Informe Final

"Sistema Configurable de Minería Web"

Alina Urquiza Pérez

207201853

Asesor: Silva López Rafaela Blanca
Departamento: Sistemas
Categoría: Titular B -Medio Tiempo

Asesor: Maricela Claudia Bravo Contreras
Departamento: Sistemas
Categoría: Titular C –Tiempo Completo

Fecha de Entrega: 28 de Noviembre del 2012

INTRODUCCION

PLANTEAMIENTO DE LA PROBLEMÁTICA QUE SE ABORDA CON EL PROYECTO TERMINAL

La World Wide Web es el repositorio más grande y ampliamente conocido de páginas Web. Estos documentos o páginas Web están escritos en una gran diversidad de idiomas y abarcan todos los tópicos del conocimiento humano. La utilización de la Web ha experimentado un crecimiento exponencial desde su aparición y resulta notorio debido al gran número de sitios que ofrecen variados servicios en línea, por ejemplo en campos como: la educación, la investigación, las consultas de servicios, el entretenimiento, el comercio electrónico, entre otros.

Este proyecto fue pensado para resolver principalmente la problemática de encontrar información en la Web que es interesante y que a simple vista no es evidente ni fácil de entender.

Existen diferentes dificultades a los que se enfrentan los usuarios debido al crecimiento exponencial de recursos y contenidos de información en la Web. Entre los más importantes se encuentra: la baja precisión en las búsquedas y la escasa cobertura. La baja precisión en los resultados de las búsquedas se refiere a que la información encontrada es irrelevante con respecto a las necesidades del usuario.

La escasa cobertura se debe a que no todos los buscadores tienen la suficiente capacidad de indexar la Web, debido a varios factores; el ancho de banda, el espacio de disco duro etc.

La minería Web resuelve este tipo de problemas al descubrir patrones interesantes. Ofreciendo métodos y algoritmos para recuperar información relevante. La Minería de contenido de la Web juega un papel importante al extraer de información útil del contenido de los documentos Web. Esta información puede ser texto, imágenes, audio, vídeo, o registros estructurados, tales como listas y tablas.

MARCO CONCEPTUAL

Este proyecto se enfocó en el diseño y creación de un sistema configurable denominado “Crawler” que permite especificar una URL, el directorio local de almacenamiento de los documentos y condiciones de paro, para recopilar documentos de la Web. El objetivo de esta etapa es recuperar automáticamente los documentos HTML, indexándolos para optimizar la búsqueda. El proceso de indexación es complejo debido a la gran cantidad de páginas Web, además que estas cambian continuamente. Un Crawler es un programa permite inspeccionar

páginas y recopilar información sobre su contenido; es decir visita las páginas de manera recursiva a partir de un conjunto de hipervínculos de páginas iníciales. En particular, se encarga de recorrer las páginas Web de Internet, descargarlas al ordenador local, parsearlas y procesarlas.

Por lo general, un Crawler dispone de un conjunto inicial de URL´s, conocidas como semillas, va descargando las páginas Web asociadas a las semillas y buscando dentro de éstas otras URL´s. Cada nueva URL encontrada se añade a la lista de URL´s que el Crawler debe visitar. A este proceso se le denomina recolección de URL´s.

Posteriormente el Crawler accede a una nueva URL, la página web asociada es descargada al ordenador local. Una vez ahí, éstas son parseadas y procesadas.

Cuando el Crawler parsea una página Web, lo que hace es decidir qué partes de ésta son de utilidad. Por ejemplo, puede quedarse sólo con los enlaces, imágenes o texto. Cabe mencionar algunas de las dificultades a los que los Crawlers se deben enfrentar: enormes cantidades de páginas que recorrer, elevado número de actualizaciones de páginas existentes, páginas que crean su contenido de forma dinámica etc.

El parser es un proceso que comprende solo la cabecera del documento HTML. El parser es el proceso de analizar de forma sintáctica, el documento HTML, en busca de determinadas etiquetas para su futuro análisis o manipulación, en este caso identifica el contenido de la cabecera de toda página HTML, etiquetas “TITLE” y “META”.

El Clustering es una técnica de exploración de datos para resolver problemas de clasificación. Su objeto consiste en ordenar objetos (personas, cosas, animales, plantas, variables, etc.) en grupos (conglomerados o clúster) de forma que el grado de asociación/similitud entre miembros del mismo clúster sea más fuerte que el grado de asociación/similitud entre miembros de diferentes clústers.

Los resultados de un análisis de clúster pueden contribuir a la definición formal de un esquema de clasificación tal como una taxonomía para un conjunto de objetos, sugerir modelos estadísticos para describir poblaciones, asignar nuevos individuos a las clases para diagnóstico e identificación, etc.

Lograr encontrar las palabras y cadenas más convenientes, con lleva un estudio de las keywords más buscadas en internet y de los buscadores más conocidos. Desafortunadamente, resulta complicado, encontrar estadísticas con las keywords adecuadas para nuestra web, esto resulta todavía más complicado si tratamos de encontrar estos resultados en español.

Para poder encontrar keywords de utilidad es necesario usar StopWords. Un StopWords son palabras que pueden ser consideradas irrelevantes para el conjunto de resultados a ser exhibidos en una búsqueda realizada en un motor de búsqueda Ejemplos: las, y, los, de, para, con, sin, fue. Claro que irrelevantes, depende de la búsqueda realizada, pues el contexto de la búsqueda hará toda la diferencia para cada palabra usada en la investigación realizada.

Para poder relacionar una lista de keywords y un conjunto de archivos .txt se construyó una matriz de referencia y de pesos. Una matriz de pesos es una matriz cuyos elementos representan las distancias entre los datos, tomando pares de un conjunto. Se trata por lo tanto, de una matriz simétrica de tamaño NxN (dado un conjunto de N puntos en el espacio euclídeo) conteniendo números reales no negativos como elementos

Las matrices de pesos están relacionadas con las matrices de referencias, diferenciándose en que las últimas sólo informan sobre qué vértices están conectados, pero no especifican costos o distancias entre los vértices. Además cada elemento de una matriz de pesos es más pequeño cuanto más cercanos se encuentren los puntos, mientras que vértices cercanos (conectados) producen elementos mayores en una matriz de referencia. Debido a la gran dimensión de dicha matriz se utiliza una técnica llamada Descomposición de valor singular (SVD) la cuál es un medio de descomposición de matriz en un producto de tres matrices simples.

El algoritmo para poder realizar el clustering fue Indexación semántica latente (LSI) es un método de indexación y recuperación de datos que utiliza una técnica matemática llamada descomposición de valor singular (SVD) para identificar patrones en las relaciones entre los términos y conceptos contenidos en una colección estructurada de texto. LSI se basa en el principio de que las palabras que se utilizan en los mismos contextos tienden a tener un significado similar. Una característica clave de LSI es su capacidad para extraer el contenido conceptual de un cuerpo de texto mediante el establecimiento de asociaciones entre los términos que aparecen en contextos similares.

TRABAJOS RELACIONADOS.

Los proyectos contienen ciertas similitudes con este proyecto ya que utilizan Minería Web.

- Daedalus es una empresa ubicada en Madrid, España. Daedalus se ha ido consolidando como un referente tecnológico en diversas áreas: las tecnologías de la lengua, la minería de datos, la tecnología Web y la inteligencia de negocio. Daedalus es una empresa fuertemente comprometida con la investigación, el desarrollo y la innovación. Participa en diversos proyectos de

minería Web a nivel tanto internacional como nacional algunos de sus proyectos son:

a) WMA (Web Mining Analytics) es un proyecto dedicado al desarrollo de herramientas que facilitan la extracción y el análisis de información estratégica disponible en Internet. WMA es capaz de extraer automáticamente datos específicos de distintas fuentes (Internet, bases de datos, etc.) y de proporcionar información muy relevante sobre los mismos. Esta solución integra minería de contenido de la Web y minería del uso de la Web

b) MOWGLI es un proyecto dedicado a la generación de perfiles en comercio electrónico. Utiliza la minería Web para generar perfiles de usuario. Para llevar esto a cabo utilizan Minería del uso de la Web.

- Julio García Seminario y David Casanova presentan la propuesta "Implementación De Una Web Mining sobre reconocimiento de patrones de comportamiento de usuarios para la caja municipal de santa (Caja de Ahorro)". En la Universidad San Pedro, Escuela de Informática y Sistemas del Área Facultad de Ingeniería ubicada en Perú. Este proyecto consiste en determinar el patrón de comportamiento de los usuarios de páginas Web, por lo cual desarrollaron un modelo de solución en el cual se describen los procedimientos para determinar un patrón de comportamiento. Desarrollaron un prototipo en base al modelo de solución y que tiene como finalidad determinar un parámetro que identifique un patrón de comportamiento.

- Francisco Manuel Rangel Pardo presenta la tesis "Clasificación de Páginas Web en Dominios Específicos", para obtener el grado de Maestro en Lenguajes y Sistemas Informáticos: Tecnologías de la Lengua en la Web. En la Universidad Nacional de Educación a Distancia. Este trabajo consiste en clasificar las páginas Web en dominios determinados para ello se centra en obtener una representación formal de la intención del autor para transmitir información acerca de la pagina que se crea y que se plasma de la meta-información de la misma en la estructura de los enlaces y en la URL.

- Hernán Merlino presenta la tesis "Ambiente de integración de herramientas para exploración de datos centrados en la Web" para obtener el grado de en Ingeniería del Software. En el Instituto Tecnológico de Buenos Aires. Este trabajo propone una herramienta para exploración de datos Web que permite estructurar todo el proceso de exploración. Esta herramienta utiliza diversas técnicas de exploración, además de permitir la reutilización de procesos ya ejecutados con anterioridad y la combinación de los mismos para su posterior comparación. Este proyecto se basa en los tres tipos de Minería Web.

TABLA COMPARATIVA

Proyecto	Características					
	Minería de contenido de la Web	Minería de la estructura de la Web	Minería del uso de la Web	Datos no Estructurados	Datos semi-estructurados	Datos Estructurados
Daedalus	SI	NO	SI	SI	NO	SI
Implementación De Una Web Mining sobre reconocimiento de patrones de comportamiento de usuarios para la caja municipal de santa	NO	NO	SI	NO	NO	SI
Clasificación de Páginas Web en Dominios Específicos	NO	SI	NO	NO	SI	SI
Ambiente de integración de herramientas para exploración de datos centrados en la Web	SI	SI	SI	SI	SI	SI
Esta propuesta	**SI**	**NO**	**NO**	**NO**	**NO**	**SI**

DESARROLLO

A) DISEÑO

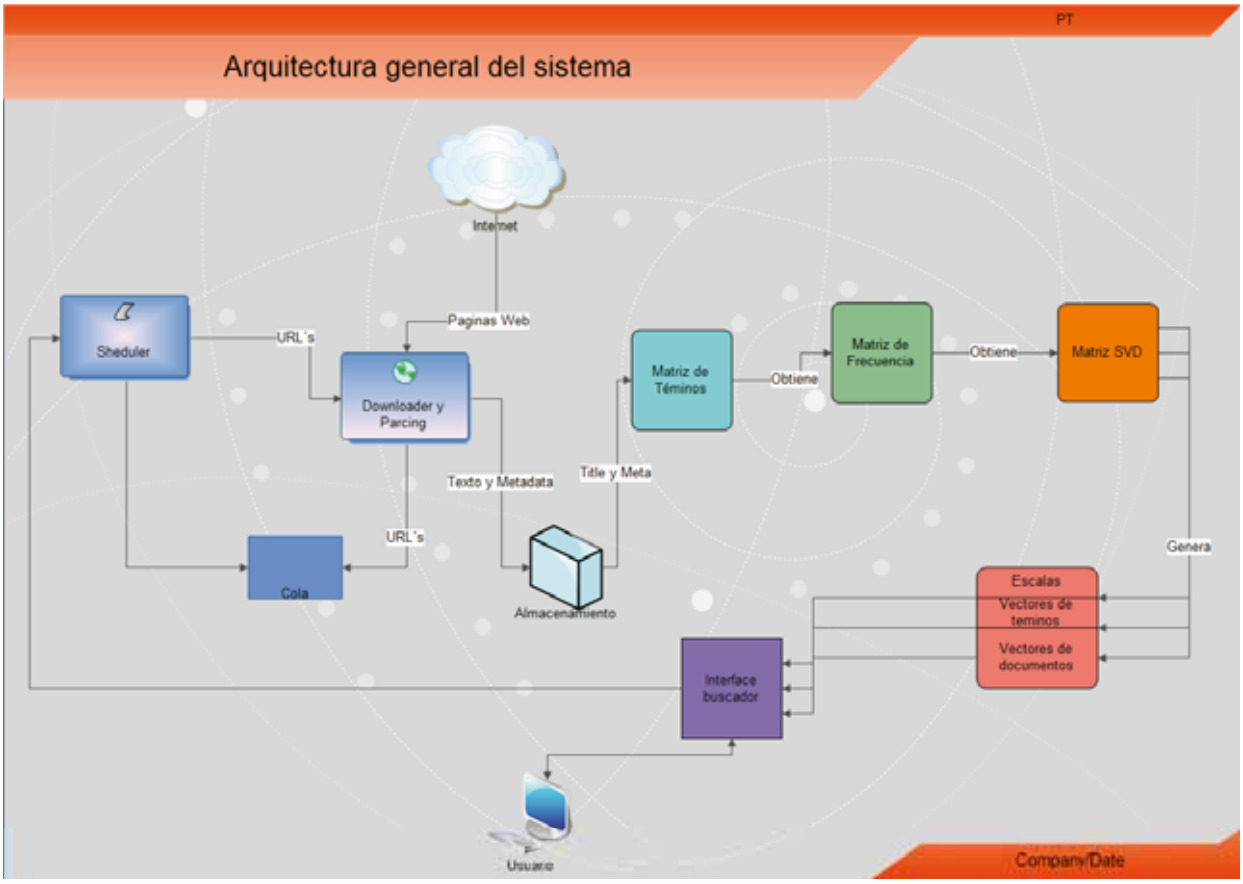


Diagrama de Actividades (Crawler)

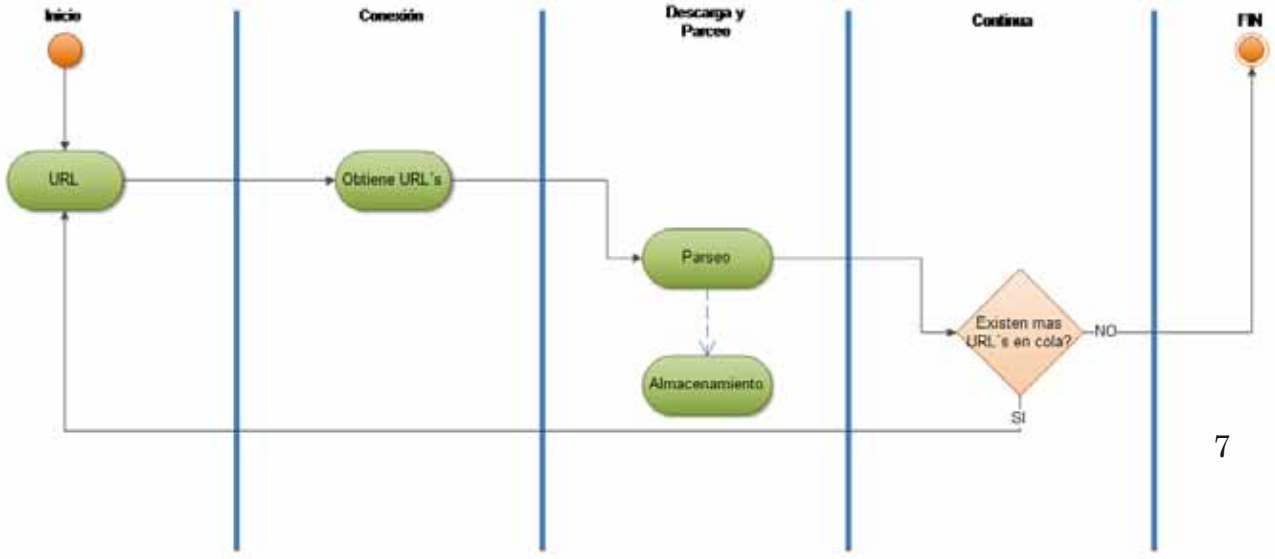
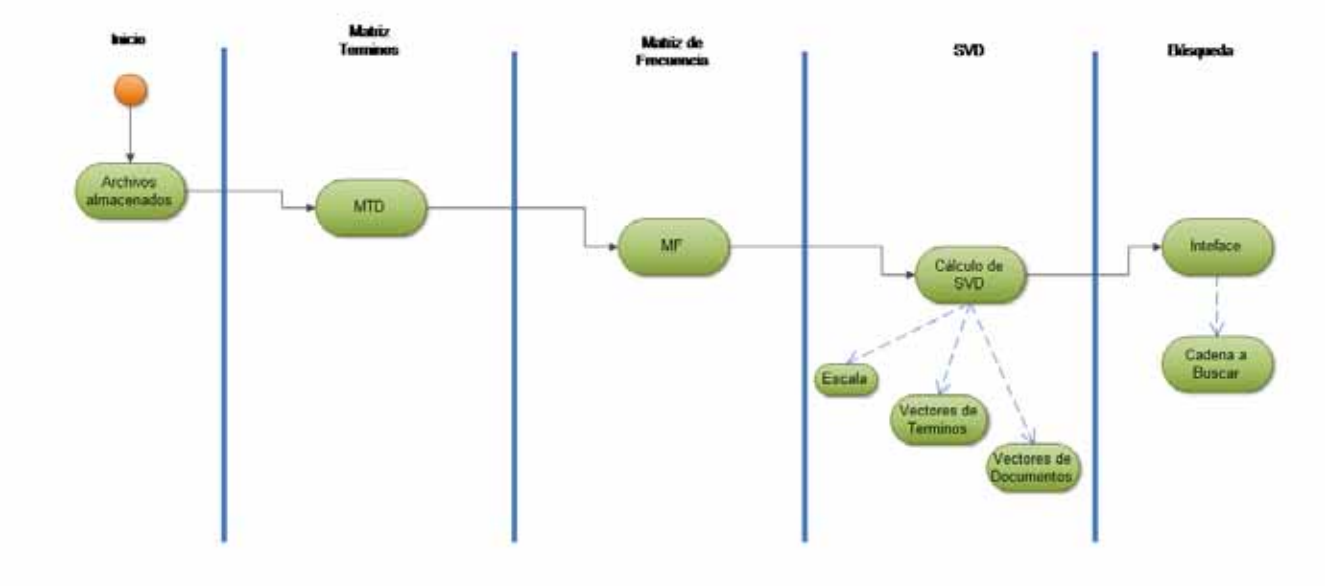
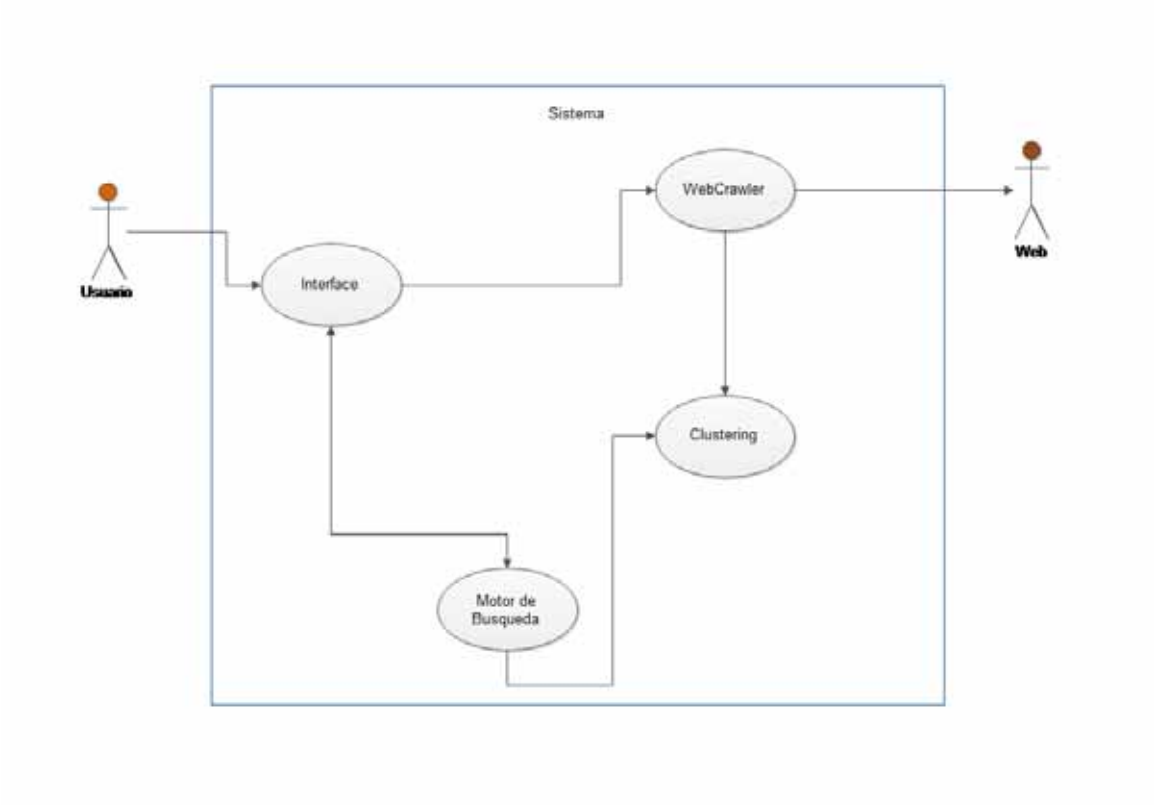


Diagrama de Actividades (Clustering)



Casos de Uso



B) IMPLEMENTACIÓN

- Entorno:

Para garantizar la portabilidad del sistema, este sistema fue desarrollado sobre plataforma Java, utilizando servicios Web mediante el uso del entorno de desarrollo "*NetBeans"* utilizando el compilador JDK 1.7. El desarrollo se vio favorecido por esta herramienta ya que facilita mucho la programación en Java haciendo más ágil el desarrollo del sistema.

- API´s:

Para el desarrollo del Web Crawler se utilizó las siguiente API:

- Jericho

Esta API fue de suma importancia para el desarrollo del sistema ya que con ella se hace el parseo de los archivos descargados con el WebCrawler, con esta API se obtuvieron las etiquetas TITLE y META de las URL´s visitadas. Se optó por utilizar esta API después de distintas pruebas con otras, el problema se encontró en la implementación de las otras API´s estas se centraban en parseo de archivos XML, y para el desarrollo del sistema no servían, por eso Jericho se adaptó al desarrollo del sistema ya que se centra en el parseo de archivos HTML.

Para el desarrollo del Clustering se utilizó la siguiente API:

- LingPipe

Esta API se utilizó para el procesamiento de los archivos de texto que contienen los TITLE y las METAS que se obtuvieron con el WebCrawler. Gracias a esta API se pudo obtener la matriz SVD la cual sirve en el desarrollo del algoritmo de aprendizaje no supervisado. Con la matriz SVD se obtienen los patrones que existen entre todas las palabras y documentos que se tienen, y así se puede implementar el motor de búsqueda el cual recibe de entrada una cadena y con solo esto se obtienen los resultados donde esta cadena encuentra palabras con significado similar. Se eligió esta API por su fácil uso y su eficiencia para generar la matriz SVD ya que las otras API´s que se probaron (*JAMA, Efficient Java Matrix Library*), eran muy ineficientes al momento de generar la matriz SVD. Por ejemplo,

la API *JAMA* la matriz se obtuvo en más de 20 minutos, y con la API *EJML* la matriz SVD no se pudo obtener puesto que el programa generaba un error, gracias a esto se optó por utilizar la API LingPipe ya que la matriz SVD se obtuvo en menos de 4 minutos.

- Requerimientos de Sistema

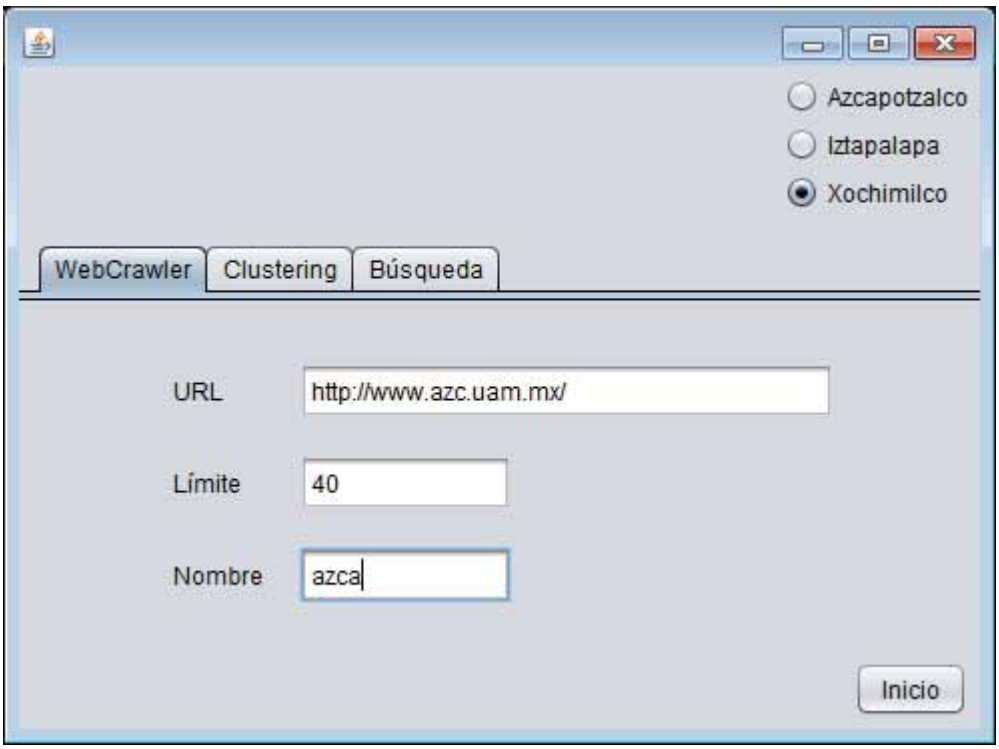
El sistema se desarrolló en una computadora con procesador Intel Dual Core 2.0 GHz y memoria RAM 3 GB. Cualquier computadora con iguales características o superiores podrá ejecutar el sistema sin problemas.

C) PRUEBAS

Objetivo 1: Diseñar y crear un sistema (Crawler) configurable que permita especificar una URL ó IP, dominio, el directorio local de almacenamiento de los documentos y condiciones de paro, para recopilar documentos de la Web.

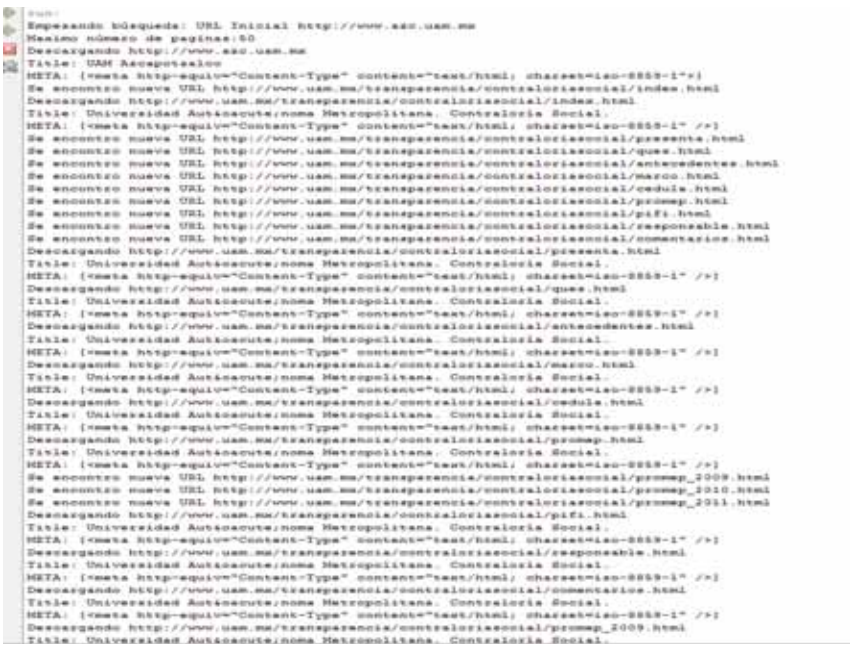
Nombre de la prueba: Crawler

Procedimiento: Se ejecuta la aplicación



En el campo de **URL** se pone el sitio a escanear.
En el campo **Límite** se pone número de páginas máximas descargar.
En el campo nombre se coloca el nombre final que tendrán los archivos descargados. Por último se da clic en el botón Inicio.

Resultado:



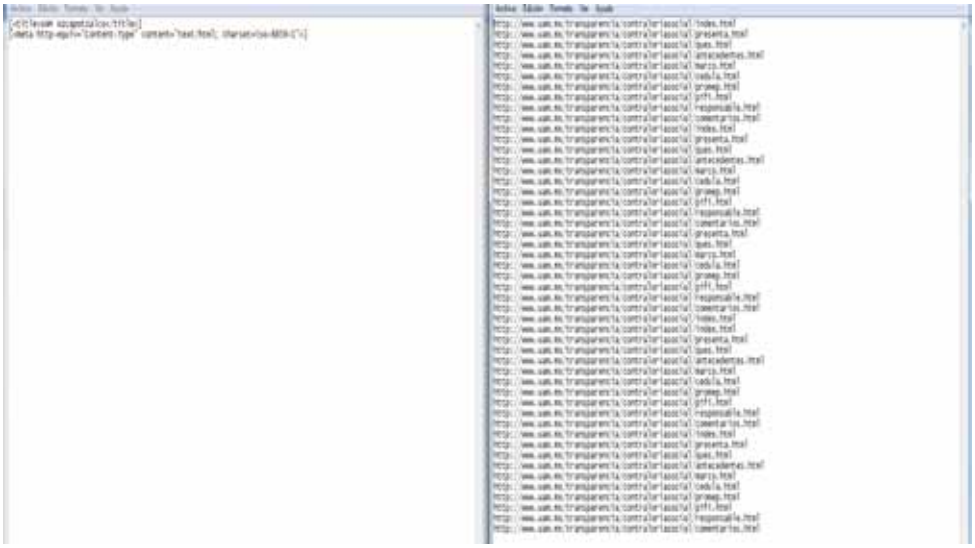
Como se puede observar se rescata la URL dada por el usuario y se comienza la búsqueda de archivos tipo HTML si se encuentran se descargan, se les hace parseo para obtener los TITLE y METAS. Esto nos lleva al Objetivo 2, el cuál es realizado al mismo tiempo que el Objetivo 1.

Objetivo 2: Diseñar e implementar un módulo que permita convertir los documentos extraídos durante el proceso de recuperación de información, en documentos libres de imágenes y objetos tipo animación de tal forma que consistan solamente de texto para facilitar el procesamiento anterior.

Nombre de la prueba: Parseo

Procedimiento: Estos son los archivos generados por el WebCrawler, los cuales contienen las URL´s que fueron visitadas y descargadas, y de lado izquierdo podemos ver un ejemplo de un archivo generado después del parseo, donde se obtienen las etiquetas requeridas.

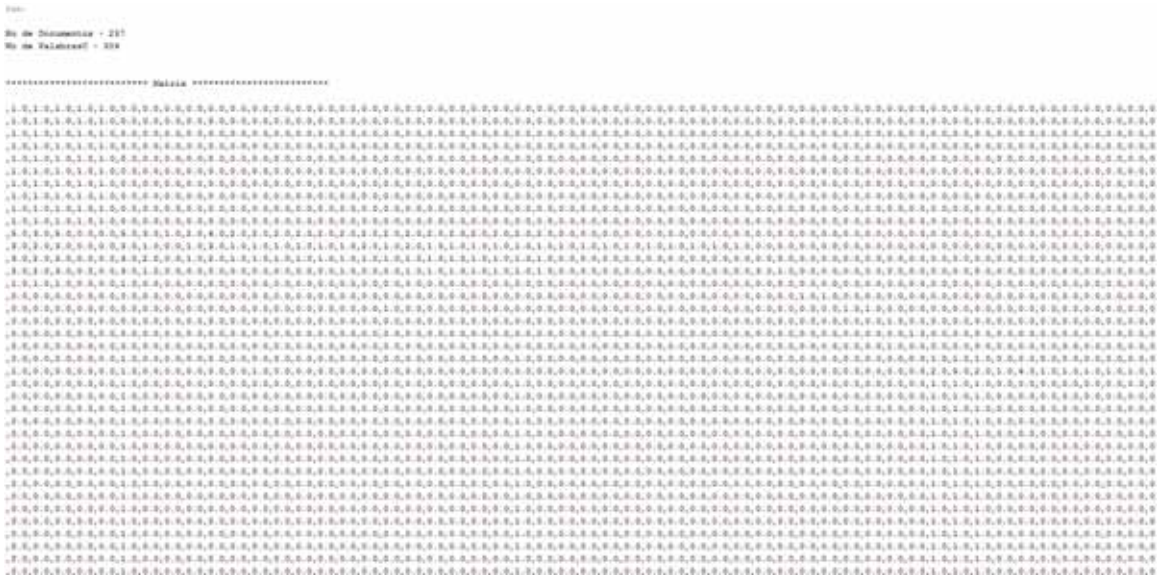
Resultado:



Objetivo 3: Implementar un algoritmo de aprendizaje no supervisado de agrupamiento con el propósito de descubrir automáticamente patrones comunes entre todas las páginas recopiladas.

Nombre de la prueba: Agrupamiento de datos

Procedimiento: Al ejecutar el programa se despliega en la consola La matriz de Referencia



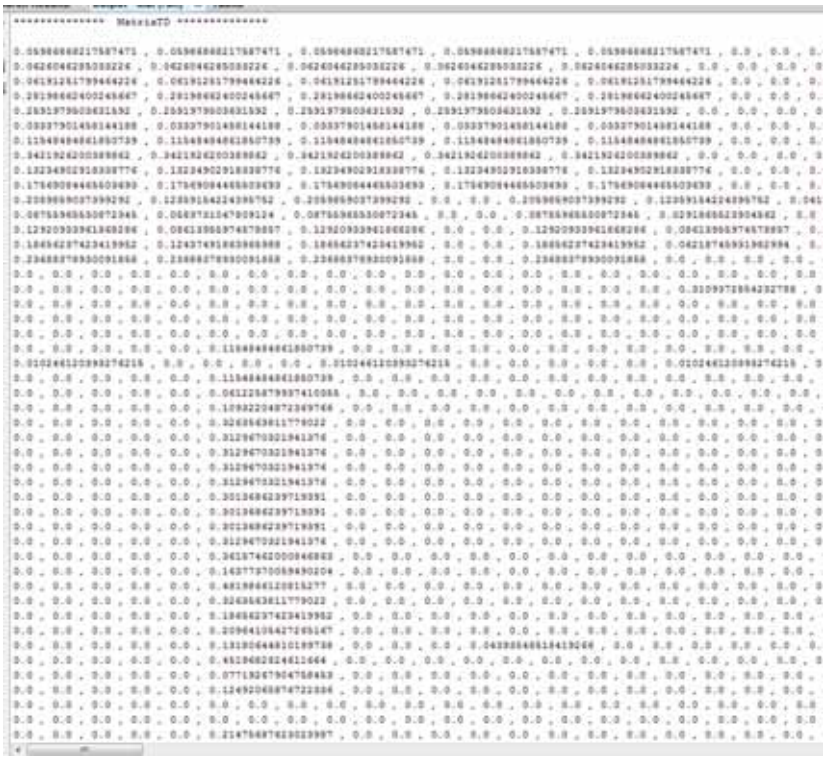
Después el número de palabras por documento.

```
,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
No total de palabras en el documento : 0 -> 5
No total de palabras en el documento : 1 -> 5
No total de palabras en el documento : 2 -> 5
No total de palabras en el documento : 3 -> 5
No total de palabras en el documento : 4 -> 5
No total de palabras en el documento : 5 -> 5
No total de palabras en el documento : 6 -> 5
No total de palabras en el documento : 7 -> 5
No total de palabras en el documento : 8 -> 5
No total de palabras en el documento : 9 -> 5
No total de palabras en el documento : 10 -> 23
No total de palabras en el documento : 11 -> 31
No total de palabras en el documento : 12 -> 22
No total de palabras en el documento : 13 -> 15
No total de palabras en el documento : 14 -> 4
No total de palabras en el documento : 15 -> 2
No total de palabras en el documento : 16 -> 3
No total de palabras en el documento : 17 -> 1
No total de palabras en el documento : 18 -> 1
No total de palabras en el documento : 19 -> 1
No total de palabras en el documento : 20 -> 5
No total de palabras en el documento : 21 -> 50
No total de palabras en el documento : 22 -> 5
No total de palabras en el documento : 23 -> 5
No total de palabras en el documento : 24 -> 5
No total de palabras en el documento : 25 -> 5
No total de palabras en el documento : 26 -> 5
No total de palabras en el documento : 27 -> 5
No total de palabras en el documento : 28 -> 5
No total de palabras en el documento : 29 -> 5
No total de palabras en el documento : 30 -> 5
No total de palabras en el documento : 31 -> 5
No total de palabras en el documento : 32 -> 5
No total de palabras en el documento : 33 -> 5
No total de palabras en el documento : 34 -> 5
No total de palabras en el documento : 35 -> 5
No total de palabras en el documento : 36 -> 5
No total de palabras en el documento : 37 -> 5
No total de palabras en el documento : 38 -> 5
No total de palabras en el documento : 39 -> 5
No total de palabras en el documento : 40 -> 48
No total de palabras en el documento : 41 -> 4
No total de palabras en el documento : 42 -> 8
No total de palabras en el documento : 43 -> 5
No total de palabras en el documento : 44 -> 3
No total de palabras en el documento : 45 -> 4
No total de palabras en el documento : 46 -> 9
No total de palabras en el documento : 47 -> 9
```

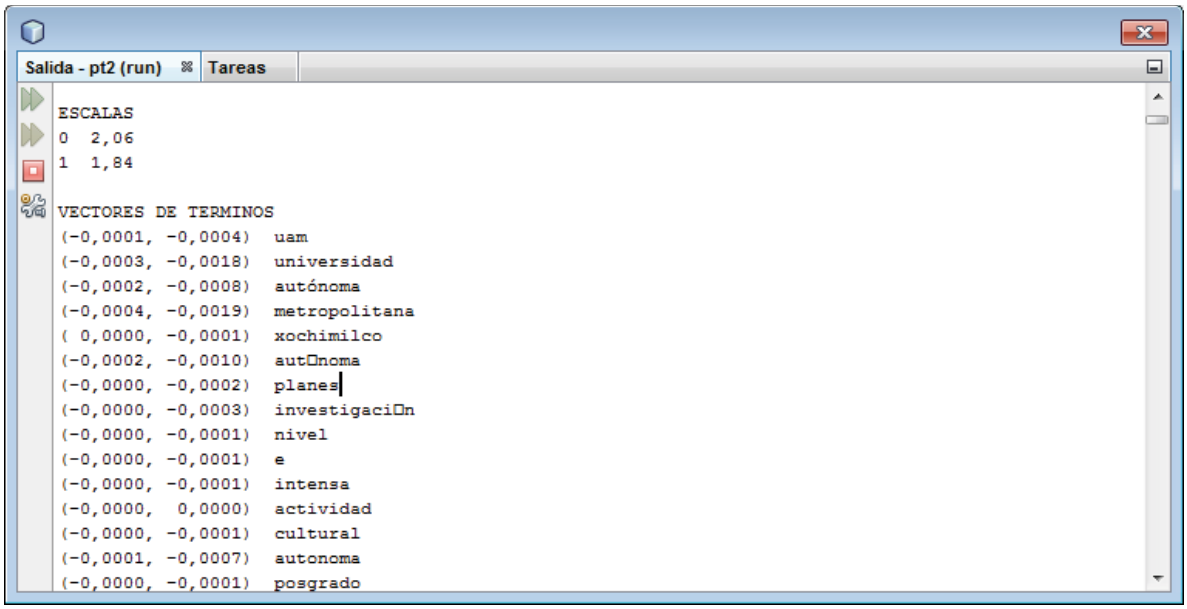
En cuantos documentos se encuentra x palabra

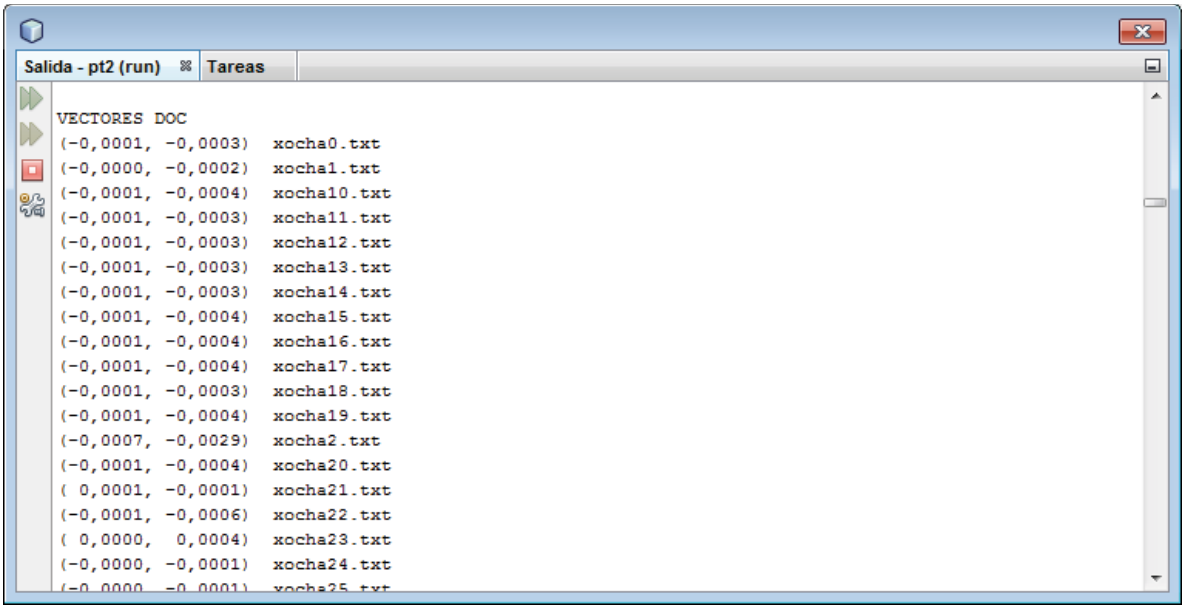
```
palabra : 0 esta en 129 documento(s)
palabra : 1 esta en 125 documento(s)
palabra : 2 esta en 126 documento(s)
palabra : 3 esta en 10 documento(s)
palabra : 4 esta en 13 documento(s)
palabra : 5 esta en 175 documento(s)
palabra : 6 esta en 68 documento(s)
palabra : 7 esta en 5 documento(s)
palabra : 8 esta en 56 documento(s)
palabra : 9 esta en 34 documento(s)
palabra : 10 esta en 29 documento(s)
palabra : 11 esta en 32 documento(s)
palabra : 12 esta en 29 documento(s)
palabra : 13 esta en 30 documento(s)
palabra : 14 esta en 29 documento(s)
palabra : 15 esta en 82 documento(s)
palabra : 16 esta en 30 documento(s)
palabra : 17 esta en 41 documento(s)
palabra : 18 esta en 63 documento(s)
palabra : 19 esta en 67 documento(s)
palabra : 20 esta en 68 documento(s)
palabra : 21 esta en 79 documento(s)
palabra : 22 esta en 68 documento(s)
palabra : 23 esta en 127 documento(s)
palabra : 24 esta en 73 documento(s)
palabra : 25 esta en 6 documento(s)
palabra : 26 esta en 7 documento(s)
palabra : 27 esta en 7 documento(s)
palabra : 28 esta en 7 documento(s)
palabra : 29 esta en 7 documento(s)
palabra : 30 esta en 8 documento(s)
palabra : 31 esta en 8 documento(s)
palabra : 32 esta en 8 documento(s)
palabra : 33 esta en 7 documento(s)
palabra : 34 esta en 4 documento(s)
palabra : 35 esta en 39 documento(s)
palabra : 36 esta en 1 documento(s)
palabra : 37 esta en 6 documento(s)
palabra : 38 esta en 30 documento(s)
palabra : 39 esta en 23 documento(s)
palabra : 40 esta en 2 documento(s)
palabra : 41 esta en 4 documento(s)
palabra : 42 esta en 62 documento(s)
palabra : 43 esta en 61 documento(s)
palabra : 44 esta en 61 documento(s)
palabra : 45 esta en 3 documento(s)
palabra : 46 esta en 3 documento(s)
palabra : 47 esta en 3 documento(s)
palabra : 48 esta en 3 documento(s)
```

Con estos datos se genera la matriz de pesos…



La matriz de pesos es necesaria puesto que es un parámetro de entrada para el método que genera la SVD. Con esto se obtienen: Las escalas, los vectores de términos y los vectores de documentos.



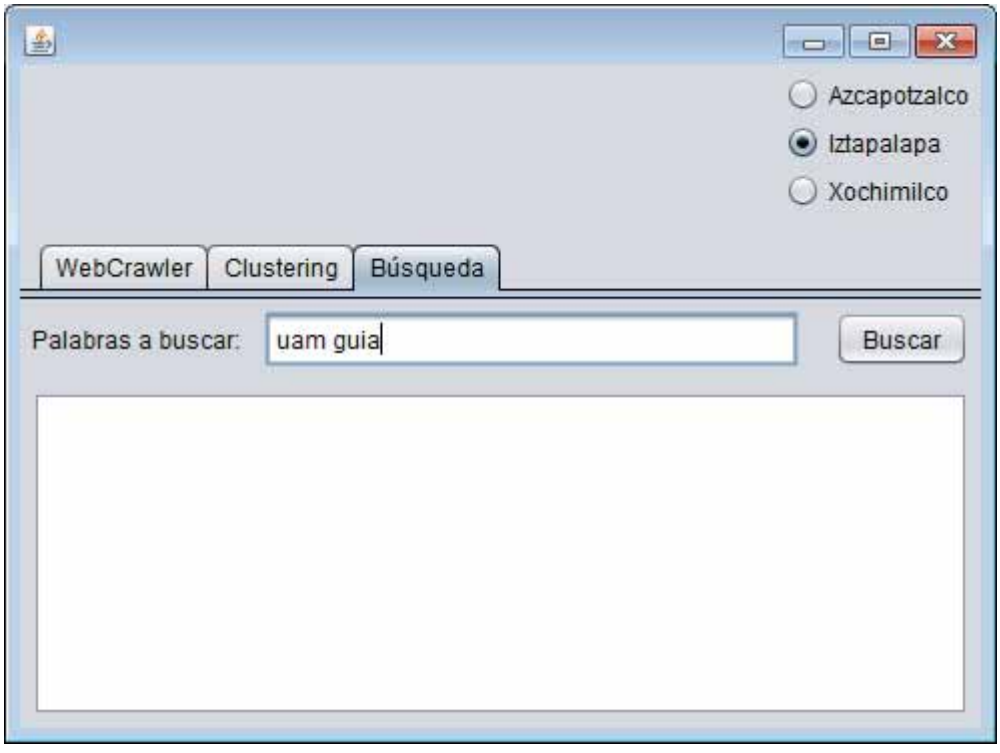


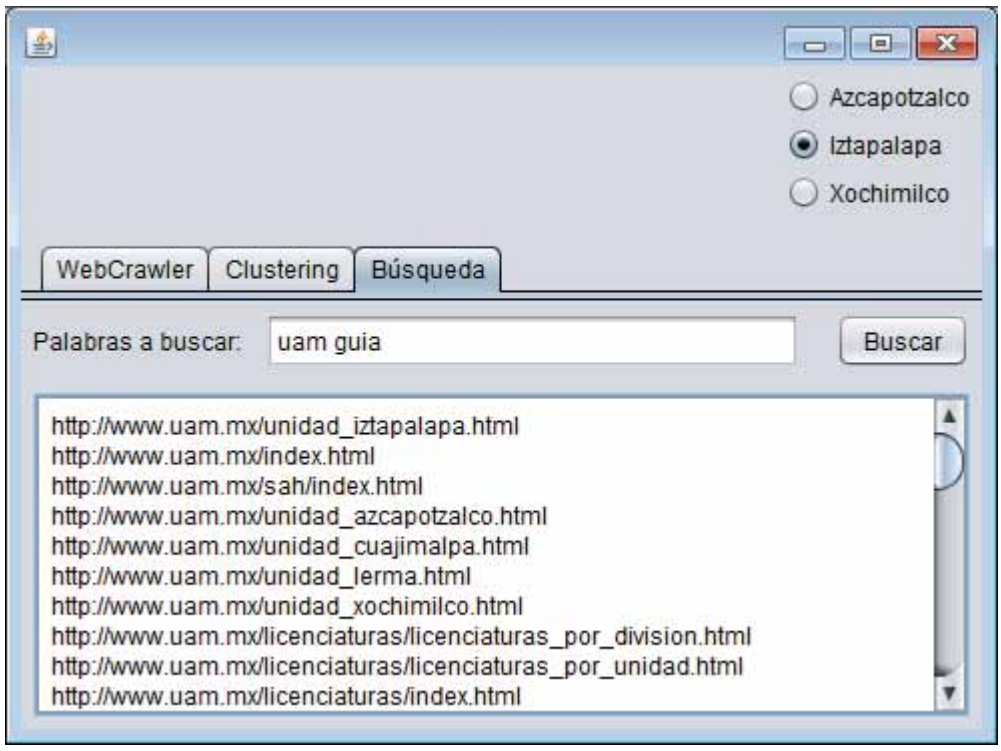
Objetivo 4: Diseñar e implementar una interfaz gráfica para visualizar e interpretar los patrones descubiertos.

Nombre de la prueba: Interfaz Gráfica

Procedimiento: Con los datos obtenidos, se optó por hacer un buscador para esto se debe elegir a que unidad se le desea hacer alguna búsqueda, se pueden introducir "N" número de palabras.

Por último se le da buscar y el resultado es una lista de las páginas Web a las que están relacionadas dichas palabras. Donde el primer link de la lista es el que tiene más relación con la búsqueda y asi sucesivamente.





CONCLUSION

En la Actualidad la World Wide Web es un inmenso repositorio en donde podemos observar y consultar información. Sin embargo, esta información es difícil de procesar o manipular para asimilarla adecuadamente o adquirir nuevos conocimientos. La minería Web resuelve este tipo de problemas al descubrir datos importantes.

Para poder realizar este proyecto el primer paso fue crear un sistema (Crawler) configurable que permita especificar una URL ó IP, dominio, el directorio local de almacenamiento de los documentos y condiciones de paro, para recopilar documentos de la Web. En base al estudio y análisis de varios documentos que contienen información relacionada con los crawlers, puedo agregar que un Crawler es un programa que recorre las páginas del World Wide Web de forma metódica y automatizada, las descarga y procesa, después visita una URL, identifica los links en dichas páginas y los añade a la lista a visitar de manera recurrente de acuerdo a determinado conjunto de reglas.

Para convertir los documentos extraídos durante el proceso de recuperación de información, en documentos libres de imágenes y objetos tipo animación. Se llevo a cabo un proceso llamado parseo que consiste en analizar una secuencia de símbolos a fin de determinar su estructura gramatical con respecto a una gramática formal dada. Formalmente es llamado análisis de sintaxis. El parseo

transforma una entrada de texto en una estructura de datos (usualmente un árbol) que es apropiada para ser procesada. Generalmente un parser primero identifica los símbolos de la entrada y luego construye el árbol de parseo para esos símbolos.

Ahora bien la implementación de un algoritmo de aprendizaje no supervisado de agrupamiento con el propósito de descubrir automáticamente patrones comunes entre todas las páginas recopiladas no fue una tarea fácil.

Se utilizó el algoritmo Indexación semántica latente (LSI) este algoritmo se enfoca en el proceso de indexar documentos, además de registrar las palabras clave que contiene un documento, el método examina la colección de documentos en su conjunto, para ver qué otros documentos contienen algunas de las mismas palabras. LSI considera que los documentos que tienen muchas palabras en común están semánticamente cerca, y las palabras que no tienen mucho en común son semánticamente lejanas.

LSI no requiere una coincidencia exacta para obtener resultados útiles. Cuando hace una búsqueda con una palabra clave muy simple esta no funcionará si no hay ninguna coincidencia exacta. Una gran ventaja de LSI es que tiene un enfoque estrictamente matemático, a él no le importa el significado de los documentos o palabras que analiza. Esto hace que sea una técnica poderosa y genérica capaz de indexar cualquier colección de documentos en cualquier idioma.

El Sistema de Minería Web deriva su potencialidad al ser atractivo para varios rubros ya que se puede minar cualquier página Web. Por ejemplo en el comercio electrónico la minería Web es utilizada en las agencias gubernamentales para clasificar las amenazas y la lucha contra el terrorismo.

La capacidad de predicción de la aplicación de la minería puede beneficiar a la sociedad mediante la identificación de actividades delictivas. Las compañías pueden establecer una relación mejor atención al cliente, dándoles exactamente lo que necesitan. Las empresas pueden entender las necesidades del cliente y las satisfacen más rápido.

Las empresas pueden encontrar, atraer y retener a los clientes, ya que pueden ahorrar en los costos de producción mediante la visión adquirida en los requerimientos del cliente.

Por último a las dificultades que me enfrente a lo largo del proyecto fueron varias pero entre las más importantes fue el proceso de entendimiento de los conceptos básicos, en la Web existe mucha información sobre este tema pero para lograr entenderla se deben escoger los documentos más adecuados conforme al proyecto. Otra dificultad fue en la implementación del algoritmo del clustering las

supere consultando libros, y leyendo más acerca del algoritmo, para que sirve y que aplicaciones se han desarrollado con este.

REFERENCIAS

- G. Villanueva Baschwitz (2012, Noviembre 7). Programación, depuración del algoritmo SSVD en FORTRAN para el cálculo de valores y vectores propios de una matriz simétrica con alta precisión relativa. [En línea]. Disponible en: http://gauss.uc3m.es/web/personal_web/molera/research/memoria%20pfc%20gvb%20final.pdf

- Amón (2012, Octubre 31). Funciones de Similitud sobre Cadenas de Texto: Una Comparación Basada en la Naturaleza de los Datos. [En línea]. Disponible en: ttp://www.bdigital.unal.edu.co/2033/4/71644758.20104.pdf

- R. Cruz Reyes (2012, Octubre 31). Comparación de Tres Modelos de Texto para la Generación Automática de Resúmenes Web. [En línea]. Disponible en: http://scfi.uaemex.mx/~yledeneva/Publicaciones_archivos/ComparacionDeTresModelosResumenAutom%C3%A1tico%20FINAL.pdf

- E. Carrera. (2012, Octubre 30). Un Algoritmo Simple y Eficiente para la Clasificación Automática de Páginas Web. [En línea]. Disponible en: http://www.evcarrera.net/papers/andescon08.pdf

- I. López Arévalo (2012, Noviembre 11). Línea de Investigación. [En línea]. Disponible en: http://www.tamps.cinvestav.mx/~ilopez/proyectos/lineaInvestigacion-mineria-datos.pdf

- J. Ortega (2012, Noviembre 11). Minería del uso de webs. [En línea]. Disponible en: http://www.elprofesionaldelainformacion.com/contenidos/2009/enero/03.pdf

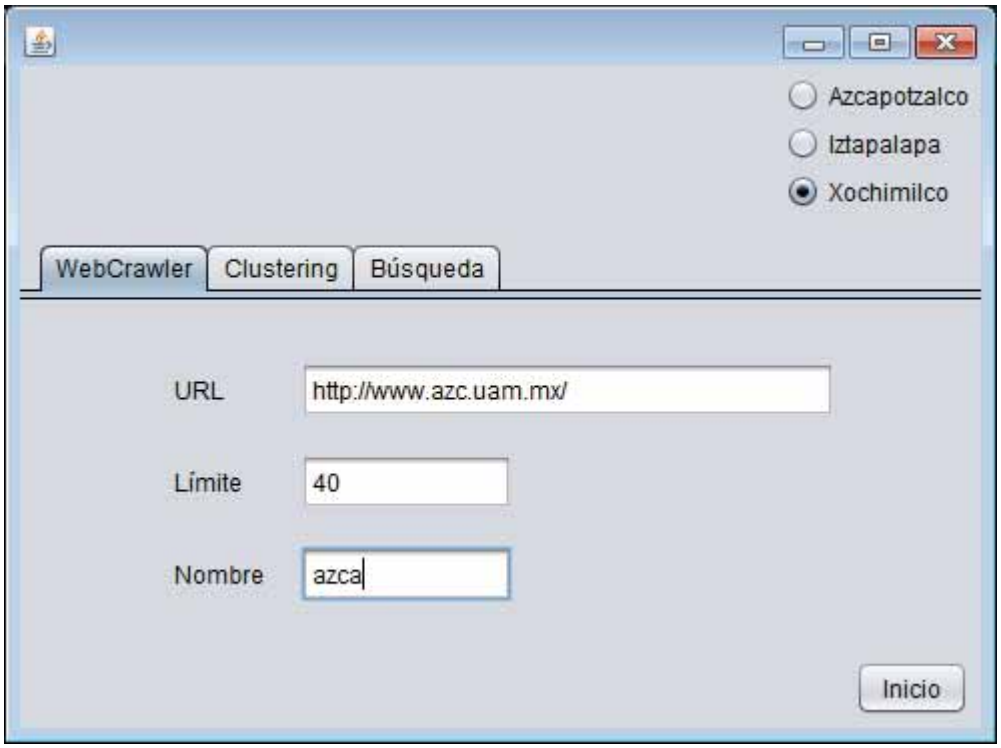
- Daedalus. (2011, Octubre 28). Daedalus-Data. . [En línea]. Disponible en: http://www.daedalus.es/

MANUAL DE USUARIO

PASO 1: Diseñar y crear un sistema (Crawler) configurable que permita especificar una URL ó IP, dominio, el directorio local de almacenamiento de los documentos y condiciones de paro, para recopilar documentos de la Web.

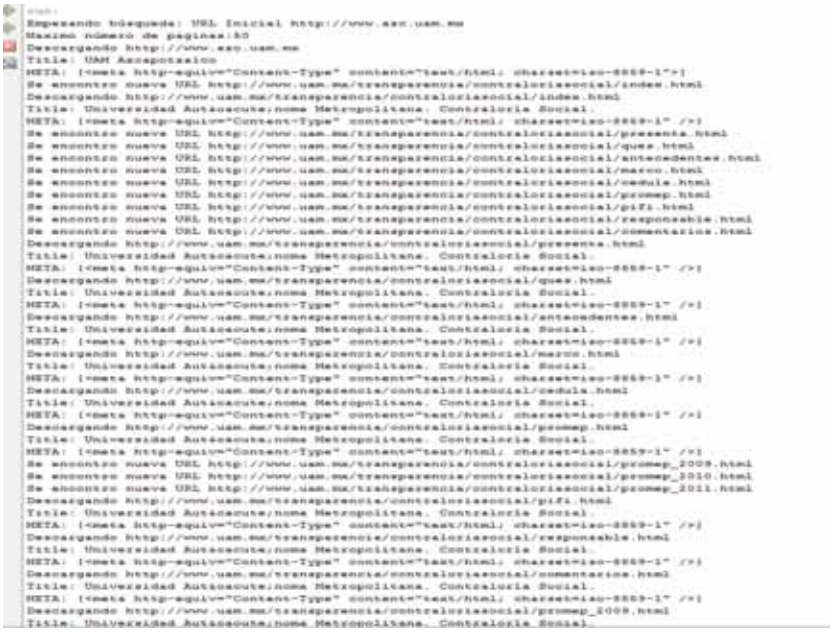
Nombre de la prueba: Crawler

Procedimiento: Se ejecuta la aplicación



En el campo de **URL** se pone el sitio a escanear.
En el campo **Límite** se pone número de páginas máximas descargar.
En el campo nombre se coloca el nombre final que tendrán los archivos descargados. Por último se da clic en el botón Inicio.

Resultado:



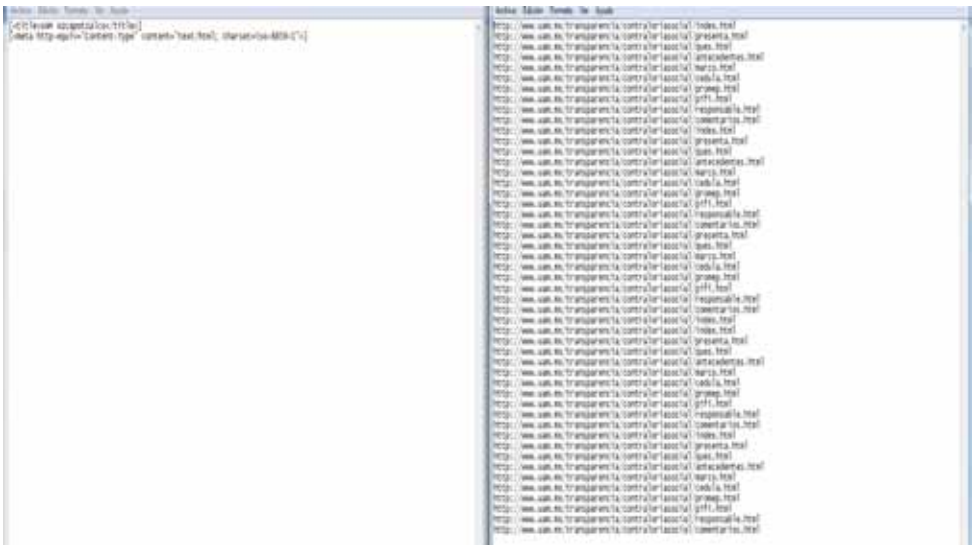
Como se puede observar se rescata la URL dada por el usuario y se comienza la búsqueda de archivos tipo HTML si se encuentran se descargan, se les hace parseo para obtener los TITLE y METAS. Esto nos lleva al Objetivo 2, el cuál es realizado al mismo tiempo que el Objetivo 1.

PASO 2: Diseñar e implementar un módulo que permita convertir los documentos extraídos durante el proceso de recuperación de información, en documentos libres de imágenes y objetos tipo animación de tal forma que consistan solamente de texto para facilitar el procesamiento anterior.

Nombre de la prueba: Parseo

Procedimiento: Estos son los archivos generados por el WebCrawler, los cuales contienen las URL´s que fueron visitadas y descargadas, y de lado izquierdo podemos ver un ejemplo de un archivo generado después del parseo, donde se obtienen las etiquetas requeridas.

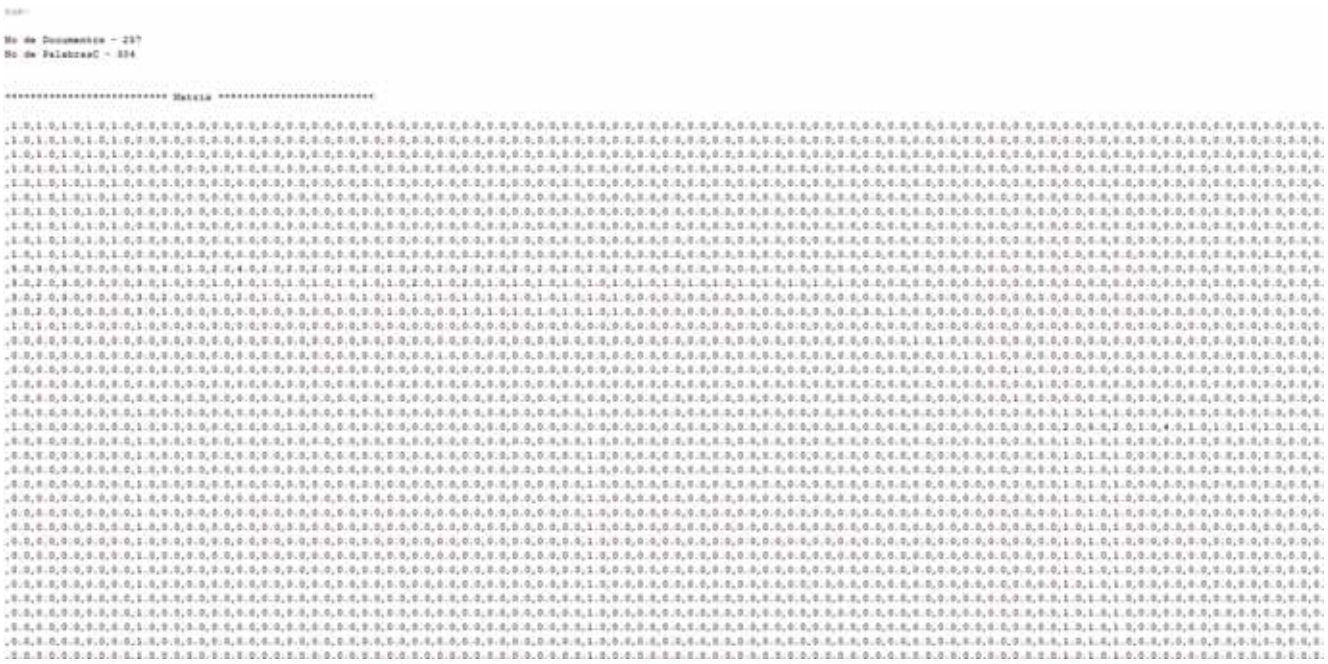
Resultado:



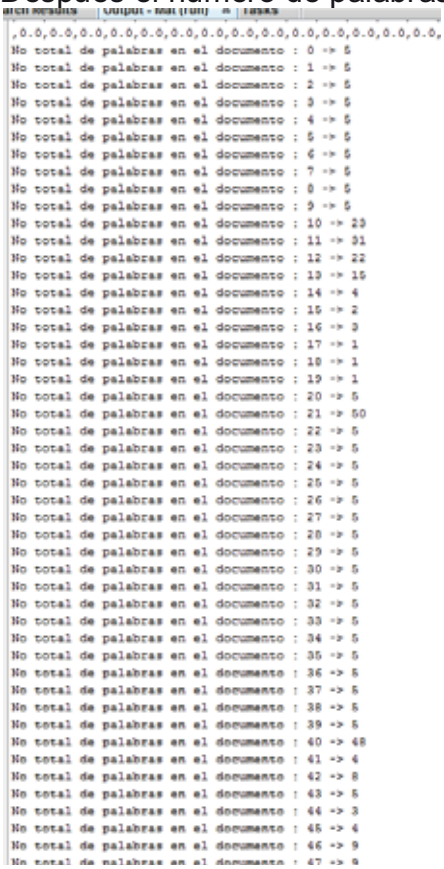
PASO 3: Implementar un algoritmo de aprendizaje no supervisado de agrupamiento con el propósito de descubrir automáticamente patrones comunes entre todas las páginas recopiladas.

Nombre de la prueba: Agrupamiento de datos

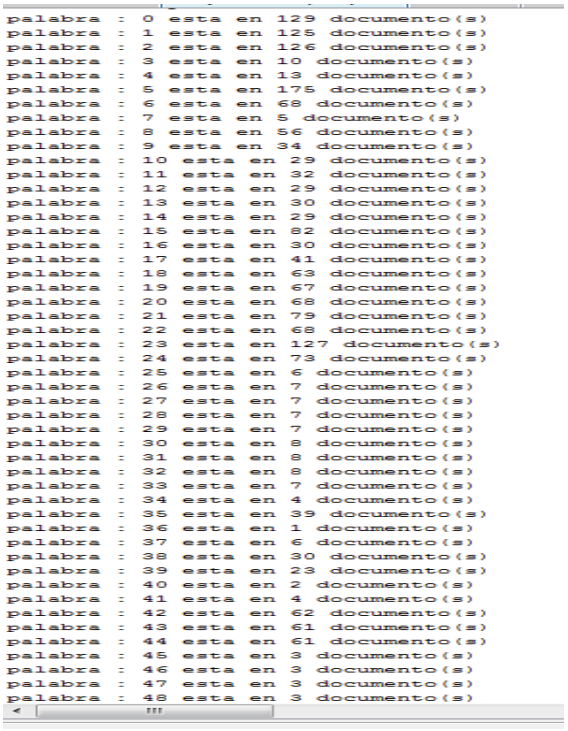
Procedimiento: Al ejecutar el programa se despliega en la consola La matriz de Referencia



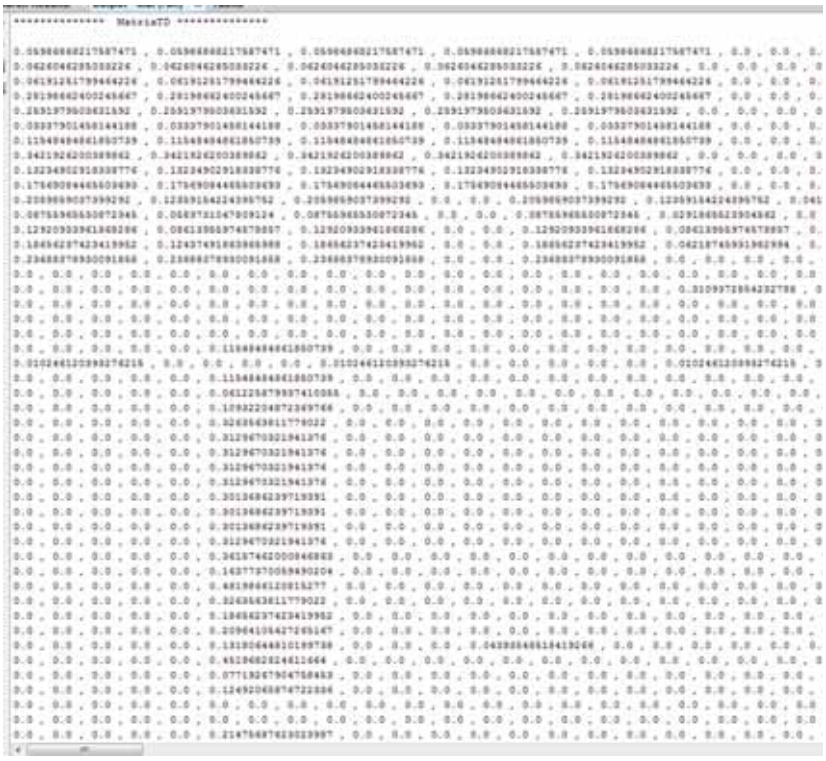
Después el número de palabras por documento.



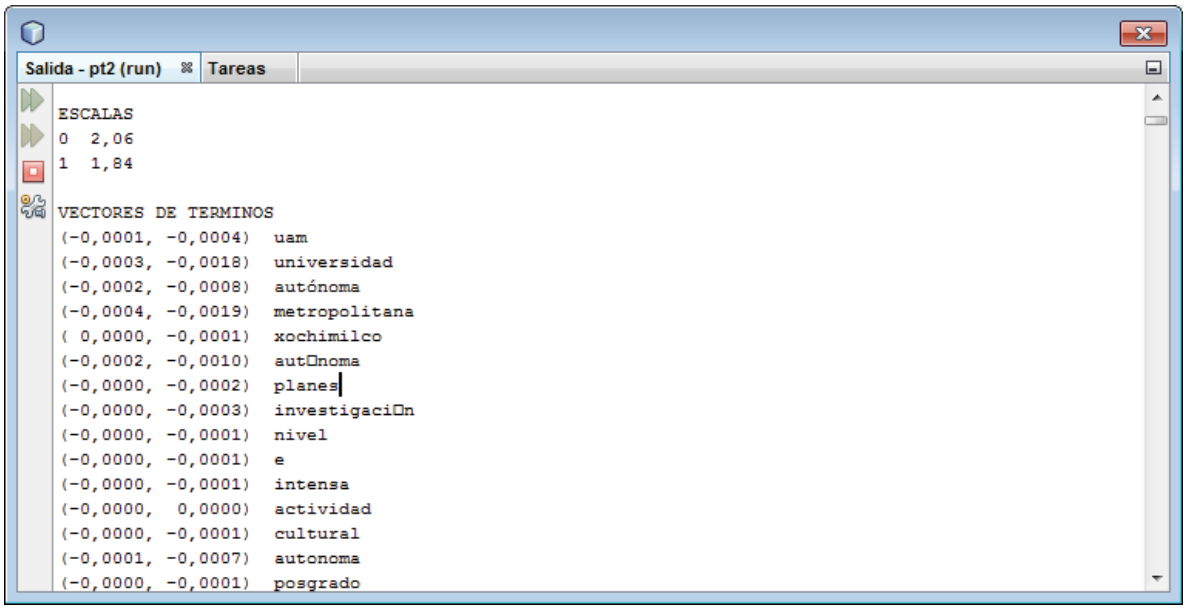
En cuantos documentos se encuentra x palabra

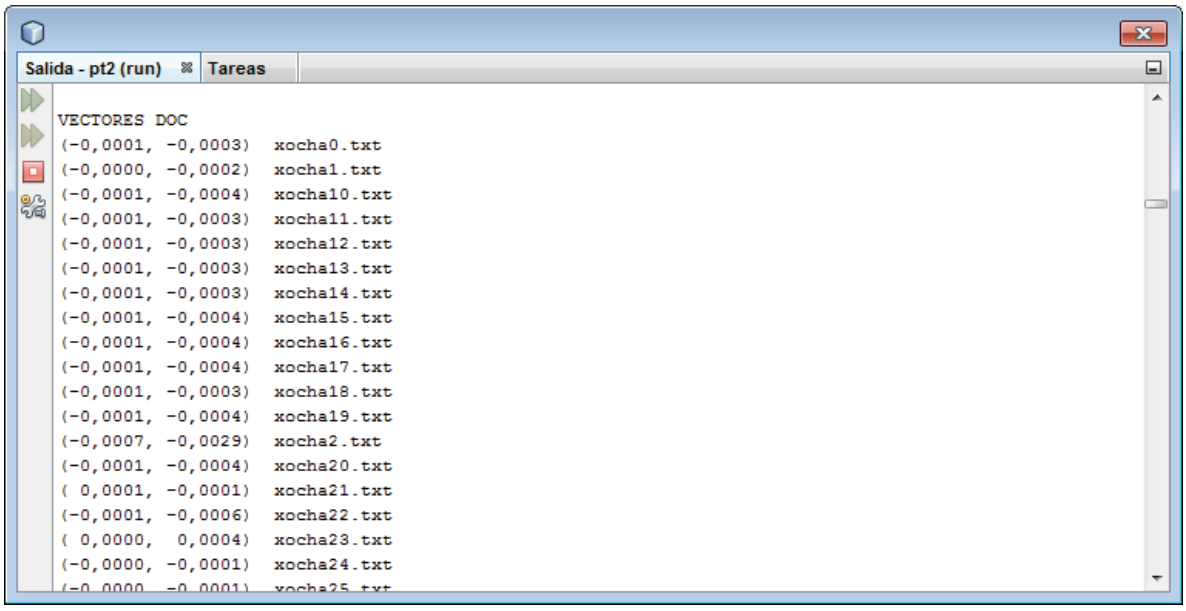


Con estos datos se genera la matriz de pesos…



La matriz de pesos es necesaria puesto que es un parámetro de entrada para el método que genera la SVD. Con esto se obtienen: Las escalas, los vectores de términos y los vectores de documentos.

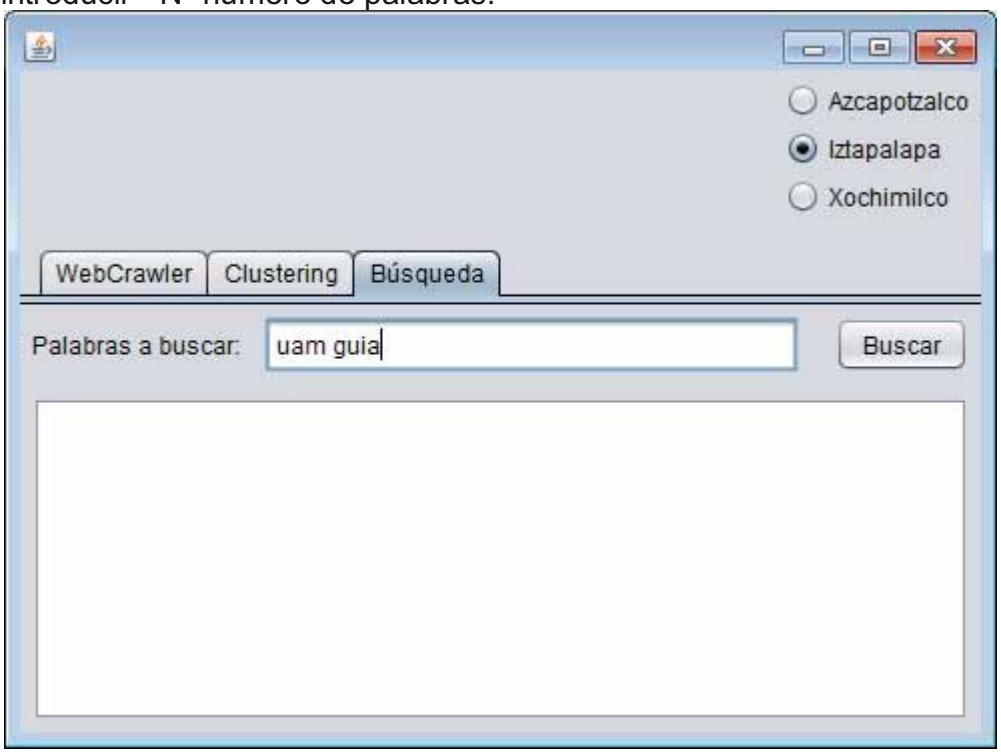




PASO 4: Diseñar e implementar una interfaz gráfica para visualizar e interpretar los patrones descubiertos.

Nombre de la prueba: Interfaz Gráfica

Procedimiento: Con los datos obtenidos, se optó por hacer un buscador para esto se debe elegir a que unidad se le desea hacer alguna búsqueda, se pueden introducir "N" número de palabras.



Por último se le da buscar y el resultado es una lista de las páginas Web a las que están relacionadas dichas palabras. Donde el primer link de la lista es el que tiene más relación con la búsqueda y asi sucesivamente.