



División de Ciencias de la Comunicación y Diseño

Licenciatura en Tecnologías y Sistemas de Información

Proyecto:

Sistema Automático para la Visualización de Relaciones de Opiniones Producidas en Twitter

Por:

Griselda Velázquez Villar

Supervisado por:

M. C. Gabriela Ramírez de la Rosa

Dr. Esaú Villatoro Tello

Julio, 2015

Contenido

1. Introducción	5
2. Planteamiento del Problema	6
3. Objetivos	7
3.1 Objetivo General	7
3.2 Objetivos Específicos	7
4. Marco Teórico	8
4.1 Representación de los Documentos	9
4.2 Agrupamiento de documentos por temática similar	10
4.2.1 Pesado Booleano	10
4.2.2 Pesado de frecuencia de término	11
4.2.3 Pesado tf-idf	12
4.3 Medición de Similitud	12
4.4 Algoritmos de Agrupamiento	13
4.4.1 Agrupamiento Jerárquico	14
4.4.2 Agrupamiento de Partición	15
5. Trabajo Relacionado	16
5.1 Herramientas Especializadas en el Agrupamiento de Datos	17
5.1.1 Cluto & gCluto	17
5.1.2 Weka	18
5.2 Herramientas Dedicadas al Análisis de Información en Redes Sociales	19
5.2.1 Spot	19
5.2.2 AnaliticPro	20
5.2.3 Socialmention*	21
5.3 Cuadro Comparativo de las Herramientas de Visualización de Datos	23
6. Desarrollo del Sistema de Visualización	25
6.1 Esquema General del Sistema	25
6.2 Módulo 1: Recuperar Tweets en Tiempo Real	26
6.3 Módulo 2: Agrupamiento de Tweets por Temática Similar	27
6.4 Modulo 3: Representación Visual de los Resultados de Agrupamiento	29
7. Conclusiones y Trabajo Futuro	30
8. Referencias	31

9. Anexos	33
9.1 Anexo A: Mapa de Navegación	33
9.2 Anexo B: Diagramas de Casos de Uso.....	34
9.3 Anexo C: Diagramas de Secuencia.....	41
9.4 Anexo D: Diagramas de Clases	46
9.5 Anexo E: Diseño de la Base de Datos	52
9.6 Anexo F: Manual de Usuario.....	55

Resumen

El Sistema Automático para la Visualización de Relaciones de Opiniones Producidas en Twitter, es el resultado del análisis y agrupamiento por temática similar que existe dentro de un conjunto de tweets. Dado que la información que se publica en Twitter es variada y relacionada a una infinidad de temas, en el sistema se implementó la recuperación de tweets por medio de una consulta que permite identificar tweets que contengan al menos la cadena de caracteres que se socita en dicha consulta, y que además, sean tweets que estén escritos en idioma español. Para el agrupamiento de los tweets se incorporó al sistema la librería de Weka, la cual contiene los métodos y algoritmos de agrupamiento para llevar a cabo el proceso de agrupamiento por temática similar. Los resultados obtenidos del agrupamiento de los tweets son los siguientes: el número de grupos que se forman, los tweets que conforman cada grupo, las cinco palabras más mencionadas en cada grupo y las cinco palabras que tienen en común dos grupos, es decir las palabras que coinciden exclusivamente entre pares de grupos. Las palabras que se mencionan entre los pares de grupos, determinan las posibles relaciones que existen en el agrupamiento por temática similar. Finalmente la información se muestra en una gráfica de burbujas diseñada con la librería D3js, en la cual se visualizan los resultados de forma intuitiva y fácil interpretación.

1. Introducción

El surgimiento de las redes sociales en Internet han propiciado que un mayor número de personas tenga la posibilidad de publicar libremente opiniones y comentarios acerca de una gran variedad de temas sociales, culturales, deportivos, científicos e incluso opiniones sobre productos y servicios.

Gracias a la popularidad que han adquirido estas redes sociales, actualmente es de gran interés para muchas empresas o figuras públicas conocer lo que se dice de ellas dentro de este mundo digital. Al mismo tiempo, a través de estos medios de comunicación, es posible tener un acercamiento con distintos usuarios, mismos que aprovechan esta interacción para dar a conocer de forma específica su opinión sobre determinados temas, productos o servicios. Una de las redes sociales más utilizadas para este fin es Twitter que permite enviar mensajes cortos llamados tweets, con una longitud máxima de 140 caracteres [1].

Con la finalidad de aprovechar la información obtenida de las interacciones entre usuarios y empresas, estas últimas han creado la figura de un Analista de Reputación en Línea [2]. Entre las actividades principales de este analista están: monitorear, clasificar y analizar información generada en torno a una empresa o figura pública en un determinado periodo de tiempo.

Debido a que las publicaciones crecen de manera acelerada, el análisis manual de esta información resulta complicado y desgastante para el analista de reputación en línea. En consecuencia, surge la necesidad de contar con sistemas automáticos que permitan realizar este análisis de forma más sencilla y oportuna. Recientemente diversos grupos de investigación están interesados en esta problemática y se han dado a la tarea de desarrollar sistemas enfocados al análisis de opiniones generadas en Twitter. En este contexto surge Replab [3] como una competencia internacional para evaluar sistemas automáticos enfocados en el análisis de información producida en Twitter.

Hasta el momento, los diferentes grupos que han participado en Replab, se han enfocado en desarrollar métodos automáticos para tareas como: i) selección de tweets relevantes para una entidad¹, ii) identificación de implicaciones negativas, positivas o neutrales hacia una entidad, iii) agrupamiento de opiniones por temática similar, y iv) la identificación de líderes de opinión dentro de una comunidad. Sin embargo, el problema de cómo mostrar el resultado del análisis automático a un analista de reputación en línea de forma que se le facilite la toma de decisiones, ha sido poco explorado.

¹ Por entidad se entiende el nombre de una organización, o nombre de alguna figura pública.

En este proyecto se explora una alternativa de visualización de los resultados producidos por sistemas de análisis de contenido en Twitter, en particular sistemas desarrollados en el marco de la competencia de Replab. El sistema propuesto se enfoca específicamente en la visualización de resultados del agrupamiento de opiniones por temática similar y la forma en cómo se relacionan.

2. Planteamiento del Problema

Se ha observado que los usuarios de las redes sociales en Internet juegan un papel muy importante en relación al prestigio o estima que se tiene de un producto o servicio. Son ellos los que por medio de opiniones y comentarios que publican en algunas redes sociales como Twitter, contribuyen a definir el posicionamiento que tiene una entidad particular dentro del mercado.

Actualmente Twitter, se encuentra entre las principales redes sociales más utilizadas del mundo, con 255 millones de usuarios activos mensualmente y 500 millones de tweets enviados por día [1]. Esta red social permite a sus usuarios enviar y publicar mensajes de texto plano de corta longitud con un máximo de 140 caracteres. La información que allí se publica es de gran utilidad a las empresas, permite conocer sus fortalezas y debilidades ante la competencia, conocer el posicionamiento en el mercado y en general, aporta información útil para la toma de decisiones.

Debido al gran volumen de información que se genera constantemente, las empresas han creado nuevos perfiles profesionales específicos de Internet. Estos profesionales son expertos en utilizar herramientas que permiten nuevas vías para llegar al usuario y nuevas formas de mercadotecnia. Uno de ellos es el de Analista de Reputación en Línea. El trabajo de este profesional pasa por tres fases principalmente: la primera fase consiste en el monitoreo, misma que permite conocer en todo momento lo que se está publicando con relación a la empresa, producto o figura pública. La segunda fase consiste en la identificación de temas de interés dentro de la comunidad de usuarios de Twitter, priorizando los mensajes con implicaciones negativas hacia la entidad en cuestión. Finalmente, la tercera fase consiste en proponer estrategias de mercado que permitan revertir los efectos negativos previamente identificados o incluso fortalecer los aspectos positivos de la entidad en revisión.

Dado que la tarea de identificación de temas y análisis de información es compleja, se requiere de software que procese grandes volúmenes de

información en tiempo real². Tradicionalmente el análisis de información, en nuestro caso tweets, se lleva a cabo manualmente por varias personas, invirtiendo gran cantidad de tiempo en la detección de información que pudiera ser relevante en un determinado conjunto de opiniones. Hoy en día gracias al avance en el desarrollo de algoritmos de agrupamiento, es posible contar con sistemas automáticos que permiten agrupar tweets por temáticas similares, los cuales, generan informes textuales que pudieran ser útiles al analista de reputación en línea.

Sin embargo, la lectura de los resultados de un informe de este tipo, es poco intuitiva y en consecuencia difícil de interpretar. Es por esto, que surge la necesidad de desarrollar sistemas que aprovechen los resultados de algoritmos de agrupamiento para generar representaciones visuales. Nuestra motivación para utilizar una representación visual es debido a que se ha mostrado que una de las mejores técnicas para hacer comprensibles los datos es a través de imágenes o graficas que sean intuitivas y fáciles de interpretar [4].

Por todo lo anterior, en este proyecto se desarrolló una aplicación Web que recupera tweets en tiempo real, por medio de una consulta relacionada a un tema en particular, posteriormente se aplicaron los algoritmos de agrupamiento a esos tweets y finalmente, se visualizan de forma gráfica los resultados de agrupamiento así como las posibles relaciones que existan entre los grupos obtenidos.

3. Objetivos

3.1 Objetivo General

“Desarrollar un sistema automático de visualización de las relaciones existentes entre las opiniones generadas en Twitter”.

3.2 Objetivos Específicos

Para cumplir con el objetivo general, se plantean los siguientes objetivos específicos:

² Tiempo real: sistema que interacciona con su entorno físico y responde a los estímulos del entorno dentro de un plazo de tiempo determinado. En este caso los usuarios interactúan con Twitter y el sistema de visualización recupera los tweets que se van generando en un determinado tiempo.

1. Utilizar algoritmos de agrupamiento que han mostrado ser eficaces en el área de procesamiento de lenguaje natural y que permitan encontrar relaciones entre textos de opinión, en particular tweets.
2. Proponer y desarrollar una representación visual que utilice los resultados de los algoritmos de agrupamiento para mostrar las relaciones entre opiniones.

4. Marco Teórico

El agrupamiento o Clustering, es una técnica que permite identificar de manera automática agrupaciones o clústeres de elementos de acuerdo a una medida de similitud entre ellos. Los elementos para este proyecto son tweets (principalmente el texto plano que consta de 140 caracteres). La tarea principal es dividir un conjunto de tweets en grupos de contenido similar, tomando en cuenta que tweets similares contendrán palabras similares.

Esta técnica es utilizada en la recuperación de información, mediante el agrupamiento de documentos, la cual, se encarga de localizar documentos relevantes de acuerdo a una petición de información o consulta. Para agrupar los documentos por temática similar se tiene en cuenta el contenido de dichos documentos, es decir las palabras, que son una representación razonable del contenido de cada documento.

Se puede definir formalmente el agrupamiento como:

Sea $D = \{d_1, d_2, \dots, d_N\}$ un conjunto de documentos. Se define un m -agrupamiento de D como una partición de D en grupos o clases c_1, c_2, \dots, c_m de forma que se cumplan las siguientes condiciones

- $c_i \neq \emptyset, i = 1, \dots, m$
- $\bigcup_{i=1}^m c_i = D$
- $c_i \cap c_j = \emptyset, i \neq j, i, j = 1, \dots, m$

Además se busca que los elementos de un grupo c_i sean más similares entre sí y menos similares a los elementos de otros grupos.

A continuación veremos los pasos básicos en un proceso de agrupamiento, los cuales constan de: la construcción de la representación de los tweets, la medición de similitud entre todos los tweets de la colección y finalmente la aplicación de algoritmos de agrupamiento.

4.1 Representación de los Documentos

El modelo de representación más utilizado en el agrupamiento de documentos es el modelo de espacio vectorial (VSM) propuesto por Salton en 1975 [16], mismo que utilizaremos para la agrupación de los tweets. La idea principal del modelo de espacio vectorial reside en la construcción de una matriz de palabras y documentos³, donde las columnas y las filas representan las palabras y los documentos respectivamente.

Un conjunto de elementos en el modelo de espacio vectorial puede verse como una matriz P , como se muestra en la figura 1. En donde cada componente $p_{i,j}$ de la matriz representa el peso de la palabra j en el documento i .

		Palabras o vocabulario de la colección	w_1	w_2	\dots	w_{k-1}	w_k
		d_1	$p_{1,1}$	$p_{1,2}$	\dots	$p_{1,k-1}$	$p_{1,k}$
		d_2	$p_{2,1}$	$p_{2,2}$	\dots	$p_{2,k-1}$	$p_{2,k}$
Documentos	
		d_{n-1}	$p_{n-1,1}$	$p_{n-1,2}$	\dots	$p_{n-1,k-1}$	$p_{n-1,k}$
		d_n	$p_{n,1}$	$p_{n,2}$	\dots	$p_{n,k-1}$	$p_{n,k}$

Figura 1. Ejemplo de una matriz palabra-documento utilizada en el modelo de espacio vectorial.

³ Para este proyecto el concepto de documentos se reemplaza por tweets.

4.2 Agrupamiento de documentos por temática similar

Para poder agrupar documentos por temática similar, es necesario calcular el peso $p_{i,j}$ de cada término o palabra. Para esto se utilizan distintas medidas de peso como son: valores binarios (0,1), frecuencia de términos (tf) y la combinación de la frecuencia de palabras (tf) con la frecuencia de palabras en la colección de elementos ($tf-idf$).

Cabe mencionar que en este proyecto se están considerando estos tres tipos de pesado y se da la opción para que el usuario sea el que determine el tipo de pesado que quiera utilizar en el agrupamiento de la colección.

A continuación se describe cada uno de los tipos de pesado utilizados en este proyecto.

4.2.1 Pesado Booleano

En un caso simple, se aplican exclusivamente valores binarios, de tal forma que si en el documento d_i aparece la palabra w_j , el valor de $p_{i,j}$ será 1 y en caso contrario 0. Como lo muestra la figura 2.

Palabras o vocabulario de la colección

	w_1	w_2	...	w_{k-1}	w_k
d_1	1	0	...	1	1
d_2	0	1	...	1	0
Documentos
d_{n-1}	0	0	...	0	1
d_n	1	1	...	1	1

Figura 2. Matriz del modelo de espacio vectorial con pesado booleano.

Es común que algunas palabras aparezcan más de una vez en el mismo documento, y que además unas palabras pueden considerarse como más significativas que otras, por lo tanto el valor de pesado $p_{i,j}$ se calcula con métodos más sofisticados.

4.2.2 Pesado de frecuencia de término

En el pesado de frecuencia de término (tf) a cada palabra en un documento se le asigna un peso proporcional a la cantidad de veces que ésta aparece en el documento.

En la figura 3 se muestran los valores obtenidos a través de la aplicación de la expresión $\log(1+f_{ij})$, donde f_{ij} es la frecuencia del término i en el documento j . Los valores de la matriz están normalizados entre 0 y 1, con el objetivo de que el peso asociado no dependa de la frecuencia relativa de esta palabra con otras palabras y poder evitar favorecer documentos más largos que otros.

		Palabras o vocabulario de la colección				
		w_1	w_2	...	w_{k-1}	w_k
Documentos	d_1	0	0.7	...	0.2	0
	d_2	0.5	0	...	0.6	0

	d_{n-1}	0	0	...	0	0.2
	d_n	0.6	0.4	...	0	0.4

Figura 3. Matriz del modelo de espacio vectorial con pesado tf .

4.2.3 Pesado tf-idf

El pesado *tf-idf* combina la frecuencia de la palabra, con la frecuencia de la palabra en la colección de documentos, representado por el factor *idf*. El factor *idf* da un estimado de la importancia de la palabra para describir a un documento en función de cuántos documentos en la colección contienen dicha palabra; es decir, entre menos documentos contengan a la palabra mayor será la utilidad de ésta palabra para describir a un documento.

Para calcular el peso *tf-idf* se aplica la siguiente formula:

$$p_{i,j} = tf(w_j, d_i) \times idf(w_j), idf(w_j) = \log \frac{N}{df(w_j)}$$

4.3 Medición de Similitud

La medición de similitud es un paso previo para poder utilizar los algoritmos de agrupamiento, nos permite determinar el grado de asociación entre todas las parejas de la colección. Se pueden utilizar distancias o medidas de similitud y disimilitud, entre las que se encuentran: la medida de similitud coseno [17], coeficiente Dice o Jaccard [17].

En este proyecto utilizamos la medida de similitud coseno, ya que esta medida es ampliamente utilizada en el trabajo de agrupamiento de documentos, y su popularidad se debe a que es posible obtener una representación geométrica del modelo vectorial, como se muestra en la figura 4. La idea principal de esta medida, es calcular el coseno del ángulo entre el vector del documento d_1 y el vector del documento d_2 , la fórmula es la siguiente:

$$s_{1,2} = \cos(d_1, d_2) = \frac{\sum_{l=1}^k p_{l,d_1} \times p_{l,d_2}}{\sqrt{(\sum_{l=1}^k p_{l,d_1}^2)} \times \sqrt{(\sum_{l=1}^k p_{l,d_2}^2)}}$$

Donde d_1 y d_2 representan los documentos a comparar, p_{l,d_1} representan el peso de la palabra l en el documento d_1 ; y p_{l,d_2} representa el peso de la palabra l en el documento d_2 .

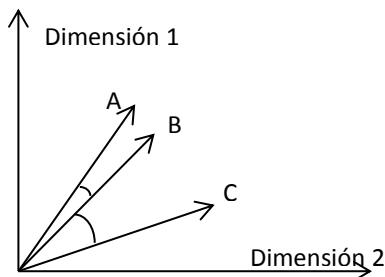


Figura 4. Matriz del modelo de espacio vectorial: Similitud entre vectores

Los valores obtenidos por la similitud coseno oscilan entre 0 y 1. Un valor cero (ángulo de 90°) indica que los documentos comparados son completamente distintos, de acuerdo a las características con las que fueron representados. El valor 1 (ángulo de 0°) indica que los documentos coinciden en los valores de todas sus características.

En una colección con N documentos, se genera una matriz de similitud s de tamaño $N \times N$, como se muestra la figura 5.

		Documentos				
		d_1	d_2	...	d_{n-1}	d_n
Documentos	d_1	1				
	d_2	$s_{2,1}$	1			
	
	d_{n-1}	$s_{n-1,1}$	$s_{n-1,2}$...		1
	d_n	$s_{n,1}$	$s_{n,2}$...	$s_{n,n-1}$	1

Figura 5. Matriz de similitud de N documentos

Se puede observar que en la matriz de similitud los elementos de la diagonal son iguales al máximo valor obtenido para la medida de similitud, es decir 1. La parte superior derecha de la figura 5 se muestra en blanco, porque la matriz es simétrica al igual que la función de similitud, es decir, $(s_{i,j} = s_{j,i})$

4.4 Algoritmos de Agrupamiento

Actualmente existen varios algoritmos de agrupamiento, los cuales pueden clasificarse en dos tipos principales que dependen de la estructura de agrupamiento que generan. Los dos grupos de algoritmos son [18]: algoritmos jerárquicos y algoritmos de partición, como lo muestra la figura 6.

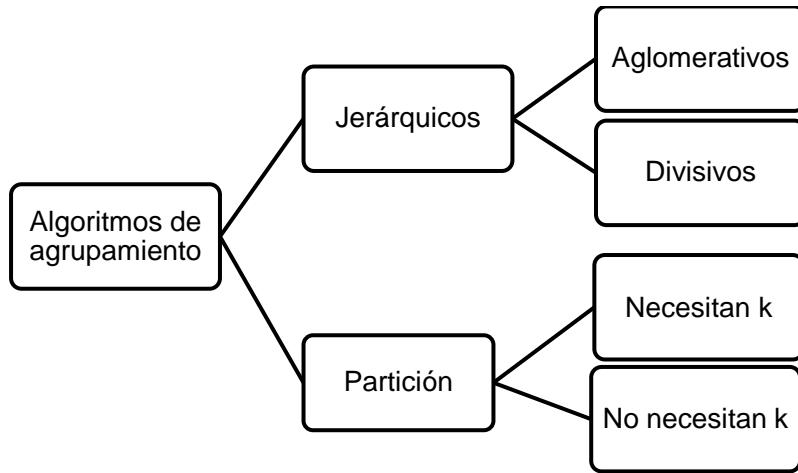


Figura 6. Clasificación de los algoritmos de Agrupamiento

4.4.1 Agrupamiento Jerárquico

Los algoritmos jerárquicos se caracterizan por generar una estructura de árbol, llamada dendrograma, en la que cada nivel del árbol es un posible agrupamiento de los objetos de la colección. Cada vértice o nodo del árbol es un grupo de documentos. La raíz del árbol, que es el primer nivel, se compone de un único grupo que contiene todos los elementos. Cada hoja del último nivel del árbol es un grupo compuesto por un único documento, hay tantas hojas como documentos tenga la colección. En los niveles intermedios, cada nodo del nivel n es dividido para formar sus hijos del nivel $n+1$.

En la figura 7 se muestra un dendograma generado por una colección de elementos, así como el agrupamiento obtenido en cada nivel.

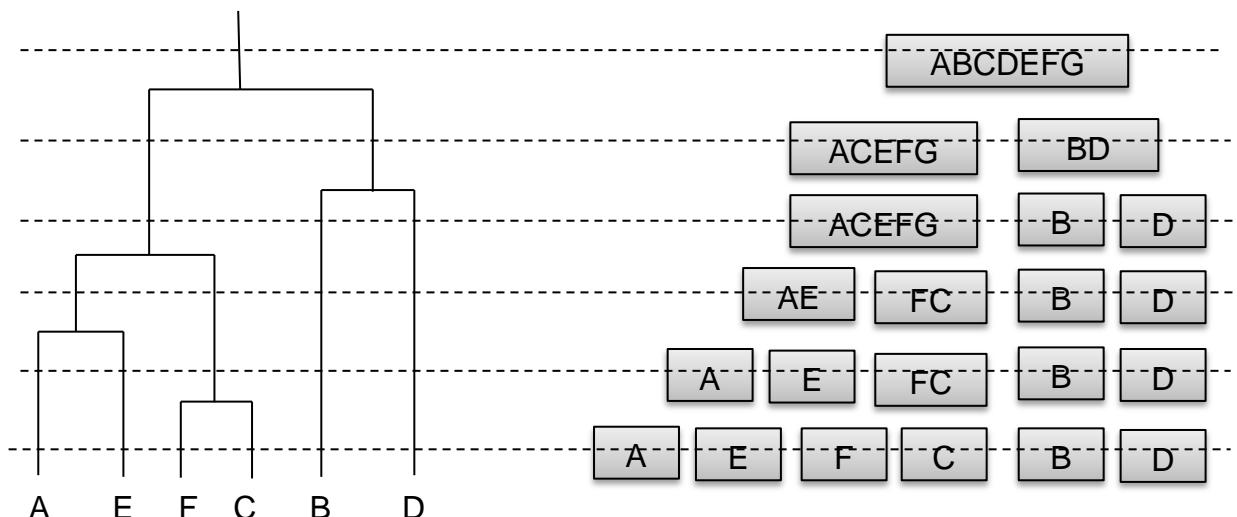


Figura 7. Ejemplo de dendograma.

Los algoritmos jerárquicos pueden dividirse en aglomerativos y divisivos:

- Aglomerativos: Estos algoritmos parten de las hojas del árbol, considerando a cada elemento como un grupo. De forma iterativa se van uniendo en grupos más cercanos. Este procedimiento continúa hasta que todos los documentos se encuentran dentro de un grupo o hasta que se cumplan ciertas condiciones de paro.
- Divisivos: El algoritmo de división supone que en un primer paso todos los documentos conforman un primer grupo, es decir, inician desde la raíz del árbol. Este grupo se va dividiendo sucesivamente en grupos más pequeños hasta que cada elemento constituye un grupo o hasta que se cumplan las condiciones de paro.

En este proyecto se utiliza como algoritmo jerárquico, el algoritmo denominado Hierarchical Clustering y es aglomerativo.

4.4.2 Agrupamiento de Partición

En el agrupamiento de partición se agrupan los elementos entorno a elementos centrales llamados centroides en cada grupo. Los algoritmos particionales se dividen en dos grupos: los que necesitan que se les indique la cantidad de grupos a formar y los que no la requieren esta información. Los algoritmos que no necesitan que se les especifique el número de grupos inducen de manera natural el número de grupos a formar.

Entre los algoritmos particionales se encuentra el algoritmo k-means y es el algoritmo de partición que se está utilizando en este proyecto. Este algoritmo si requiere que se le especifique el número de grupos a formar.

Se eligió el algoritmo k-means porque es un algoritmo que se usa en ambientes de agrupamiento de documentos y ha obtenido buenos resultados dado la simplicidad de su uso [19]. En la figura 8 se describe el desarrollo del algoritmo de partición k-means.

Algoritmo K-means

1. Indicar el número de grupos K a formar.
2. Determinar los centros para cada uno de los K grupos iniciales. La selección se puede determinar de distintas maneras incluso aleatoriamente.
3. Determinar la distancia de las instancias en relación a los centros y asignar cada instancia al centro más cercano
4. Recalcular el centro de cada grupo y reasignar las instancias de acuerdo a los nuevos centros.
5. Se repite este proceso hasta que se han cumplido ciertas condiciones de para o cuando se alcanza un número de iteraciones y los centros ya no cambian.

Figura 8. Desarrollo de Algoritmo K-means.

5. Trabajo Relacionado

Actualmente se tienen disponibles en Internet aplicaciones enfocadas al agrupamiento, análisis y visualización de información; estas aplicaciones las podemos dividir en dos categorías: la primera consiste en herramientas especializadas en el agrupamiento de grandes volúmenes de datos contenidos en archivos. Estas herramientas emplean un lenguaje muy técnico y como consecuencia se dificulta la interpretación de los resultados. Dentro de esta categoría se encuentran Cluto y Weka, herramientas multiplataforma que pueden ser incluidas en el desarrollo de sistemas de visualización de información.

En la segunda categoría están las herramientas que se dedican al análisis de datos que se publican exclusivamente en redes sociales, y que permiten una fácil interpretación de los resultados incluso a usuarios con poca experiencia en el análisis de información, tal es el caso de Spot, AnaliticPro y Socialmention*. Estas herramientas proporcionan la visualización del análisis de datos por medio de métricas e indicadores que permiten evaluar y conocer la reputación que se tiene acerca que un producto o tema en particular.

A continuación describiremos la tarea de agrupamiento, análisis y visualización de información que emplean las herramientas antes mencionadas, así como la forma visual en que presentan los resultados.

5.1 Herramientas Especializadas en el Agrupamiento de Datos

5.1.1 Cluto & gCluto

Cluto, es una aplicación multiplataforma, enfocada al agrupamiento de datos, permite analizar las características de los diferentes grupos que se forman. Una propiedad importante de esta aplicación es que trata el problema de agrupamiento como un proceso de optimización que busca maximizar o minimizar una función de criterio en particular, es decir, que los elementos que integran los grupos tengan una fuerte relación de similitud entre ellos y que a su vez la relación entre grupos sea mínima. Esta herramienta puede trabajar con grandes volúmenes de información y bases de datos que contengan miles de objetos así como un gran número de características [9].

gCluto, es una aplicación gráfica que proporciona las herramientas de visualización de resultados obtenidos del agrupamiento de Cluto. Su función principal es hacer que las capacidades de agrupamiento de Cluto se representen de manera gráfica. Para ello, gCluto proporciona varias maneras de visualizar, de forma interactiva, los resultados del agrupamiento, como se muestran en las figuras 9 y 10.

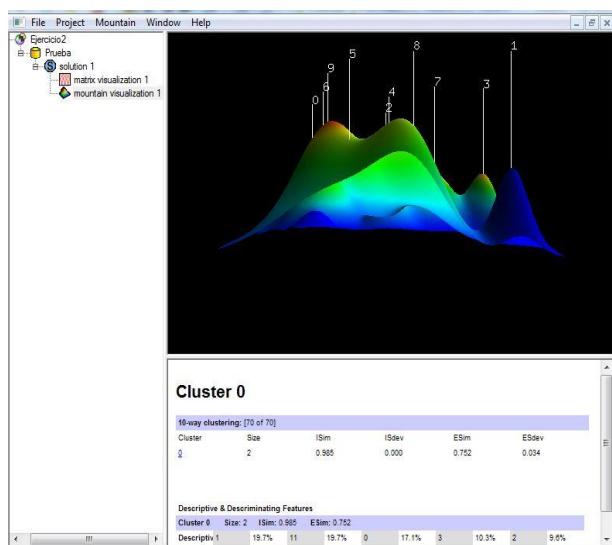


Figura 9: Visualización de montaña.

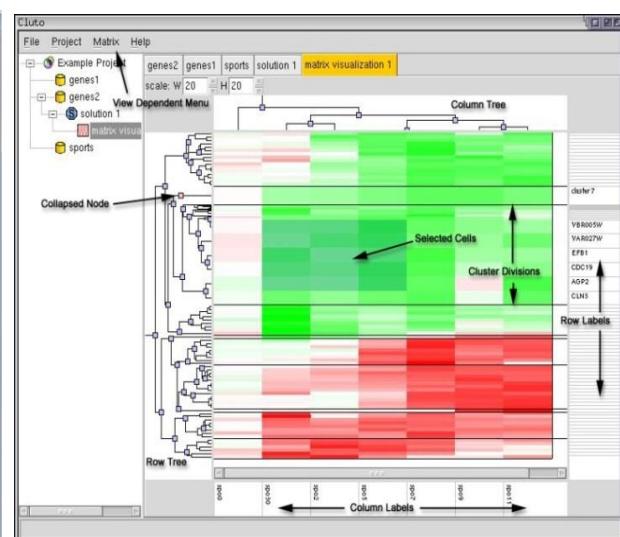


Figura 10: Visualización de matriz.

5.1.2 Weka

Es una herramienta que se usa en el aprendizaje automático y la minería de datos, emplea un conjunto de librerías en Java para la extracción de información desde grandes bases de datos o archivos de texto [11]. Este paquete contiene una colección de herramientas de visualización y algoritmos para el análisis de datos. A través de una interfaz gráfica, el usuario puede acceder fácilmente a las funcionalidades que ofrece esta herramienta.

Entre las funcionalidades que realiza Weka para el análisis y el estudio de datos se encuentran las siguientes:

- Preprocesos: permite realizar manipulaciones sobre los datos aplicando filtros.
- Agrupamiento: diferentes algoritmos de agrupamiento generan grupos de instancias con características similares.
- Asociación: Algoritmos para encontrar relaciones de asociación entre características.
- Clasificación: aplicar algoritmos de clasificación a un conjunto de datos.
- Selección de atributos: Una vez cargados los datos, Weka es capaz de buscar las mejores características del modelo que se vaya a utilizar.
- Visualizar: la herramienta de visualización de datos permite representar graficas en 2D que relacionan pares de atributos; se pueden seleccionar diferentes formas de visualizar los datos, por ejemplo de la figura 11, se muestra la visualización de relaciones entre dos atributos.

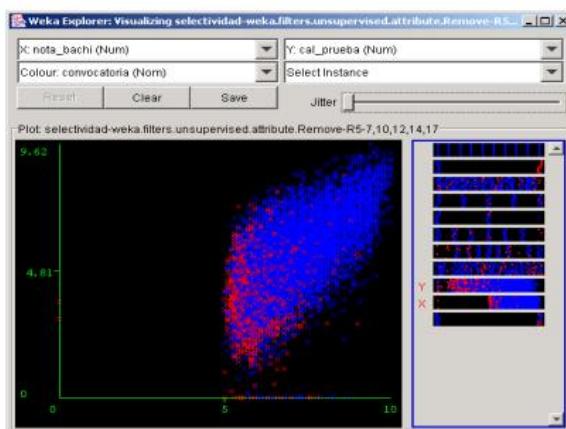


Figura 11: Visualización de la relación entre dos atributos

5.2 Herramientas Dedicadas al Análisis de Información en Redes Sociales

5.2.1 Spot

Es una aplicación que permite la visualización interactiva de lo que se está publicando en Twitter en tiempo real [12]. La idea principal es mostrar rápidamente las opiniones que se generan sobre un tema en particular.

La forma de presentar los tweets, es a través de grupos contenidos en burbujas. Las burbujas se organizan y visualizan de diferentes maneras para proporcionar información sobre el tema de interés. Al realizar la búsqueda del tema, se recuperaran los últimos 200 tweets para la visualización. Hay que tomar en cuenta que los resultados de búsqueda de Twitter sólo se remontan alrededor de una semana. Por lo que a veces la búsqueda sólo podrá regresar unos pocos tweets. La figura 12 muestra de la interfaz de usuario de la aplicación Spot.



Figura 12: Visualización de los cinco grupos principales del tema seleccionado.

5.2.2 AnaliticPro

Es una aplicación que procesa grandes volúmenes de información producida en distintas redes sociales, el análisis de datos es de forma automática [13]. Se realizan mediciones con criterios personalizados, se identifica el sentido de los comentarios, se relacionan y construyen frases para conocer lo que en realidad dicen las personas sobre algún tema público o privado.

Algunas de las características que integran AnaliticPro son las siguientes:

- Medir en tiempo real lo que se dice de una marca, competencia o algún tema de interés. Esto, a través de la recopilación, clasificación y análisis de datos que se está publicando en distintas redes sociales.
- Permite obtener informes estratégicos, personalizados y procesados por expertos. Se personalizan los requerimientos y se entregan las respuestas que se necesitan los clientes.
- Visualización de resultados a través de gráficas, se muestra en la figura 13.



Figura 13: Visualización de la evaluación de diferentes métricas, de lo que se publica en las redes sociales.

5.2.3 Socialmention*

Es una aplicación que monitorea y analiza la información que se está generando en distintas redes sociales de Internet en tiempo real [14]. Además, permite seguir y medir fácilmente lo que se opina sobre alguna persona, empresa, producto, etc. Las búsquedas se realizan en más de 80 medios sociales incluyendo las más visitadas como son: Twitter, Facebook, friendFeed, YouTube, Digg, Google, etc.,

Para poder analizar la información, esta herramienta clasifica las publicaciones en medidas de:

- *Strength*, este parámetro muestra los mensajes publicados sobre la marca en las últimas 24 horas.
- *Sentiment*, se refiere al contenido de los mensajes en relación a su polaridad, es decir, identifica comentarios positivos y negativos.
- *Passion*, muestra la cantidad de autores que hablan de la marca. Si los mismos autores comentan repetidamente sobre el tema la pasión será alta, pero si por el contrario los autores comentan una sola vez, la pasión será baja.
- *Reach*, indica la influencia de la marca, es el número de autores únicos que hacen referencia a la marca, dividido por el número total de publicaciones.

Adicionalmente Socialmention* permite ver cuáles son las palabras claves más utilizadas en las publicaciones, muestra los Hashtags asociados a la marca, lista las redes sociales en donde se generaron las publicaciones, como se muestra en figura 14.

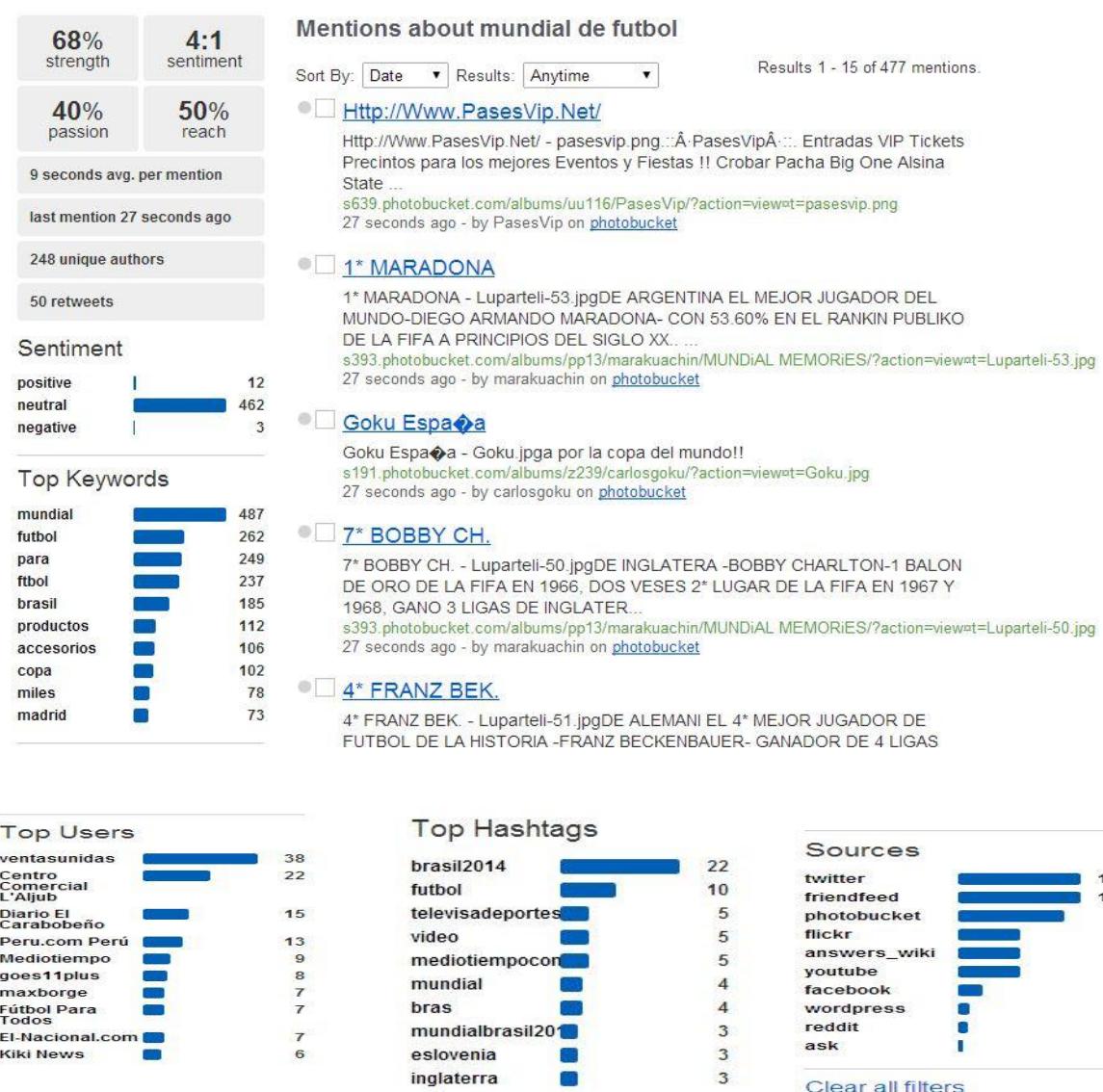


Figura 14: Visualizaci n de las m tricas que utiliza Socialmention* para el an lisis de informaci n.

5.3 Cuadro Comparativo de las Herramientas de Visualización de Datos

En la tabla 1 se muestra una comparación de las herramientas utilizadas en el agrupamiento y visualización de información.

	Herramienta libre	Recuperación de datos en tiempo real	Búsquedas en diferentes redes sociales de internet	Búsquedas exclusivamente en Twitter	Agrupamiento de documentos en base a contenido	Implementan Múltiples métodos de agrupamiento	Visualización de resultados con datos estadísticos	Visualización amigable e intuitiva	Identifica relaciones entre grupos
Cluto & gCluto	✓	X	X	X	✓	✓	✓	X	X
Weka	✓	X	X	X	✓	✓	✓	X	X
Spot	✓	✓	X	✓	✓	X	X	✓	X
AnaliticPro	X	✓	✓	X	✓	X	✓	X	X
Socialmention	✓	✓	✓	X	✓	X	✓	X	X
Sistema de visualización de relaciones de opiniones producidas en Twitter	✓	✓	X	✓	✓	✓	✓	✓	✓

Tabla 1. Comparación de herramientas de agrupamiento y visualización de información.

A continuación se describen las características consideradas en el cuadro de comparación de las herramientas de visualización de datos:

- Herramienta libre: Aplicaciones que no tienen un costo para su uso y que proporcionan los programas de instalación para ser incorporados al desarrollo de otros sistemas.
- Recuperación de datos en tiempo real: Obtener los datos casi al mismo tiempo que se están publicando en las redes sociales.
- Búsqueda en diferentes redes sociales de Internet: Recuperar datos en más de una red social.

- Búsqueda exclusivamente en Twitter: Localizar y recuperar datos solo en Twitter.
- Agrupamiento de documentos en base a contenido: Agrupar documentos que contengan palabras o datos de temática similar.
- Agrupamiento de documentos en base a contenido: Agrupar documentos que contengan palabras o datos de temática similar.
- Implementar múltiples métodos de agrupamiento: Aplicar métodos de agrupamiento, utilizando distintos algoritmos que realizan esta tarea en un determinado conjunto de documentos.
- Visualización de resultados con datos estadísticos: Mostrar cantidades en porcentaje de cada una de las métricas aplicadas al grupo.
- Visualización amigable e intuitiva: Que la interfaz cumpla con las técnicas de usabilidad y la representación visual proporcione información relevante al usuario.
- Identifica relaciones entre grupos: Usar conectores gráficos que conecten e indiquen la relación de contenido que existe en los diferentes grupos.

De acuerdo a la tabla 1, el sistema de visualización de relaciones de opinión producidas en Twitter integra algunas funcionalidades que se realizan en los programas mencionados pero sobretodo brinda una mejor visualización de los resultados de agrupamiento por temática similar.

6. Desarrollo del Sistema de Visualización

El sistema automático para la visualización de relaciones de opiniones producidas en Twitter es un sistema web. La implementación del mismo consistió en el desarrollo de tres módulos que se trabajaron de forma secuencial para ir cubriendo las tareas requeridas. Los programas y librerías que se utilizaron son los siguientes:

- Lenguaje de programación Java.
- Librería twitter4j para la conexión con Twitter.
- Librería de Weka para el agrupamiento por temática similar.
- Librería json para generar archivos de visualización.
- Librería de bases de datos MySQL.
- Primefaces para la interfaz de usuario.
- D3js para la visualización de información.

A continuación se muestra el esquema general del sistema así como cada uno de los módulos que lo integran.

6.1 Esquema General del Sistema

El esquema general del sistema contempla los tres módulos que se desarrollaron en la implementación. Cada módulo realiza una tarea específica, como se muestra en la figura 15.

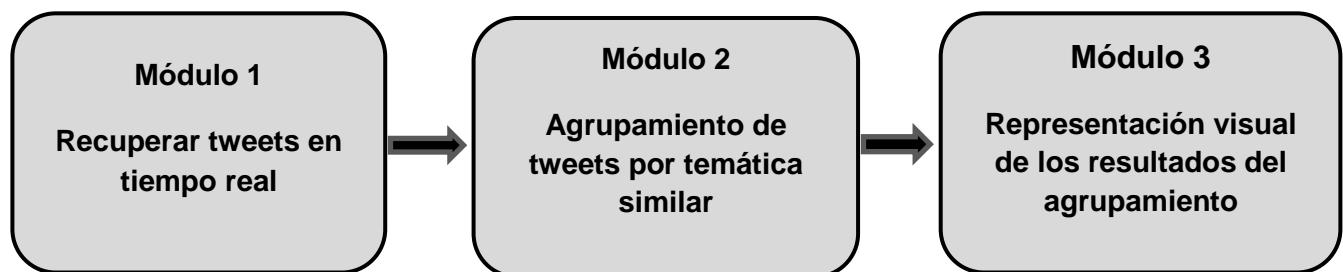


Figura 15. Esquema general del sistema.

6.2 Módulo 1: Recuperar Tweets en Tiempo Real

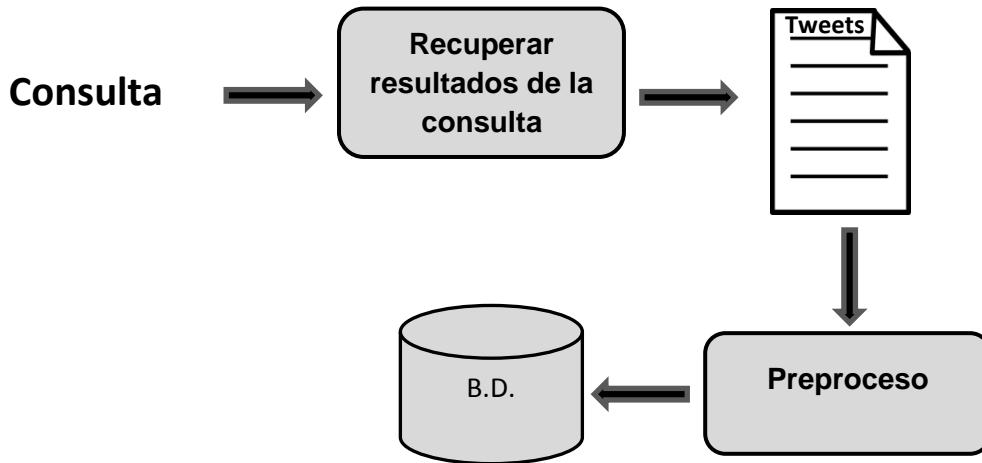


Figura 16. Módulo 1: Recuperar tweets en tiempo real.

La recuperación de los tweets comienza con una consulta, como se muestra en la figura 16, en donde se especifica el tema a buscar. La consulta consiste en una cadena de caracteres, y el número de tweets que se desean recuperar. Para poder realizar esta tarea se utilizó la librería twitter4j [5], la cual permite la conexión con Twitter, esto, previo registro en la página de desarrolladores de Twitter, en donde se obtienen las llaves de acceso que permiten hacer uso de la información que se genera en esa red social.

Si la conexión a Twitter es exitosa, se recuperan los tweets y se muestran en el sistema, en caso contrario de despliega un mensaje de error y se cierra la sesión.

Una vez recuperados los tweets se almacenan en la base de datos en MySQL denominada *bdtwitter*. Cabe mencionar que el sistema guarda las consultas que realiza el usuario para su posterior uso.

Una parte importante en el agrupamiento de tweets por temática similar es el preprocesamiento[15] de información, ya que permite tener un texto con palabras significativas que ayuden a identificar el tema de cada tweet.

El preprocesamiento de información considera las siguientes tareas:

- Convertir el texto a minúsculas.
- Reemplazar vocales acentuadas por vocales sin acento.

- Eliminar las palabras vacías, que es el nombre que reciben las palabras que carecen de significado como artículos, pronombres y preposiciones.
- Eliminar símbolos
- Eliminar hashtags
- Eliminar url

6.3 Módulo 2: Agrupamiento de Tweets por Temática Similar

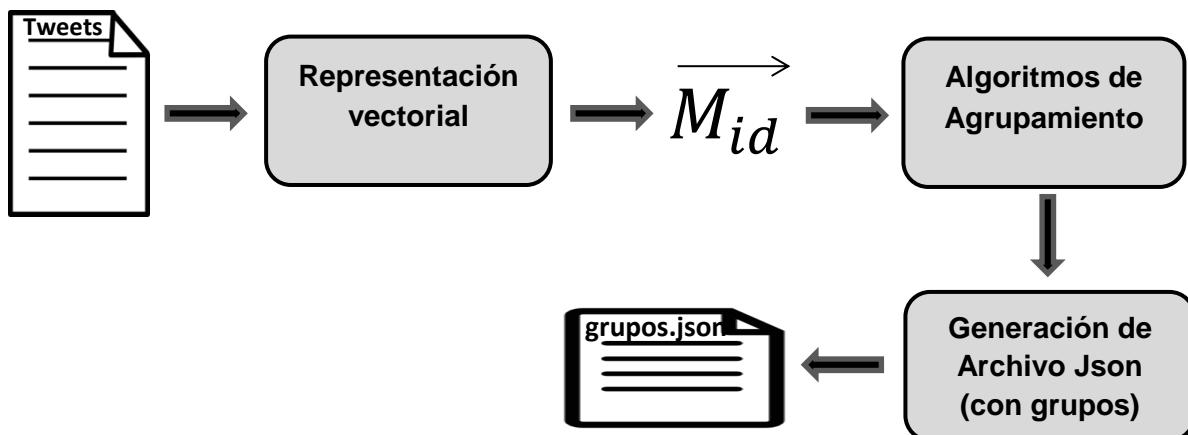


Figura 17. Módulo 2: Agrupamiento de tweets por temática similar

La figura 17 muestra el proceso para agrupar los tweets por temática similar, para lo cual se hace uso de la librería de Weka[11], ésta librería contiene las funciones necesarias para realizar el agrupamiento de documentos por temática similar, como se menciona en la sección 4 del presente documento.

La función *StringToWordVector* de la librería de Weka genera la representación vectorial del conjunto de elementos y al mismo tiempo recibe como parámetro el tipo de peso que se desea utilizar (binario, tf, if-idf). El sistema contempla los tres tipos de peso que se han descrito anteriormente, y será el usuario el que elija el tipo de peso que desea aplicar.

Weka también cuenta con los algoritmos de agrupamiento jerárquico y de partición. El sistema integra dos algoritmos de agrupamiento el primero de partición, kmeans, y el segundo jerárquico, Hierarchical Cluster. Se tiene la opción de elegir cualquiera de los dos.

Finalmente en este módulo se genera un archivo de tipo json⁴, que contiene el resultado del agrupamiento con la siguiente información:

- El número de grupos que se formaron.
- Los tweets que contiene cada grupo.
- Las cinco palabras que describen a cada grupo, estas palabras se obtienen de la siguiente manera:
 1. Se crea un diccionario con las palabras que contienen los tweets (esto asegura que solo se tenga una vez cada palabra).
 2. Se toman las palabras que se tienen en el diccionario y se comparan con las palabras que contienen cada uno de los tweets, esto permite hacer un conteo del número de veces que aparece la palabra en la colección de tweets.
 3. Se ordena el diccionario de palabras con los valores de mayor a menor.
 4. Se extraen las primeras cinco palabras del diccionario.
- Las cinco palabras que más se repiten entre grupos. Estas palabras nos permiten tener las relaciones entre grupos y se obtienen de la siguiente manera:
 1. Se crea un sub-grupo de palabras con las cinco palabras que se obtuvieron en la descripción de cada grupo.
 2. Se toman dos sub-grupos de palabras y se extraen las palabras que son comunes entre ambos sub-grupos.

⁴ Archivo Json (JavaScript Object Notation), es una notación para definir datos basada en lenguaje javascript. Su mayor ventaja es que puede ser leído por cualquier lenguaje de programación.

6.4 Modulo 3: Representación Visual de los Resultados de Agrupamiento

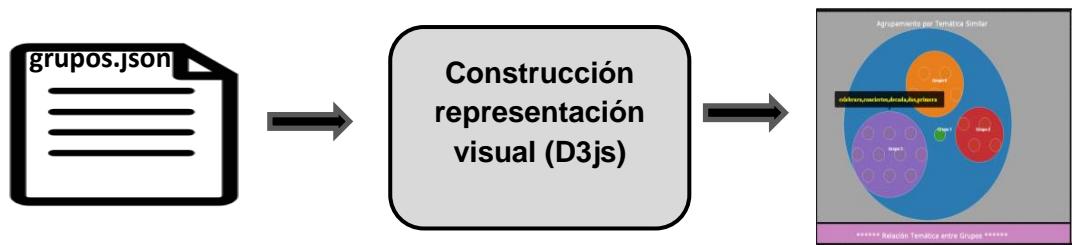


Figura 18. Módulo 3: Visualización de agrupamiento temático

Para la representación visual de los resultados de agrupamiento temático, que se tienen en el archivo de tipo json, se hace uso de la librería D3js⁵, propiamente de la gráfica *Bubble Chart*. Esta gráfica permite visualizar los resultados del agrupamiento como son: los grupos que se formaron, los tweets de cada grupo, las cinco palabras que representan a cada grupo y las cinco palabras que se tienen en común entre dos grupos.

⁵ D3JS (Data-Driven Documents) es una librería JavaScript que permite manipular y visualizar documentos basados en datos.

7. Conclusiones y Trabajo Futuro

En el sistema de visualización de relaciones de opiniones producidas en Twitter, se logró obtener los tweets en tiempo real y almacenarlos en una base de datos. Esto da la posibilidad de realizar el análisis de información en determinados periodos de tiempo y así poder detectar en un momento dado las variantes, en cuanto a opiniones, que van surgiendo en torno a un tema en específico. En cuanto al agrupamiento de información por temática similar, se emplearon dos algoritmos, uno partición y otro jerárquico que trabajan con tres tipos de peso, esto permite visualizar gráficamente diferentes formas de cómo los tweets se agrupan de forma automática. Adicionalmente se crearon tres graficas más, ya que en la recuperación de tweets no solo se obtuvo el texto del mismo sino que además se recuperó información relacionada al tweet como es la fecha de creación, el lugar o ciudad donde fue publicado, el número de veces que ha sido marcado como favorito, el número de veces que ha sido retuiteado y desde que plataforma fue publicado. También se recuperó información relacionada al usuario como el nombre de usuario con el que se registró, el nombre completo, la frase o descripción de su perfil y la url de la foto o imagen que tiene en su cuenta. Las gráficas muestran el agrupamiento por: la plataforma en donde fueron publicado el tweet, el número de veces que el tweet ha sido marcado como favorito y el número de veces que el tweet fue retuiteado.

Como trabajo futuro se plantea la realización de las siguientes actividades:

- Utilizar algoritmos de agrupamiento que tengan diferente forma de crear los grupos por ejemplo con el algoritmos estrella, cobweb, EM, entre otros.
- Realizar gráficas con mayor aportación de información, esto es posible dado que se cuenta información adicional relacionada al usuario y al tweet.
- Explotar la parte gráfica, ya que la librería D3js es robusta y flexible en visualización de datos.

8. Referencias

- [1] Acerca de Twitter. Recuperado el 10 de abril de 2014 de
<https://about.twitter.com/>
- [2] Responsable de la comunidad de internet. Recuperado el día 15 de abril de 2014, de
http://es.wikipedia.org/wiki/Responsable_de_comunidad_de_internet
- [3] RepLab, Recuperado el día 15 de abril, 2014 de
<http://www.limosine-project.eu/events/replab2012>
- [4] Naciones Unidas, Comisión Económica para Europa. Cómo hacer comprensible los datos. Parte 2. Una guía para representar estadísticas. Págs. 7-12, Ginebra, 2009.
- [5] Twitter4J. Recuperado el día 12 de junio, 2014 de
<http://twitter4j.org/en/index.html>
- [6] Processing 2. Recuperado el día 16 de junio, 2014 de
<https://www.processing.org/>
- [7] OpenGL. Recuperado el día 20 de junio, 2014 de
<http://www.opengl.org/>
- [8] D3 Data-Driven Documents. Recuperado el día 22 de junio, 2014 de
<http://d3js.org/>
- [9] KarpisLab. Recuperado el día 11 de mayo, 2014 de
<http://glaros.dtc.umn.edu/gkhome/views/cluto>
- [10] gCluto Documentation. Recuperado el día 12 de mayo, 2014 de
<http://glaros.dtc.umn.edu/gkhome/fetch/sw/gcluto/OLD/manual/index.html>
- [11] WEKA The University of Waikato. Recuperado el día 15 de mayo, 2014 de
<http://www.cs.waikato.ac.nz/ml/weka/>
- [12] Spot. Recuperado el día 25 de mayo, 2014 de
<http://neofromix.com/2012/IntroducingSpot.html>
- [13] Analitic. Recuperado del día 30 de mayo, 2014 de
<http://www.analiticpro.cl/caracteristicas.php>

[14] Socialmention*. Recuperado el día 05 de junio, 2014 de
<http://www.socialmention.com/>

[15] El Procesamiento del Lenguaje Natural en la Recuperación de Información Textual y áreas afines.. Recuperado el día 25 de febrero, 2014 de
<http://www.upf.edu/hipertextnet/numero-5/pln.html>

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

[17] Ricardo Baeza-Yates and Berthier Ribeiro-Neto. Modern Information Retrieval. Addison Wesley, 1999.

[18] M. Steinbach, G. Karypis and V. Kumar, "A Comparison of Document Clustering Techniques," Proc. KDD-2000 Workshop TextMining, Aug. 2000

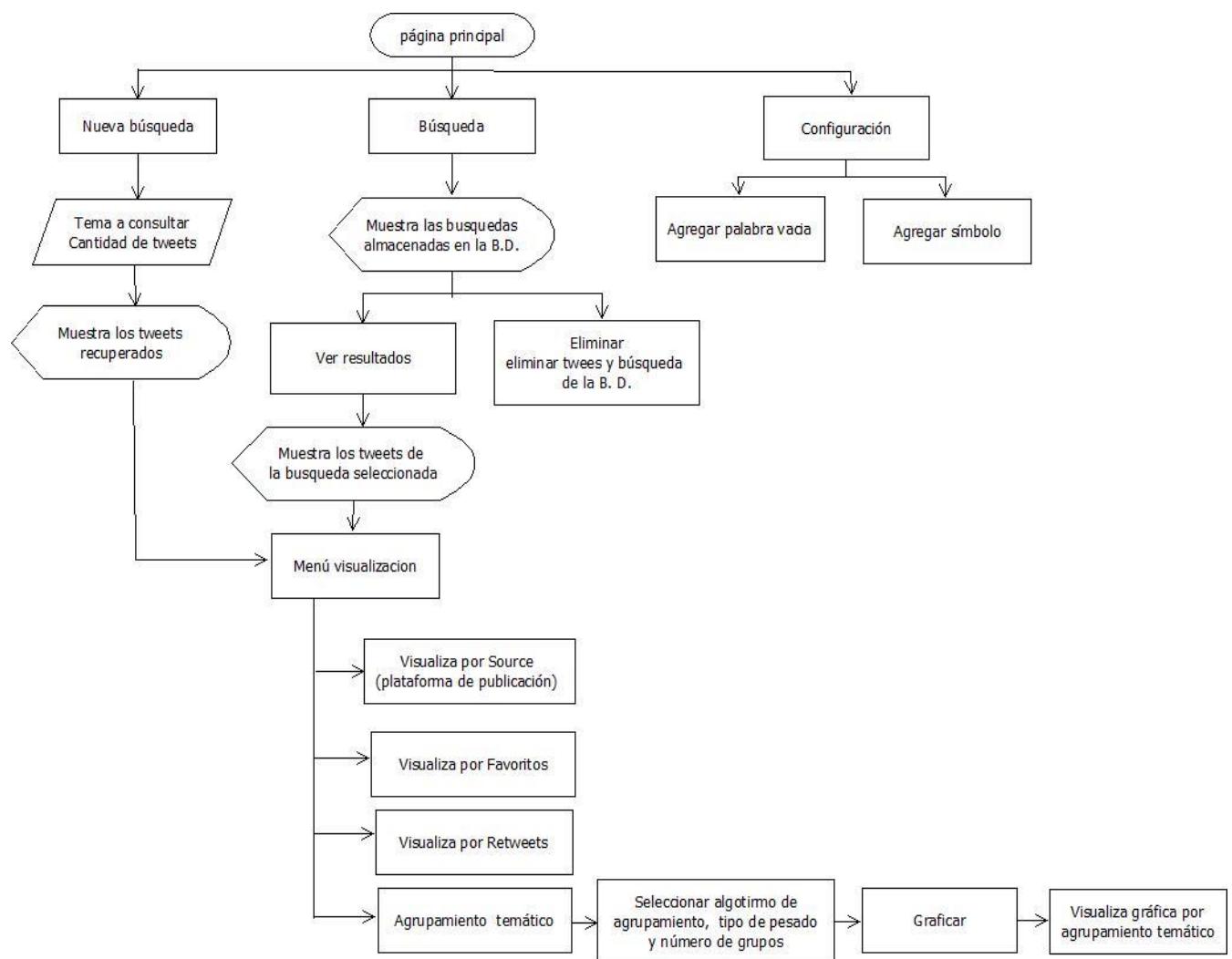
[19] Ejemplo de agrupamiento k-means. Recuperado el día 05 de mayo, 2015 de
<http://people.revoledu.com/kardi/tutorial/kMean/EjemploNumerico.htm>

9. Anexos

Se incluyen cinco anexos relacionados al mapa de navegación, diagramas de casos de uso, diagramas de secuencia, diagramas de clases, diseño de la base de datos y el manual de usuario del sistema.

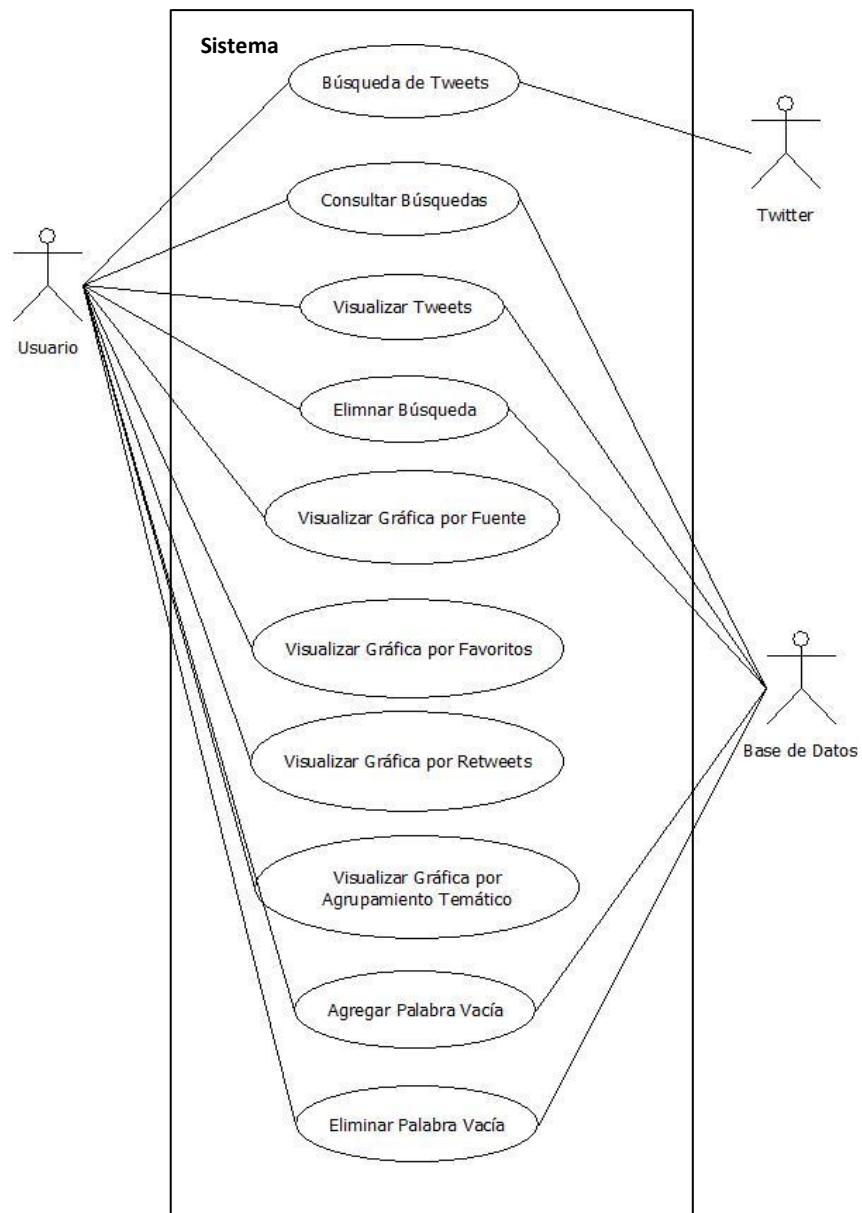
9.1 Anexo A: Mapa de Navegación

En este mapa de navegación se muestra el esquema de las páginas web del sistema, así como las interacciones con el usuario.



9.2 Anexo B: Diagramas de Casos de Uso

Este diagrama muestra las acciones que realizan los actores en el sistema y la forma en cómo interactúan con los diferentes casos de uso.



A continuación se describen la secuencia de los diez casos de uso que el usuario realiza en el sistema.

Diagrama de caso de uso 1: Búsqueda de tweets

Caso de Uso 1	Búsqueda de tweets
Descripción:	Permite realizar la búsqueda y recuperación de tweets.
Actores:	Usuario
Precondiciones:	El usuario debe de ingresar el tema a buscar y el número de tweets que desea recuperar.
Flujo Normal:	
1	El usuario escribe el tema de los tweets a recuperar.
2	El usuario selecciona de la lista desplegable el número de tweets a recuperar.
3	El usuario da clic en “Búscar”
Flujo Alternativo:	
2 ^a	El usuario puede escribir el número de tweets a recuperar
Poscondiciones:	
	El sistema muestra en una tabla los tweets que se recuperaron.

Diagrama de caso de uso 2: Consultar búsquedas

Caso de Uso 2	Consultar búsquedas de la B.D.
Descripción:	Permite visualizar las búsquedas o consultas que se hayan realizado con anterioridad y que se encuentran almacenadas en la base de datos.
Actores:	Usuario
Precondiciones:	Levantar servidor MySQL y Apache
Flujo Normal:	
1	El usuario selecciona la opción “Búsquedas”.
2	El sistema despliega una tabla con las búsquedas almacenadas en la B.D. La tabla muestra el número de búsqueda, el tema, el número de tweets que cada búsqueda, la fecha y hora en que se realizó la búsqueda.
3	El usuario selecciona la búsqueda que desea consultar.
Flujo Alternativo:	
2 ^a	
Poscondiciones:	
	El sistema mostrará los tweets de la búsqueda seleccionada. Previa selección de la opción “Ver Resultados”.

Diagrama de caso de uso 3: Visualizar tweets

Caso de Uso 3	Visualizar tweets a
Descripción:	Se muestran los tweets que se encuentran dentro de la búsqueda.
Actores:	Usuario
Precondiciones:	Hacer clic en la opción “Búsquedas” y seleccionar búsqueda
Flujo Normal:	
1	El usuario selecciona la opción “Ver Resultados”.
2	El sistema despliega una tabla con los tweets de la búsqueda. Se muestran los campos id, texto y fecha de creación para cada tweet.
3	El usuario puede dar clic en un círculo del lado izquierdo de cada tweet y el sistema despliega otra tabla con campos relacionados con el tweet y con el usuario que publica el tweet. Datos del tweet: id lenguaje source o plataforma desde donde se publicó Favorito las veces que ha sido marcado como favorito Retweet las veces ha sido retuiteado
	Datos del usuario: id usuario nombre completo lenguaje localización url de la foto o imagen de perfil
Flujo Alternativo:	
2 ^a	
Poscondiciones:	
	El sistema despliega la opción “Menú Visualización”.

Diagrama de caso de uso 4: Eliminar búsqueda

Caso de Uso 4	Eliminar búsqueda .
Descripción:	Permite eliminar una búsqueda con todos los tweets que contiene.
Actores:	Usuario
Precondiciones:	Seleccionar la opción “Búsquedas”
Flujo Normal:	
1	El usuario da clic en la opción “Eliminar”.
Flujo Alternativo:	
2 ^a	
Poscondiciones:	El sistema actualiza la lista de las búsquedas

Diagrama de caso de uso 5: Visualizar gráfica por source

Caso de Uso 5	Visualizar gráfica por source
Descripción:	Se visualiza la gráfica por source o plataforma, la cual agrupa las plataformas desde donde se publicaron los tweets además muestra el número de veces que se utilizó la plataforma en la publicación del conjunto de tweets.
Actores:	Usuario
Precondiciones:	Seleccionar la opción “Búsquedas” seguido de la opción “Ver Resultados” y finalmente la opción “Menú Visualización”
Flujo Normal:	
1	El usuario da clic en la opción “Visualiza Source”.
Flujo Alternativo:	
2 ^a	
Poscondiciones:	El sistema despliega la gráfica y se muestra el nombre de la plataforma de publicación y el número de veces que se utilizó.

Diagrama de caso de uso 6: Visualizar gráfica por favoritos

Caso de Uso 6	Visualizar gráfica por favoritos
Descripción:	Se visualiza la gráfica por favoritos, la cual agrupa los tweets de acuerdo al número de veces que han sido marcados como favoritos.
Actores:	Usuario
Precondiciones:	Seleccionar la opción “Búsquedas” seguido de la opción “Ver Resultados” y finalmente la opción “Menú Visualización”
Flujo Normal:	
1	El usuario da clic en la opción “Visualiza Favoritos”.
Flujo Alternativo:	
2 ^a	
Poscondiciones:	El sistema despliega la gráfica y muestra el texto del tweet y el número de veces que ha sido marcado como favorito.

Diagrama de caso de uso 7: Visualizar gráfica por Retweets

Caso de Uso 7	Visualizar gráfica por retweets
Descripción:	Se visualiza la gráfica por reweets, la cual agrupa los tweets de acuerdo al número de veces que han sido retuiteados.
Actores:	Usuario
Precondiciones:	Seleccionar la opción “Búsquedas” seguido de la opción “Ver Resultados” y finalmente la opción “Menú Visualización”
Flujo Normal:	
1	El usuario da clic en la opción “Visualiza Retweets”.
Flujo Alternativo:	
2 ^a	
Poscondiciones:	El sistema despliega la gráfica y muestra el texto del tweet y el número de veces que ha sido marcado como retuiteado.

Diagrama de caso de uso 8: Visualizar gráfica por agrupamiento temático

Caso de Uso 8	Visualizar gráfica por agrupamiento temático
Descripción:	Se visualiza la gráfica por agrupamiento de temática similar, la cual muestra los grupos o clusters con las cinco palabras que predominan en el grupo. También se muestra en un rectángulo con las relación de palabras que existen entre grupos.
Actores:	Usuario
Precondiciones:	Seleccionar la opción “Búsquedas” seguido de la opción “Ver Resultados” y finalmente la opción “Menú Visualización”
Flujo Normal:	
1	El usuario da clic en la opción “Agrupamiento Temático”.
2	Seleccionar algoritmo de agrupamiento
3	Seleccionar tipo de peso
4	Seleccionar número de grupos a formar
5	Hacer clic en el botón “Graficar”
Flujo Alternativo:	
2 ^a	
Poscondiciones:	El sistema despliega la gráfica con los grupos que se pidieron así como las palabras que predominan en cada grupo y las palabras que se comparten entre grupos.

Diagrama de caso de uso 9: Agregar palabra vacía

Caso de Uso 9	Agregar palabra vacía
Descripción:	Se agrega una palabra denominada vacía (artículos, preposiciones, conjunciones, etc.) a la tabla de palabras, la cual será eliminada cuando aparezca en el texto del tweet. Esto con la finalidad de tener un texto con palabras significativas.
Actores:	Usuario
Precondiciones:	Seleccionar la opción “Configuración”
Flujo Normal:	
1	El usuario da clic en la opción “Agregar”.
2	Escribir la palabra
3	Guardar
Flujo Alternativo:	
2 ^a	
Poscondiciones:	El sistema da de alta la palabra en la base de datos.

Diagrama de caso de uso 10: Eliminar palabra vacía

Caso de Uso 10	Eliminar palabra vacía
Descripción:	Se elimina una palabra denominada vacía (artículos, preposiciones, conjunciones, etc.) a la tabla de palabras.
Actores:	Usuario
Precondiciones:	Seleccionar la opción “Configuración”
Flujo Normal:	
1	El usuario da clic en la opción “Eliminar”.
2	
3	
Flujo Alternativo:	
2 ^a	
Poscondiciones:	El sistema da de alta la palabra en la base de datos.

9.3 Anexo C: Diagramas de Secuencia

Los siguientes diagramas muestran la forma en cómo interactúan los objetos en cada caso de uso del sistema.

Diagrama de secuencia 1: Búsqueda de tweets

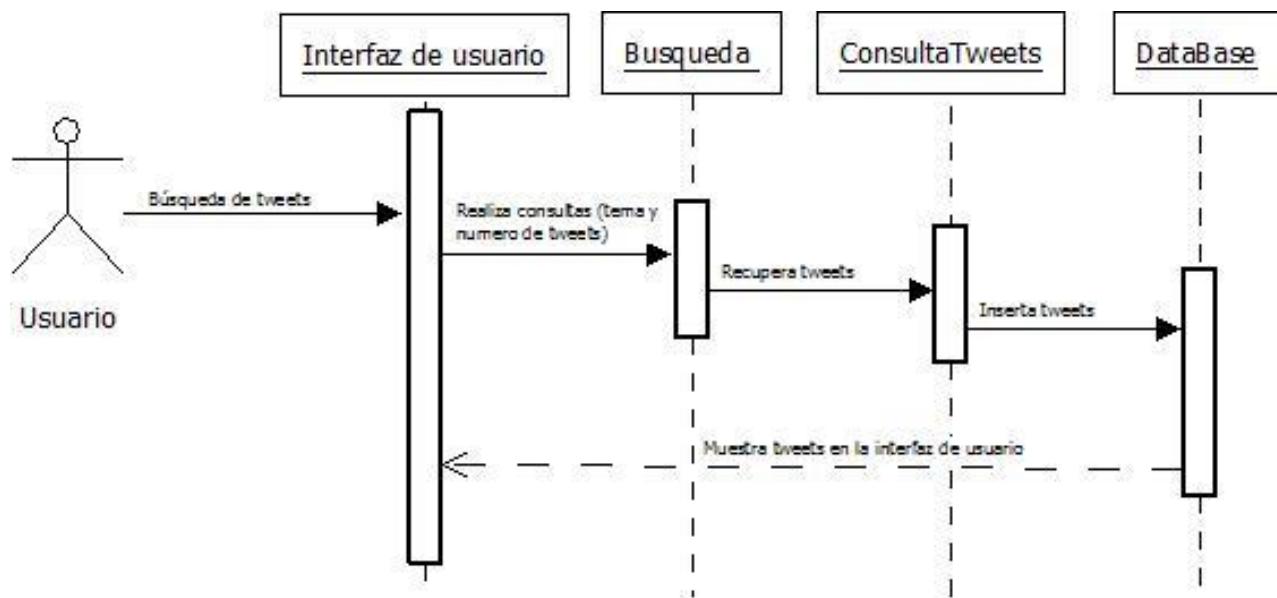


Diagrama de secuencia 2: Consultar búsquedas

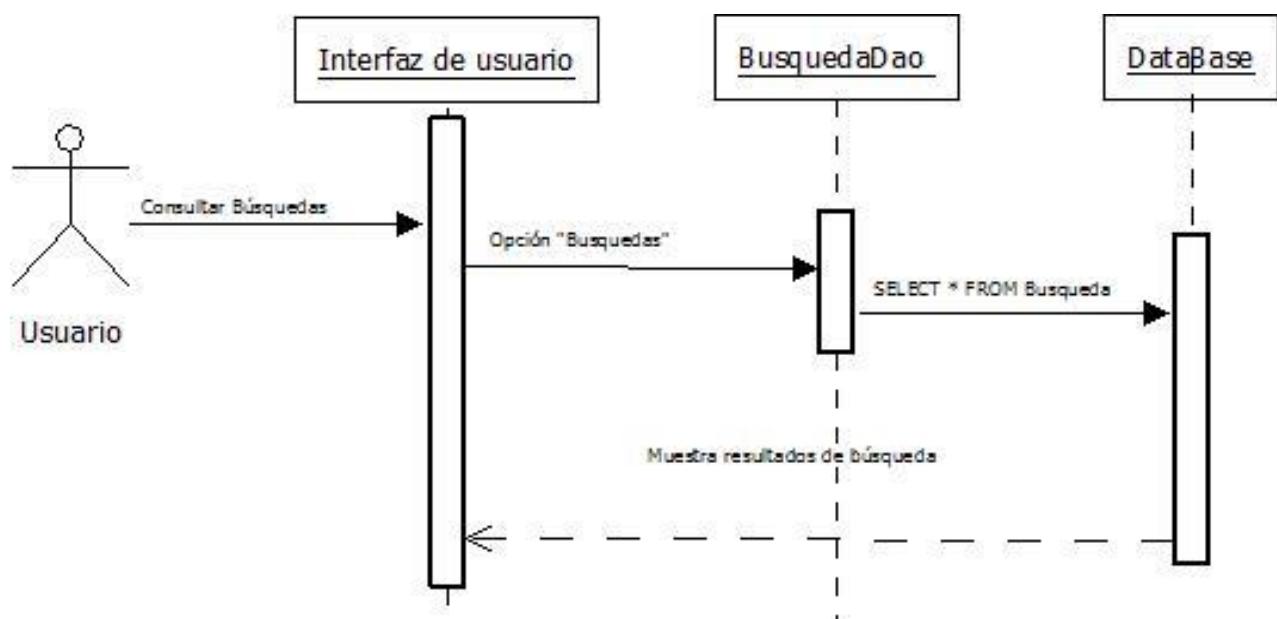


Diagrama de secuencia 3: Visualizar tweets

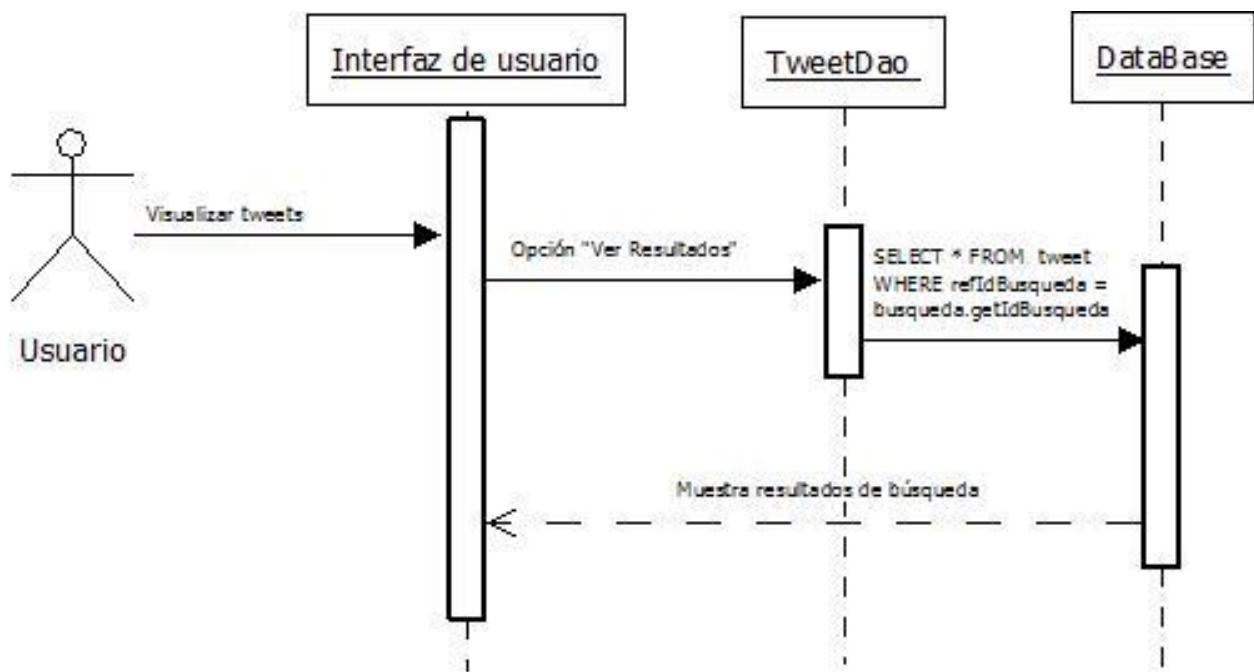


Diagrama de secuencia 4: Eliminar búsqueda

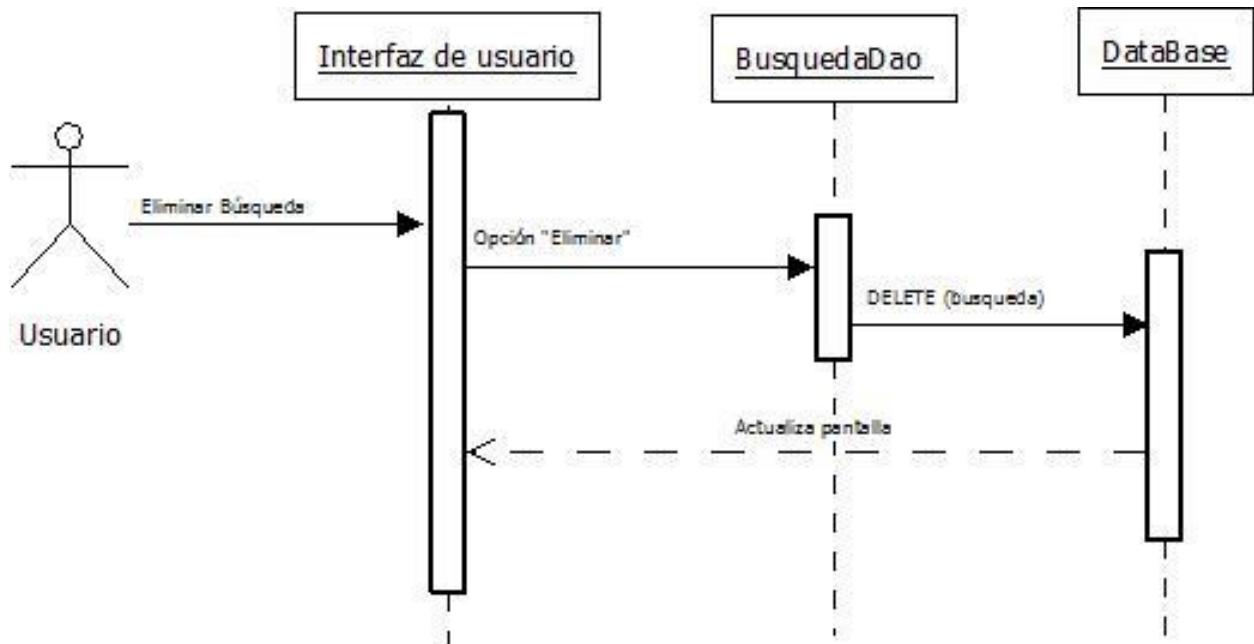


Diagrama de secuencia 5: Visualizar gráfica por Source

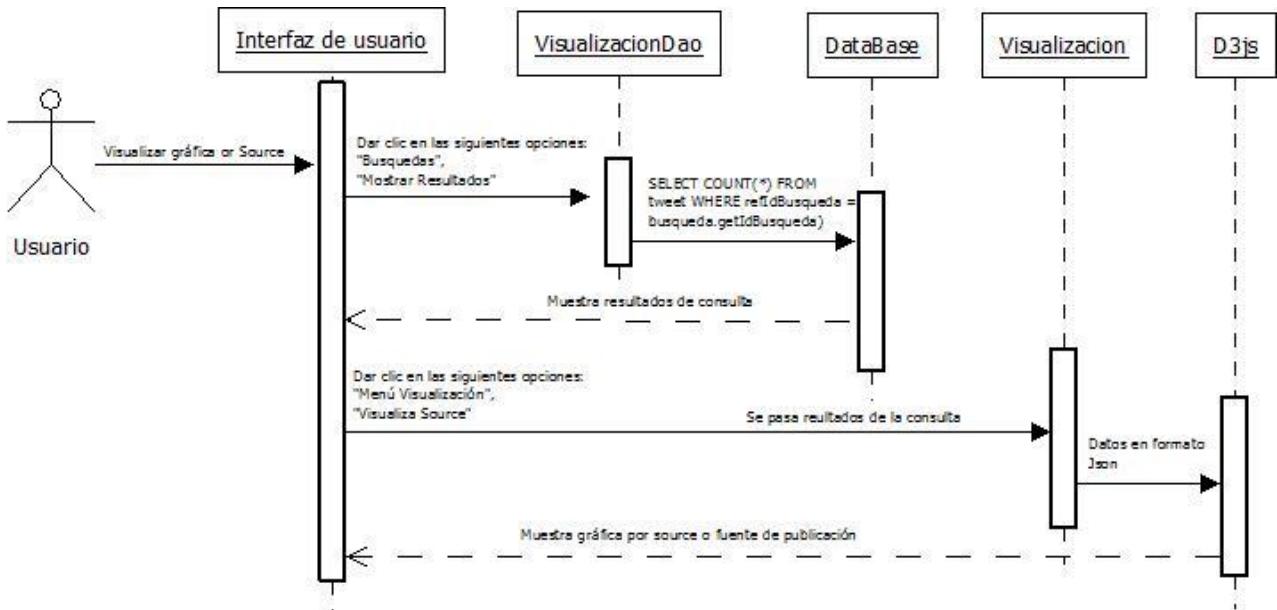


Diagrama de secuencia 6: Visualizar gráfica por Favoritos

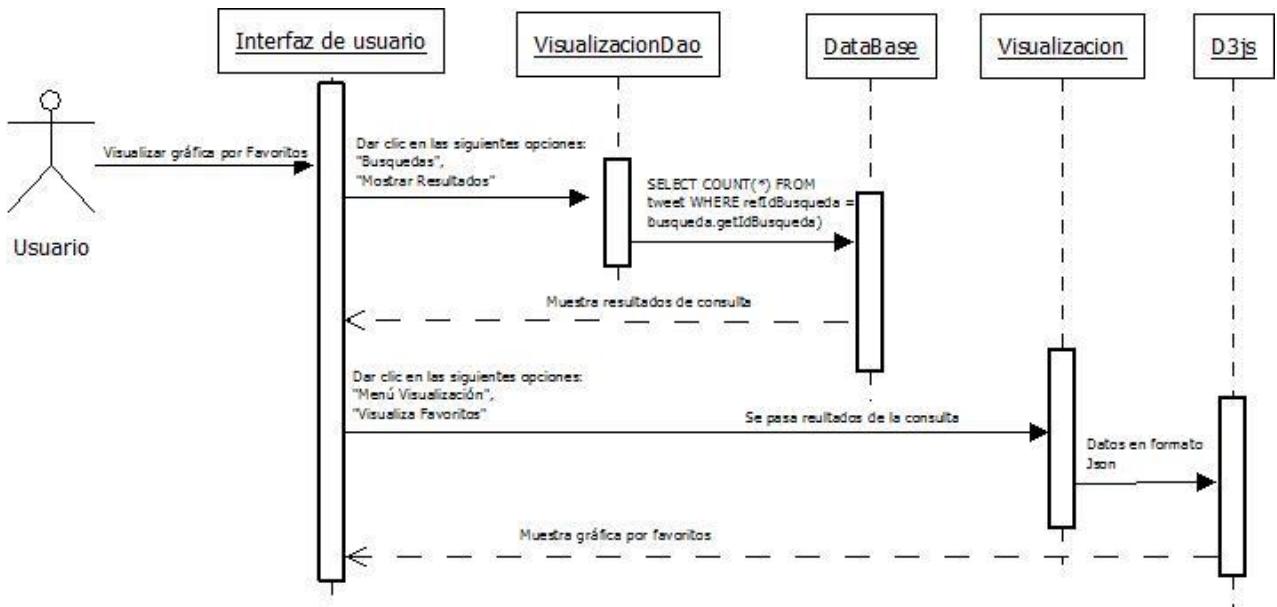


Diagrama de secuencia 7: Visualizar gráfica por Retweets

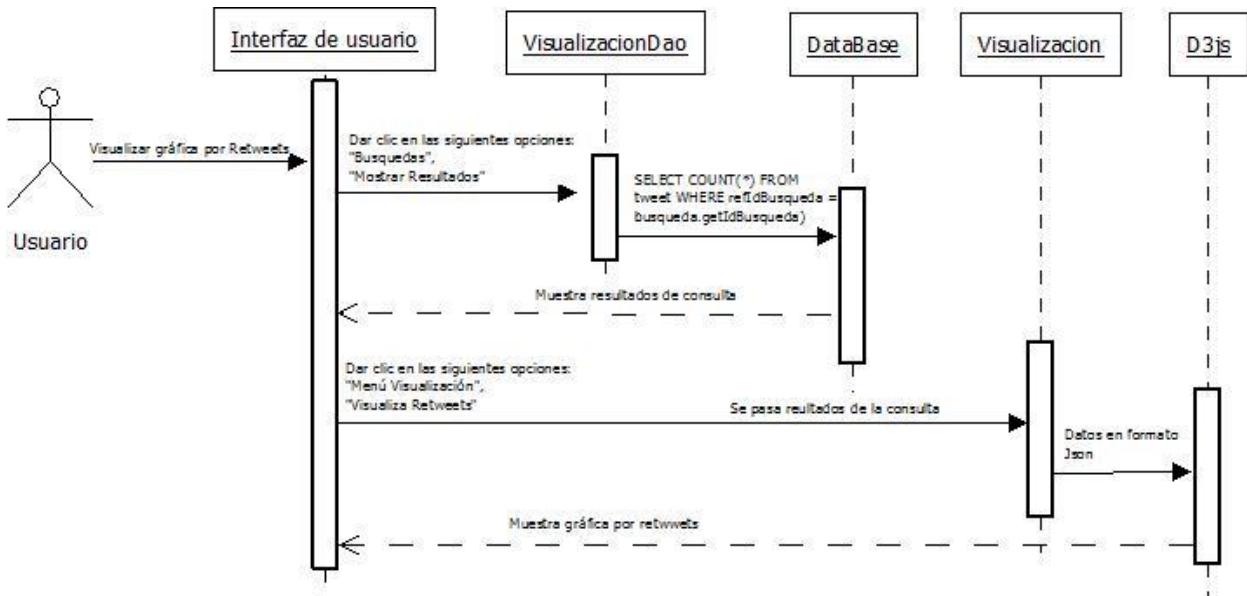


Diagrama de secuencia 8: Visualizar gráfica por Temática Similar

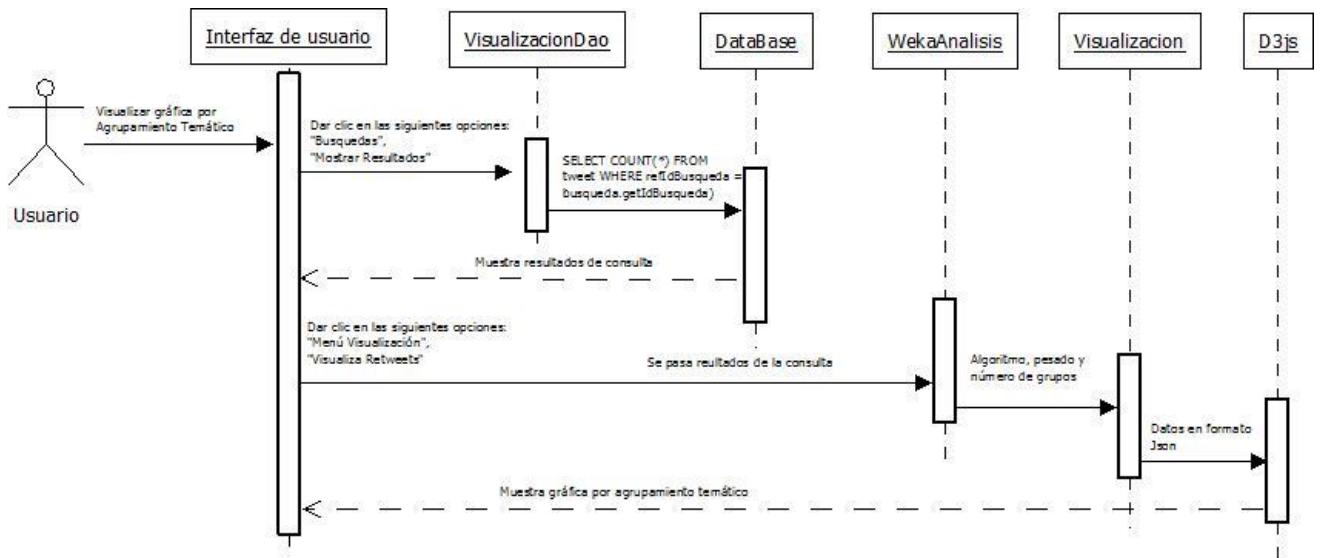


Diagrama de secuencia 9: Agregar palabra vacía

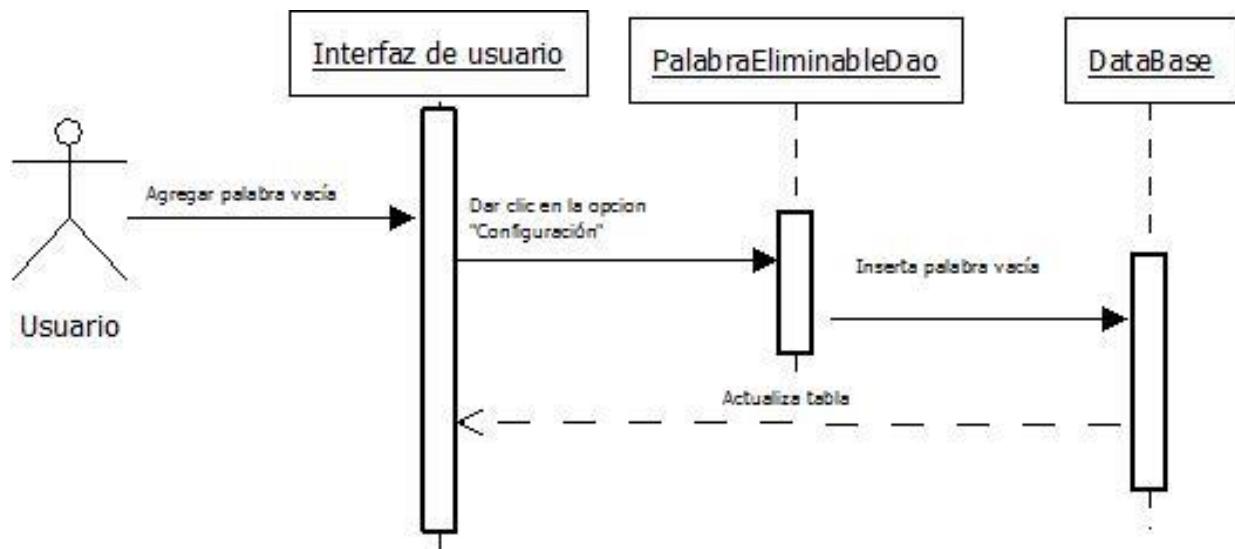
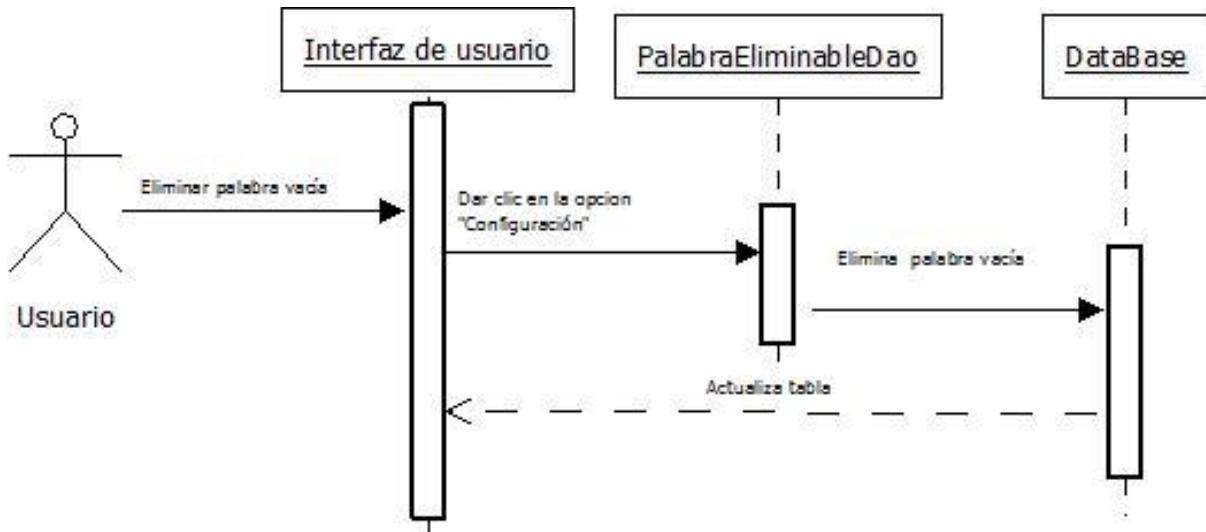
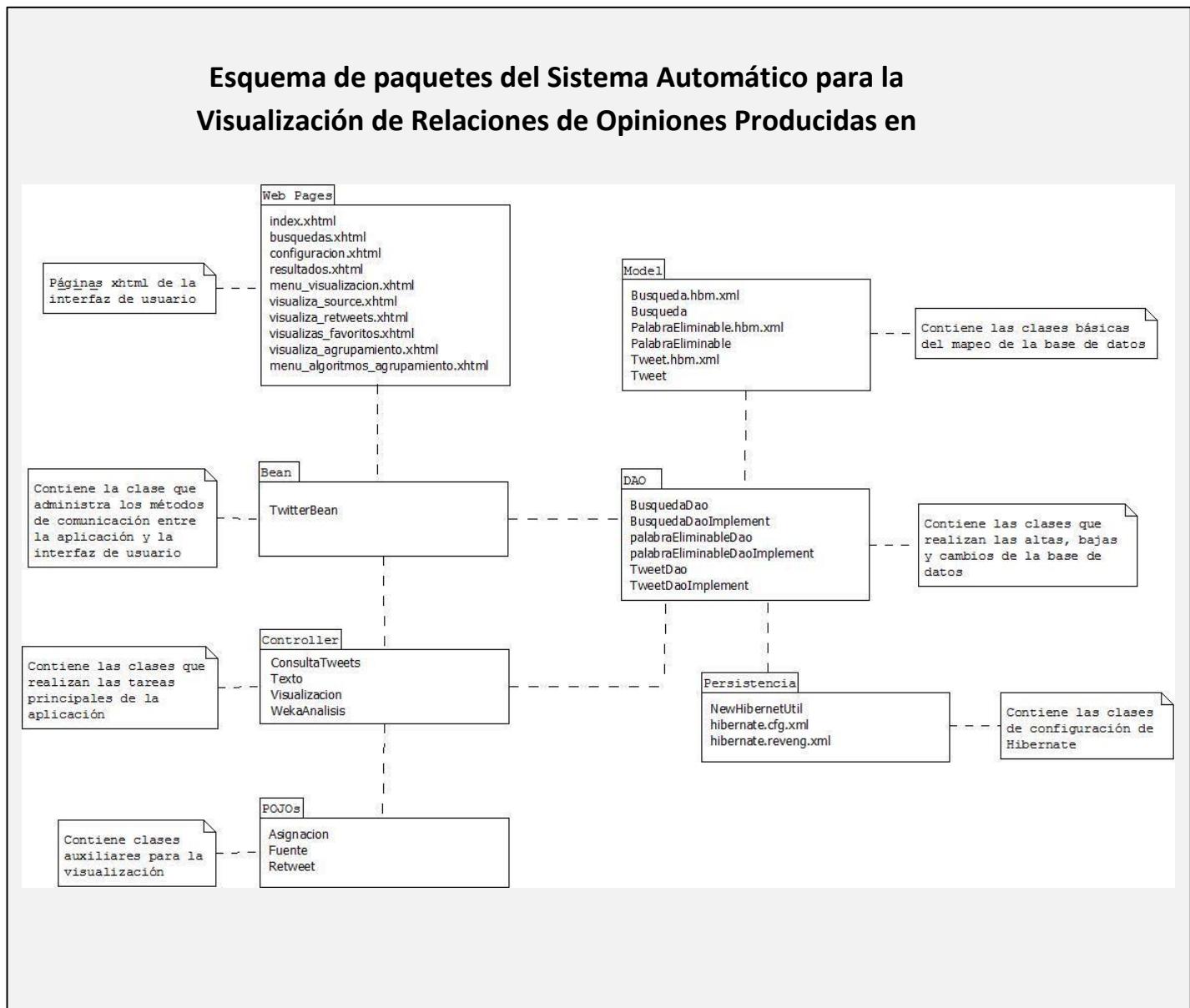


Diagrama de secuencia 10: Eliminar palabra vacía



9.4 Anexo D: Diagramas de Clases

En este anexo se describen los paquetes y las clases que se implementaron en el desarrollo del sistema.



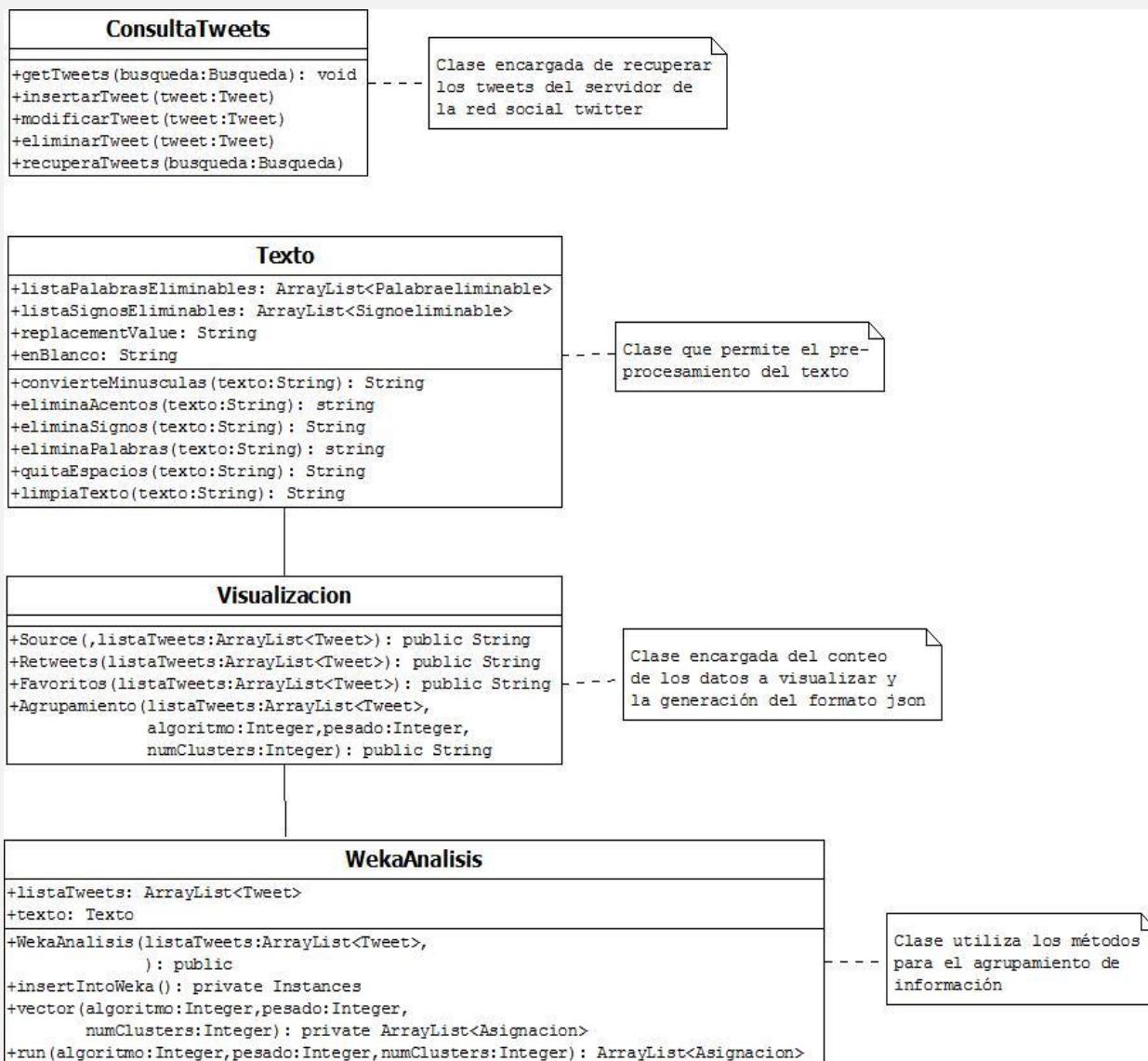
Package Bean

TwitterBean

```
+listaBusquedas: private ArrayList<Busqueda>
+listaPalabrasEliminables: private ArrayList<Palabraeliminable>
+listaTweets: private ArrayList<Tweet>
+busqueda: private Busqueda
+palabraEliminable: private Palabraeliminable
+tweet: private Tweet
+numClusters: private Integer
+algoritmo: private Integer
+pesado: private Integer
+init(): public void
+btnBuscar(): public String
+btnBusquedas(): public String
+btnConfiguracion(): public String
+btnNuevaBusqueda(): public String
+btnVerResultados(): public String
+btnEliminarBusqueda(): public void
+btnMenuVisualizacion(): public void
+btnVisualizaSource(): public String
+btnVisualizaFavoritos(): public String
+btnVisualizaRetweets(): btnVisualizaRetweets
+btnAgrupamientoTematico(): public String
+btnVisualizaAgrupamiento(): public String
+insertarBusqueda(): public void
+eliminarBusqueda(): public void
+insertarPalabraEliminable(): public void
+eliminarPalabraEliminable(): public void
+insertarTweet(): public void
+eliminarTweet(): public void
+Getter and Setter()
+()
```

Métodos de comunicación entre la aplicación y la interfaz de usuario

Package Controller



Package Model

Busqueda	
<pre>+idBusqueda: Integer +temaConsultar: String +cantidadDeTweets: Integer +fecha: Date +hora: Date +tweets: Set<Tweet> +Busqueda(): (String temaConsultar, Integer cantidadDeTweets, Date fecha, Date hora, Set<Tweet> tweets)</pre>	<p>Clase básica que permite el mapeo con la tabla búsqueda de la base de datos</p>
Palabraeliminable	
<pre>+idEliminarPalabra: private Integer +palabra: private String +getIdEliminarPalabra(): public Integer +setIdEliminarPalabra(idEliminarPalabra:Integer): public void +getPalabra(): public String +setPalabra(palabra:String): public void +toString(): String</pre>	<p>Clase básica que permite el mapeo con la tabla Palabraeliminable de la base de datos</p>
Tweet	
<pre>+localId: Integer +busqueda: Busqueda +createdAt: Date +favoriteCount: Integer +retweetCount: Integer +tweetId: Long +userId: Long +screenName: String +source: String +text: String +textModif: String +tweetLanguage: String +userDescription: String +userLanguage: String +userLocation: String +userName: String +originalProfileImageUrl: String +Tweet(busqueda:Busqueda,createdAt:Date, favoriteCount:Integer,retweetCount:Integer, tweetId:Long,userId:Long,screenName:String, source:String,,text:String,,textModif:String, tweetLanguage:String,userDescription:String, userLanguage:String,userLocation:String, userName:String,originalProfileImageUrl:String): public +toString(): String</pre>	<p>Clase básica que permite el mapeo con la tabla Tweet de la base de datos</p>

Package Dao

BusquedaDao
+mostrarBusquedas (): ArrayList<Busqueda>
+insertarBusqueda (busqueda:Busqueda): void
+eliminarBusqueda (busqueda:Busqueda): void

Interfaz que permite las altas,
bajas y cambios de la tabla
Busqueda de la base de datos

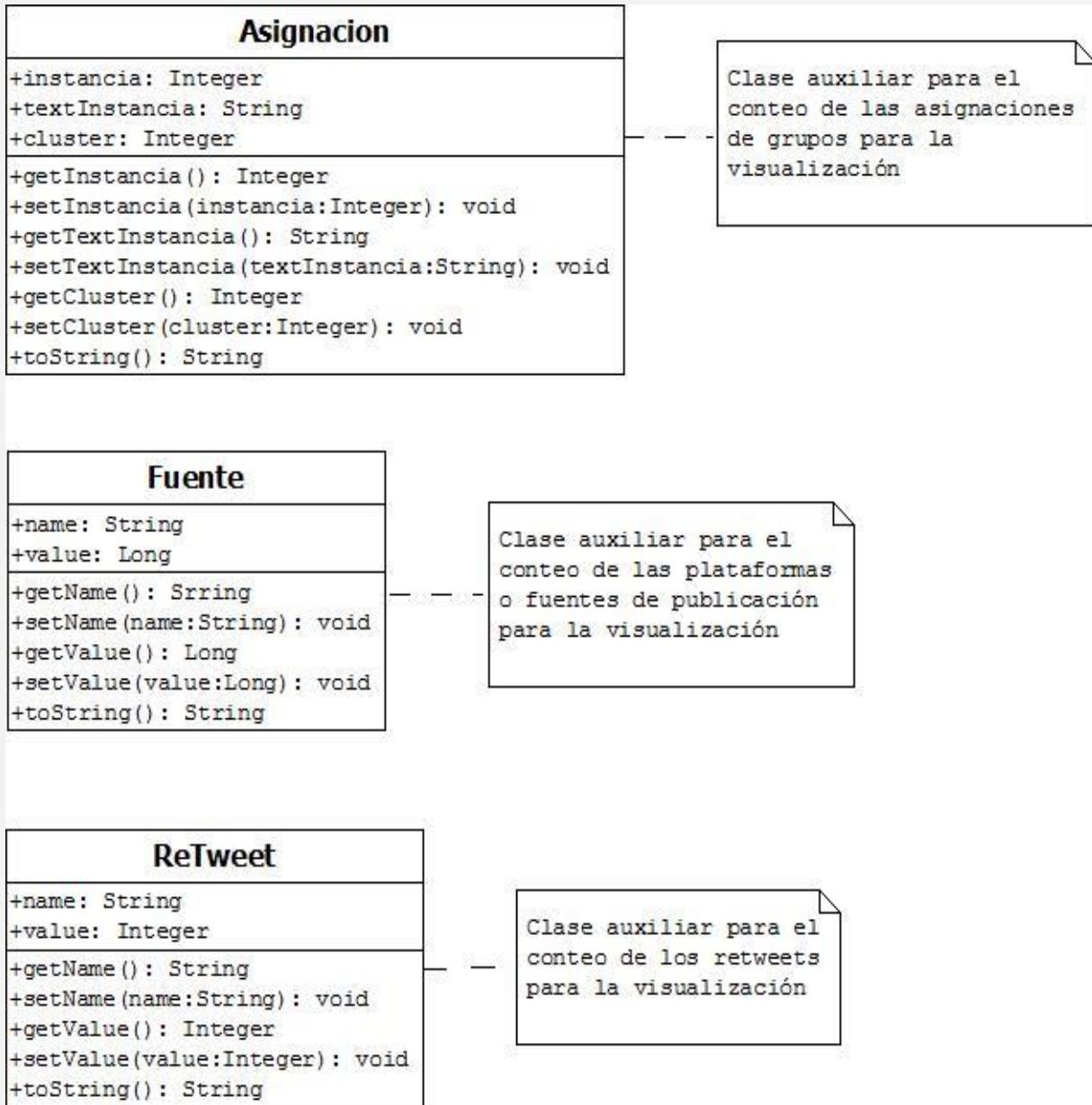
PalabraEliminableDao
+mostrarPalabrasEliminables (): ArrayList<Palabraeliminable>
+insertarPalabraEliminable (palabraEliminable:Palabraeliminable): void
+eliminarPalabraEliminable (palabraEliminable:Palabraeliminable): void

Interfaz que permite las altas,
bajas y cambios de la tabla
Palabraeliminable

TweetDao
+mostrarTweets (busqueda:Busqueda) : ArrayList<Tweet>
+insertarTweet (tweet:Tweet): void
+eliminarTweet (tweet:Tweet): void

Interfaz que permite las altas,
bajas y cambios de la tabla
Tweet de la base de datos

Package POJOs



9.5 Anexo E: Diseño de la Base de Datos

En este anexo se muestra el esquema de la base de datos y se describe cada una de las tablas.

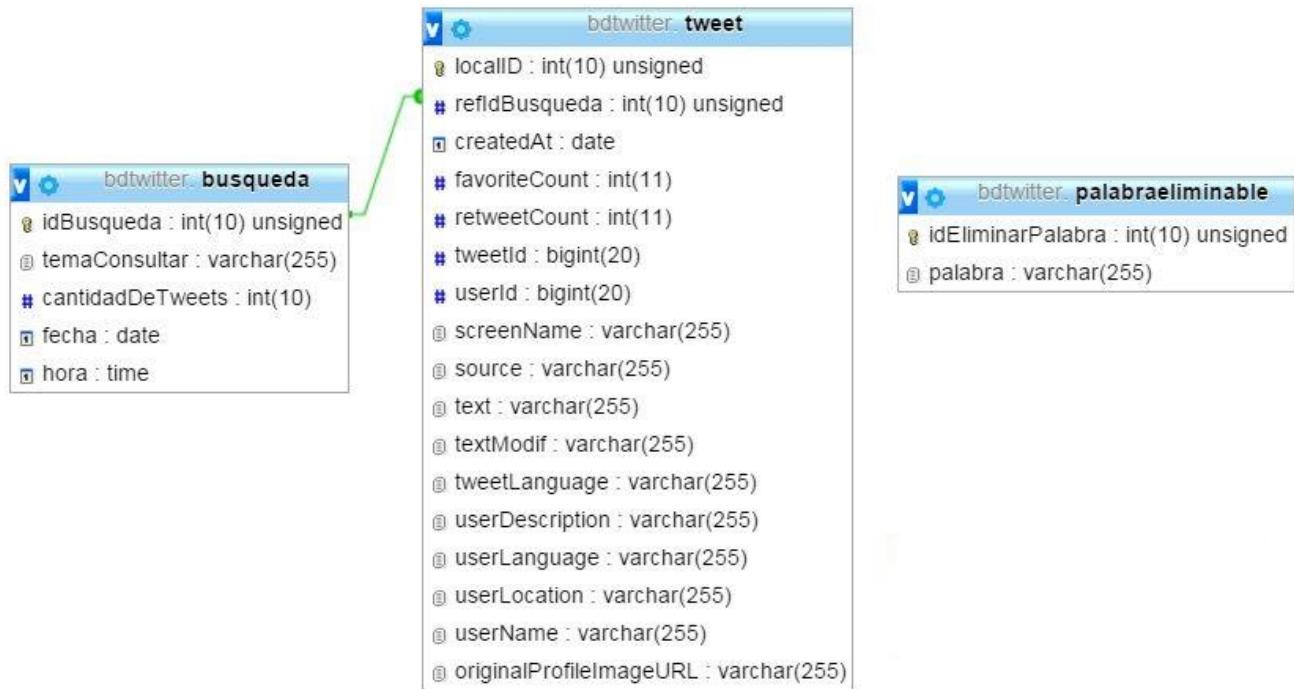


Tabla busqueda

Campo	Tipo	Llave Primaria	Descripción	Llave Foránea
<code>idBusqueda</code>	Int	PK	Id de la búsqueda	
<code>temaConsultar</code>	varchar		Tema a consultar	
<code>cantidadDeTweets</code>	Int		Cantidad de tweets a recuperar	
<code>Fecha</code>	Date		Fecha en que se realiza la consulta aaaa/mm/dd	
<code>Hora</code>	Time		Hora en que se realiza la consulta hh:mm:ss	

Tabla tweet

Campo	Tipo	Llave Primaria	Descripción	Llave Foránea
localID	Int	PK	Id que el sistema asigna al registro en la tabla	
refIdBusqueda	Int		Id de la búsqueda	PF
createdAt	Date		Fecha en que se publicó el tweet	
favoriteCount	Int		Número de veces que el tweet ha sido marcado como favorito	
retweetCount	Int		Número de veces que el tweet ha sido retuiteado por otros usuarios	
tweetId	bigint		Id del tweet, es una serie de números que asigna Twitter al momento en que se publica el tweet.	
userId	Bigint		Id del usuario, es una serie de números que se asigna al usuario cuando es dado de alta en Twitter	
screenName	Varchar		Usuario o alias con que se registra el usuario en Twitter	
Source	Varchar		Es la plataforma desde donde se publica el tweet.	
Text	Varchar		Es el texto del tweet (texto plano de una longitud máxima de 140 caracteres)	
TextModif	Varchar		Es el texto del tweet preprocessado, es decir, se han eliminado las palabras vacías (artículos, preposiciones, conjunciones, etc.)	
tweetLanguage	Varchar		Es el lenguaje en el que está publicado el tweet	
UserDescription	Varchar		Es un texto que el usuario quiere mostrar en su perfil	
UserLanguage	Varchar		Es el lenguaje que el usuario registra al momento de darse de alta en Twitter	
UserLocation	Varchar		País o ciudad donde reside el usuario	
userName	Varchar		Nombre completo del usuario	
originalProfileImageURL	Varchar		url de la imagen o fotografía que el usuario tiene en su cuenta de Twitter	

Tabla palabraeliminable

Campo	Tipo	Llave Primaria	Descripción	Llave Foránea
idEliminarPalabra	Int varchar	PK	Id de la palabra vacía Palabra o caracteres a eliminar	

9.6 Anexo F: Manual de Usuario

El contenido de este manual considera lo siguiente:

- Introducción
 - Página principal
 - Realizar una búsqueda
 - Consultar una búsqueda
 - Mostrar tweets de una búsqueda en particular
 - Visualizar menú de gráficas
 - Visualizar gráfica por Source o plataforma
 - Visualizar gráfica por Favoritos
 - Visualizar gráfica por Retweets
 - Visualizar gráfica por Agrupamiento Temático
 - Agregar y eliminar palabras vacías
-

Introducción

En este manual se describirán los pasos necesarios que se requieren para utilizar el Sistema Automático para la Visualización de Relaciones de Opiniones producidas en Twitter.

El sistema fue creado con el objetivo de brindar una herramienta de apoyo en la toma de decisiones a personas dedicadas al análisis de información que se genera en Twitter, o a personas que tengan inquietud en conocer lo que se dice de algún tema en específico en esta red social.

Para esto, el sistema forma grupos de tweets, los cuales comparten una temática similar y muestra las palabras que describen a los grupos así como las palabras que comunes entre grupos.

La información es mostrada al usuario de forma gráfica para facilitar su interpretación, es decir, que el usuario tenga una perspectiva general del contenido temático que prevalece en la colección de tweets.

Se incluyen impresiones de pantalla del sistema para apoyar al usuario en la interacción con el mismo.

Página principal

La página principal es la que despliega el sistema al momento de abrirlo, como se muestra en la figura E.1. De inicio la página tiene activada la opción de “Nueva Búsqueda” en donde el usuario podrá realizar la búsqueda de tweets.

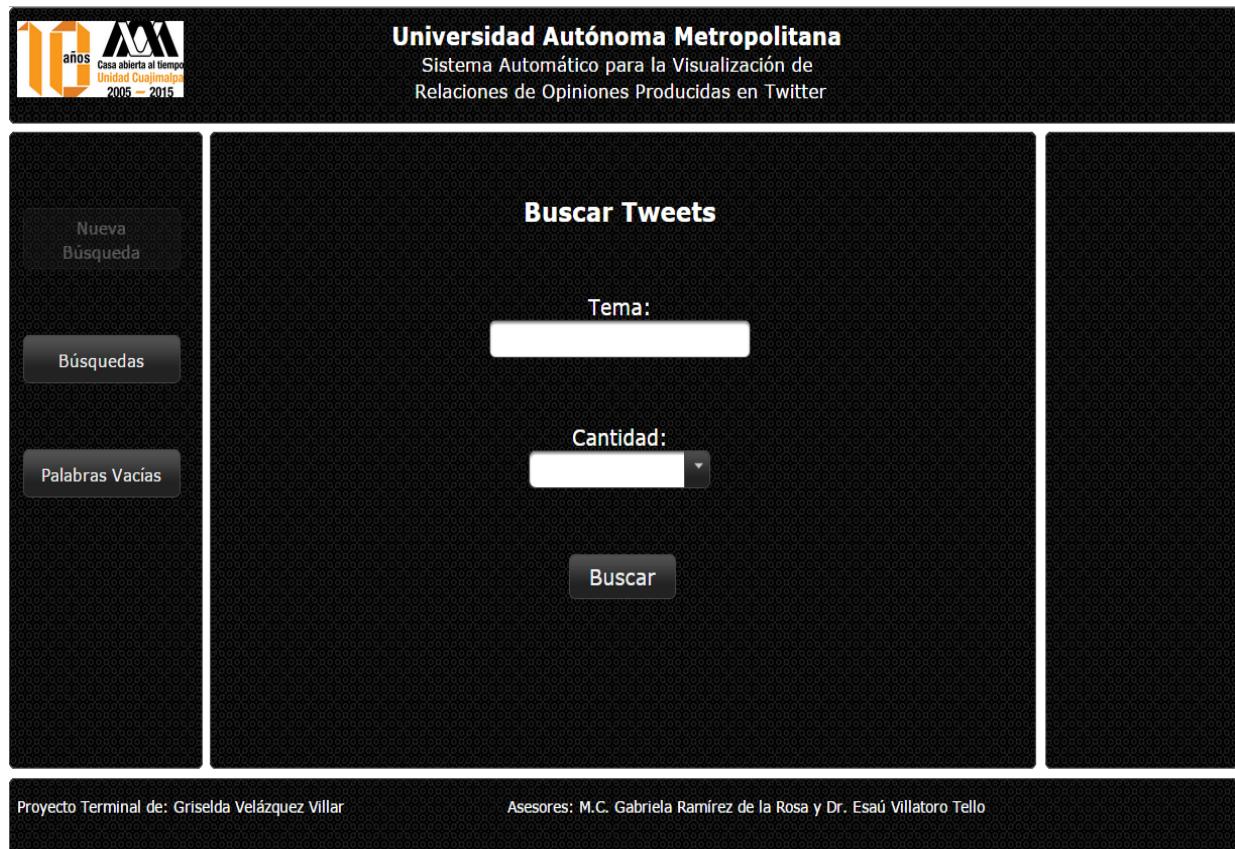


Figura E.1 Página de inicio del sistema.

Realizar una búsqueda y recuperación de tweets

Ir al rectángulo de tema y escribir el tema a buscar, se puede aplicar la estrategia de búsquedas⁶. En el espacio de cantidad, el sistema sugiere seis cantidades: diez, veinte, cincuenta, cien, doscientos y quinientos, de las que se puede seleccionar alguna. Si no se desea seleccionar ninguna de las cantidades sugeridas, el usuario podrá escribir la cantidad de tweets que desea recuperar, como lo muestra la figura E.2

The screenshot shows a search interface for tweets. At the top, there is a logo for the Universidad Autónoma Metropolitana (UAM) Cuajimalpa unit, which is celebrating its 10th anniversary (2005-2015). Below the logo, the text reads: "Universidad Autónoma Metropolitana Sistema Automático para la Visualización de Relaciones de Opiniones Producidas en Twitter". On the left side of the interface, there is a sidebar with three buttons: "Nueva Búsqueda", "Búsquedas", and "Palabras Vacías". The main area is titled "Buscar Tweets" and contains two input fields: "Tema:" with the value "uam cuajimalpa" and "Cantidad:" with the value "50". A blue "Buscar" button is located below these fields. At the bottom of the interface, there is footer text: "Proyecto Terminal de: Griselda Velázquez Villar" and "Asesores: M.C. Gabriela Ramírez de la Rosa y Dr. Esau Villatoro Tello".

Figura E.2 Página de búsqueda y recuperación de tweets.

Dar clic en el botón “Buscar” y comienza la búsqueda.

El sistema despliega, en una tabla, los tweets que se recuperaron, como se muestra en la figura E.3

⁶ Una estrategia de búsqueda se define como el conjunto de procedimientos y operaciones que un usuario realiza con el fin de obtener la información lo más exacto a sus necesidades.

Universidad Autónoma Metropolitana
Sistema Automático para la Visualización de
Relaciones de Opiniones Producidas en Twitter

#	Text	Created At
1	Inauguración d la Exposición Gráfica d Catedráticos, Intercambio cultural ntre la DERNyAP d #Cuajimalpa y UAM Santa Fe http://t.co/q6WYQ0d44e	2015-06-28
2	RT @hojaderutamexic: Celebrará @uamcuajimalpa primera década con dos conciertos del Cepromusic @Conaculta @UAM_Comunidad http://t.co/sQWs6H...	2015-06-28
3	4. Me gusta más la uam cuajimalpa que les acatan.	2015-06-28
4	#Cultura:Celebrará UAM-Cuajimalpa primera década con dos conciertos del #Cepromusic http://t.co/oOPNsQhwX1	2015-06-27
5	#Cultura:Celebrará UAM-Cuajimalpa primera década con dos conciertos del #Cepromusic http://t.co/oOPNsQhwX1	2015-06-26
6	RT @elgolfoveracruz #Cultura:Celebrará UAM-Cuajimalpa primera década con dos conciertos del #Cepromusic http://t.co/J2jlzgMoxf	2015-06-26
7	#Cultura:Celebrará UAM-Cuajimalpa primera década con dos conciertos del #Cepromusic http://t.co/oOPNsQhwX1	2015-06-26
8	RT @hojaderutamexic: Celebrará @uamcuajimalpa primera década con dos conciertos del Cepromusic @Conaculta @UAM_Comunidad http://t.co/sQWs6H...	2015-06-26
9	La Unidad Cuajimalpa de la #UAM realiza la primera edición del Libro tránsito 2015. http://t.co/S8fQObvBge	2015-06-26
10	Celebrará UAM-Cuajimalpa primera década con dos conciertos del Cepromusic http://t.co/wsPuaTILu	2015-06-26
11	Celebrará UAM-Cuajimalpa primera década con dos conciertos del Cepromusic http://t.co/1MIZxTn49e	2015-06-26
12	Celebrará UAM-Cuajimalpa primera década con dos conciertos del Cepromusic http://t.co/DAKOaAqCgu	2015-06-26
13	Celebrará @uamcuajimalpa primera década con dos conciertos del Cepromusic @Conaculta @UAM_Comunidad http://t.co/sQWs6HEBlU	2015-06-25
14	RT @Redes_Morena: La unidad Cuajimalpa de la UAM realiza la primera edición del libro tránsito 2015 http://t.co/ciuQdyeMNx	2015-06-25

20 en total.

Proyecto Terminal de: Griselda Velázquez Villar Asesores: M.C. Gabriela Ramírez de la Rosa y Dr. Esaú Villatoro Tello

Figura E.3 Tweets recuperados.

De entrada se despliegan solo los campos: id o número consecutivo, el texto y la fecha de creación de cada tweet como se observa en la figura E.3.

Al hacer clic en el círculo localizado en el lado izquierdo los tweets, figura E.4, se despliega todos los campos que conforman el registro de cada tweet, como se observa en la figura E.5.

#	Text	Created At
1	Inauguración d la Exposición Gráfica d Catedráticos, Intercambio cultural ntre la DERNyAP d #Cuajimalpa y UAM Santa Fe http://t.co/q6WYQ0d44e	2015-06-28
2	RT @hojaderutamexic: Celebrará @uamcuajimalpa primera década con dos conciertos del Cepromusic @Conaculta @UAM_Comunidad http://t.co/sQWs6H...	2015-06-28
3	4. Me gusta más la uam cuajimalpa que les acatan.	2015-06-28
4	#Cultura:Celebrará UAM-Cuajimalpa primera década con dos conciertos del #Cepromusic http://t.co/oOPNsQhwX1	2015-06-27
5	#Cultura:Celebrará UAM-Cuajimalpa primera década con dos conciertos del #Cepromusic http://t.co/oOPNsQhwX1	2015-06-26
6	RT @elgolfoveracruz #Cultura:Celebrará UAM-Cuajimalpa primera década con dos conciertos del #Cepromusic http://t.co/J2jlzgMoxf	2015-06-26

Figura E.4 Opción para desplegar información completa del tweet.

#	Text	Created At
1	Inauguración d la Exposición Gráfica d Catedráticos, Intercambio cultural ntre la DERNyAP d #Cuajimalpa y UAM Santa Fe http://t.co/q6WyQ0d44e	2015-06-28
	Tweet Id: 615297210712875008 User Id: 1227956420 Screen Name: ALEJANDRO_S_C User Name: DesiertoDeLosLeones User Description: Desierto de los Leones Parque Recreativo y Cultural #DelegaciónCuajimalpa Cuajimalpa #CDMX #CompromisoQueSeSienteUnidosParaProgresar J.U.D User Language: es User Location: Tweet Language: es Source: Twitter for Android Favorite Count: 0 Retweet Count: 0 OPI-URL: http://pbs.twimg.com/profile_images/555579634314915840/lFvc7U	
2	RT @hojaderutamexic: Celebrará @uamcuajimalpa primera década con dos conciertos del Cepromusic @Conaculta @UAM_Comunidad http://t.co/sQWs6H...	2015-06-28
3	4. Me gusta más la uam cuajimalpa que fes acatlan.	2015-06-28
4	#Cultura: Celebrará UAM-Cuajimalpa primera década con dos conciertos del #Cepromusic http://t.co/oOPNsQhwX1	2015-06-27
20 en total.		

Proyecto Terminal de: Griselda Velázquez Villar Asesores: M.C. Gabriela Ramírez de la Rosa y Dr. Esaú Villatoro Tello

Figura E.5 Página que muestra el registro completo del tweet.

A continuación se describen los campos que integran el registro de un tweet:

Campo	Descripción
localID	Id que el sistema asigna a cada registro (tweet).
refIdBusqueda	Id de la búsqueda al que pertenece el tweet.
createdAt	Fecha en que se publicó el tweet.
favoriteCount	Número de veces que el tweet ha sido marcado como favorito.
retweetCount	Número de veces que el tweet ha sido retuiteado por otros usuarios.
tweetId	Id del tweet, es una serie de números que asigna

	Twitter al momento en que se publica el tweet.
userId	Id del usuario, es una serie de números que se asigna al usuario cuando es dado de alta en Twitter.
screenName	Usuario o alias con que se registra el usuario en Twitter.
Source	Es la plataforma desde donde se publica el tweet.
Text	Es el texto del tweet (texto plano de una longitud máxima de 140 caracteres).
TextModif	Es el texto del tweet preprocesado, es decir, se han eliminado las palabras vacías (artículos, preposiciones, conjunciones, etc.).
tweetLanguage	Es el lenguaje en el que está publicado el tweet.
UserDesctrion	Es un texto que el usuario quiere mostrar en su perfil.
UserLanguage	Es el lenguaje que el usuario registra al momento de darse de alta en Twitter.
UserLocation	País o ciudad donde reside el usuario.
userName	Nombre completo del usuario.
originalProfileImageURL	url de la imagen o fotografía que el usuario tiene en su cuenta de Twitter.

Consultar una búsqueda

Al hacer clic en el botón “Búsquedas”, el sistema despliega la lista de las búsquedas o consultas que se han realizado y que están almacenadas dentro la base de datos, como se muestra en la figura E.6.

The screenshot shows a web application interface for managing search queries. At the top, there is a header with the logo of the Universidad Autónoma Metropolitana (UAM) and the text "Universidad Autónoma Metropolitana" and "Sistema Automático para la Visualización de Relaciones de Opiniones Producidas en Twitter". On the left side, there are three buttons: "Nueva Búsqueda", "Búsquedas", and "Palabras Vacías". The main content area is titled "Búsquedas Realizadas" and contains a table with the following data:

#	Tema	Tweets	Fecha	Hora
1	Queretaro	80	2015-05-24	23:05:26
2	trump	2000	2015-06-17	00:38:29
3	jeb bush	5000	2015-06-17	00:44:00
4	uam cuajimalpa	20	2015-06-28	18:07:47
5	reyes de españa	500	2015-07-01	01:22:16
6	chapo	1000	2015-07-13	21:24:54
7	chapo	3000	2015-07-13	21:37:22

Below the table, a message states "7 búsquedas en total.". To the right of the table, there are two buttons: "Ver Tweets" and "Eliminar Búsqueda". At the bottom of the page, there are credits: "Proyecto Terminal de: Griselda Velázquez Villar" and "Asesores: M.C. Gabriela Ramírez de la Rosa y Dr. Esaú Villatoro Tello".

Figura E.6 Página que lista las consultas almacenadas en la base de datos.

El sistema guarda las búsquedas que el usuario realiza, se crea un registro por cada consulta, el cual contiene los siguientes campos:

Campo	Descripción
idBusqueda	Id que el sistema asigna a cada búsqueda.
temaConsultar	Tema que el usuario seleccionó para la búsqueda.
cantidadDeTweets	Cantidad de tweets que el usuario recuperó.
Fecha	Fecha en que se realizó la búsqueda aaaa/mm/dd
Hora	Hora en que se realizó la búsqueda hh:mm:ss

Mostrar tweets de una búsqueda en particular

Estando dentro de la página de consultar búsqueda, se selecciona una búsqueda de la lista de búsquedas y se da clic en el botón “Ver Tweets”. Figura E.7.

The screenshot shows a search interface for the Universidad Autónoma Metropolitana. At the top, there's a banner for '10 años UAM Casa abierta al Tiempo Unidad Cuajimalpa 2005 - 2015'. On the left, there are three buttons: 'Nueva Búsqueda', 'Búsquedas', and 'Palabras Vacías'. In the center, a table titled 'Búsquedas Realizadas' lists seven search entries with columns for '#', 'Tema', 'Tweets', 'Fecha', and 'Hora'. The last entry, '#4 uam cuajimalpa', has a blue link. To the right of the table is a large button labeled 'Ver Tweets' with a red arrow pointing to it. Below this button is another button labeled 'Eliminar Búsqueda'.

#	Tema	Tweets	Fecha	Hora
1	Queretaro	80	2015-05-24	23:05:26
2	trump	2000	2015-06-17	00:38:29
3	jeb bush	5000	2015-06-17	00:44:00
4	uam cuajimalpa	20	2015-06-28	18:07:47
5	reyes de españa	500	2015-07-01	01:22:16
6	chapo	1000	2015-07-13	21:24:54
7	chapo	3000	2015-07-13	21:37:22

Figura E.7 Página que muestra el botón para visualizar los tweets.

El sistema despliega los tweets que se recuperaron en dicha búsqueda. Como se muestra en la figura E.8

The screenshot shows the same search interface as Figure E.7, but now displaying the results for the selected search query '#4 uam cuajimalpa'. The central area is a table titled 'Tweets' with columns for '#', 'Text', and 'Created At'. It lists 14 tweets from various users (@hojaderutamexic, @UAM_Comunidad, @Redes_Morena, etc.) with their respective URLs and creation dates. Below the table, it says '20 en total.' On the right side, there's a 'Menú Visualización' button. At the bottom, there are footer links for 'Proyecto Terminal de: Griselda Velázquez Villar' and 'Asesores: M.C. Gabriela Ramírez de la Rosa y Dr. Esaú Villatoro Tello'.

#	Text	Created At
1	Inauguración d la Exposición Gráfica d Catedráticos,Intercambio cultural ntre la DERNyAP d #Cuajimalpa y UAM Santa Fe http://t.co/q6WYQ0d44e	2015-06-28
2	RT @hojaderutamexic: Celebrará @uamcuajimalpa primera década con dos conciertos del Cepromusic @Conaculta @UAM_Comunidad http://t.co/sQWs6H...	2015-06-28
3	4. Me gusta más la uam cuajimalpa que les acatan.	2015-06-28
4	#Cultura:Celebrará UAM-Cuajimalpa primera década con dos conciertos del #Cepromusic http://t.co/oOPNsQhwX1	2015-06-27
5	#Cultura:Celebrará UAM-Cuajimalpa primera década con dos conciertos del #Cepromusic http://t.co/oOPNsQhwX1	2015-06-26
6	RT @elgolfoveracruz #Cultura:Celebrará UAM-Cuajimalpa primera década con dos conciertos del #Cepromusic http://t.co/12jzqMoxf	2015-06-26
7	#Cultura:Celebrará UAM-Cuajimalpa primera década con dos conciertos del #Cepromusic http://t.co/oOPNsQhwX1	2015-06-26
8	RT @hojaderutamexic: Celebrará @uamcuajimalpa primera década con dos conciertos del Cepromusic @Conaculta @UAM_Comunidad http://t.co/sQWs6H...	2015-06-26
9	La Unidad Cuajimalpa de la #UAM realiza la primera edición del Libro tránsito 2015, http://t.co/S8fQObvBge	2015-06-26
10	Celebrará UAM-Cuajimalpa primera década con dos conciertos del Cepromusic http://t.co/wshuaFTlLu	2015-06-26
11	Celebrará UAM-Cuajimalpa primera década con dos conciertos del Cepromusic http://t.co/1MzXTr49e	2015-06-26
12	Celebrará UAM-Cuajimalpa primera década con dos conciertos del Cepromusic http://t.co/DAKOaAqCgu	2015-06-26
13	Celebrará @uamcuajimalpa primera década con dos conciertos del Cepromusic @Conaculta @UAM_Comunidad http://t.co/sQWs6HEBiU	2015-06-25
14	RT @Redes_Morena: La unidad Cuajimalpa de la UAM realiza la primera edición del libro tránsito 2015 http://t.co/ciuQdyeMNx	2015-06-25

20 en total.

Figura E.8 Página que muestra los tweets de la consulta seleccionada.

Visualizar menú de gráficas

Estando en la página en donde se muestran los tweets de una consulta, hacer clic en el botón “Menú Visualización” Figura E.9.

The screenshot shows a web application interface for the Universidad Autónoma Metropolitana. At the top left is a logo for '10 años Casa abierta al tiempo Unidad Cuajimalpa 2005 - 2015'. The main title is 'Universidad Autónoma Metropolitana Sistema Automático para la Visualización de Relaciones de Opiniones Producidas en Twitter'. On the left sidebar, there are three buttons: 'Nueva Búsqueda', 'Búsquedas', and 'Palabras Vacías'. The central area displays a table titled 'Tweets' with 10 rows of tweet data, including text and creation date. On the right side, a vertical menu has a blue button labeled 'Menú Visualización' which is highlighted with a red arrow pointing to it.

#	Text	Created At
1	Inauguración d la Exposición Gráfica d Catedráticos,Intercambio cultural ntre la DERNyAP d #Cuajimalpa y UAM Santa Fe http://t.co/q6WYq0d44e	2015-06-28
2	RT @hojaderutamexic: Celebrará @uamcuajimalpa primera década con dos conciertos del Cepromusic @Conaculta @UAM_Comunidad http://t.co/sQWs6H...	2015-06-28
3	4. Me gusta más la uam cuajimalpa que fes acatlan.	2015-06-28
4	#Cultura:Celebrará UAM-Cuajimalpa primera década con dos conciertos del #Cepromusic http://t.co/oOPNsQhwX1	2015-06-27
5	#Cultura:Celebrará UAM-Cuajimalpa primera década con dos conciertos del #Cepromusic http://t.co/oOPNsQhwX1	2015-06-26
6	RT @elgolfoveracruz #Cultura:Celebrará UAM-Cuajimalpa primera década con dos conciertos del #Cepromusic http://... http://t.co/J2jZgMoxf	2015-06-26
7	#Cultura:Celebrará UAM-Cuajimalpa primera década con dos conciertos del #Cepromusic http://t.co/oOPNsQhwX1	2015-06-26
8	RT @hojaderutamexic: Celebrará @uamcuajimalpa primera década con dos conciertos del Cepromusic @Conaculta @UAM_Comunidad http://t.co/sQWs6H...	2015-06-26
9	La Unidad Cuajimalpa de la #UAM realiza la primera edición del Libro tránsito 2015. http://t.co/S8fQObvBge	2015-06-26
10	Celebrará UAM Cuajimalpa primera década con dos conciertos del Cepromusic https://t.co/afufQbuzSTIu...	2015-06-26

Figura E.9 Página que muestra la ubicación den botón “Menú Visualización”.

En seguida el sistema despliega el menú de visualización, que contiene cuatro gráficas, como se muestra en la figura E.10.

The screenshot shows the 'Menú de Visualización' page. At the top left is a logo for '10 años Casa abierta al tiempo Unidad Cuajimalpa 2005 - 2015'. The main title is 'Universidad Autónoma Metropolitana Sistema Automático para la Visualización de Relaciones de Opiniones Producidas en Twitter'. On the left sidebar, there are three buttons: 'Nueva Búsqueda', 'Búsquedas', and 'Palabras Vacías'. The central area displays a menu with four options: 'Visualiza Source', 'Visualiza Favoritos', 'Visualiza Retweets', and 'Agrupamiento Temático'. Below the menu, it says 'Tema de la Búsqueda: uam cuajimalpa' and 'Cantidad de Tweets: 20'. At the bottom, it credits 'Proyecto Terminal de: Griselda Velázquez Villar' and 'Asesores: M.C. Gabriela Ramírez de la Rosa y Dr. Esaú Villatoro Tello'.

Figura E.10 Página que muestra el menú de visualización.

A continuación se describen cada una de las cuatro gráficas que contiene el menú de visualización.

Visualizar gráfica de Source o plataforma

Esta gráfica agrupa los la colección de tweets por source, es decir, por la plataforma en donde fue publicado cada tweet de la colección. El sistema hace uso del campo source que está en el registro de los tweets y muestra una gráfica de burbujas con los grupos que formo, como se muestra en la figura E.11.

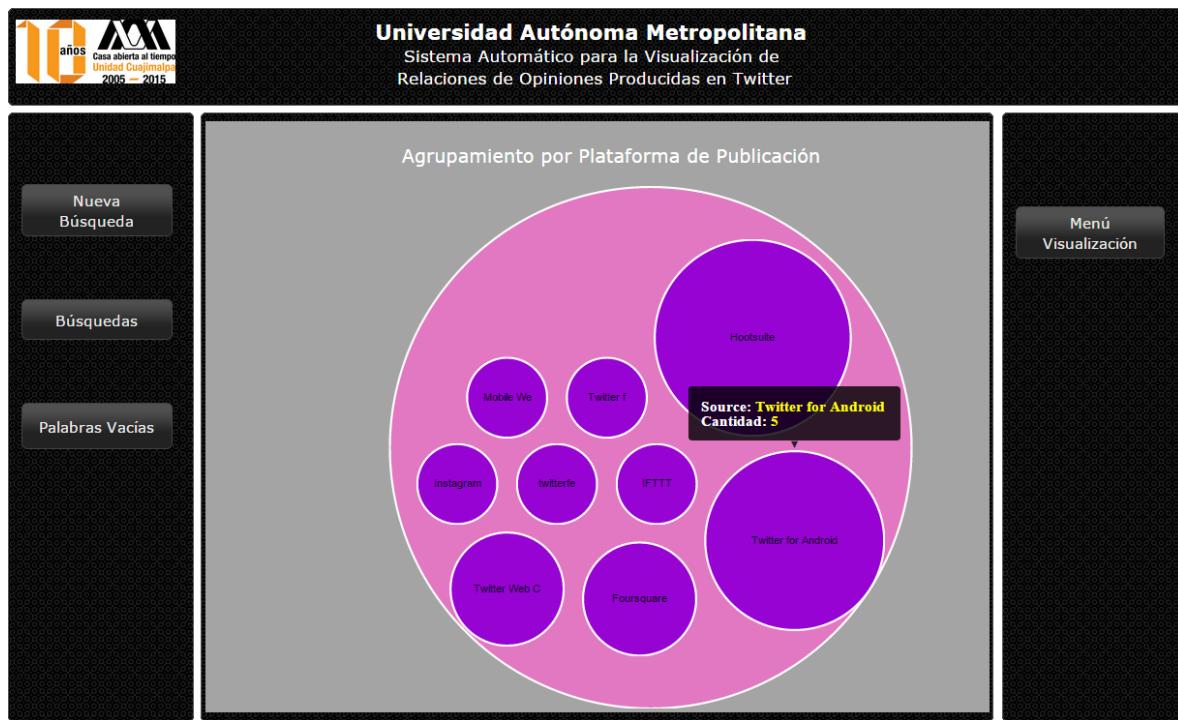


Figura E.11 Página que muestra la gráfica por source o plataforma.

Cada burbuja representa una plataforma diferente de donde se publicaron los tweets, así como la cantidad tweets que se publicaron en dicha plataforma.

Visualizar gráfica por Favoritos

La gráfica por favorito agrupa la colección de tweets por la cantidad de veces que los tweets han sido marcados como favoritos. Se agrupan de mayor a menor como se muestra en la figura E.12.

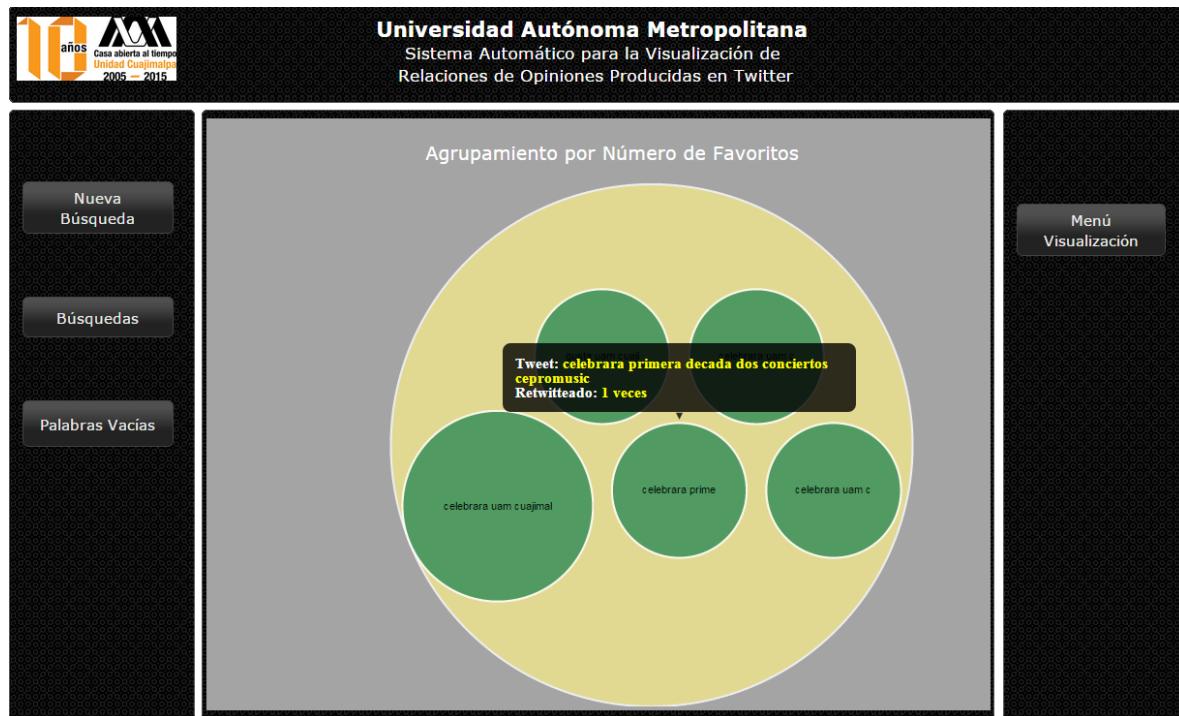


Figura E.12 Página que muestra la gráfica por favoritos.

Al acercar el mouse a cada burbuja de la gráfica, se mostrara el texto del tweet y la cantidad de veces que ha sido marcado como favorito.

Visualizar gráfica por Retweets

Esta gráfica agrupa los tweets por retweets y los muestra de mayor a menor, como se muestra en la figura E.13.

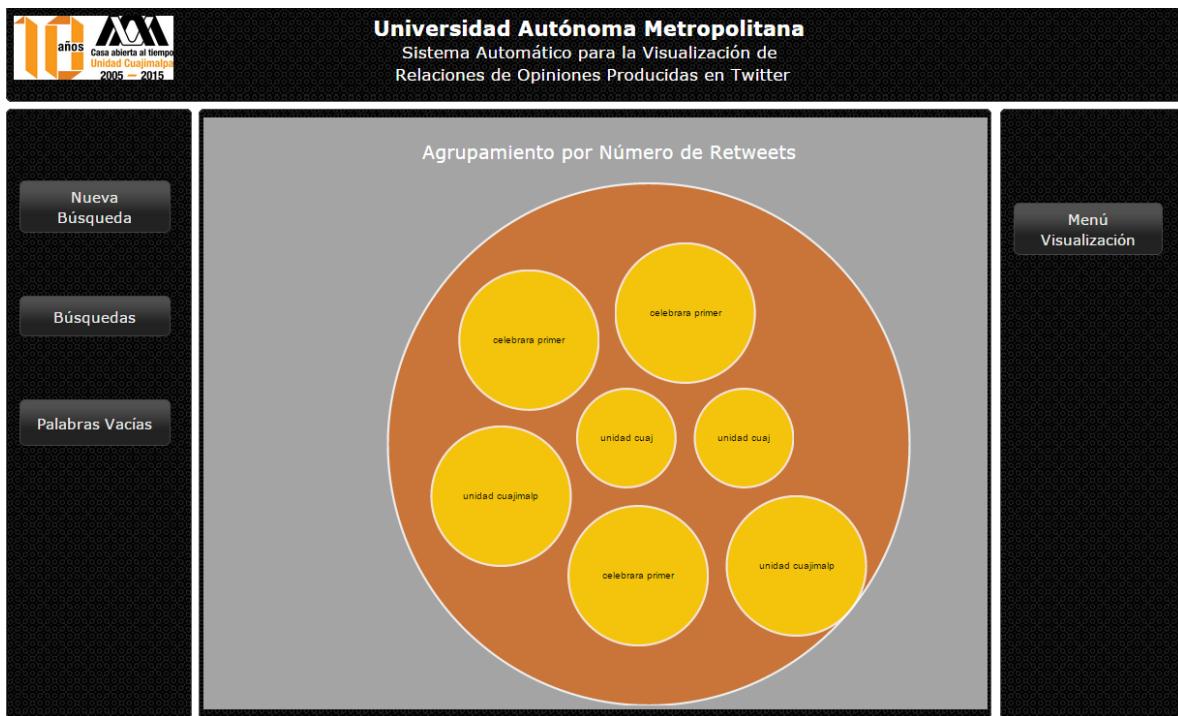


Figura E.13 Página que muestra la gráfica por retweets.

Cada burbuja representa el tweet que ha sido retwitteado por diferentes usuarios de Twitter y además se muestra el número de veces que ha sido retwitteado.

Visualizar gráfica por Agrupamiento Temático

La gráfica por agrupamiento temático permite hacer uso de los algoritmos de agrupamiento y el tipo de pesado que se mencionan en el marco teórico de este documento

Para poder visualizar la gráfica por agrupamiento temático es necesario seleccionar algunos datos como son:

El tipo de algoritmo: el sistema contempla el algoritmo K-means y el Jerárquico.

El tipo de pesado: binario, tf, tf-idf

El número de grupos: el sistema proporciona hasta nueve grupos y da por default cuatro.

Esta información se puede seleccionar de la página de algoritmos de agrupamiento, como se muestra en la figura E.14.

The screenshot shows a web interface for clustering algorithms. At the top right, the header reads "Universidad Autónoma Metropolitana" and "Sistema Automático para la Visualización de Relaciones de Opiniones Producidas en Twitter". On the left sidebar, there are buttons for "Nueva Búsqueda", "Búsquedas", and "Palabras Vacías". The main content area is titled "Algoritmos de Agrupamiento" and contains three dropdown menus: "Algoritmo" set to "kmeans", "Pesado" set to "binario", and "Número de Grupos" set to "4". Below these is a "Graficar" button. On the far right, there is an "Ayuda" button. At the bottom, it says "Proyecto Terminal de: Griselda Velázquez Villar" and "Asesores: M.C. Gabriela Ramírez de la Rosa y Dr. Esaú Villatoro Tello".

Figura E.14 Página de algoritmos de agrupamiento.

Al momento de dar clic en el botón “Graficar” se despliega la gráfica de agrupamiento temático, como se observa en la figura E-15.

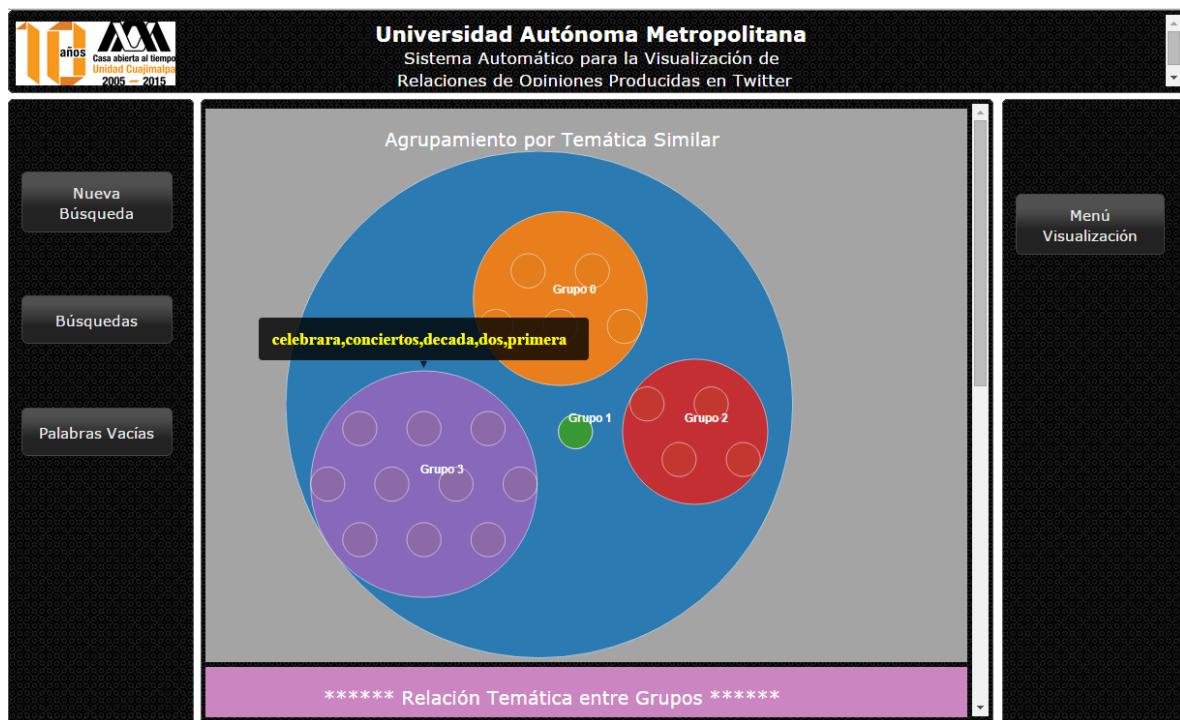


Figura E.15 Página que muestra la gráfica por temática similar.

En esta grafica se visualizan la siguiente información:

- El número de grupos que se formaron.
- Los tweets que integran cada grupo.
- Las cinco palabras que predominan en cada grupo
- Las cinco palabras que se comparten entre un par de grupos.

Agregar o eliminar palabras vacías

El sistema integra la opción de agregar palabras vacías que se almacenan en una tabla de la base de datos y que se hace uso de ellas al momento de preprocesar la información para obtener un texto con palabras significativas.

Al dar clic al botón “Palabras Vacías” se despliega la pantalla en donde podemos agregar o eliminar una palabra vacía, como se observa en la figura E.16.

The screenshot shows a web interface for managing empty words. At the top, there's a logo for the Universidad Autónoma Metropolitana (UAM) and text indicating it's the 'Sistema Automático para la Visualización de Relaciones de Opiniones Producidas en Twitter'. On the left sidebar, there are buttons for 'Nueva Búsqueda', 'Búsquedas', and 'Palabras Vacias'. The main content area has a title 'Palabras Vacias' with a table showing 8 entries:

#	Palabras
1	a
2	aca
3	ahi
4	ahora
5	ajena
6	ajeno
7	ajenos
8	ajenas

Below the table, it says '238 en total.' and there are 'Eliminar' and 'Editar' buttons. To the right, there are fields for 'Agregar Palabra:' and 'Editar Palabra:' with 'Guardar' and 'Editar' buttons respectively. At the bottom, it credits 'Proyecto Terminal de: Griselda Velázquez Villar' and 'Asesores: M.C. Gabriela Ramírez de la Rosa y Dr. Esaú Villatoro Tello'.

Figura E.16 Página para agregar o eliminar palabra vacía.