Universidad
de La Laguna

Minería de texto para el análisis de la colaboración en SIENA

Text mining for collaboration analysis in SIENA

Eduardo Ezequiel Barrio Pareja
Departamento de Ingeniería de Sistemas y Automática y Arquitectura y Tecnología de Computadores
Escuela Técnica Superior de Ingeniería Informática

Trabajo de Fin de Grado

La Laguna, 6 de Septiembre de 2014

Dña. **Carina Soledad González González,** con N.I.F. **54.064.251-Z** profesor Titular de Universidad adscrito al Departamento de Ingeniería de Sistemas y Automática y Arquitectura y Tecnología de Computadores de la Universidad de La Laguna

Y

D. **Lorenzo Moreno Ruiz,** con N.I.F **50.400.691-D** profesor Titular de Universidad adscrito al Departamento de Ingeniería de Sistemas y Automática y Arquitectura y Tecnología de Computadores de la Universidad de La Laguna

C E R T I F I C A N

Que la presente memoria titulada:

"Minería de texto para el análisis de la colaboración en SIENA."

Ha sido realizada bajo su dirección por D. Eduardo Ezequiel Ezequiel Barrio Pareja, con N.I.F. 78.717.782-Y.

Y para que así conste, en cumplimiento de la legislación vigente y a los efectos oportunos firman la presente en La Laguna a 6 de Septiembre de 2014.

Agradecimientos

A la Profesora Dña. Carina Soledad González González, Profesora Titular adscrito al departamento de Ingeniería de Sistemas y Automática y Arquitectura y Tecnología de Computadores de la Universidad de La Laguna por su disponibilidad constante, su alta adaptabilidad y por encima de todo, su gran calidad humana.

Al profesor D. Lorenzo Moreno Ruíz, Profesor Titular adscrito al departamento de Ingeniería de Sistemas y Automática y Arquitectura y Tecnología de Computadores de la Universidad de La Laguna, por su buen humor, buen trato y paciencia conmigo.

A todo el grupo de desarrollo de SOBEK, en especial a Daniel Epstein, ya que sin su ayuda y colaboración en este proyecto el resultado hubiera sido otro.

A la empresa MHP Software, por las facilidades que me han dado para poder compaginar el trabajo con este proyecto.

A Beatrice Popescu, por haber estado siempre disponible para cualquier duda que me surgiera y permitirme trabajar sobre su aplicación, SIENA.

A Víctor Plaza por compartir sus conocimientos conmigo y este proyecto.

A Patricia Pareja, por su asesoramiento acertado y su buen hacer.

Resumen

Este proyecto surge a partir de dos proyectos anteriores, SIENA (Sistema Integrado de Enseñanza Aprendizaje), desarrollado por Beatrice Popescu, y SOBEK, desarrollado por la universidad de Rio Grande do Sul(Brasil).

El proyecto se basa en la ampliación de las actuales funcionalidades de la herramienta informática SIENA, en concreto en el análisis de chats grupales en la resolución de test evaluativos. Para ello se hará uso de la herramienta de minería de texto SOBEK, la cual nos permite, a partir de un texto, generar un grafo el cual muestra los conceptos más relevantes del texto, así como sus interrelaciones con otros conceptos, permitiendo al profesorado realizar una evaluación rápida y concisa de lo tratado en ese chat.

Palabras clave: Proyecto, herramienta informática, chats, test, minería de texto, grafo, conceptos, interrelaciones, usuarios, profesorado.

Abstract

This project springs from two previous projects, SIENA (Sistema Integrado de Enseñanza Aprendizaje), developed by Beatrice Popescu, and SOBEK, developed by Rio Grande do Sul University (Brazil).

The project is based on the improvement of the current functions of SIENA computing tool, in particular in the analysis of group chats in the resolution of evaluative test. The text mining tool SOBEK will be used, which allows the generation of a network from a text, that shows the most relevant concepts, and its relationships, allowing teachers to make a quick and concise evaluation about the chat matters.

Keywords: Project, computing tool, chats, evaluative test, text mining, network, concepts, relationships, teachers.

Índice General

Índice de figuras

Capítulo 1. Introducción

Antes de la invención de internet y la creación de la Web, la gran mayoría de las comunicaciones se realizaban de manera oral, no quedando constancia de ningún tipo de este intercambio de información. Esta situación ha cambiado drásticamente en las últimas dos décadas.

Debido a la gran revolución de Internet los datos de las conversaciones se están almacenando a un ritmo alarmante. Las personas y las organizaciones participan en estos intercambios de información, ya sea a través de correos electrónicos, reuniones virtuales, chats, mensajes de texto en blog o foros. Los avances en el procesamiento del lenguaje natural ofrecen amplias oportunidades para que estos documentos sean analizados [1].

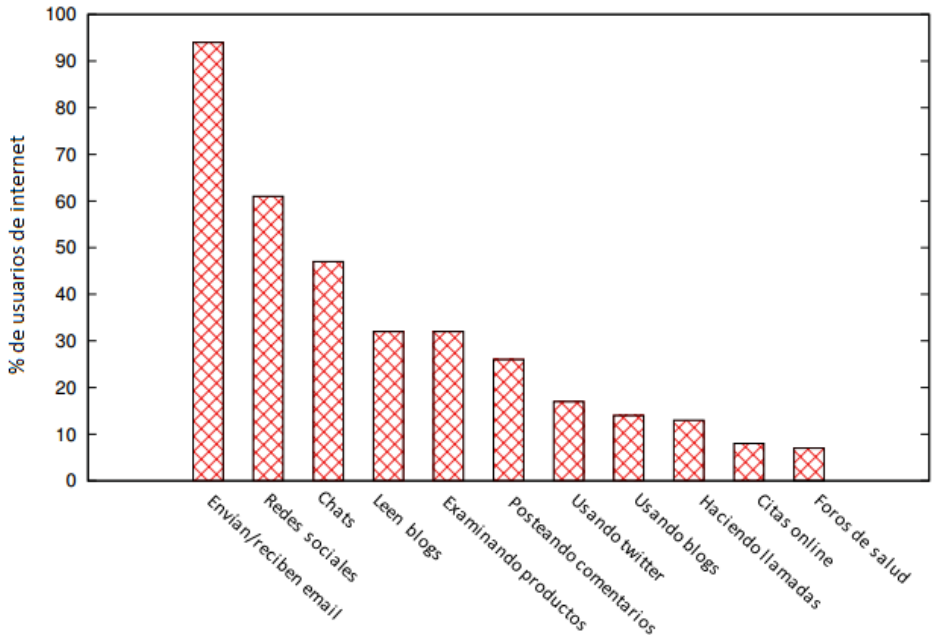

Figura 1. Estado actual de las conversaciones online.

1.1 Minería de datos en el Ámbito educativo.

En esta era digital las oportunidades que ofrece la Web permiten que ámbitos como el educativo experimenten grandes avances en cortos períodos de tiempo.

Esto ha motivado la creación de sistemas informáticos para la enseñanza, e-learning, así como una gran cantidad de herramientas y procesos. Estas herramientas y procesos aplicadas al e-learning buscan guiar a los estudiantes durante su aprendizaje para maximizarlo[2].

Uno de estos procesos es la minería de datos o descubrimiento de conocimiento a través del análisis de grandes volúmenes de datos.

Las principales aplicaciones de las técnicas de minería de datos en educación son:

- Sistemas de personalización. Se trata de obtener la mayor cantidad de datos posible de nuestros usuarios con el fin de mejorar el servicio que se les ofrece[3].

- Sistemas recomendadores. Su utilidad se basa en ofrecer al usuario distintos tipos de temas que son de su interés comparando su perfil con alguna de las características de los temas[4].

- Sistemas de modificación. Cuya base se asienta sobre la detección de necesidades, realizando las modificaciones necesarias para mejorar la experiencia del usuario[5].

- Sistemas de detección de irregularidades. Sistemas que se basan en la detección de errores.

La aplicación de estas técnicas de minería de datos en la educación puede ser observada desde dos principales puntos de vista:

- **Orientado hacia los autores**. Con el objetivo fundamental de proporcionar ayuda a los profesores/autores de los sistemas de e-learning para que puedan mejorar el funcionamiento y/o rendimiento de estos sistemas a partir de la información recabada a través de los usuarios. Sus principales aplicaciones son:

    - Obtener una mayor realimentación de la enseñanza.
    - Conocer más sobre cómo los estudiantes aprenden en la Web.
    - Evaluar a los estudiantes mediante sus patrones de navegación.
    - Poder reestructurar los sitios web con el fin de mejorar la experiencia educativa.
    - Clasificar a los estudiantes en grupos.

- **Orientado hacia los estudiantes**. Su principal objetivo es ayudar o realizar recomendaciones a los alumnos durante su interacción con el sistema de e-learning.

Es importante poner de manifiesto que, aunque el área que relaciona la minería de datos con el sector educativo es reciente, ya cuenta con un buen número de

investigadores y muestra de ello son las múltiples contribuciones publicadas en diferentes congresos a nivel internacional como EDM (Educational Data Mining), ICCE, ICALT, ITS, Elearn, PAKDD, GECCO, UM, AH, WISE, ISDE y revistas tales como IJEL, IEEE Education, UMUAI[6].

1.2 Minería de texto.

Poder descubrir conocimiento en grandes volúmenes de datos que se generan de todas estas conversaciones es un reto en sí mismo. Es por ello que nace la minería de texto[7], siendo el área de investigación más reciente del procesamiento automático de textos.

En efecto, se define como el proceso automático de descubrimiento de patrones interesantes en una colección de textos. Estos patrones no deben de existir explícitamente en ningún texto de la colección, y deben surgir de relacionar el contenido de varios de ellos.

El proceso de minería de texto consiste en dos etapas principales: una etapa de pre procesamiento y una etapa de descubrimiento. En la primera etapa, los textos se transforman a algún tipo de representación estructurada o semiestructurada que facilite su posterior análisis, mientras que en la segunda etapa las representaciones intermedias se analizan con el objetivo de descubrir en ellas algunos patrones interesantes o nuevos conocimientos[8].

1.3 SIENA

La herramienta informática SIENA (Sistema Integrado de Enseñanza Aprendizaje) que se encuentra dentro del ámbito educativo, es una aplicación web para detectar los conocimientos previos de un alumno o como ayuda para el autoaprendizaje y la autoevaluación, con objeto de realizar un aprendizaje centrado en el alumno (aprendizaje significativo).

Figura 2. Siena en la Universidad de La Laguna.

Esta herramienta web está pensada para trabajar con mapas conceptuales. El mapa conceptual, que será creado por el profesor, dispondrá los conceptos organizados en el mapa partiendo desde los conceptos objetivos hasta los conocimientos previos. Es decir, en un mapa dado, un concepto A aparecerá antes de un concepto B, si para entender el concepto B se precisa del conocimiento previo del concepto A[9].

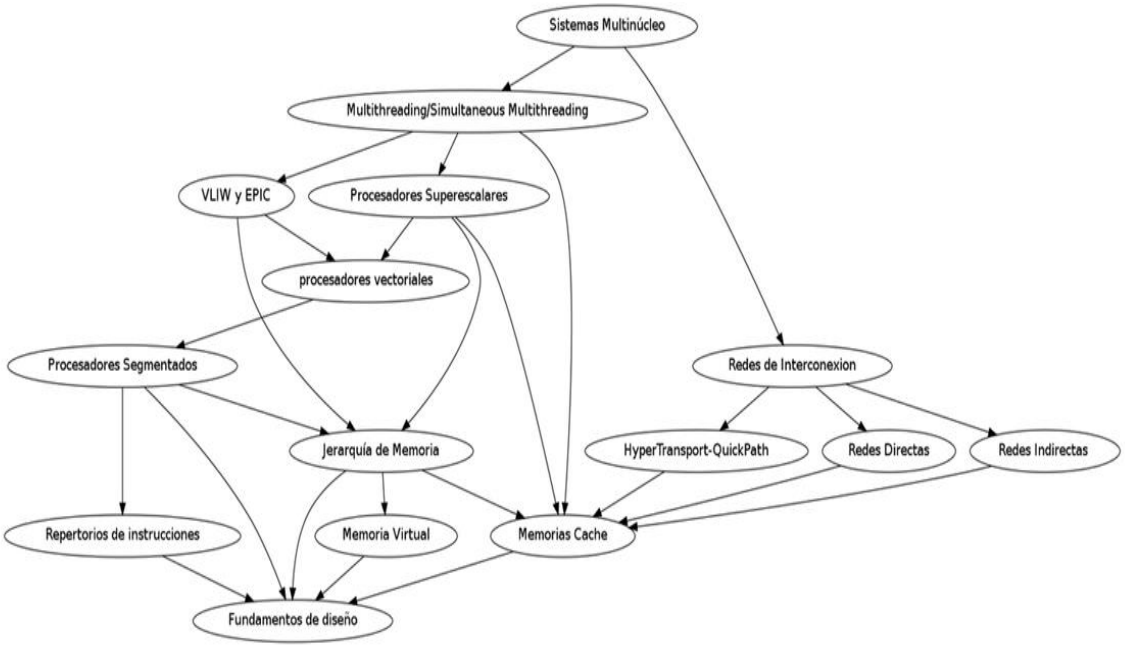

Figura 3. Mapa conceptual de la asignatura Arquitectura e Ingeniería de Computadores. ULL.

Los profesores son los encargados de la creación de tests evaluativos que posteriormente son asignados y realizados por los alumnos.

Nota: 0.821

#	Respuesta	Respuesta correcta	Tiempo(antes de que se acabe)	Pregunta	Dificultad / Adivinanza	Puntos antes	Puntos después	
0	0	false	139	¿Cual no es la estrategia de Administración de Memoria Virtual?	0.24 / 0.1	0.50000	0.21053	alu191
1	0	true	1199	Na figura temos dois semicírculos de diâmetros PS, de medida 4, e QR, paralelo a PS. Além disso, o semicírculo menor é tangente a PS em O. Qual é a área destacada?	0.18 / 0.15	0.21053	0.59313	alu191
2	0	true	118	siena4	0.54 / 0.1	0.59313	0.87023	alu191
3	2	true	138	Memoria Segmentada	0.24 / 0.1	0.87023	0.98076	alu191
4	0	false	538	A reta que passa no ponto de intersecção das retas -2x-y+3=0 e x-y+3=0 e pelo ponto (-2,-1) é	0.27 / 0.2	0.98076	0.94506	alu191
5	1	false	139	¿Cual esquema NO es de los 3 esquemas principales de traducción memoria virtual - memoria física:?	0.24 / 0.1	0.94506	0.82101	alu191

Figura 4. Ejemplo de test en Siena.

La realización de estos test por parte de los alumnos puede ser realizada tanto de manera individual como de manera colaborativa.

Estos test colaborativos cuentan con un chat para que los alumnos se comuniquen entre ellos, de manera que pueden estar respondiendo las preguntas conjuntamente a la vez que se resuelven dudas o se transmiten los conocimientos necesarios para poder resolver estas preguntas a través de estos chats.

From	Body	Question	Keywords	Tiempo	Comentario	Borrar
Rubén	creo que es de salto condicional, pero no tengo claro si tambien podria ser de inmediato	2		17:59:02	(Change)	Borrar
Rubén	por que el inmediato podria estar en la direccion de 16 bits	2		17:59:23	(Change)	Borrar
Sandro	y siendo inmediata para que te dan la direccion?	2		17:59:26	(Change)	Borrar
Rubén	pero claro no seria inmediato entonces no?	2		17:59:34	(Change)	Borrar
Rubén	voy a poner salto condicional solo	2		18:00:10	(Change)	Borrar

Figura 5. Ejemplo test/chat SIENA

1.4 Justificación

Dado que el número de usuarios en las universidades españolas aumenta año tras año, y ya eran cerca de un millón y medio en dos mil trece, parece importante dar más facilidades a los profesores para analizar de manera rápida y eficiente las tareas, trabajos, test y demás formatos evaluativos.

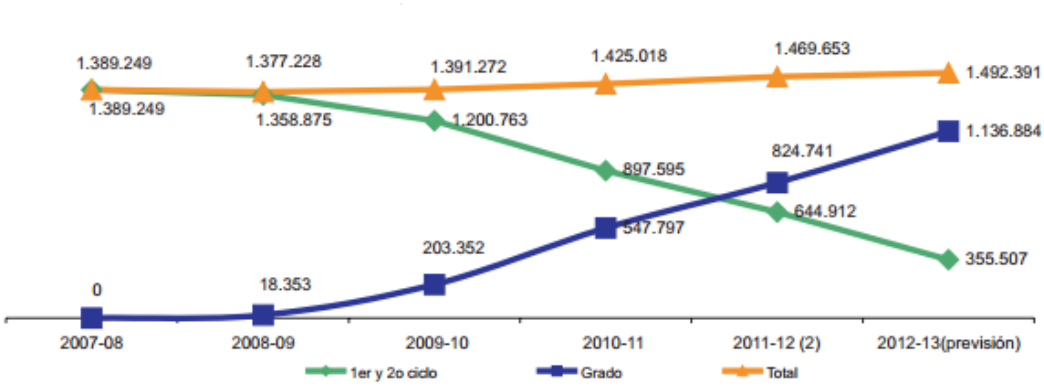

Figura 6. Evolución alumnado universitario en España.

Y contando con la gran cantidad de posibilidades que ofrece la informática en el entorno educativo, con la ayuda de esa gran herramienta informática de este mismo entorno, que es SIENA, nos proponemos realizar un aporte a esta misma herramienta.

1.5 Ideas y Objetivos

Se pretende ampliar la oferta de servicios que actualmente están disponibles en la herramienta informática SIENA de ámbito educativo que se encuentra actualmente operativa en la Universidad de La Laguna.

La parte que nos ha parecido más interesante de analizar es la correspondiente a los test evaluativos que se realizan de manera colectiva, es decir, por más de un alumno.

El servicio que se implementará permitirá al profesorado analizar los test evaluativos con los que cuenta actualmente SIENA. De esta manera, se necesita poder realizar un análisis de los mensajes que intercambian los alumnos para cada una de las preguntas que se les realiza en test.

Este servicio debe tener las siguientes características para que su aplicación y uso sea lo más eficiente posible:

- Debe ser usable, de manera que el profesorado no requiera de ningún curso de instrucción acerca de su uso.

- Debe ser rápido, de manera que no se tarde más tiempo usando este servicio que realizando la

misma tarea de forma manual o usando aplicaciones exteriores a SIENA.

- Debe estar totalmente integrado en SIENA, respetando en todo momento la estructura que actual del proyecto y realizando únicamente los cambios indispensables para su uso.

- Debe mejorar la rapidez y eficiencia de la tarea evaluativa del profesorado.

- Debe tratarse de una herramienta de minería de texto, capaz de mostrar la información transmitida por los alumnos. De esta manera el profesorado puede realizar una evaluación de manera rápida y visual de los contenidos transmitidos.

- Este servicio debe contar con algún tipo de respaldo o garantía de buen uso. De esta manera el funcionamiento de la herramienta debe estar en parte avalado por la entidad desarrolladora.

Capítulo 2. Estado actual de herramientas de Text Mining

A continuación se analizarán diferentes herramientas de minería de texto con el fin de comprobar si es posible hacer uso de alguna de ellas en el ámbito educativo, para ello esta herramienta debe contar con las siguientes características:

- Ser de libre uso, u Open-Source, de manera que no se tenga que abonar una licencia.
- Debe ofrecer la posibilidad de usarse en la Web, no sólo como aplicación de escritorio.
- Se debe poder interactuar con la herramienta de una manera sencilla y rápida.
- Debe poderse adherir a la estructura actual de SIENA.
- Debe de ser eficiente y no consumir demasiados recursos.
- Debe de estar adaptado (o poderse adaptar) a un entorno Web, donde los datos de entrada puedan variar en cada petición del análisis.

Con los anteriores requisitos se han contemplado las siguientes herramientas de minería de texto:

- ***Authonomy:*** Este software ofrece herramientas para realizar text mining, clustering y categorización a través de la búsqueda y el procesamiento de texto tomando datos estructurados e información humana no estructurada. Es de pago y pertenece a la compañía HP.

- ***SAS Text Analytics:*** Este software se utiliza para descubrir patrones y tendencias en cualquier texto. Esta herramienta se utiliza tanto para la minería de textos como para el análisis de sentimientos y clasificación de contenidos. Es de pago y pertenece a la compañía SAS[10].

- ***SOBEK:*** Esta herramienta está diseñada para el ámbito educativo. No es open-source, pero su uso es libre y gratuito. Permite ver los conceptos más relevantes de un texto de manera rápida y visual.

- ***PolyAnalist:*** Esta herramienta permite destilar el significado de un texto de forma concisa, ver resúmenes de textos, navegar de forma eficiente a través de grandes bases textuales. Para poder hacer uso de la herramienta se requiere de una licencia de pago.

- ***WordStat:*** Esta herramienta permite trabajar con grandes volúmenes de datos. Permite el análisis de contenido, inteligencia de negocio y análisis competitivo de sitios web. Es capaz de procesar hasta veinte millones de palabras por minuto e identificar conceptos. Para poder hacer uso de la herramienta se requiere de una licencia de pago. Pertenece a la compañía Provalis Research[11].

- ***Attensity Analyze:*** Esta herramienta permite analizar conversaciones online y redes sociales. Analiza y monitorea conversaciones en foros, blogs, sitios de discusión. Para poder hacer uso de la herramienta se requiere de una licencia de pago. Pertenece a la compañía Attensity[12].

- ***Orange:*** Herramienta de minería de datos. Tiene disponible un add-on para minería de texto. Es open-source, interfaz amigable y escrita en Python[13].

Capítulo 3. Elección de herramienta de minería de texto

Tras haber analizado las herramientas anteriormente expuestas, la herramienta que hemos seleccionado para realizar estas tareas ha sido **SOBEK**, ya que cuenta con las siguientes ventajas:

1. Se enfoca en el ámbito educativo, dado que nace con esta intención. Esto ofrece la ventaja de que está pensada para ofrecer ayudas a este sector en concreto.

2. Actualmente cuenta con un grupo de desarrollo. Es una aplicación que sigue en constante evolución, por lo que las mejoras que se realicen en esta herramienta serán fácilmente integrables en futuras modificaciones.

3. Se ha presentado en congresos de renombrado prestigio; *World Congress in Computer Science, Computer Engineering, and Applied Computing, 2011*, así como en *International Journal of Computer Information Systems and Industrial Management, 2011*.

4. Su uso es libre, de manera que no hay que abonar licencias de ningún tipo.

5. Pueden exportarse los resultados obtenidos en formato .xml de manera que pueden visualizarse de manera posterior en otros equipos.

6. Se puede interactuar con la aplicación, de manera que no se limita a ofrecer resultados, sino que permite que los usuarios realicen modificaciones sobre los resultados obtenidos por la herramienta.

7. No se requiere una configuración de los equipos donde se pretenda utilizar.

8. Funciona en los navegadores web más comunes (Google Chrome, Mozilla Firefox, Internet Explorer y Opera).

9. Permite realizar minería de texto independientemente de la longitud del mensaje.

10. Cuenta con una versión de escritorio, cuyos resultados son los mismos que la versión Web. De esta manera, dado que es una herramienta conocida en el ámbito educativo, puede resultar familiar al personal docente.

11. Dado que SIENA está pensada para trabajar con mapas conceptuales, es interesante que la funcionalidad que aporte SOBEK tenga similitud con el formato de uso que plantea SIENA.

Capítulo 4. SOBEK

SOBEK es una herramienta de minería de texto capaz de identificar los conceptos relevantes en un texto a partir del análisis de los términos y sus relaciones y propone aplicarse en el entorno educativo.

La principal diferencia de la minería de datos y de texto se puede observar de manera meridiana en SOBEK. Mientras que la minería de datos realiza sus búsquedas y análisis en bases de datos de registros formales, la minería de texto emplea fuentes de datos no estructuradas.

De esta manera SOBEK es capaz de encontrar los conceptos más relevantes en un texto y su interrelación con otros de estos conceptos.

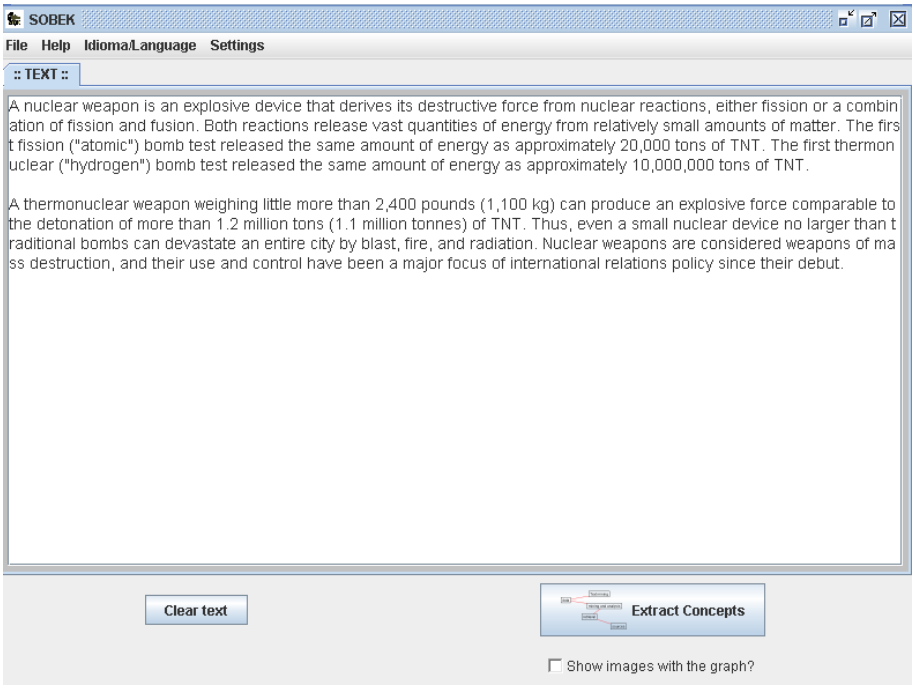

Figura 7. Interfaz de Sobek Escritorio.

El método utilizado por SOBEK para realizar esta minería de texto se basa en el modelo *"n-simple distance graph model"*, en el que los nodos representan los términos más importantes que se encuentra en el texto y líneas que enlazan estos nodos muestran la información de adyacencia asociada a cada término. Por lo que en conjunto este gráfico muestra la manera en que se relacionan los términos más importantes del texto analizado[14].

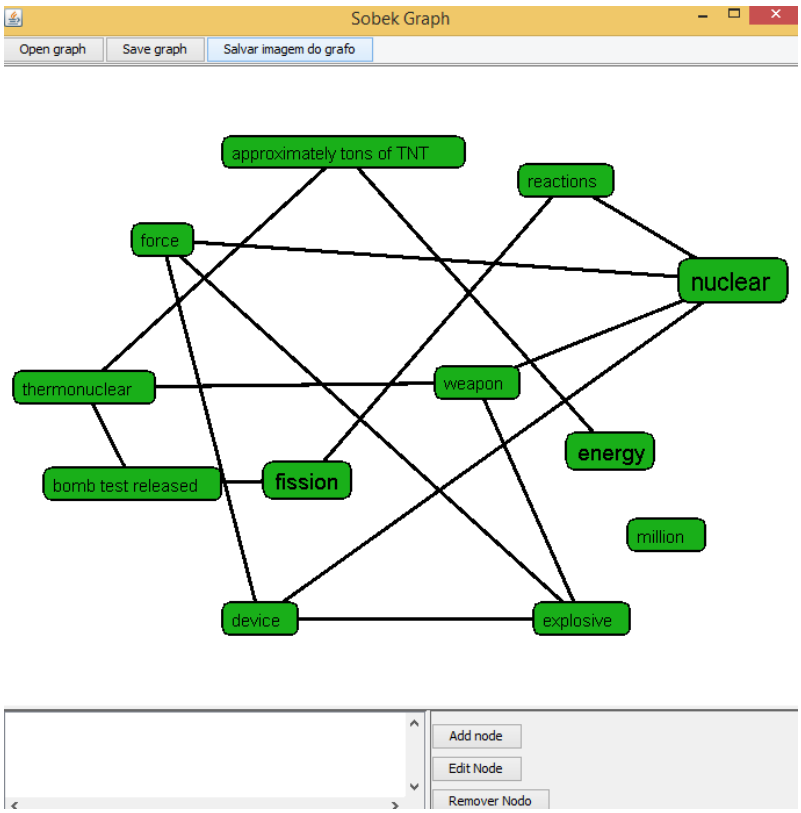

Figura 8. Ejemplo de grafo, por SOBEK.

De esta manera se obtiene una visualización rápida y concreta de lo expuesto en el texto. Para poder analizar qué palabras tienen más relevancia que otras (mayor número de repeticiones), se muestran en un tamaño de letra mayor y un color más oscuro, para su

mejor distinción, así como una descripción del número de veces que aparece cada término y su disposición en el texto.

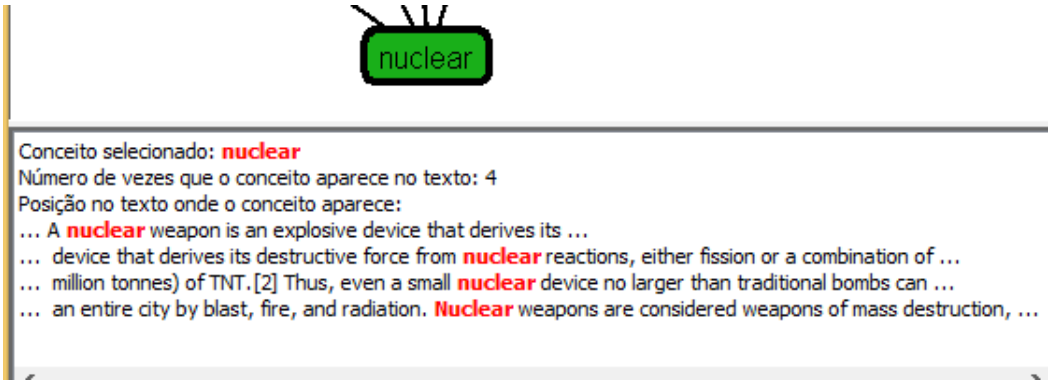

Figura 9. Número de repeticiones en SOBEK.

Mientras que otros enfoques de la minería de texto se basan en el análisis morfosintáctico de patrones relevantes (como "Sustantivo Sustantivo", "Sustantivo Preposición Sustantivo", "Adjetivo Sustantivo", etc.), SOBEK se basa en un método más simple, la frecuencia en que aparecen los términos en el texto[15].

Se establece una frecuencia mínima Θ, el cual indica el número mínimo de ocurrencias de un concepto para que se muestre en el gráfico final. Por lo que un concepto que se presente un número de veces menor que el valor de Θ no se mostrará en el gráfico.

Seguidamente se utiliza un analizador lingüístico para eliminar las inflexiones y las terminaciones de las palabras con el fin de reducir las palabras a una raíz común.

El método del que hace uso SOBEK se basa en un parámetro *n* para extraer los conceptos compuestos por más de una palabra. Según éste parámetro se crean combinaciones de la palabra actual con las palabras *n*

posteriores. De esta forma lo que se trata es de crear una amplia combinación de palabras, encontrando así el grupo más frecuente de palabras que aparecen en el texto.

De esta manera la secuencia *"AA BB CC DD EE"* con *n=3* generaría las siguientes combinaciones:

- *AA, AA BB, AA BB CC.*
- *BB, BB CC, BB CC DD.*
- *CC, CC DD, CC DD EE.*
- *DD, DD EE.*
- *EE*

Con el fin de evitar secuencias a partir de artículos y preposiciones se utilizan filtros específicos de palabras (STOP Words).

Después de haber identificado las combinaciones más frecuentes de palabras, las cuales pasamos a llamar conceptos, el siguiente paso es calcular la similitud entre estos conceptos.

Consideremos los conceptos *x="AA DD BB" y z="BB CC DD EE FF AA"*. Se calcula el coeficiente de similitud entre ellos, que es el producto escalar. Este coeficiente de similitud, que es representado por *S*, calcula la cantidad de palabras presentes en ambos conceptos, representado por *P*, y el número de palabras del concepto más amplio, representado por *B*. Por lo que tenemos:

$$P = 3.\ B = 6.$$

$$S = P / B.$$

$$S = 1/2, \text{ en este caso.}$$

Luego de haber calculado S, se calcula el coeficiente de relevancia, *R*, que se calcula para cada concepto. En el proceso de cálculo de *R*, se utilizan el número de palabras del concepto, *C*, y la frecuencia absoluta, *F*, aparte de *S*, usando la siguiente fórmula.

$$R = S * C + F.$$

El concepto con el valor de R mayor se mantiene de base y al final del proceso se incluye en el gráfico. De esta manera y considerando que *F(x)=3 y F(z)=2*, el concepto que formará parte del grafo será el *z*, aunque aparezca menos veces, puesto que la idea es mantener el concepto que *"dice más"*[16][17].

Posteriormente se analiza la relación entre los conceptos que forman el gráfico. Para ello, tenemos que saber qué conceptos vienen antes y después de este concepto en el texto. De esta manera si un concepto es "vecino" de otro (aparece antes o después en el texto), se aumentará un contador que almacena estas ocurrencias, así, sólo se mostrarán las interrelaciones entre conceptos con las más altas ocurrencias de "vecindad".

Dado que un grafo completamente conectado no proporciona ninguna información acerca de cómo cada concepto está relacionado con el siguiente, se tiene un número máximo de posibles conexiones Ω.

El número de conexiones para cada concepto es llamado Con_i. El número de apariciones del concepto en el texto es llamado $NumOc_i$. El número del concepto con mayores apariciones es llamado MaxOc. Así, el número de conexiones para cada concepto es calculado haciendo uso de la siguiente fórmula:

$$Con_i = \frac{NumOc_i * \Omega}{MaxOc}$$

Capítulo 5. Planteamiento de Procedimiento

5.1 Restricciones Iniciales.

Tras realizar un análisis en conjunto del proyecto, las restricciones iniciales que se han encontrado, son las siguientes:

- SIENA se encuentra desarrollado bajo el lenguaje de programación Ruby, más en concreto bajo el framework Ruby on Rails, por lo que el desarrollo del proyecto tendrá que tener en consideración este factor.

- Dado que SIENA se encuentra desplegado en la WEB, la nueva funcionalidad a implementar debe estar pensada para este aspecto.

- Las restricciones por parte de la herramienta SOBEK son desconocidas, por lo que se analizarán las restricciones encontradas en SOBEK luego de su implementación en SIENA.

En primer lugar, se analizará el funcionamiento y el código de SIENA para ver hasta qué punto pueden realizarse las mínimas modificaciones necesarias para que la implementación de SOBEK sea posible.

Se analizará el estado actual de SIENA, de manera que debe ser posible acceder a la información referente a los mensajes de los usuarios de una manera fácil y eficiente.

5.2 Planteamiento Inicial

Se entablará contacto con los desarrolladores de SOBEK, comunicándoles nuestra intención de utilizar su software para fines educativos.

Se analizará la viabilidad de implementación de la herramienta de minería de texto SOBEK en la plataforma SIENA.

Dado que vamos a obtener un beneficio con la implementación de SOBEK en SIENA, se propondrá al equipo de SOBEK algún tipo de ayuda que permita mejorar de alguna manera su proyecto.

En resumen, se implementará SOBEK en SIENA para poder realizar así minería de texto en los chats de los tests colaborativos para que el profesor de manera rápida y visual pueda saber qué conceptos se transmiten los alumnos, facilitando así la evaluación de estos chats.

Capítulo 6. Descripción del trabajo realizado.

6.1 SIENA

Se ha analizado SIENA para comprender cómo está estructurada la herramienta. De esta manera nos hemos encontrado con la siguiente información relevante para el proyecto:

- SIENA ha sido desarrollada por Beatrice Popescu, la cual, amablemente, nos ha facilitado el código de la herramienta para poder trabajar con ella. También nos ha proporcionado un backup (copia de seguridad de base de datos) para que contáramos con datos reales con los que poder trabajar y hacer comprobaciones.

- SIENA se encuentra desarrollada en Ruby, bajo la versión 1.8.7, aunque puede utilizarse con versiones superiores. Hace uso del framework Ruby on Rails en su versión 4.0.4.

- Se ha hecho un despliegue de SIENA de manera local con el fin de poder trabajar con una copia de manera sencilla.

- Todo el proceso de desarrollo de este proyecto ha sido llevado a cabo bajo control de versiones, haciendo uso del software de control de versiones Git, bajo la plataforma Github. En enlace al

repositorio se encuentra en la página web: https://github.com/EduardoBarrio/TFG-SIEBEK

- Dado que se contaba con una copia de seguridad de una base de datos real se ha hecho uso del sistema de gestión de bases de datos(SGBD) PostgreSQL. Nos hemos decantado por este SGBD dado que su código fuente está disponible libremente y su uso es gratuito[18].

- La información referente a los chats realizados por los alumnos está totalmente accesible, de manera que se pueden tratar de una forma sencilla, prácticamente sin modificar la estructura de la herramienta SIENA.

6.2 SOBEK

6.2.1 Contacto con el equipo de SOBEK

Dado que SOBEK ha sido desarrollado por la Universidad de Rio Grande do Sul, hemos tenido que ponernos en contacto con su grupo de desarrollo, para explicarles las intenciones de este proyecto.

El contacto con el equipo de SOBEK se ha realizado a través de uno de los desarrolladores del proyecto SOBEK, Daniel Epstein.

Los primeros mensajes por parte de este proyecto trataron sobre una posible colaboración, de manera que este proyecto pudiera hacer uso de su herramienta SOBEK para un uso educativo. Todas estas comunicaciones se realizaron mediante correo electrónico.

Desde un primer momento el entorno de desarrollo de SOBEK se mostró receptivo a una colaboración con este proyecto. Comentaban que estaban participando en otros proyectos similares, por lo que contaban en la actualidad con un applet, escrito en lenguaje Java, (en el cual siguen trabajando y desarrollando), que se encontraba operativo y disponible para su uso.

Así, se decidió por parte de SOBEK y el entorno de este proyecto que SOBEK proporcionaría el applet anteriormente citado y el entorno de este proyecto realizaría la traducción, al castellano, de la lista de palabras las cuales utiliza SOBEK para realizar sus funciones.

A parte, nos ha parecido interesante aportar un poco más al proyecto SOBEK, de esta manera se ha incluido un *"feedback"* para la mejora de SOBEK.

6.2.2 Feedback a SOBEK.

Este documento se incluye como anexo al proyecto.

6.2.3 Analizando el applet.

De esta manera pusieron en nuestras manos el applet que posteriormente se implantaría en SIENA. Este applet en primera instancia no estaba firmado (su licencia de firma había caducado), pero estaban en fase de volverlo a firmar.

Como acercamiento se realizó un pequeño script en php en el cual se comprobó que el applet cumplía con los requisitos que se esperaban de él para este proyecto.

6.3 Implementación

Una vez analizados tanto SIENA como SOBEK se ha realizado un pequeño boceto, modificando lo menos posible la estructura ya creada de SIENA, de cómo debería ser la interfaz de análisis de chats de SIENA, con SOBEK implementado.

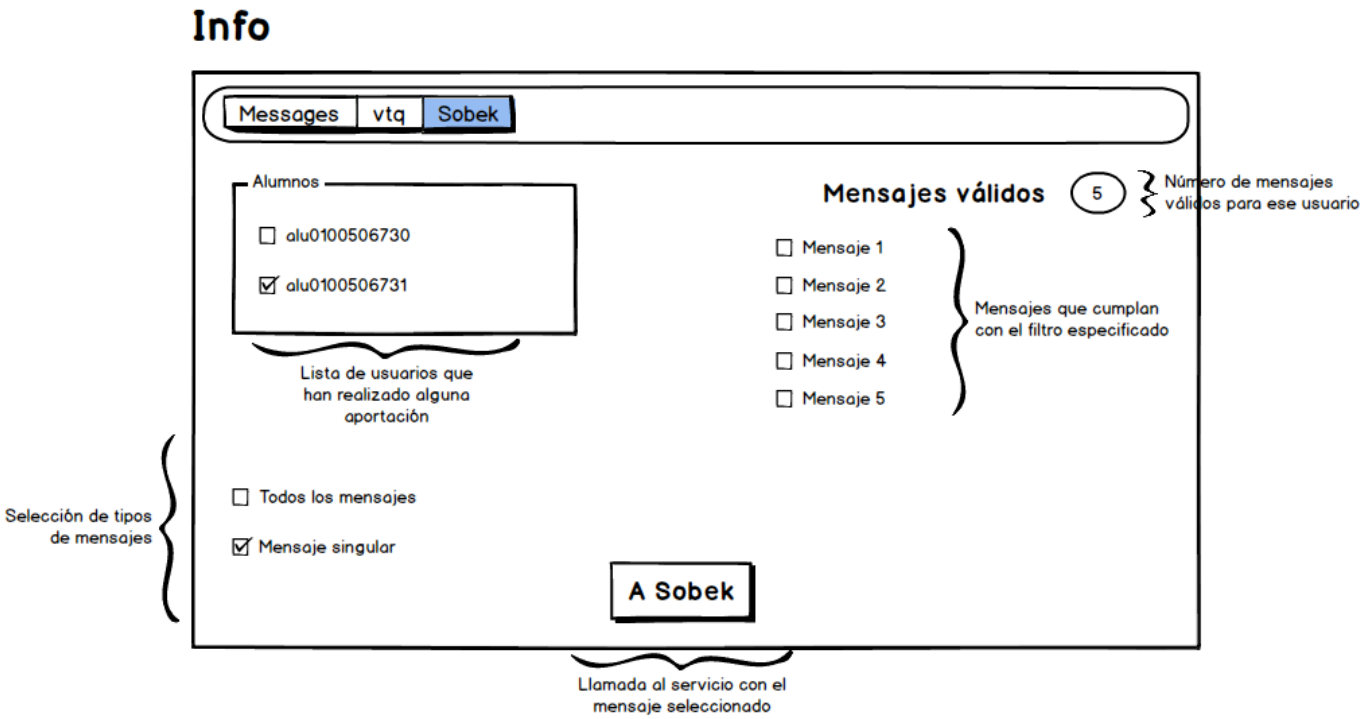

Figura 10. Boceto inicial Sobek en Siena.

De esta forma se ha realizado la estructura de la interfaz de SOBEK en SIENA:

- En esta interfaz aparecen los usuarios que participan en el chat del test. De esta manera sólo aquellos usuarios que realicen aportaciones vía chat aparecerán en esta interfaz.

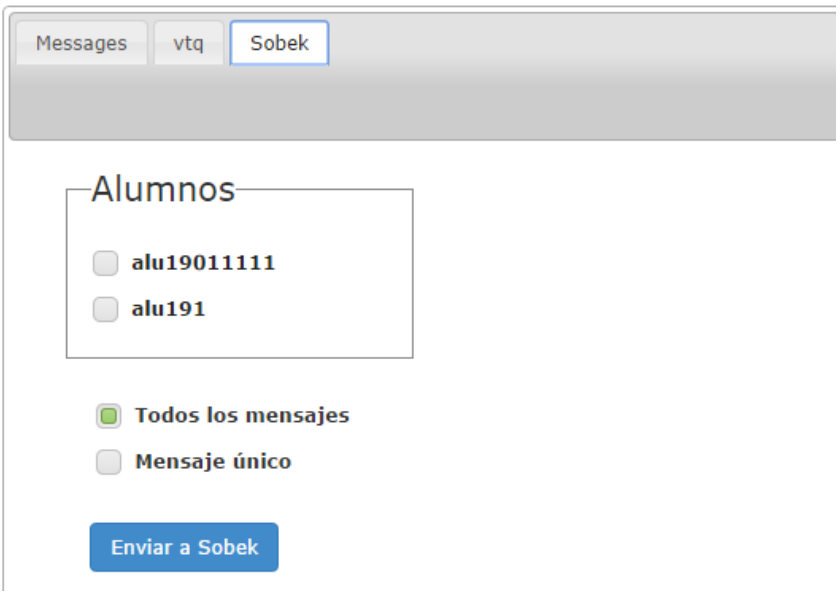

Figura 11. Alumnos en SIENA.

- Aquellos usuarios que hayan participado en el test, pero no hayan aportado información al chat quedarán excluidos, ya que se entiende que no aportan información relevante de ser analizada.

- Los chats se mostrarán de manera individualizada para cada uno de los usuarios. De esta forma, el profesorado podrá realizar consultas acerca de usuarios individuales.

- Cada chat estará asociado a una pregunta del test que se ha realizado. Gracias a esto podemos seleccionar todos los mensajes de una misma pregunta para un usuario en particular. Se permite hacer una selección rápida por preguntas. Las preguntas tendrán un formato diferente a los mensajes para su mejor reconocimiento.

Figura 12. Mensajes por preguntas en SIENA.

- Se permite que se envíen todos los mensajes de un mismo usuario, de manera que se puede comprobar de manera rápida todo lo que ha aportado al chat. Para ello basta con seleccionar el usuario y la opción "Todos los mensajes".

- Se permite que se envíen todos los mensajes de todos los usuarios que han participado en el chat. Con ello se consigue ver toda la información que se ha transmitido en ese test. Para ello se deben de seleccionar todos los usuarios que han participado en el chat y la opción, "Todos los mensajes".

Figura 13. Analizando todos los mensajes de los usuarios.

- Se permite seleccionar los mensajes de diferentes tests de manera conjunta.

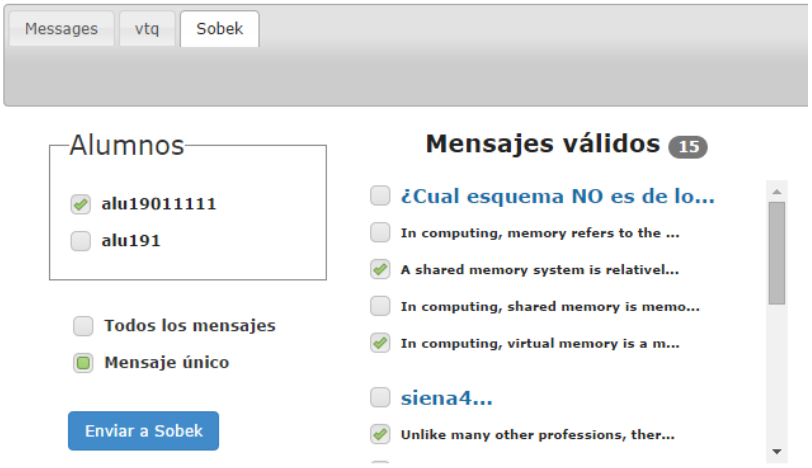

Figura 14. Mensajes de diferentes preguntas.

- Sólo aquellos mensajes cuya cantidad de palabras sea mayor de cinco aparecerán en esta interfaz, puesto que de esta manera al menos el 40%(Ver

figura 15) de los chats que se generan en SIENA podrán ser analizados. Se entiende que los mensajes con menos de cinco palabras no aportarán demasiada información, por lo que no tiene sentido que sean analizados.

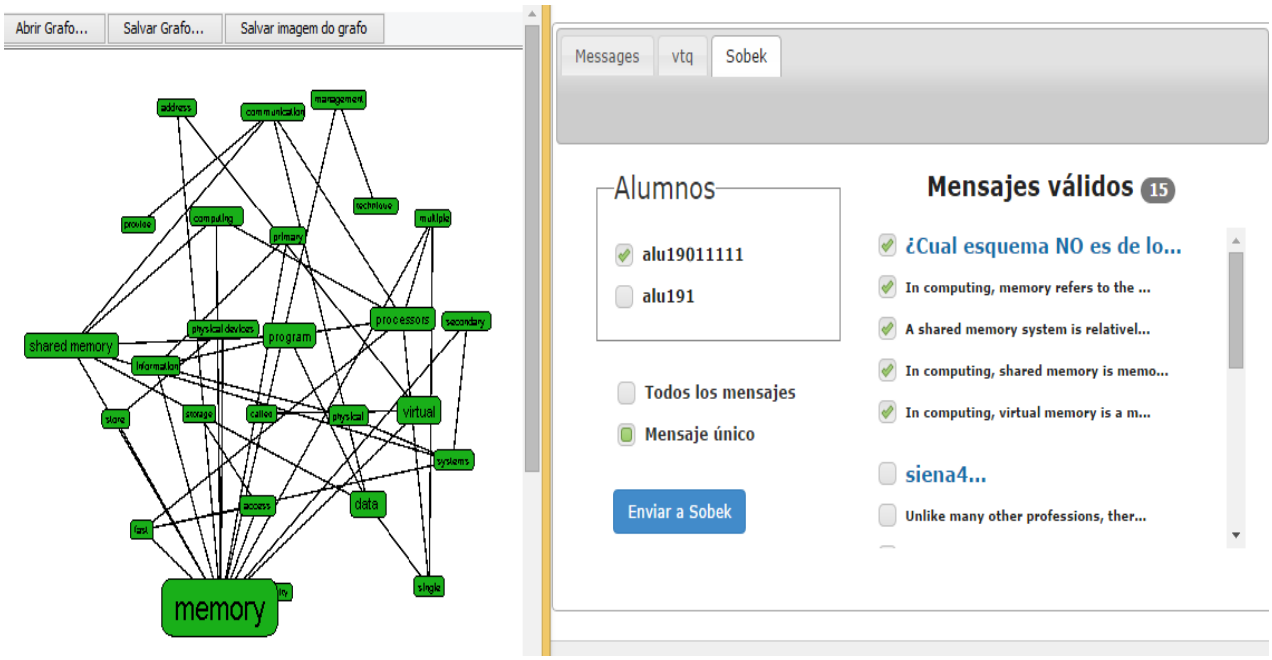

Figura 15. Ejemplo de SOBEK en SIENA.

Dado que el applet está pensando para textos cuyas dimensiones son bastante mayores que los que nos encontramos en un chat, el número de repeticiones para que un concepto aparezca en el gráfico es de tres.

Por ello, hemos decidido duplicar el mensaje, con el fin de poder extraer más información que si lo hiciéramos con el texto sin duplicar.

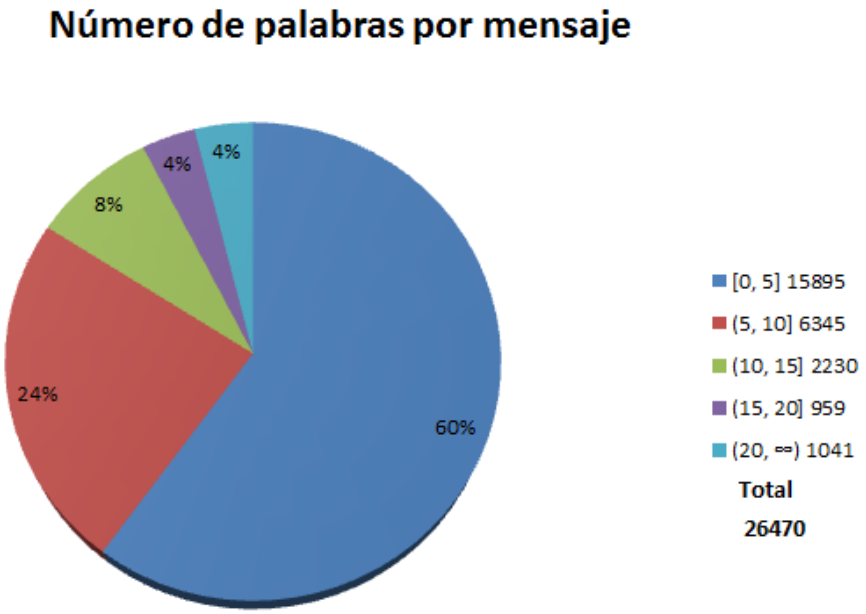

Figura 16. Número de palabras por mensaje (Datos obtenidos del análisis del backup de SIENA).

6.4 Limitaciones encontradas

Las restricciones finales que se han encontrado en el proyecto han sido las siguientes:

- Las propias restricciones impuestas por el applet. Dado que la opción que nos proporciona SOBEK es hacer uso de un applet, hay que adaptar el proyecto a esta restricción.

- El applet se encuentra disponible en dos lenguajes; portugués e inglés, de manera que actualmente su funcionamiento óptimo sólo está disponible para estos lenguajes.

- Las propias restricciones de SIENA, ya que el trabajo que se realice ha de hacerse sobre esta plataforma.

Como se puede observar son muchos los condicionantes de este proyecto, dado que las dependencias externas son fuertes.

Capítulo 7. Conclusiones

Durante el desarrollo de este proyecto se ha realizado un análisis de varias herramientas de minería de texto, así como el análisis de diferentes algoritmos. Se ha aportado a la herramienta SIENA una nueva funcionalidad, la cual permite realizar un análisis rápido y visual de chats colaborativos. Esta nueva funcionalidad ha sido posible gracias a la implantación de una herramienta de minería de texto, SOBEK.

Para poder contar con SOBEK se ha tenido que entablar conversaciones con sus desarrolladores, de esta manera se ha estado tratando con diferentes equipos (SIENA y SOBEK) en paralelo. Toda la comunicación con los desarrolladores de SOBEK se ha realizado en inglés y español.

Como aportación a SOBEK se ha realizado la traducción al castellano de los términos que utiliza SOBEK para su funcionamiento, además de un *feedback*, con el fin de la continua mejora de la herramienta.

Capítulo 8. Summary and Conclusions

During the development of this project an analysis of several mining text tools has been made. A new function for SIENA has been implemented, allowing a quick and visual analysis of collaborative chats. This new function is based on the implementation of SOBEK text mining tool.

In order to be able to use SOBEK, conversations with its developers have been established, discussing with different teams (SIENA and SOBEK) simultaneously. The languages used with SOBEK developers were English and Spanish.

As a contribution to SOBEK, a Spanish translation of the terms used by SOBEK has been done, as well as a feedback communication, for the future improvement of the tool.

Capítulo 9. Trabajos futuros

Una vez realizado este proyecto y viendo las posibilidades que ofrecen las herramientas SIENA y SOBEK al mundo educativo, algunas de las posibles líneas de trabajos futuros en al ámbito educativo podrían ser las siguientes:

- Implantación de SOBEK en otras plataformas, con el fin de realizar resúmenes de textos, trabajos, redacciones… Con ello se conseguiría un agilizamiento del proceso de evaluación por parte del profesorado.

- Evaluaciones automáticas de textos en SIENA mediante el análisis de los grafos resultantes del uso de SOBEK. Con esto se conseguiría que fuera la propia SIENA la que de manera automática ofreciera al profesorado una evaluación del alumnado.

- Sería bastante interesante poder realizar un análisis de interacción entre los usuarios, parecido a un SNA, de manera que se visualizaran las interacciones entre los diferentes usuarios. De esta manera podríamos analizar en un mismo grafo los conceptos y a la vez su interacción por usuarios.

- Mejora de la interfaz de SIENA con el fin de hacerla más atractiva al estudiantado.

Capítulo 10. Anexo 1.

Feedback a SOBEK

En este anexo se recoge el *feedback* enviado al equipo de desarrollo de SOBEK, con el fin de contribuir al proyecto.

Documento enviado a SOBEK.

Antes de nada quería dar las gracias al equipo de desarrollo de SOBEK, en especial a Daniel Epstein, sin el cual este trabajo fin de grado no podría haberse realizado.

SOBEK ha sido implantando en la Universidad de La Laguna (http://www.ull.es/) de manera satisfactoria y, aparte de la lista de "StopWords" que ha sido traducida enviamos este feedback con el fin de contribuir a la mejora de SOBEK.

Este feedback contiene algunas sugerencias (técnicas en su mayoría) para el crecimiento de SOBEK.

- *La llamada al applet se realiza enviando solamente una cadena de texto de manera que devuelve un grafo con los conceptos más relevantes, de esta manera creemos que una posible mejora es que el applet reciba dos parámetros:*
    - *Cadena de texto (como actualmente).*
    - *Un número que represente la cantidad de incidencias de una palabra para que aparezca en el gráfico. Puesto que así podría utilizarse en diferentes entornos,*

ya que no es lo mismo que en un chat se repitan 3 veces la misma palabra, a que se repita 3 veces la misma palabra en un libro.

- *Sería interesante que se pudiera elegir el formato de salida del applet. De manera que al enviar un texto a SOBEK, se le indique en qué formato queremos que nos devuelva el gráfico (png, jpeg…).*

- *Podría pensarse en migrar el applet a un servicio web o API de manera que la dependencia del applet desaparezca y todo el cómputo se realice en un mismo lugar, centralizando toda esta información. Esto puede tener grandes beneficios dado que se pueden realizar estadísticas, análisis y gran cantidad de estudios con estos datos.*

Bibliografía

[1] Carenini G, Murray G, Ng R. *Methods for Mining and Summarizing Text Conversation*, Morgan & Claypool Publisher, 2011; p.1, 5.

[2] Arabie P, Hubert J, De Soete G. *Clustering y Classification. World Scientific Publishers*. 1996.

[3] EFE: fundéu BBVA[sede Web]. Madrid: Serrano J; 2011 [acceso 30 de Agosto de 2014]. *La personalización web, santo grial del comercio electrónico.* Disponible en http://www.fundeu.es/escribireninternet/la-personalizacion-web-santo-grial-del-comercio-electronico/.

[4] Ricci F, Rokach L, Shapira B. Introduction to Recommender Systems Handbook, Springer, 2011, pp. 1 – 35.

[5] Perkowitz M, Etzioni O. Adaptive web sites: Automatically synthesizing web pages. National Conference on Artificial Intelligence. WI. 1998.

[6] Romero C, Ventura S, Hervás C. Estado actual de la aplicación de la minería de datos a los sistemas de enseñanza basada en web. III Taller Nacional de Minería de Datos y Aprendizaje. 2005; pp.49-56.

[7] Mitra S, Acharya T. *Data mining: multimedia, soft computing and bioinformatics.* John Wiley & Sons, 2003.

[8] Montes G. Minería de texto empleando la Semejanza entre Estructuras Semánticas. *Resumen de Tesis Doctoral.* 2005, pp. 64-65.

[9] ULL: Universidad de La Laguna[sede Web]. La Laguna: Universidad de La Laguna [acceso 31 de Agosto de 2014]. *Sobre la herramienta SIENA.* Disponible en http://sienasocial.ull.es/.

[10] SAS [Sede Web]. Cary, USA. [acceso 6 de Julio de 2014]. *SAS Text Analytics. https://www.sas.com/offices/latinamerica/mexico/text-analytics/index.html.*

[11] Provalis Research [Sede Web]. Montreal, Canada. Provalis [acceso 5 de Julio de 2014]. *Productos. http://provalisresearch.com/es/productos/software-de-analisis-de-contenido/.*

[12] Attensity [Sede Web]. California, USA. Attensity [acceso 5 de Julio de 2014]. *Analyze the Voice of the Customer Across Multiple Channels. http://www.attensity.com/products/attensity-analyze/.*

[13] Orange [Sede Web]. [acceso 6 de Julio de 2014]. *Orange* http://orange.biolab.si/.

[14] Schenker A, Bunke H, Last M, Kandel A. *Graph-Theoretic Techniques for Web Content Mining.* World Scientific. 2003; pp.31-36.

[15] Feldman E, Fresko M, Kinar Y, Lindell Y, Liphstat O, Rajman M, Schler Y, Zamir O. *In Proceedings of the Second European Symposium on Principles of Data Mining and Knowledge Discovery.* Springer-Verlag. 1998: pp.65-73.

[16] Reategui E, Klemann M, Epstein D, Lorenzatti A. *World Congress in Computer Science, Computer Engineering,and Applied Computing.* 2011: pp.1-3.

[17] Sobek [Sede Web]. Federal University of Rio Grande do Sul, Brasil. [acceso 6 de Julio de 2014]. *Sobek text Mining* http://sobek.ufrgs.br/.

[18] PostgreSQL-es [Sede Web]. Rafael Martínez, Madrid. [acceso 3 de Septiembre de 2014]. *Sobre PostgreSQL http://www.postgresql.org.es/sobre_postgresql.*