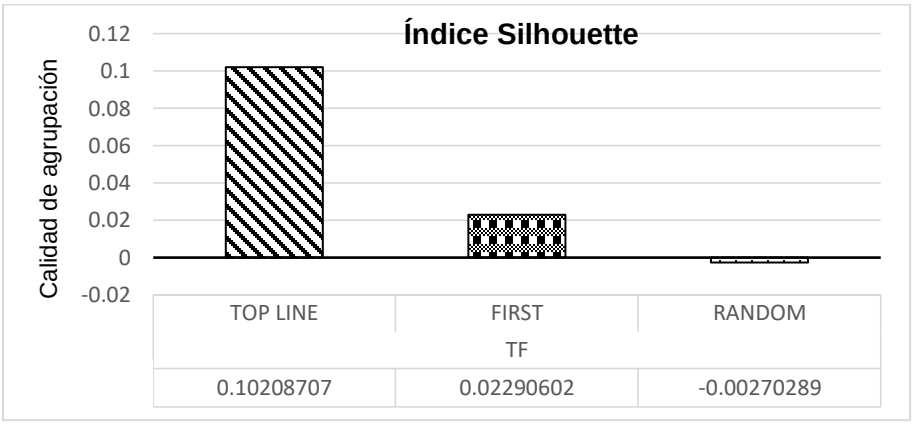
De acuerdo a la gráfica 5.24 podemos ver que el índice Davies Bouldin puede evaluar correctamente la relación entre la calidad de los grupos de oraciones generados y la calidad de los resúmenes utilizados usando el pesado de frecuencia inversa del documento (*idf*). Las evaluaciones muestran mejor calidad en la línea base tope (*top line*), seguido de la línea base (first) de primeras oraciones y (random) de oraciones aleatorias. Al aplicar preprocesamiento y usar bigramas, las calidad de las agrupaciones aumenta considerablemente.



Gráfica 5.24 Evaluación de calidad de agrupación con el índice Davies Bouldin usando bigramas con el pesado de frecuencia inversa del documento en el corpus DUC-2002.

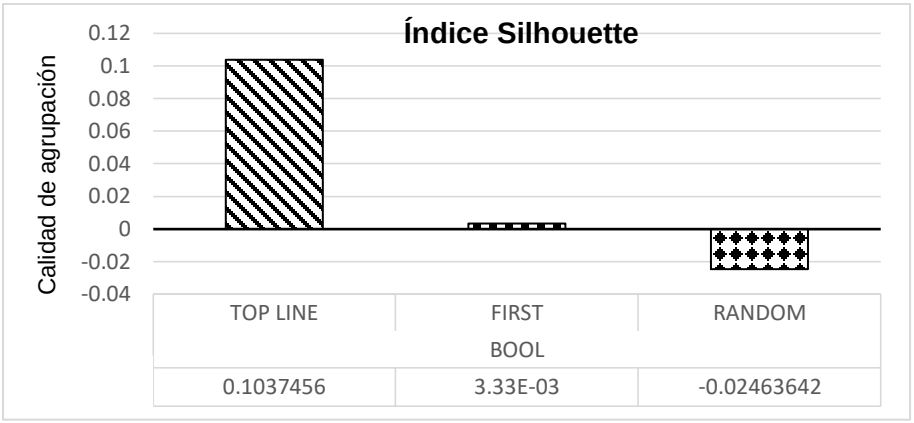
5.7.2.3 Índice Silhouette

Las evaluaciones por parte de este índice también fueron mejores al aplicar preprocesamiento a los documentos. La gráfica 5.25 muestra como el índice Silhouette sigue manteniendo el comportamiento esperado respecto a las líneas base. Las evaluaciones muestran mejor calidad en la línea base tope (*top line*), seguido de la línea base (*first)* de primeras oraciones y *(random)* de oraciones aleatorias*.* Al aplicar preprocesamiento a los documentos se obtiene una mejoría en la calidad de los grupos, es decir, la evaluación de la calidad de las agrupaciones es mejor. Este experimento se hizo con el pesado por frecuencia de término (*tf*).



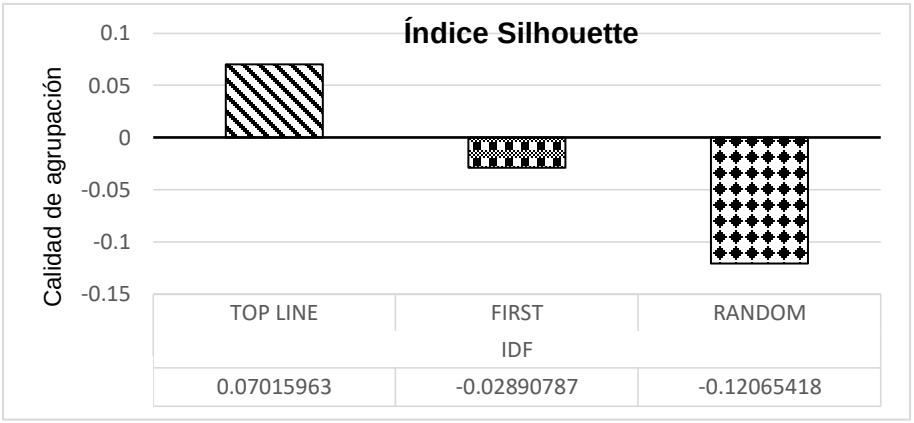
Gráfica 5.25 Evaluación de calidad de agrupación con el índice Silhouette usando bigramas con el pesado por frecuencia de término en el corpus DUC-2002.

La gráfica 5.26 muestra que este índice es capaz de evaluar la relación en la calidad de los grupos de oraciones generados y la calidad de los resúmenes obtenidos. Se usa el pesado por booleano (*bool*) el cual muestra buen acoplamiento con este índice. Las evaluaciones de este índice permiten distinguir claramente la calidad de una línea base de otra. Las evaluaciones muestran mejor calidad en la línea base tope (*top line*), seguido de la línea base (*first)* de primeras oraciones y *(random)* de oraciones aleatorias. Al aplicar preprocesamiento a los documentos se obtiene una mejoría en la calidad de los grupos, es decir, la evaluación de la calidad de las agrupaciones es mejor.



Gráfica 5.26 Evaluación de calidad de agrupación con el índice Silhouette usando bigramas con el pesado booleano en el corpus DUC-2002.

La gráfica 5.27 muestra que existe un problema entre el índice Silhouette y el pesado de término de frecuencia inversa del documento (*idf*), ya que las evaluaciones de calidad de agrupación se ven afectadas, es decir, disminuye la calidad del agrupamiento. Las evaluaciones de este índice permiten distinguir claramente la calidad de una línea base de otra. Las evaluaciones muestran mejor calidad en la línea base tope (*top line*), seguido de la línea base (*first)* de primeras oraciones y *(random)* de oraciones aleatorias.



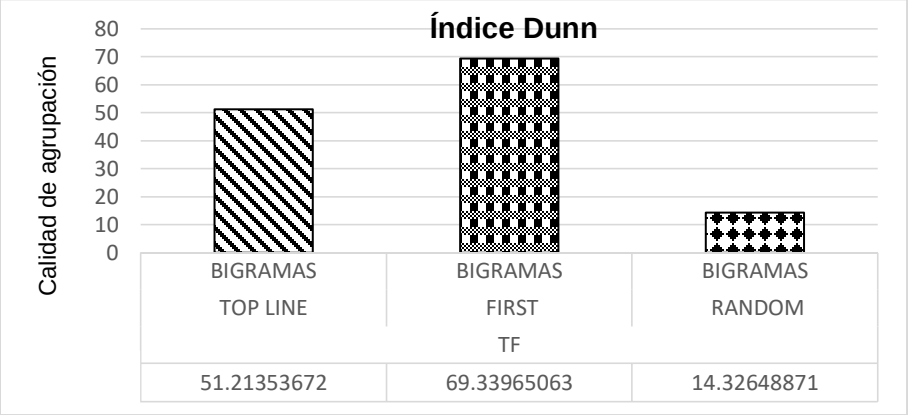
Gráfica 5.27 Evaluación de calidad de agrupación con el índice Silhouette usando bigramas con el pesado de frecuencia inversa del documento en el corpus DUC-2002.

5.8 Experimentos con preprocesamiento en DUC-2001

Los siguientes experimentos se realizaron sobre el corpus DUC-2001 y solo se hicieron en documentos a los cuales se les aplicó preprocesamiento. Posteriormente se usaron dos modelos de representación de texto los cuales son bolsa de palabras y n-gramas y un pesado de términos que es frecuencia del término.

5.8.1 Índice Dunn

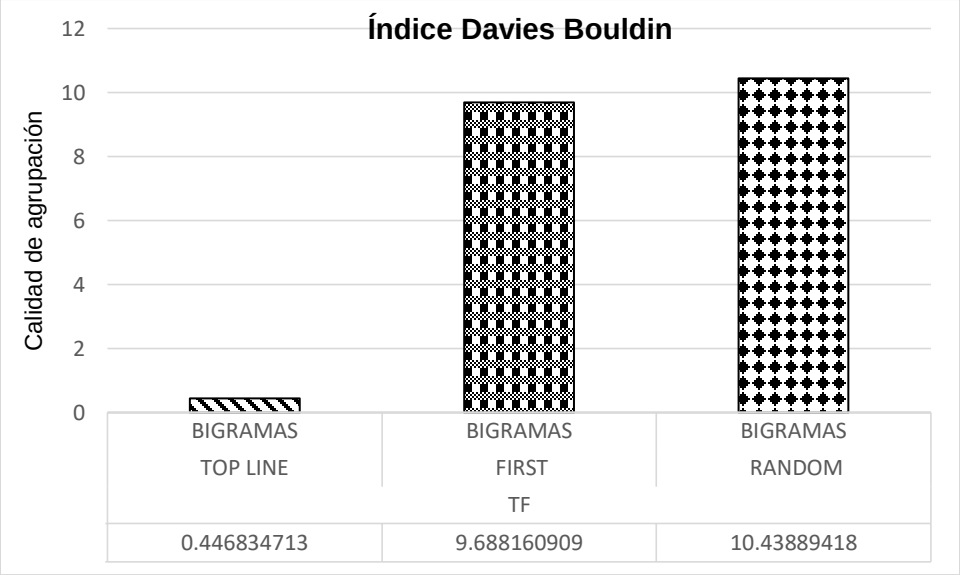
Como podemos ver en la gráfica 5.28 el índice de Dunn da la mejor evaluación a la línea base primeras oraciones (*first*), a lo largo de todos los experimentos el índice de Dunn ha mostrado un mal desempeño en la evaluación de la calidad de las agrupaciones generadas a partir de oraciones clave. Este índice queda descartado para la evaluación de calidad de grupos de oraciones. Este experimento se realizó con el modelo de bolsa de palabras y el pesado por frecuencia de término.



Gráfica 5.28 Evaluación de calidad de agrupación con el índice Dunn usando bolsa de palabras con el pesado de frecuencia del término en el corpus DUC-2001.

5.8.2 Índice Davies Bouldin

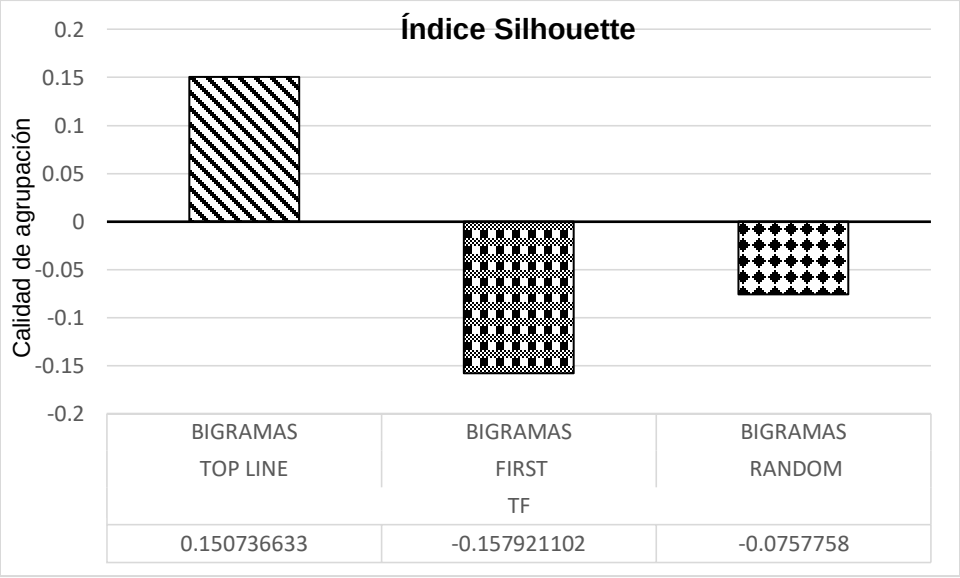
De acuerdo con Ledeneva [Ledeneva, 2014] se sabe que la línea base donde se toman las primeras oraciones es buena para noticias y artículos, pero cuando se ocupa una colección de documentos con temas variados como DUC-2001, se espera obtener mejores evaluaciones en la línea base oraciones aleatorias (*random*) que en la línea base primeras oraciones (*first*). A pesar de esto el índice de Davies Bouldin sigue manteniendo mejores evaluaciones en *top line.*



Gráfica 5.29 Evaluación de calidad de agrupación con el índice Davies Bouldin con preprocesamiento usando bigramas con el pesado de frecuencia del término en el corpus DUC-2001.

5.8.3 Índice Silhouette

El gráfico 5.30 muestra como el índice Silhouette puede evaluar coherentemente la relación de calidad de los grupos de oraciones de las tres líneas base, pero además, también evalúa de acuerdo a lo esperado con esta colección de documentos. Debido a que contiene documentos con temas variados, se esperaba mejor calidad en la línea base aleatoria (*random*) que en la línea base de primeras oraciones (*first*), y eso es lo que sucedió.



Gráfica 5.30 Evaluación de calidad de agrupación con el índice Silhouette con preprocesamiento usando bigramas con el pesado de frecuencia del término en el corpus DUC-2001.

5.9 Resumen

De acuerdo con las gráficas anteriores *top line* muestra mejores resultados con dos índices que son: Silhouette y Davies Bouldin, así mismo las mejores evaluaciones son con el modelo de representación de texto de n-gramas donde n=2. Con esto comprobamos que efectivamente las línea base tope (*top line*) es mejor que las líneas base de oraciones aleatorias (*random*) y primeras oraciones (*first*). En cuanto al índice de Dunn pudimos comprobar que no es posible evaluar la calidad de conjuntos de oraciones clave de resúmenes con este índice.

Capítulo 6:
Conclusiones

6.1 Aportaciones

Las aportaciones obtenidas de este trabajo son las siguientes:

- Se caracterizó el comportamiento de oraciones clave de resúmenes mediante medidas internas de calidad.
- Se demostró que dar preprocesamiento al texto es de gran utilidad para mejorar la calidad de las evaluaciones por parte de las medidas internas de calidad de agrupación.
- Se demostró que los índices Silhouette y Davies Bouldin pueden evaluar correctamente grupos de oraciones, mientras que Dunn falla al evaluar estos grupos.
- El correcto modelado y pesado de las oraciones ayudan a obtener un mejor agrupamiento.
- Se demostró que el modelo de representación de texto ayuda a obtener mejores evaluaciones sobre los grupos de oraciones.

6.2 Conclusiones

La necesidad de evaluar la calidad de grupos generados mediante algoritmos de aprendizaje no supervisado ha permitido la creación de métodos que permitan evaluar estas agrupaciones.

Se han desarrollado medidas que permiten evaluar la calidad de las agrupaciones y son dividas en medidas de calidad externas e internas.

Las medidas de calidad externas de agrupación solo pueden aplicarse cuando se tiene conocimiento previo del conjunto de datos con el que se va a trabajar pero no pueden ser aplicadas a conjuntos de datos desconocidos, es decir, no se pueden aplicar a cualquier conjunto de datos que queramos evaluar.

Por otra parte las medidas de calidad internas de calidad de agrupación ya han sido probadas con éxito para la evaluación de grupos de documentos y conjuntos de datos sin tener conocimiento previo del conjunto de datos que se va a evaluar.

Las medidas de calidad internas de agrupación permitieron evaluar la relación que existe entre la calidad de los grupos generados con la calidad de los resúmenes utilizados. Es decir que, entre mejor sea la calidad de las agrupaciones, mejor es la calidad de los resúmenes obtenidos.

Por otra parte, hasta ahora se desconocía si estas medidas de calidad internas de agrupación eran capaces de evaluar correctamente grupos de oraciones.

A partir de esta investigación se concluye que dos medidas de calidad interna de agrupación que son; índice Davies Bouldin e índice Silhouette pueden ser usadas para tareas de evaluación de conjuntos de oraciones de un texto.

Estas medidas pueden evaluar coherentemente grupos de oraciones independientemente del idioma y el dominio con el que se esté trabajando. Debido a esto podemos ocupar estas medidas en trabajos donde sea necesario evaluar la calidad de grupos de oraciones como el trabajo realizado por García [García, 2002], en el cual propone un enfoque de generación de resúmenes automáticos extractivos por agrupación independiente del idioma y del lenguaje. En este trabajo se generan grupos de oraciones con un algoritmo de aprendizaje no supervisado, pero los grupos son evaluados mediante medidas de calidad externas de agrupación, por lo que podemos aplicar medidas internas de calidad de agrupación para evaluar los grupos de oraciones generados en su trabajo.

6.3 Trabajo futuro

Ahora que sabemos que se pueden evaluar grupos de oraciones con medidas internas de calidad de agrupación, podemos estudiar problemas en donde sea viable agrupar oraciones, por ejemplo la evaluación de grupos de resúmenes que propone Alexander Gelbuckh [Gelbukh, 2005], para determinar las ideas principales de un texto, evaluación de grupos de frases clave e incluso para tareas poco comunes como la clasificación de poemas y canciones de acuerdo a su contenido.

Anexos

Calidad de agrupación de cada documento sobre el corpus DUC-2002 con tres líneas base con el modelo de representación de texto bolsa de palabras y pesado de frecuencia de término

TOP LINE			BASELINE FIRST			BASELINE RANDOM		
DB	SILHOUETTE	DUNN	DB	SILHOUETTE	DUNN	DB	SILHOUETTE	DUNN
1	0.01961	2.342466	4	0.082543	0.008114	1	-0.0345	23
1	0.088435	10.28571	1	-0.0741	-0.01203	1	0.071253	24.66667
1	0.431243	4.441558	1	0.025542	0.024246	3	0.135569	13.66667
2.5	0.062392	7.376471	2	0.040353	0.028184	2	0.048235	6.339623
1	0.187067	16.33663	1.5	-0.09298	-0.00526	2	0.009632	29
0.857143	0.157561	14.89975	1.25	-0.0119	-0.00123	0.625	-0.00387	21.2
1	0.152043	13.76209	2	0.044481	0.009092	1.571429	0.038112	14.72727
0.846154	0.300922	6.847552	0.25	0.341254	0.312733	1.333333	0.006506	4.596208
0.5	0.119512	12.49841	2	0.042371	0.010571	1.333333	-0.01304	15.10499
1.666667	0.042608	11.01333	0.5	0.019196	8.20E-04	0.75	0.020737	9.103211
1.5	0.063371	19.95221	4	-0.04486	-0.00606	2	0.023636	21.5
1.25	1.05E-03	23.4	0.571429	-0.01484	-0.00608	0.857143	0.055839	27.71429
0.222222	0.017827	9.346899	1	-0.00213	-0.00143	0.75	0.019834	18
2	0.167144	23	4	-0.48036	-0.01445	0.5	0.028161	9.655172
0.333333	0.065251	18.9313	2	0.063513	0.004633	3	0.093279	6.135693
1	0.266302	6	0.5	0.070064	0.004408	7	-0.0739	8.941799
0.25	-0.07073	27	3	-0.01155	6.10E-04	0.428571	0.054298	30.85714
2.333333	-0.10801	18	1.5	-0.00105	-0.00706	1.5	0.062405	16
1	0.289942	16	1.2	-0.01453	-0.00462	1.6	0.12819	10.12941
0.416667	1.184913	8.194178	0.333333	0.127114	0.008797	8.75	0.001244	11
1.166667	0.038933	19.28571	0.181818	0.00815	-0.00572	19	0.011293	17
6	-0.00752	9	3	-0.93333	-0.00708	1.333333	-0.05148	25.5
0.666667	0.101253	22.61279	6	0.016338	-0.00196	1.2	-0.0908	14.5045
0.380952	0.072871	13.86578	4.5	6.38E-03	-0.00178	7	0.028896	15.30495
1.5	0.038054	14.74699	1.333333	-0.07058	-0.00337	0.727273	0.013004	23.86272
2	-0.00546	27.25	1.333333	-0.15011	-0.0032	1.8	-0.01011	29.66667
0.5625	0.182867	9.311176	0.3	0.002685	0.003344	0.666667	0.001232	12.14352
0.75	-0.03932	7.819588	1.666667	-0.04381	-0.01225	1	0.036019	9.572072
0.25	0.013064	6.285714	1.333333	-0.03458	-0.0387	0.75	-0.02733	25.75
2.5	-0.06145	13.03968	1.333333	-0.01856	-0.00648	0.666667	0.004288	20.5
0.769231	0.017551	16.18322	1.571429	-0.08685	-0.01059	1	0.087036	19.71564
1	0.175349	11.57661	1.5	-0.93	0.006744	0.833333	0.031359	17.4

1.333333	-0.01213	8.618012
2.25	0.111487	17.10586
1	0.106105	10.7563
3	0.035733	8.484848
0.909091	0.934475	14.8
0.666667	0.056236	12.7572
2	0.044104	9.18
1.666667	0.429486	10.24681
1.166667	0.497267	14.19807
2.4	0.03702	17.62895
0.5	0.06716	9.274809
0.2	0.058358	16.78661
1.333333	0.40393	12.92
0.571429	0.082331	16.3414
2	-0.03453	12.5
1	-0.03453	14
0.176471	0.067659	25.82353
2	-0.02148	21
1	-0.01903	23.5102
3.75	0.045967	12.78939
0.166667	0.101466	8.871391
1.666667	0.197442	15.16667
0.875	0.250954	14.85714
0.428571	0.135558	10.28125
1.428571	-0.06924	15.71429
0.769231	0.093059	19.8
0.916667	0.307528	13.25
1.428571	-0.1312	12.03497
2.25	0.366851	9.375
5	0.005884	16.625
1.666667	0.360273	11.06667
0.333333	0.010437	3
3	0.246679	6.666667
0.416667	0.084416	19.2
2.25	0.378624	24.11111
0.5	0.300338	13.375
0.333333	0.068358	14.05512
0.5	0.366053	15.02825
1.5	-0.26445	14.68429
1.6	0.149313	7.823529
1.333333	-0.02142	24.47917
0.25	-0.11479	17.83486
0.8	-0.07415	23.5
0.166667	-0.01097	2.61E-04
3.333333	-0.02952	-0.01392
1.333333	0.064297	0.017711
1.5	0.492716	0.018388
0.666667	0.021176	0.015562
1	0.008182	-0.0038
2.333333	0.010037	9.47E-05
0.833333	0.010838	-0.00158
5	0.15896	0.018967
0.857143	-0.00267	-0.00599
1.5	0.065089	0.016119
0.2	-0.10683	-0.00356
1.666667	0.005047	0.004247
0.5	0.0077	6.06E-04
0.75	-0.10357	-0.00449
0.333333	-0.07319	-0.0158
0.75	0.060197	0.003204
1	-0.04496	-0.00177
3	-0.01954	-0.00177
1.125	-0.01882	-0.00146
4	0.010371	-0.00366
2	0.004949	-0.00281
0.777778	0.001786	0.003527
3.333333	-0.0132	-0.00845
1.875	-0.00598	-0.00556
1.384615	-0.0013	-0.00502
3	0.088887	0.043531
0.615385	-0.0104	-0.00444
0.444444	0.047012	0.005457
0.5	0.170283	0.007659
0.545455	0.013322	0.011824
1	-0.15006	-0.00567
2	0.030825	0.015947
1	0.002675	-0.00257
3	0.031956	0.005811
0.875	0.076153	0.014
2	0.021321	-0.00692
1	0.095455	0.001432
8	0.021272	-0.00502
1.25	-0.00149	-0.01449
0.2	-0.02058	-0.00453
2	-0.19641	-0.00279
4	-0.14396	-0.01967
1.75	-0.00849	23.75
6	0.039142	20.5
0.5	0.013728	6.217391
0.25	0.129033	15.84091
1.25	0.062099	12.38804
0.625	0.056231	14.3
0.909091	3.98E-05	17.7471
1.142857	0.065902	24.625
3.5	0.040567	24.85714
6.5	0.005785	18
2	0.16764	9.041096
1.875	0.007268	15.875
1.4	0.053847	14.23874
2.4	0.023327	19.8
0.6	-0.01084	20.66667
1	0.162204	4.222222
0.083333	0.013266	4
0.5	0.006914	20.27676
0.5	0.073529	13.125
1.25	0.014808	23.75
2	0.038464	21.5
2.6	0.015928	13.64765
0.869565	-0.01827	19.52174
1.222222	-0.04597	21.22222
0.888889	0.097695	26.33333
0.625	0.014549	27
1.571429	0.044625	14.18255
2	0.005667	15.20556
1.090909	0.052215	18.93871
5	0.054674	16
0.916667	-0.03068	8.000581
0.714286	-0.0165	19.71429
1.333333	0.01363	3.888889
1.75	0.085594	22.2
0.75	-0.05302	28.91667
12	0.01101	9.5
4	0.049431	8.285714
1	0.021997	9.391304
4	0.132023	11.52778
2	0.048004	5.430485
1.333333	0.177545	31.75
3.5	0.034466	13.5
4	0.056575	8.07874

1	0.350087	19.125
0.666667	0.003212	17.75
0.666667	0.095205	16.77539
2	0.06143	8.955224
0.333333	-0.21199	12
2	-0.04923	5.365385
0.666667	-0.23259	12.2293
1	0.271492	2.571429
1.461538	0.053536	17.21053
1.666667	-9.46E-03	22.68613
1.285714	0.018571	22.88889
1	-0.09153	11.00971
0.454545	0.105722	16.81818
1.5	-0.01566	9.166667
1	-0.03991	9
1.285714	-0.05287	8.413866
1	0.103202	11.36877
4	-0.08267	3
1.5	-0.24777	12.7875
0.8	0.353555	12.375
2	0.471574	12.4
0.833333	0.023251	13.4
0.333333	0.302766	8.329412
3	-0.08829	16.22727
1	0.018064	17.37882
0.666667	-0.02454	11.75
3	0.205216	9
1.5	0.035782	19.33333
5	0.05022	8.405735
4	0.08345	13.69444
4.25	0.042106	12.48971
0.5	0.417479	15.33333
1.2	-0.08735	19.83333
1.857143	0.078688	14.63926
1.25	0.841377	11.93782
7	0.215778	4.032258
4	0.366313	2.649083
0.5	0.035936	8.369565
0.8	0.006685	20.22899
0.571429	-0.09291	14.26829
6	0.229966	11
1.4	0.020573	6.102687
4.5	0.072507	15.44444
1.4	-8.60E-04	-0.0049
1	-0.94662	-0.00988
1.666667	-0.06472	-0.00106
1	0.035807	6.54E-04
0.5	-0.06858	-0.00809
2	-0.05228	-0.01289
1.75	-0.00376	-0.01108
2	0.134811	0.016953
0.785714	-0.00849	-0.00461
1.25	-0.01128	-0.00824
0.454545	-0.01054	-0.00433
0.5	-0.18768	-0.01632
0.75	-0.01754	-0.00281
1.5	0.005694	-0.00631
0.333333	0.155404	0.014496
0.833333	-0.97126	-0.02192
1.142857	0.013453	-0.00332
1.5	-0.92248	-0.01377
1.333333	-0.19459	-0.0225
0.666667	0.229305	0.066336
0.5	0.042525	0.016479
4	0.003051	0.00119
4	0.149206	0.003479
1.5	0.110313	-0.00175
1	0.085093	0.004033
0.2	-0.05023	-0.00467
0.5	0.114669	-0.00202
3	0.011853	-0.00182
10	0.018447	-0.00699
3	-0.97712	-0.03164
0.583333	-0.02383	-0.00202
0.777778	0.076262	0.005271
1	-0.02267	-0.02166
0.8	-0.09687	-0.00739
1	0.034552	0.028724
0.375	-0.08237	-0.00813
1.333333	0.212174	0.028247
0.25	-0.01788	0.001469
1.666667	0.005615	-0.00112
5.5	-0.984	-0.00403
1.8	0.066868	0.022633
5	-0.06642	-0.00185
1.75	-0.00691	-0.00486
0.833333	-0.00295	15.9695
1.333333	0.043258	12.33333
0.8	0.042229	14.59908
1.5	0.195057	11.51613
1	-0.1208	23.33333
1.25	0.018283	13.56522
2	0.082569	6.120219
2	-0.07078	15
1.5	0.037264	19.46089
2.6	0.009965	24.69231
0.75	0.003079	25.08333
0.25	0.025457	6
1.5	0.054969	13.00333
2	0.043231	9.625
0.5	0.238513	2.689655
5	0.047751	9.141622
0.75	0.011923	16.54878
5	-0.03966	4.568421
1	0.078322	8.064516
1	0.019002	10.25
3.5	0.0259	6.024845
1.666667	-0.01249	28
0.5	0.020107	7.886598
2	0.062608	18.5
2	0.162056	15
1	0.082714	10.91329
4	-0.00749	20.96635
0.266667	-0.00491	27.25
0.214286	-0.02171	20
1.166667	0.009909	11.1811
0.823529	-0.00413	16.78571
0.076923	-0.00129	12
0.666667	0.097114	7.518519
1	-7.21E-04	21.53333
0.727273	0.022454	11.78045
3	0.021852	3.379353
0.8	0.019334	9.246226
3.5	0.016897	13.74675
1.1	0.04236	24.00785
2.333333	0.0094	14.45957
0.727273	-0.01478	10.1654
0.166667	0.021651	3.25
0.666667	0.099039	7.687198

1	0.065867	17.10601	2	0.006252	-0.00502	2.714286	0.010619	17.85714
0.5	-0.23358	18	9	-0.04886	-0.00719	1.2	-0.01889	34.4
1.777778	0.025377	11.18681	1.714286	-0.03409	-0.00572	2.6	0.02455	12.66307
1	0.054681	14.46809	1	0.065169	0.027732	0.75	0.038893	17
1	-0.23189	14.10612	1.2	-0.96942	-0.02065	0.5	0.066542	14.25
0.666667	0.190052	20	0.2	-0.01416	-1.39E-04	1	0.129522	16.64083
1	0.212114	12.04518	0.434783	0.002063	-8.04E-04	1.769231	-0.01017	20.30769
0.333333	0.088462	16.88596	2	0.045865	0.00755	1.2	0.013623	10.1514
0.583333	0.699188	13.75	1	0.478952	0.105174	2.25	-0.02234	13
5	0.234538	9.053571	0.75	-0.00277	-7.73E-05	0.285714	0.068654	14.64498
1	0.151468	10.70554	1.142857	-0.00653	0.004597	1.25	-0.02365	7.1775
0.428571	0.124754	22.66667	1.8	-6.88E-03	-0.0061	0.363636	0.031365	16
2	0.097459	14.92593	3	-0.04431	-0.01603	1	0.018106	7.312441
1	0.180921	6.169492	2	0.010567	0.00853	3.5	0.224279	22.71429
0.666667	0.404548	13.93976	0.181818	-0.94853	0.005473	0.333333	0.052343	5.661558
0.538462	0.070285	9.761273	2.4	-0.00957	-0.00568	2.75	0.035544	12.54087
0.714286	0.460977	11.8	2	0.014609	0.017064	0.461538	0.115576	14.40045
1.571429	0.288163	7.818182	0.5	0.040127	0.027729	1.25	0.042764	4.931564
1	-7.03E-02	12.2	1.2	-0.02949	-0.00394	2	0.059092	19
1.625	0.18	10.46154	1.5	0.048349	0.019826	0.8	0.045555	12.2
0.846154	0.170563	17.07692	3	0.305143	0.050877	2.5	0.089011	14.69118
0.823529	0.985626	13	2.4	0.020104	0.011258	1.05	0.012456	12.14286
0.4	0.015553	14.84774	1	-0.0289	-0.00424	1.5	0.020624	13.328
1	-0.01842	4.271845	0.666667	0.028986	-0.00406	0.333333	-0.02797	17
1.666667	0.048412	14.58	1	-0.01463	-0.0055	2.333333	0.07328	19.42857
1	-0.03793	2.304878	1	0.074617	2.85E-04	4.5	0.032721	9
3.5	0.01503	10.31837	0.454545	-0.98127	-0.00941	1.285714	-0.06777	25.11111
0.8	0.488527	17.54904	1.75	0.015893	0.007915	1	0.045328	19.2
0.333333	-0.11318	25.66667	1.4	-0.0289	-0.00564	11	0.013689	6
1.75	0.086981	12.75	6	-0.01386	-0.00272	0.75	0.009052	8.822917
1.454545	-0.07506	8.916045	0.6	-0.05375	-0.00546	1.625	-0.01896	19.01655
2.666667	0.00486	16.8398	1.4	-0.01403	-0.00712	0.777778	0.009742	21.33333
1.1	-6.91E-04	20.3174	0.090909	-0.03131	-0.00142	2	0.003833	23.125
1.666667	0.069934	17.98588	3.5	0.010224	-8.53E-04	5.333333	0.017036	23.33333
2	0.152698	16.70625	2	-0.00951	-6.87E-04	3.4	-0.0061	15.75899
1	-0.31099	11.4	1.5	0.024855	-0.00173	4	0.057941	7.789474
2.666667	-0.08801	9.666667	1	-0.07534	-0.02034	0.833333	-0.01673	10.65672
1.666667	-0.13933	17.2	2	0.255859	0.014809	10	0.021931	13
0.714286	0.313065	14.42857	1.166667	0.042705	0.028106	0.705882	0.01871	12
1	-0.02767	16.42273	0.636364	-0.11477	-0.01033	1.2	-0.00612	18.33333
1	-0.02767	16.53453	2.142857	-0.11477	-0.00915	0.888889	0.019638	14.1928
0.5	-0.04738	9.165829	2	-0.01497	-0.00187	0.5	-0.0164	17.75391

1.333333	-0.04738	22.66667
1.5	0.013137	19.5
0.25	0.113965	15.90323
0.8	0.048597	6.194805
1	0.050661	12.83333
2	0.348739	6.757764
2	0.006617	26.25
1.3	0.006617	20.23077
0.75	0.669198	10.33333
0.2	0.018172	22.08
1	0.115502	20.92308
0.333333	-0.03315	5.333333
0.5	0.070713	13.3599
1.285714	0.166365	10.59696
1.8	0.030622	15.72998
1	-0.05663	16.28571
2	0.041453	8.16
1.666667	-0.02556	9.089109
1	-0.18365	9.454545
1	0.108653	16
2	0.007611	15.13925
2.25	0.116274	9.76822
2.5	0.111914	9.927273
0.666667	-0.08621	20.125
1.666667	0.025501	16.6
3	-0.15886	5.586207
0.333333	-0.62025	14.97416
1	0.169752	13.55556
2.333333	-0.07184	18.33333
2	0.176747	8.941667
3	0.035583	13.02128
0.714286	0.040145	14.73774
1.285714	0.080299	15.37582
1	0.076635	19.5
1	0.084643	16.00966
1.8	0.359798	10.88831
0.666667	0.228341	13
2.5	0.24147	16
1.166667	0.755	13.08333
0.875	0.736439	14.56041
1.714286	0.086757	17.71429

2	-0.01497	-0.00713
0.5	0.013108	-0.00428
2	0.016003	-0.00559
3	-0.01333	-0.00511
3	0.005218	-2.84E-04
1	0.106624	0.007413
0.333333	-0.00719	5.82E-04
1	-0.00719	-8.16E-04
0.181818	0.246424	0.018297
0.6	-0.10815	-0.04704
3	-0.0102	0.005086
1	0.024131	-3.48E-06
0.6	0.013236	-0.00764
0.75	0.02106	0.002977
0.416667	-0.00983	-0.00277
1	-0.02383	-0.00361
2	-0.01803	-0.00117
2.5	-0.02005	-0.01027
3	0.016177	-0.00413
0.2	0.111124	0.006063
0.636364	-0.00724	-8.31E-04
4	-0.0124	1.39E-04
0.5	-0.06451	-1.04E-04
1	-0.01559	-0.01754
2	-0.05508	-0.00531
1	-0.02239	-0.0058
1	-0.00638	-0.00477
6	-0.94215	-4.41E-04
2	-0.09293	-0.00213
5	0.027122	0.006637
0.666667	0.019939	0.009039
3	0.037038	0.005568
1.142857	0.009627	0.006161
1.25	0.013084	0.006523
0.785714	-0.01023	-0.00466
1	0.137726	0.050133
0.333333	0.039305	0.003037
1.166667	0.041363	0.015287
1.666667	0.029877	0.018743
2.333333	0.036033	0.016213
0.714286	0.007387	5.07E-04

1.5	0.138882	14.09859
0.571429	-3.61E-04	10.34911
1.111111	-0.07342	16.98047
1.666667	0.035459	12.38095
2.166667	0.017936	10.88533
0.5	0.081043	3.543307
1.6	-0.0051	21.61818
1	0.215421	27
4	0.045731	8.567308
1	0.03565	13.69014
0.833333	0.002157	23
1.5	0.12203	22.5
1.6	0.086006	16.96559
1.166667	0.009244	8.430451
1	7.84E-04	28.84615
1.2	-0.16821	9.52
0.857143	0.031688	11.70868
1.2	-0.01304	15.86207
0.285714	-0.06882	26
1	0.069456	21
0.625	-0.00312	15.8125
0.6	-3.20E-02	16.4382
0.857143	0.022781	17
1.75	0.015251	18.5864
1.125	0.021467	24.125
1	0.036947	3.266667
0.333333	0.027399	13.52657
3	0.001358	22.04206
3.333333	-0.01531	19
1	0.00158	13.3413
0.9	-0.04008	20.37587
8	-0.00268	14.24202
1.5	0.097943	17.27647
1.2	0.020106	7.938075
1	0.185509	18.01959
1.625	0.01547	11.8696
1.857143	0.005064	15.0696
0.5	-0.01822	15
1.222222	0.110676	16.11111
2	0.025418	14.88889
1.333333	0.019342	12.45662

2.428571	0.086757	12.16919
0.75	-0.04229	10.76646
3	-0.04957	11
1	0.398661	16.90909
0.666667	0.026603	13.44
2	0.103014	17.5
1.2	0.037862	20.2
1.625	-0.11954	15.86741
1.8	-0.06575	12.51953
0.888889	0.04046	19.75
2.25	0.00237	20.24099
2.25	-0.21135	20.44444
0.333333	0.194431	8.470588
0.6	-0.0187	11.25255
1.333333	0.249714	11.75
1	0.25434	14.4105
3.333333	0.247162	12.025
3	0.105881	7.518072
1.5	0.062917	14
0.666667	0.383223	5.621622
1.666667	0.285807	16
0.857143	-0.07519	17.16667
5	-0.06783	10.26207
0.5	0.134459	11.91667
0.666667	0.228393	10.75269
1.25	-0.0586	10.05161
4	0.183721	7.171315
3.333333	0.017646	14.81932
2	0.082097	6.980989
0.5	0.018078	10.53837
0.5	-0.38397	1.657895
1	0.714488	13.88889
2	-0.01524	20.49102
0.692308	-0.06218	20.33333
4	-0.14496	6.495727
1.333333	-0.14619	19.25
1.666667	-6.99E-03	16.06858
0.5	-0.32575	13.37759
0.6	0.018127	22.8
2.333333	-0.03882	16.2004
1	0.113004	13.16832
1.666667	0.058384	13.5633
0.5	0.226945	17.25
1.2	0.006701	-0.00273
1	-0.03473	-0.00173
1.666667	-0.03354	-0.01479
4.5	-0.01135	0.006885
3	0.018198	0.001053
0.5	-0.10256	-0.00546
1.25	-0.01838	-0.0059
1	-0.04	-0.0061
1.166667	-0.03173	-0.00807
1.375	-0.05664	-0.00543
0.25	0.025182	-0.00241
0.166667	-0.00846	-0.03873
1	0.058636	0.033441
2.5	-7.91E-04	-0.01156
0.714286	0.002813	0.003363
1.333333	0.027284	0.013495
5	0.04407	0.004458
0.666667	0.047091	0.005683
0.5	0.442589	0.04262
1	0.056801	0.021743
1.5	0.059325	0.008831
0.5	-0.02598	-0.0066
3	-0.16251	-0.00366
5	0.097924	0.014151
2.666667	0.03139	0.003188
2	0.019778	-0.00597
0.625	0.016667	-0.00304
4	-0.0161	-0.00429
1.666667	0.009324	-0.00718
3.5	-0.01084	-0.00454
0.333333	-0.01488	-0.00719
2.666667	0.115176	0.01791
0.4	-0.12376	-0.00667
1.142857	-0.02568	-0.00298
1	-0.00885	5.81E-04
0.666667	-0.00384	-0.00719
2.571429	0.014216	-4.86E-04
1.166667	-0.00688	-0.0035
1	-0.01729	-0.00133
4	-0.01363	-0.00409
0.75	0.026664	9.14E-04
0.5	-0.00797	-0.0081
1.2	0.030331	0.008115
1.266667	-0.00884	16.76426
1	0.002046	13.06849
7	2.74E-02	11
1.428571	0.022456	12.11657
0.666667	0.018061	21.5
3	-0.00969	20.94949
2.5	0.026817	12.6
1	-0.03416	23.4
0.75	0.019274	17.11704
2.5	0.031026	25.1
0.25	-8.16E-04	10.27419
2	0.150519	26.5
0.285714	0.041796	15.12088
3	0.036674	21.10538
0.705882	5.24E-05	19.91667
1.230769	-0.03086	5.106001
0.545455	0.008556	10.51865
1.25	-0.01315	11.66139
2	-0.01774	16.82609
0.666667	0.112386	16.5
2	0.067815	11.79394
1	0.062947	7.261905
1	0.052782	28
8	0.064398	7.727273
1.2	0.079952	11.43839
1.5	0.018756	14.96444
0.2	0.033798	14.5
1.571429	-0.00884	17.54909
4	0.096175	27.375
0.230769	-0.00559	6.333333
0.333333	0.068828	8.658854
16	0.034894	13.04688
0.428571	-0.01793	32.66667
11	0.005546	19.07661
2	0.073728	11.43855
0.714286	0.042245	19.83333
0.571429	2.45E-02	15.44041
0.5	0.039761	16
1.5	0.012004	14.86452
0.75	0.009746	12.60417
0.571429	0.038786	13.25
2	0.033518	19.81756
1.375	-0.02535	15.90909

1.5	-0.08855	5.066667
2.333333	0.33738	11.10906
1.25	0.067896	13.775
1	0.170378	18.125
3	0.545186	13.66667
0.571429	0.117902	19.31782
0.75	0.096828	19
1.5	0.140252	11.875
1.285714	0.16732	12.29141
0.875	0.16732	12.1935
1	0.502636	6.487233
4	0.195419	25.5
0.785714	0.228296	8.688676
1	0.228296	16.85714
0.857143	0.376883	14.09576
1.083333	0.376883	17.30769
10	0.146016	11
1.428571	0.190457	16.5
0.666667	0.107036	16.60535
6	0.134504	14.93007
1.25	0.09883	20
1.5	0.095708	21.66667
0.666667	0.310365	8.928571
11	0.136986	20
2	-0.00919	7.4375
1	0.329508	2.526316
1	-0.11784	25
0.333333	-0.0334	12.29167
1	0.063912	12.92298
0.625	0.062541	13.82452
2.2	0.011476	5.251192
0.571429	-0.01359	18.83955
1.181818	0.428124	12.39802
0.583333	0.072544	16.22613
1	0.792791	16
4	-0.00616	20
1.666667	-0.00616	16.22951
0.2	-0.08775	21
0.2	-0.11386	13
0.2	-0.10475	13.5
0.166667	0.04071	11.44
5	0.03675	5.870064
0.25	-0.14188	-0.01372
0.4	0.034089	0.006922
2.5	0.014676	0.011365
0.545455	-0.01053	-0.00304
0.166667	0.010806	0.010586
0.8	-0.00973	-0.00151
1.8	-0.00973	-0.00428
1.5	0.054272	0.010401
4.5	0.004661	-0.00536
2.666667	0.00441	-0.00778
0.5	0.145819	0.007315
0.857143	-0.01203	0.006398
0.733333	-0.01671	-8.68E-04
2	-0.01532	-0.00149
1.428571	-0.00877	-0.00175
0.636364	-0.00772	-0.00408
3	0.094916	0.002352
3.4	-0.98018	-0.00229
0.533333	-8.42E-04	-0.00534
0.5	0.01583	5.87E-04
1	-0.00932	-7.68E-04
1	-0.00932	-0.00282
2.666667	0.043313	0.006698
1	0.084775	0.008179
2.5	-0.10097	-0.01253
3	0.060774	0.026318
2	-0.01005	-0.0058
3.5	-0.03891	-0.00259
0.6	0.030015	0.006343
0.3125	-0.01708	-8.66E-04
1	-0.05171	-0.00425
0.4	-0.03915	-0.0115
1.666667	0.059926	0.040449
0.888889	-0.03441	-0.00691
2	0.184801	0.081216
3	-0.01802	-0.01219
0.2	-0.01802	-0.0011
0.2	0.017588	0.003143
0.666667	0.071429	2.52E-04
0.142857	0.071429	1.58E-04
11	-0.00434	-0.00656
0.25	-0.01063	-0.00391
1.666667	-7.18E-04	5.865922
2.4	0.020318	19.9619
0.857143	0.001744	16
1.5	0.025443	17.1174
0.727273	-0.0118	17.4881
3.5	0.03343	8.5
3	0.058976	13.49333
5	0.029781	6.624658
0.8	0.020018	16.70649
1.428571	-0.00359	19.21331
1.142857	0.11566	16.52261
1.428571	-0.00139	28.42857
1.083333	0.036972	15.04369
1.555556	-0.00683	15.98687
1	0.010746	11.31333
1	0.01546	14.97368
0.8	0.015224	14.74298
0.444444	0.083539	18.66667
0.9375	-0.02738	13.69619
0.666667	0.0254	23.5
0.692308	-9.29E-05	23.22383
0.916667	0.006159	18.26982
0.375	-0.00909	12.44095
0.571429	0.047657	17.58553
2	0.053778	5.248227
0.666667	0.071802	10.81002
0.333333	-0.03312	13.87619
0.5	0.02703	13.38679
3	0.031666	20
1.181818	0.01568	21.66735
2.333333	0.088107	12.33555
0.333333	0.095222	27
1.8	0.022818	13.2
1.181818	-0.01893	19.36364
3	0.029353	12
1	0.030689	21.66667
1.25	0.04926	24.4
0.2	0.027312	24.5
1	4.59E-03	16.44231
6	0.032059	14.10309
0.6	0.018163	13.66337
2.75	-0.00162	18.25

1.333333	0.210595	10.99259	2	0.010775	2.79E-04	1	0.138475	7.91038
6	-0.00812	15.66245	1	-0.01145	-0.00326	2	-0.00717	18.8
4	0.002413	5.917808	1.25	-0.00993	-5.64E-04	5	-0.00223	14.50709
0.285714	-0.02073	12.96282	0.4	-0.01621	-9.99E-04	0.9	0.028269	24.39875
6	-0.13347	10.41176	1	-0.01362	-0.00772	0.333333	-0.04651	23.88889
4	-0.13347	22.5	7	-0.01362	-0.00941	3	0.144802	19
0.666667	0.331438	12.20764	0.571429	0.041937	0.008517	1.5	0.026723	8.584337
3	-0.15134	7.407407	4	-0.14775	-0.02265	0.5	0.054423	18.56838
1.166667	0.577501	5.666667	3	2.415997	0.165095	4	0.533938	0.65
1	-0.01683	3.148148	0.75	0.069206	0.011491	2	-0.04188	7.617188
1	-0.08033	18.08889	0.75	-0.00147	-0.01228	4	0.050501	5.733871
1	0.040972	3.467153	0.5	0.023068	-0.00148	0.5	0.078508	14.46319
1	-0.02149	17.26522	4	0.013559	0.006337	1.5	-0.00682	8.046624
1.666667	-0.01495	12.12731	2	0.019317	0.001362	1	0.018358	18
1	0.036433	10	0.5	7.36E-04	0.002513	1.25	-0.03069	25.6
6	0.053176	8.770213	8	-0.00463	-0.00586	2.166667	-0.06761	22.30769
12	-0.13807	7	0.3	-0.06885	-1.51E-04	0.1	-0.00718	8
1.1	0.043833	15.14655	1.285714	0.001659	-0.00467	0.916667	0.050477	17.7457
0.916667	0.200365	15.14197	0.5	0.014124	0.002284	1.571429	0.011075	16.32772
0.347826	0.113072	14.625	0.192308	0.059689	0.001791	0.107143	-0.00656	18
2.333333	-0.0417	13	0.333333	0.161861	-0.0031	9	0.044456	9
0.4	0.16506	16.99091	6.333333	0.154322	0.013297	4	0.037823	10.25268
0.3	0.16506	17.8	0.285714	0.154322	0.01182	0.851852	-0.03393	16.10771
1	0.229519	8.409091	2.333333	-0.97674	-0.00782	0.5	0.068059	24.10904
2.75	0.369049	17.5	1.333333	0.301775	0.064472	1.3	-0.01551	19.8
1.25	0.238085	13.875	5	-0.88571	0.024377	0.714286	0.071796	18.14286
1.25	0.199289	14.6	1.5	0.060417	0.021952	1	-0.12566	7.332857
0.9	0.110832	16.83131	0.333333	0.002672	-8.32E-04	2.571429	0.014719	12.85025
1.6	0.10427	16.63366	1.428571	0.001921	-0.00548	0.9375	-0.01924	19.55061
3	0.040881	8.985714	1	-0.93852	-0.00532	0.5	0.057512	8.389535
7	0.039209	9.166667	1	-0.93852	-0.00148	7	0.016805	13.59
0.222222	0.105477	14.5	0.888889	-0.00362	-0.00576	0.777778	-0.00641	21
1.5	-0.08732	21.59838	0.2	-0.03164	-0.0098	1.333333	0.026361	24.66667
2	0.294442	14.39486	0.666667	0.022435	0.019899	0.846154	0.013508	18.70032
0.25	0.429869	12	0.75	0.150412	0.012448	0.125	0.018034	17
0.75	0.462155	24	3	0.120874	0.012253	0.625	-0.00194	24.8
1	0.066575	14.83601	0.875	-0.00657	0.003734	1.5	0.017991	16.8
1.1	0.078718	19	1.625	0.017277	9.99E-05	6	0.031742	17.33333
1	0.373735	7.935115	3	0.07656	0.013822	2.666667	0.074317	16
1.333333	0.267892	13	3	0.037799	0.011331	0.6	0.051739	10.44174
1	0.020931	8.519149	1	-0.10844	-0.00864	2	-0.00386	19.11783
3	0.017972	11.31973	0.666667	0.145523	0.015599	0.166667	0.084564	19.39103

2.5	0.404593	8.723077
1	-0.13836	15.90184
1	-0.0861	7.166667
3	0.107931	4.707602
5	0.085666	18.90732
2	0.186977	11.45455
1.5	0.116312	17.78027
2	-0.20437	16.05096
0.8	0.065283	8.775
2.25	0.273388	14.75
5	0.385627	6.992857
1.666667	0.1153	9.333333
2	-0.05932	16.75
1	-0.185	15.33333
1.5	2.588068	10.33333
5	-0.34583	16.93171
1.222222	0.729003	10.11111
1.333333	0.303189	9.777778
0.875	1.172848	11.85714
0.545455	1.035395	17.45455
2.5	0.115739	7.146667
0.5	0.377522	10.97727
1.375	-0.00309	15.8785
4.5	-0.04308	8.035714
3.5	-0.13193	21.57143
4	0.184195	9.013699
0.166667	0.059404	16.44444
0.714286	0.258963	14.2
1.6	4.10E-02	18.75
0.416667	0.057452	15.33333
0.714286	0.0871	22.66667
4	-0.04262	16.73913
4	0.019594	11.94949
0.25	-0.31961	36.75
7	-0.00756	5.394495
1	0.01463	2.75
2.5	-0.02372	23.8
0.8	-0.04147	25.94483
0.125	0.072367	17.02215
2	0.002139	22.2
0.5	-0.01666	20.66667
14	-0.3958	10
2	-0.19554	32.5

2.333333	0.001987	0.004566
1	-0.13271	-0.00915
6	-0.03022	-0.00621
0.666667	0.028566	0.026657
1	0.044887	0.006562
4	0.02303	0.013763
0.4	0.041072	0.008688
8	0.009344	-0.00266
2.5	0.010645	-0.00298
0.6	0.174954	0.011718
1.25	9.89E-04	0.008103
1.25	-0.96377	-0.005
0.333333	0.067598	0.009522
3.5	-0.04668	-0.01378
1.666667	0.251504	0.076272
1.5	-0.02765	-0.00996
2	0.136198	0.03965
1.625	-4.30E-04	0.010406
1.5	0.221545	0.017112
1.666667	0.082346	0.053029
4	0.054455	0.005625
0.166667	0.044556	0.007233
0.666667	-0.14268	-0.0161
4.5	-0.13304	-0.00456
0.333333	-0.07971	-0.00762
1.8	0.008111	0.00678
0.833333	-0.0237	-0.00551
1.4	0.195918	0.010719
0.3	-0.00165	0.001863
0.36	0.009304	0.001064
0.342857	-0.012	7.22E-04
0.857143	-0.03915	-0.00807
1	0.014895	-0.00234
1	-0.35471	-0.0259
3.5	-0.03438	-0.02536
1	0.104938	0.013954
0.555556	-0.17914	-0.0027
3	-0.16963	-0.00383
0.833333	-0.06106	-0.01551
0.666667	-0.03776	-0.00702
5	-0.04583	-0.00872
14	-0.2466	-0.00546
2.5	-0.93464	-0.0121

1.166667	0.155905	8.713474
2	0.030119	20.31148
0.5	-0.01411	18.33333
0.333333	0.087289	11.27371
2.666667	0.185304	16.20874
1.5	0.192316	13.98148
14	-0.00466	20.20408
0.714286	0.025655	24
0.666667	0.029488	12.39498
8.5	-0.027	13
0.571429	0.063594	13.5
2	0.084298	10.33623
1.25	0.007993	21
1	0.070007	20.8
0.333333	0.011473	9.711712
3	0.075264	16.49102
0.6	0.019907	10.66667
0.666667	0.128545	9.470852
3.666667	-0.00704	16.495
0.375	0.026323	18.3125
3.333333	0.006348	17
0.833333	0.006107	10.69091
1.833333	0.098808	23.72727
6	0.142151	12.61538
1.5	0.070242	26
1.333333	0.111493	18.66667
0.75	-0.11814	13.33388
5	0.009386	24.5
5.5	0.009385	23.81818
0.096774	-0.00512	11
0.044444	1.66E-04	4.5
2	0.024216	28.5
8	-0.00355	25
1.5	0.015892	48.66667
1.2	0.056587	12.19907
1	0.085917	5.191257
2.5	0.034305	17.27709
1	0.036389	27.64865
1.4	0.001003	37.6
0.25	0.018917	31
0.4	-0.00774	16.86923
0.333333	0.021214	8.134454
2.5	0.031976	30.37215

1.5	-0.24275	9.216216
1.333333	-0.25507	16.33865
0.875	-0.13805	19.125
0.166667	-0.06076	20
2	0.025389	7.837079
1.75	-0.08736	8.920714
3.666667	0.05664	24.33333
0.333333	0.022003	14.76923
2	0.044308	18.53725
1	-0.09711	19
6	-0.04155	16.7069
0.7	-0.02737	8.378208
1.1	0.033935	18.8
1	0.061987	16.73583
4	-0.53458	9.409091
1	0.084854	10.81188
2.666667	-0.00359	10.752
1.5	-0.02386	16.31505
0.75	0.089005	14.33793
1.5	0.138835	10.87356
3	0.053222	3.38676
1	0.008864	20.25
0.333333	0.031209	16.04225
0.666667	-0.0583	21.65217
11	0.161727	19
2	-0.11163	11.99372
1.666667	-0.00681	11.46979
0.714286	-0.0688	21.6
1	0.070813	15.02222
1	-0.04021	26
3	-0.07844	7.438095
1.6	0.305333	8.873395
1	0.010799	17.09028
1.75	0.019071	23.5
1	-0.51021	24.5
0.666667	-0.07609	18.02949
0.7	0.360376	17
0.647059	0.052036	20.11765
1.111111	0.092903	19.11287
0.5	-0.0346	6.375
2.333333	0.061875	15.99095
1.5	0.100161	15.44828
0.5	-0.12577	-0.00717
0.75	-0.00818	-0.0266
0.363636	-0.02758	-4.76E-04
4	-0.10057	-0.00111
0.5	0.07319	0.019465
0.333333	-0.98881	-0.00176
3.666667	0.012044	0.001778
2.333333	0.002983	-0.01584
1.75	-0.00823	-0.00346
1	-0.13662	-0.0127
1.5	-0.04365	-0.01217
1.5	-0.00696	-0.01488
1.5	-0.95946	-4.52E-04
1	-0.01877	-0.00533
3	-0.0177	-0.01576
2	0.107843	2.49E-04
0.8	-0.11387	-0.0204
1	-0.00282	-0.01127
0.777778	0.013244	-0.00318
1.5	-0.01767	-0.01089
2.333333	0.018418	-0.00637
1.416667	-0.01644	-0.00601
1	-0.00512	-0.00101
3	0.035121	-8.96E-04
13	6.31E-02	0.007797
5	-0.01592	-0.00473
0.75	-1.96E-02	-0.0044
1	0.018938	-0.00611
0.666667	-0.02563	-0.00385
0.388889	-0.01639	-0.00103
1	-0.11008	-0.00716
2	0.038732	0.007845
1.2	-0.01942	-0.00645
1.166667	-0.01657	-0.0017
0.5	-0.02597	-0.01456
0.333333	-0.0728	-0.00314
0.4	0.059602	0.029147
0.647059	-0.01365	0.001127
1.4	0.004424	-0.00222
4	-0.00144	3.03E-04
1	0.013536	0.006984
0.75	-0.93498	0.006872
0.5	0.217102	6.695652
1	0.042224	24.20168
0.363636	-0.03876	23.45455
3.5	0.093956	29.17123
2.333333	0.116192	9.098792
0.428571	0.014881	28.66667
0.75	0.040833	28.5
9	0.026835	18
0.714286	0.029781	15.52106
1	0.064179	24.02542
0.714286	-0.01516	26.14286
1.444444	0.0209	6.465353
0.6	0.020689	15.68182
1.75	0.02805	21.95714
1.5	0.020862	23.63636
4	0.056435	3.013761
1	0.075505	11.08687
1.2	0.031884	13.97172
3.25	0.024036	17.17485
2	0.061265	12.52778
1.142857	-0.05693	19.83635
1.166667	0.094367	28
0.6	-0.00307	17.625
1	0.066043	16.23853
11	0.028168	19
3	0.014994	18.125
1.125	0.085739	19.29306
6	0.036893	31.33333
1	0.118889	27
0.388889	-0.00394	28.57143
2.5	0.053088	24.544
1.833333	0.058017	15.81818
0.2	0.268135	20.55146
1.5	0.023794	22.19973
1	-0.00228	25.5
0.8	0.025909	39
0.818182	0.00211	11.45926
0.076923	0.002727	22
1	0.022771	24.91667
1	-0.00829	7.875
1.5	0.011879	22.12364
1	0.140774	17

0.75	0.040434	12.58475
0.789474	0.648854	16.4
1.2	0.105828	14.86505
5	-0.02794	25.2
1	0.114113	9.552989
1	0.166521	1.810345
1.5	-0.16366	15.75
2	-0.02802	17.3812
0.8	-0.03417	7
1.833333	-0.03723	20.59752
2	0.306551	19
5	-0.06024	12.46737
5	0.100739	9.625751
0.923077	0.047767	8.582418
1	0.11925	0.365854
2.8	0.33822	0.239002
1.333333	0.044155	18.74689
3.2	0.031676	23.8
0.727273	0.648652	9.625
0.142857	-0.09704	18
2.75	0.16547	12.25
2.5	0.344283	12.73684
0.75	0.054618	12.57237
1.5	0.103446	8.131313
1	0.261609	1.88764
2	0.146397	7.386364
2.5	-0.27339	12.5
1.333333	-0.04829	13.13742
0.2	0.014999	17.95931
0.75	0.119002	22
0.857143	-0.26062	18.5
0.666667	0.158762	15.2
0.8	0.00959	16.55
2	0.215148	2.594595
4	0.8305	2.56295
2	-0.02183	5.416667
0.7	0.169774	17.7
6	-0.39535	19.71429
0.25	-0.01834	16
0.454545	0.427796	12.75368
2	0.432546	17
1	0.101613	4.954663
1.25	-0.97047	-9.75E-04
0.674419	0.004739	0.003509
1.076923	-0.00598	-2.06E-06
3	0.013287	-0.00577
1	-0.00424	-0.00867
18	0.032426	0.005142
2.5	-0.9589	-0.05654
1	-0.04681	-0.00346
1	0.029568	5.50E-04
0.416667	-0.00381	-0.00383
1.071429	-0.00163	-0.00763
1	0.014471	-0.00163
2	-0.03382	-0.00175
2	-0.01515	-0.00998
4	0.320014	0.013802
1.555556	-0.01439	-0.00593
1	9.43E-04	0.004591
3.2	-0.04657	-5.37E-04
1	0.255256	0.168859
0.25	-0.04906	-0.00147
1.6	7.86E-04	0.010944
1	-0.04478	-0.00746
0.833333	-0.25402	-0.00943
2.5	0.070993	-4.60E-04
0.833333	0.01691	-0.00746
3	8.18E-04	0.006928
1	-0.04352	-0.00446
2.5	-0.02398	-0.00936
1.5	-0.01876	-0.00303
0.272727	0.00717	-0.00176
0.75	-0.0299	-0.00384
1	0.005718	-0.00281
0.795918	-0.01839	-0.00203
0.428571	0.055469	5.03E-04
0.666667	0.042241	0.005778
1.333333	-0.08871	-0.02014
8	0.106428	0.001545
1	-0.08595	-0.01031
6	0.016118	-0.00461
0.666667	0.049204	0.039031
3.5	0.081208	0.003833
0.5	-0.01414	-0.00407
1.25	-0.00922	8.541463
1.069767	0.012553	16.1087
0.9	0.025665	16.6403
2	0.083561	10.90909
1.5	0.023098	15.18296
7.333333	0.020012	6.979592
1.5	0.066934	10.64815
0.333333	0.010585	27.33333
1.5	0.001262	7.75
2.166667	0.019887	21.63031
0.571429	-0.00454	22.17593
8	0.02584	12.30469
1	0.230064	11.29358
1	0.033485	7.445472
0.333333	9.29E-02	18
1.2	0.112571	3.960784
1.266667	-0.00391	20.76724
3.2	0.018405	28.4
0.818182	0.19814	9.063131
0.714286	0.03251	29.4
0.666667	-0.0385	8.119048
2.666667	-0.00493	29.25
1.2	0.001607	19.83333
1.8	-0.01614	22.4
1	0.029737	16
1.2	0.021177	10.2
7	0.044658	8
3	0.012564	19.33333
0.4	0.126357	21.2
0.666667	0.016812	17.93052
0.333333	0.031621	6.033962
0.636364	-0.02131	8.466234
1	0.002117	18.72222
0.454545	0.050319	4.306408
0.8	0.017378	11.62857
1.5	0.078947	14.7
17	0.015392	6.65308
2.666667	0.024744	25
0.714286	0.046665	12.34734
2	0.059161	16.75
2	-0.0233	20.5
1.666667	0.093762	13.35758

1	0.070223	12
8	-0.2071	13.4421
1.2	0.731316	10.4
1.363636	0.239552	5.454545
0.666667	0.075871	8.087889
6	0.035951	11.17544
6	-0.055	4
1.25	-0.01643	30.7
1.5	-0.08022	15.66667
0.8	0.85961	15.4
1.222222	0.389662	10.54545
5	0.395999	7.482353
2	0.395999	4
1.214286	0.038031	13.29783
6	-0.02764	2
1.2	-0.01411	16.4
4	-0.06338	1.73028
2	-0.11668	5.210526
1.5	0.085377	11.36219
3.5	0.279653	7.118644
0.666667	-0.00137	18.86667
0.545455	4.01E-04	25
0.888889	4.01E-04	18.66667
2	0.164351	9.345946
0.76	0.052929	18.52
0.75	0.002711	14.72139
0.416667	2.67E-01	12.6
12	-1.59E-02	8
0.636364	0.189423	15.28571
3	0.333333	12.33333
1	0.333333	7
2	-0.07424	12.84722
0.7	0.531212	18.57143
1.5	0.004636	14.76429
1	-0.37932	1.866667
0.538462	0.311677	14.69231
1.142857	0.11622	13.85578
1.125	0.103559	16.48649
0.818182	0.263622	7.75
0.333333	-0.00226	18
0.363636	0.124203	20
8	-0.00378	15.5
1.333333	-0.01414	-0.00115
8	-0.04956	-8.36E-04
2	0.025577	0.022574
1.375	-0.00628	0.005052
3	-0.01643	-0.00205
1	0.009247	-4.17E-04
1	-0.11561	-0.00259
0.789474	-0.06335	-0.00384
0.666667	-0.0485	-0.00201
1.333333	0.143179	0.029709
1.5	0.136727	0.006614
2.166667	0.136727	0.018899
0.666667	0.217725	0.019256
0.653846	-0.00475	-0.00515
1.833333	-0.0174	-0.00459
0.555556	-0.0174	-0.00272
4	0.004715	2.65E-04
1	0.025781	-0.0023
7	-0.05857	-0.00162
3.5	0.017728	-0.01176
7	-0.0304	-0.00259
0.625	-0.01991	-0.002
1.285714	-0.02034	-0.0037
1.333333	-0.04595	-0.00202
1.181818	0.003027	-0.00158
0.894737	-0.00825	-0.00499
16	0.096758	4.81E-06
0.333333	0.014929	-0.00141
1.75	0.011296	-0.00105
1	0.630353	0.182888
0.25	0.630353	0.038651
0.6	-0.03758	-0.01522
1.666667	0.080836	0.022089
1.5	-0.00799	-0.00744
4	-0.01669	-0.03283
2	0.114538	0.016413
1.166667	-0.00441	-0.00509
1.5	-0.03152	-0.0079
1.222222	0.018762	0.004046
0.666667	0.157988	0.014338
1	0.04803	0.006479
2.333333	-0.04331	-0.00805
1.25	0.023809	14.74063
0.555556	0.031064	29.55556
1.285714	-0.07116	4.714072
1.133333	0.042754	2.223598
1.333333	0.028698	15.47436
1.25	-0.05037	8.4175
1.666667	0.047273	26.2
1.125	0.019437	37.125
2	0.039193	19.25
1.333333	0.049166	5.628049
2.5	0.171343	6.302711
0.545455	0.042781	8.225108
1.142857	0.006639	9.375
1.111111	0.065232	21
1.8	0.021715	20.19225
13	0.015139	3
2	-0.05635	24.8
2.5	-0.00885	12.5
0.571429	0.071617	15
1.285714	0.022161	7.394958
1.333333	0.098747	20.04274
1.111111	-0.02645	24
0.375	0.055445	14.66667
1.142857	0.011035	23.28571
0.833333	-0.01441	14.48856
1	0.002027	23.28571
16	0.027856	2
3	-0.00649	27.91667
4.2	0.013879	11.90607
1	0.399476	11
0.666667	-0.19344	7.5
0.333333	0.015403	7
1.571429	0.013815	13.19987
0.666667	-0.00147	18.90904
3	0.121107	1.909091
1.176471	-0.01741	17.11765
0.833333	0.061833	14.61458
1	0.094955	21.33333
1.555556	-0.0288	5.885362
3	0.031488	6
0.916667	0.005594	23.91667
0.25	1.39E-01	19.68947

1.333333	0.463878	22.25
2.5	-0.12796	8.290299
0.555556	0.102765	12.92152
3	-0.06336	14.4903
1.8	0.213844	8.33
5	0.81983	7
2	-0.10028	9.219136
1	-0.09763	2.256944
0.545455	0.186689	8.5
8	0.11521	12
1	0.275842	20.25
0.233333	0.636882	13
1.166667	-0.04844	18.85714
0.5	-0.04156	13.35526
2.5	0.188619	18
2.333333	0.080289	15.14286
1.857143	0.082588	14.33021
0.571429	-0.19692	17.75
0.25	0.127776	10.23301
1.142857	0.006116	17.82012
1.428571	0.30116	16.71429
3	0.086006	17.69444
0.666667	-0.01372	15.25
1	-0.0018	3.902439
1.428571	0.602345	8.714286
3	0.134986	10.72762
0.333333	0.050776	20.33333
2.4	0.121361	7.351258
1	0.061204	19.03518
0.5	-0.10849	11.5
0.75	-0.00629	20.53333
1	0.157046	15.09615
1	0.24753	15.18359
0.916667	-0.03157	20.8861
0.235294	0.057611	23.25
1.333333	-0.1431	32.5
0.666667	0.01779	9.692308
1	0.056505	6.63
13	0.052506	3
0.333333	0.89917	3
1.5	-0.23745	23.86198
2	-1.67E-05	13
1.857143	0.013362	0.004018
0.5	-0.94024	-0.00329
2.333333	-0.02077	-0.00342
1.285714	-0.0509	-0.00606
1.142857	0.003359	-0.00176
0.333333	0.038851	0.001377
3	-0.97696	-0.00635
2	-0.05248	0.001206
2.4	0.406461	0.025882
6	-0.04782	2.64E-05
0.5	0.07241	0.005441
0.121212	0.236677	0.006652
0.444444	-0.06075	-4.93E-04
0.333333	0.101884	0.006069
0.5	0.009002	0.005646
5	0.112245	0.00187
0.75	-0.0254	-0.00511
1	0.009017	0.011222
4	0.044155	0.005735
2.75	-0.98035	-0.00672
6	0.099546	0.015005
0.5	-0.04198	-0.00559
1.5	-0.0146	-0.00175
6	0.599723	0.011481
2.166667	0.069435	0.020668
2.333333	0.037004	-0.00769
3.25	3.08E-02	0.00275
1.285714	0.007909	-0.00549
0.7	-0.98258	-0.00476
0.5	-0.03326	-0.00749
1.4	-0.04349	-0.0018
0.5	0.027822	0.003302
2.333333	0.090098	0.005794
1.428571	-0.08924	-7.24E-04
0.1875	0.055276	-6.69E-04
0.4	-0.11007	-0.00269
2	0.006396	0.011544
4	0.001069	0.003661
0.217391	-0.03309	-0.00512
0.222222	0.040146	0.009051
1.25	-0.06883	6.80E-04
3	0.014246	0.001146
1.111111	0.045782	23.75
1	-0.03776	6.579457
0.777778	-0.02828	14.14286
2	0.024231	22.61798
1.583333	0.022115	17.33333
7	0.001804	4.978534
0.8	0.01583	18.83324
0.714286	0.043974	20.6
0.214286	0.001623	6.333333
0.5	0.008732	25.48104
2	0.098968	11.5
0.681818	0.015109	15.73333
12	0.01001	17
0.5	-0.00419	7.903226
1	-8.68E-03	20.2938
2	0.016539	12.16667
1.4	0.007158	18.14286
0.4	0.043737	7.946939
1	0.042275	17.46884
3.333333	0.013523	18
1.428571	0.106573	18.85714
0.375	-0.02368	19.33333
7	0.039674	5
0.583333	0.011938	16.85714
3.875	0.011213	10.05059
0.333333	0.02316	15
0.75	0.005312	18.0099
0.375	0.037082	12.25278
0.294118	-0.02347	5.849957
0.25	0.017713	18
5	0.00576	24.59895
0.636364	0.044966	15.27476
1.1	0.01023	18.8
0.75	0.051399	20.65513
9.5	-0.01569	17
6	0.016356	20
1.666667	0.03727	22.33333
0.333333	-0.03319	20.14815
2.142857	-0.01043	35.53333
1.75	0.010036	2.689076
1.333333	0.030608	30.75
0.5	9.62E-03	19.8

0.714286	0.381301	11
1.333333	0.161288	13.22652
1.75	0.067894	19.69504
5	0.091309	3.744076
4	-0.23748	20
2	-0.00865	9.493705
1.833333	0.491576	15.33333
5	0.002877	15.51273
1.6	0.245624	18.3125
1.5	0.298607	3.041199
2	-0.32897	20.86638
0.75	0.060442	16.92537
2	0.060442	17
1.25	0.159937	4.70409
4	0.179947	14
0.647059	0.078516	6.181818
0.9	0.056989	20.17931
0.363636	0.382227	12.11102
0.5	-0.06129	17
0.5	-0.04031	22.5
0.428571	-0.19421	20.33333
2.5	0.172883	21.25
1	-0.40358	23
0.117647	0.235197	7.02439
1	0.301291	14.49061
1.4	0.256153	10.75023
5	0.097582	16.5
2	0.205478	15.66667

4.666667	0.11896	0.029886
0.4	0.088	0.0124
0.266667	-0.14898	-0.00202
1.666667	-0.0163	-0.00431
1	0.107259	0.012224
0.166667	-0.11067	-0.00556
1.142857	0.050823	0.025738
0.75	-0.01655	7.50E-04
1.444444	0.005582	-0.00872
1	0.078418	0.006109
0.666667	-0.17318	-0.01477
1	0.012163	0.00542
1.333333	0.012163	0.009204
2.5	0.021361	0.002628
1.8	-0.00968	0.008379
1.333333	-0.02504	-0.00351
1.285714	-0.09365	-0.00464
2.2	0.011893	-0.00675
1	0.002767	-0.00575
1.5	0.01236	-0.00703
1	-0.0909	-0.00253
1.181818	-0.01667	-0.00259
0.5	-0.26586	-0.01549
0.636364	0.091977	0.009728
1.153846	-0.01055	0.007454
2	0.004262	-0.00287
3	0.376601	0.001322
0.8	0.172137	-2.99E-04

0.318182	-0.00519	14.80596
0.857143	0.035366	24
8.5	-1.44E-04	11.85294
1.5	0.191675	17.75
1	0.023236	9.885496
0.5	0.019397	13.16803
1.125	0.001584	17.71267
1.5	0.019118	22.83333
0.6875	0.015903	19.45455
3.25	0.013934	8
4	-0.06694	24.52273
1.5	0.171509	15.46667
4	-0.03481	16.5
1.230769	0.016714	5.562863
1.090909	0.003952	22.09091
1	0.029554	11.85714
1.8	0.049694	20.66667
1.714286	-0.00781	23.14286
1.5	-0.02501	24.25
0.666667	-0.08289	24.83333
0.25	-0.01781	20
0.5	0.037482	20.07311
0.5	0.117243	24.5
5.5	0.011463	3.763686
0.5	0.015709	13.28856
4	0.015316	20.99069
0.111111	0.075441	13
1	0.009701	19

Fuentes consultadas

[Acero, 2001] Ignacio Acero, Matías Alcojor, Alberto Díaz, José María Gómez. (2001). Generación automática de resúmenes personalizados; Departamento de Inteligencia Artificial; Escuela Superior de Informática; Universidad Europea-CEES; Madrid, España.

[Cunha, 2007] Iria da Cunha, Silvia Fernández, Patricia Velázquez Morales, Jorge Vivaldi, Eric Sanjuán y Juan Manuel Torres-Moreno. (2007). A new hybrid summarizer based on Vector Space model, Statistical Physics and Linguistics; Institute for Applied Linguistics; Universitat Pompeu Fabra; Barcelona; España.

[Bolshakova et al, 02] N. Bolshakova and F. Azuaje. (2002). Improving Expression Data Mining Through Cluster Validation, Department of Computer Science, Trinity College Dublin, Ireland.

[Cole, 1998] Cole, R. M. (1998). Clustering with genetic algorithms. University of Western Australia.

[Dash & Liu, 01] Dash, M., and Liu, H. (2001). Efficient Hierarchical Clustering AlgorithmsUsing Partially Overlapping Partitions, Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp 495-506.

[Davies, 1979] Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. IEEE transactions on pattern analysis and machine intelligence, (2), 224-227.

[Deco, 2007] Claudia Deco, Cristina Bender, Mario Chiari. (2007). Problemas de la Traducción de la Consulta en la Búsqueda de Información Multilingüe. Departamento de Investigación Institucional; Facultad de Química e Ingeniería. Universidad Católica Argentina. http://www.infosurrevista.com.ar/biblioteca/INFOSUR>Nro1>2007>DecoBenderChiari.pdf

[DUC, 2002] DUC. Document understanding conference 2002 (2002), http://wwwnlpir.nist.gov/projects/duc

[Dunn, 1974] Dunn, J. C. (1974). Well-separated clusters and optimal fuzzy partitions. Journal of cybernetics, 4(1), 95-104.

[García, 2008] García-Hernández, R., Montiel, R., Ledeneva, Y., Rendón, E., Gelbukh, A., & Cruz, R. (2008). Text summarization by sentence extraction using unsupervised learning. MICAI 2008: Advances in Artificial Intelligence, 133-143.

[Gelbukh, 2005] Alexandrov, M., Gelbukh, A., & Rosso, P. (2005, June). An approach to clustering abstracts. In International Conference on Application of Natural Language to Information Systems (pp. 275-285). Springer Berlin Heidelberg.

[Hernández, 2006] Hernández-Reyes, E., García-Hernández, R. A., Carrasco-Ochoa, J. A., & Martínez-Trinidad, J. F. (2006). Document clustering based on maximal frequent sequences. In Advances in Natural Language Processing (pp. 257-267). Springer Berlin Heidelberg.

[Huang, 2008] Huang, A. (2008, April). Similarity measures for text document clustering. In Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand (pp. 49-56).

[Jiménez, 2005] Ernesto Jiménez Ruiz. (2005). Desarrollo de Ontologías de Forma Modular; Departamento de Lenguajes y Sistemas Informáticos; Universidad Jaume I de Castellón; España. http://www4.uji.es/~al083362/II/Doctorado/Mod_Ontologias.pdf.

[Kryscia, 2007] Kryscia Daviana Ramírez Benavides. (2007). Stemming, Lematización; Escuela de Ciencias de la Computación e Informática; Universidad de Costa Rica; Costa Rica. http://www.ecci.ucr.ac.cr/~kramirez/RI/Material/Presentaciones/Stemming.pdf

[Ledeneva, 2014] Ledeneva, Y., García-Hernández, R. A., & Gelbukh, A. F. (2014, April). Graph Ranking on Maximal Frequent Sequences for Single Extractive Text Summarization. In CICLing (2) (pp. 466-480).

[Lee06] Fu Lee Wang, Christopher C. Yang, Xiaodong Shi. (2006). Multi!document Summarization for Terrorism Information Extraction; Department of Computer Science; City University of Hong Kong; Hong Kong SAR; China.

[Lezcano, 2006] Ramón David Lezcano. (2006). Trabajo de Investigación Bibliográfica: Minería de Datos; Departamento de Informática; Universidad Nacional del Nordeste; Argentina. http://exa.unne.edu.ar/depar/areas/informatica/SistemasOperativos/MineriaDatosLezcano.pdf

[Luhn, 1957] Hans Peter Luhn. (1957). A statistical approach to mechanical encoding and searching of literary information, IBM Journal of Research and Development, pp. 309-317, 1957.

[Manning & Schütze, 1999] Christopher D. Manning and Hinrich Schütze, (1999). Foundations of Statistical Natural Language Processing, Massachusetts Institute of Technology, pp.

[Montes y Gómez, 2002] Manuel Montes y Gómez. (2002). Minería de texto empleando la semejanza entre estructuras semánticas; Tesis de doctorado; Laboratorio de Lenguaje Natural y Procesamiento de texto; Centro de Investigación en Computación; Instituto Politécnico Nacional; México.

[Mordecki, 2007] Ernesto Mordecki. (2007). Probabilidad, Programa de matemática de 2° año. Centro de Matemática; Facultad de Ciencias; Universidad de la República; Uruguay.

[Moreiro, 2002] José Antonio Moreiro González. (2002). Aplicaciones al análisis automático del contenido provenientes de la teoría matemática de la información; Departamento de Biblioteconomía y Documentación; Universidad Carlos III de Madrid; España. http://www.um.es/fccd/anales/ad05/ad0515.pdf

[Peinado, 2003] Jesús Peinado Rodríguez. (2003) Lematización para palabras médicas complejas: Implementación de un algoritmo en LISP; Facultad de Salud Pública y Administración Carlos Vidal Layseca; Universidad Peruana Cayetano Heredia; Perú. http://www.scielo.org.pe/pdf/rmh/v14n4/v14n4cc02.pdf

[Porter, 2006] Martin Porter. (2006). The Porter Stemming Algorithm. Official home page for distribution of the Porter Stemming Algorithm. http://tartarus.org/~martin/PorterStemmer/index.html

[Rendón, 2011] Rendón, E., Abundez, I., Arizmendi, A., & Quiroz, E. (2011). Internal versus external cluster validation indexes. International Journal of computers and communications, 5(1), 27-34.

[Rivero, 2010] Rivero, D., Rabuñal, J. R., Dorado, J., & Pazos, A. (2010). Introducción a los algoritmos genéticos y la programación genética. Universidade da Coruña, Servicio de Publicacións.

[Rousseeuw, 1987] Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics, 20, 53-65.

[Salton, 1975] Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. Communications of the ACM, 18(11), 613-620.

[Sidorov, 2013] Sidorov, G. (2013). Construcción no lineal de ngramas en la lingüística computacional: n-gramas sintácticos, filtrados y generalizados. Sociedad Mexicana de Inteligencia Artificial. ISBN 978-607-95367-9-4.

[Soto, 2009] Soto, R. M., García-Hernández, R. A., Ledeneva, Y., & Reyes, R. C. (2009). Comparación de Tres Modelos de Texto para la Generación Automática de Resúmenes. Procesamiento del Lenguaje Natural, 43, 303-311.

[Steinbach et al, 2000] Steinbach, M., Karypis, G., Kumar, V. (2000). A Comparison of Document Clustering Techniques, University of Minnesota, Technical Report #00-034.

[Valderrábanos, 2004] Antonio S. Valderrábanos. (2004). Recuperación de información y conocimiento lingüístico: el buscador inteligente. The bit and text company, Madrid, España. http://www.baquia.com/com/20040625/art00004.html.

[Valencia, 1997] Valencia, E. (1997, August). Optimización mediante algoritmos genéticos. In Anales del Instituto de Ingenieros de Chile (Vol. 109, No. 2, pp. 83-92).

[Van Rijsbergen, 1979] C. J. Van Rijsbergen. (1979). Information Retrieval, Butterworths, London, 2da edición cap. 7.

[Villatoro, 2006] Esaú, V. T. (2006). Generación automática de resúmenes de múltiples documentos (Doctoral dissertation, Tesis de maestría).

[Wang & Hodges, 06] Yong Wang and Julia Hodges. (2006). Document Clustering with Semantic Analysis, Proceedings of the 39th Hawaii International Conference on System Sciences, pp. 54-64.

[Xiong, 2013] Xiong, H., & Li, Z. (2013). Clustering Validation Measures.

[Zhao, 2005] Zhao, Y., Karypis, G., & Du, D. Z. (2005). Criterion functions for document clustering. Technical Report.