BENEMÉRITA UNIVERSIDAD AUTÓNOMA DE PUEBLA

FACULTAD DE CIENCIAS DE LA COMPUTACIÓN

CÁLCULO DE REPUTACIÓN DE ENTIDADES NOMBRADAS APLICADO AL DOMINIO NOTICIERO

Tesis para obtener el grado de

LICENCIADO EN CIENCIAS DE LA COMPUTACIÓN

Presenta

OSCAR PÉREZ SÁNCHEZ

ASESOR:

Dr. DAVID EDUARDO PINTO AVENDAÑO

NOVIEMBRE, 2013

Dedicatoria

Dedicado a mis padres que con su refuerzo y cariño me han apoyado por el camino que he elegido.

Agradecimientos

A mis Padres que con su cariño y afecto me guiaron a través de mi vida, que me han ayudado y apoyado en todas las decisiones que he tomado.

A mi asesor de tesis que me ha guiado y apoyado a través de este proceso y me ha dedicado parte de su tiempo.

A mis profesores que me han inculcado los conocimientos para poder concluir mis estudios, así como consejos que me sirvieron para elegir el camino a seguir.

A todas esas personas que conocí en el camino que me ayudaron y acompañaron y que me sirvieron de inspiración y con los que pase buenos momentos.

Resumen

Con el crecimiento de internet es cada vez más común que las personas busquen opiniones de otras personas para realizar algunas actividades como lo son: realizar compras, elegir libros, elegir ciudades a las cuales visitar, en tiempo de elecciones elegir a algún candidato, etc. El cálculo de reputación de entidades nombradas es un tema en el estudio del procesamiento del lenguaje natural que consiste en la tarea de analizar un texto en busca de encontrar a los sustantivos de las oraciones es decir, encontrar de que o de quien se habla en el texto y evaluar sí se habla bien, mal o neutral de ellos, si bien existen trabajos sobre esto, se enfocan principalmente en redes sociales o sitios de venta de artículos en línea.

Por ese motivo en este trabajo se enfocara en el dominio noticiero y que incluya a periódicos de toda la República Mexicana, se mostrara el proceso de recolección de noticias, la extracción de las entidades nombradas, el conteo de entidades, el análisis para el cálculo de la reputación y la creación de una página web para mostrar los resultados obtenidos.

Consideramos que este proyecto reviste un gran interés, debido a las aplicaciones prácticas que puede tener. Instituciones públicas y personajes también públicos podrían verse beneficiados de un sistema que calcule el grado de reputación de cualquier entidad nombrada que tenga visibilidad en el ambiente noticiero.

Contenido

Tabla de Figuras

CAPÍTULO I INTRODUCCIÓN

Con las fuentes de información que se tienen disponibles hoy en día, se busca que la interacción entre el hombre y la maquina sea más amigable, con el estudio del lenguaje humano y su adaptación a las computadoras se creó el área del procesamiento del lenguaje natural (PLN) para así desarrollar técnicas que nos puedan ayudar a acercar al hombre y a la máquina. Estas técnicas se encargan de procesar información ya sea en forma de texto o voz, para ello se utiliza conocimiento lingüístico.

Con el tiempo y el crecimiento de esta área se han hecho grandes avances en áreas como el reconocimiento del habla, recuperación y extracción de información, generación automática de resúmenes y traducción automática.

En este proyecto de tesis se estudiará el problema de reputación de una entidad nombrada. Es una tarea que se caracteriza por analizar las opiniones vertidas sobre dicha entidad que se encuentra dentro de un texto con la finalidad de calificarlas, por ejemplo, como positivas, negativas o neutrales. En el caso más simple, se puede solo contabilizar el número de veces que se habla sobre esta entidad en una fecha dada y determinar así un valor entero que determina su visibilidad ante la sociedad. Si podemos calcular la frecuencia relativa de dicha entidad sobre un conjunto suficientemente grande de datos, entonces este valor entero tendería a estabilizarse en un valor que llamaríamos la estimación de su probabilidad.

Para llevar a cabo el proyecto con éxito, primeramente se debe construir una colección de documentos asociados al dominio de interés, que en este caso es el de noticias de la república mexicana. Para ello se recabaran noticias de distintos periódicos que publiquen sus noticias mediante el sistema de RSS.

Una vez que este corpus sea constituido, se realizara un pre-procesamiento de las noticias para encontrar las entidades nombradas de cada una de las noticias, posteriormente se organizaran las entidades con respecto a la fecha y al contenido de la noticia, además se agruparan los datos por entidades nombradas juntando los datos que son necesarios como lo son: la fecha de la publicación, la cantidad de veces que se nombró a la entidad en esa noticia, la cantidad de veces que se nombró la entidad en un día en específico, Posteriormente, será ya posible proponer un algoritmo para el cálculo de reputación de dichas entidades nombradas el cual hará uso de un diccionario que contiene palabras (la mayoría adjetivos) que tienen asociado un valor de afinidad.

Por último se mostraran los resultados en una página web, donde se podrán consultar cada una de las entidades que se obtuvieron en el proceso, además de mostrar un gráfico en el cual se puede apreciar la frecuencia en la que las

entidades fueron apareciendo al paso de los días en los que la información fue recabada, se podrán consultar también las fechas para que aparezcan las noticias de la entidad en el día seleccionado de este modo se podrá visitar la página web original de la noticia.

1.1Objetivos generales y particulares del proyecto.

Objetivo general:

Calcular la reputación de entidades nombradas en el dominio noticiero.

Objetivos particulares:

- Construir un corpus de noticias de la república mexicana con la finalidad de usarlo en proceso de estimación de probabilidades asociadas a la reputación de una entidad nombrada.
- Identificar entidades nombradas importantes que puedan ser usadas posteriormente para calcular su grado de reputación.
- Diseñar un algoritmo, usando técnicas del estado del arte, para el cálculo de reputación de una entidad nombrada.

1.2 Metodología

Como paso inicial, definiremos las teorías, conceptos o ideas que pretendemos verificar, particularmente, la dilucidación de propuestas para la determinación del grado de reputación de una entidad nombrada en el contexto de documentos noticieros. Una tarea inicial, por supuesto, consiste en el estudio exhaustivo del estado del arte y la evaluación preliminar de las líneas de acción, a fin de establecer un proyecto de culminación factible en tiempo y forma.

Posteriormente, haremos una comprobación experimental de las hipótesis y confrontación de los resultados. En particular, usaremos métricas clásicas de evaluación que incluyen, entre otras, precisión, recall, F-Measure. En resumen, nos interesa llevar a cabo un análisis y evaluación no subjetivo, de manera tal que se corroboren adecuadamente las hipótesis planteadas. Una de las posibilidades contempladas es hacer uso del corpora ya existente y comparar nuestros resultados con aquellos reportados en la literatura bajo la misma colección de datos.

Como tercer paso, criticaremos los métodos y técnicas desarrolladas, aceptando o rechazando las hipótesis planteadas bajo la explicación de los resultados experimentales y la teoría que subyace a los métodos para la determinación del grado de reputación, así como al tipo de entidades que utilizan (textos del dominio de noticias).

Finalmente, se presentarán las ventajas y desventajas de los métodos y técnicas desarrolladas, así como sus posibles aplicaciones en otros tipos de problemas.

En particular, hemos definido las siguientes hipótesis que deben ser evaluadas:

Consideramos que las empresas, instituciones públicas y personajes públicos se verán beneficiados por el desarrollo y la aplicación de métodos y técnicas para la identificación del grado de reputación de entidades nombradas en el dominio noticiero.

Por tanto, se formulan las siguientes hipótesis:

1. Es posible calcular de manera aceptable el grado de reputación de una entidad nombrada.

2. El modelo desarrollado podrá tener aplicaciones prácticas y de interés, particularmente para entes con imagen pública.

1.3 Estado del campo del Arte

El cálculo de reputación es un área de gran interés dentro de la comunidad lingüística por sus aplicaciones en la vida real. Parte del enfoque consiste en analizar las opiniones vertidas en foros de comunicación, particularmente, en el ámbito de las redes sociales. A continuación se presentan algunos trabajos que han atacado el problema de reputación desde el punto de vista de análisis de opinión.

En el artículo [1] se propone un nuevo enfoque para extraer uniformemente objetivos de opinión explícitos e implícitos usando una teoría de centrado. Este enfoque utiliza información global en noticias, así como información contextual en oraciones adyacentes en comentarios.

En [2] indican que la mayoría de las oraciones en las opiniones tienen una estructura muy complicada y que no pueden ser representadas con los métodos actuales tales como los basados en entorno y en características. Para ello se propone una novedosa representación basada en grafos para el nivel de sentimiento de las frases. Se propone un método de aprendizaje basado en

programación enteramente lineal para producir las representaciones de grafos de las oraciones de entrada. Las evaluaciones experimentales hechas en corpus chinos etiquetados manualmente demuestran la efectividad del enfoque.

En [3] proponen un enfoque novedoso de extracción de objetivos de opiniones basado en el modelo de traducción basado en palabras (WTM). Primero se trabajó con escenarios mono lenguaje para obtener las asociaciones entre los objetivos de la opinión y las palabras de la opinión, después aplicaron un algoritmo basado en grafos para extraer el objetivo de la opinión. Usando este método se puede capturar las relaciones de la opinión de forma más precisa especialmente en relaciones que tienen una gran distancia además de evitar los errores de interpretación cuando se trata con textos informales. Los resultados experimentales en tres conjuntos de datos del mundo real de diferentes tamaños y en diferentes idiomas muestran que este enfoque es más efectivo y robusto que otros métodos.

En [4] se busca determinar la subjetividad y orientación de una opinión, para que se pueda saber si esta tiene una connotación positiva, negativa o que no tenga una connotación del todo. Para ello implementan tres diferentes variantes de un método semi-supervisado para la detección de orientación. Lo que concluyen en [6] es que es mucho más difícil determinar la subjetividad y la orientación de una opinión que solo determinar la orientación.

En [5] exploran como las características basadas en relaciones de dependencia sintáctica pueden mejorar el desempeño en la minería de opinión. Estas características las transforman en características de respaldo compuestas, concluyen que su propuesta es comparada con otras que generalizan las características de dependencia o n-gramas demostrando la utilidad de las características de respaldo compuestas.

En [6] indican que en la minería de opinión automática se ignoran a los objetivos de la opinión basados en pronombres anafóricos, por lo que se pierden un número significante de objetivos de opinión. Para resolver esto, se propone un algoritmo de resolución anáfora el cual puede mejorar el rendimiento de un sistema de minería de opinión. Los experimentos realizados sobre un corpus de análisis de películas demuestran que un algoritmo no supervisado de resolución anáfora mejora la extracción de objetivos de opinión.

En [7] indican que la tarea de extraer la opinión expresada en texto es retadora por diferentes razones, una de ellas es que la misma palabra (en particular los adjetivos) puede tener diferentes polaridades dependiendo del contexto, para ello desarrollaron un enfoque que solucione la desambiguación de los adjetivos ambiguos de opinión, dividieron esta tarea en 3 estrategias las cuales son: la evaluación de la polaridad de todo el contexto usando un sistema de minería de opinión, la valoración de la polaridad en el contexto local dado por la

combinación entre los sustantivos más cercanos y los adjetivos a ser clasificados, y por ultimo reglas orientadas a refinar la semántica local a través de la localización de los modificadores. La decisión final para la clasificación es tomada de acuerdo a la salida de la mayoría de estas estrategias logrando grandes resultados.

El análisis de sentimientos también juega un papel fundamental para la determinación de la reputación de una entidad nombrada. Diversas propuestas existen en la literatura. Algunas de estas se describen de manera general a continuación:

En [8] proponen un novedoso algoritmo de aprendizaje semi-supervisado con el cual puedan direccionar el problema de clasificación sentimental semi-supervisado usando datos etiquetados y sin etiquetar.

Después de construida la estructura es afinada mediante gradiente descendiente basada en aprendizaje supervisado con una función de perdida exponencial. Con esto se concluye que este enfoque puede ser aplicado a técnicas más avanzadas aplicadas a clasificación sentimental.

En [9] se plantea el realizar una recomendación personalizada por el usuario acerca de un sitio de noticias en línea, para ello se plantea clasificar para cada usuario los comentarios asociados a un artículo de acuerdo a las preferencias personalizadas del usuario.

Así se propone un modelo que incorpore interacciones clasificación-comentario y clasificación-autor simultáneamente. Se demuestra que este modelo es más completo e supera significativamente a otros modelos similares.

En [10] proponen un enfoque para identificar el origen de las opiniones, emociones y sentimientos, ven esta tarea como una de extracción de información, para la cual adoptaran un enfoque hibrido que combina campos aleatorios condicionales (Lafferty et al., 2001) y una variación de AutoSlog (Riloff 1996a).

Mientras que el primero se encarga de la tarea de identificación del origen mediante una tarea de etiquetado secuencial, el segundo se encarga de aprender patrones. La combinación de estos métodos realizan mejor la tarea que cada uno por separado, los resultados son de un 79.3% de precisión.

En [11] introducen un nuevo enfoque el cual consiste en que en vez de centrarse en indicadores léxicos en el análisis sentimental, se centran en el empaquetado sintáctico de las ideas, es decir en la investigación para identificar un sentimentalismo implícito.

Establecen una fuerte conexión predictiva entre características lingüísticamente bien formadas y el sentimentalismo implícito. Con esto ellos muestran como aproximaciones computacionales de estas características pueden ser usadas para mejorar los resultados de clasificación de sentimientos.

1.4 Resultado esperados.

- Un sistema automático para el cálculo de reputación para entidades nombradas en el dominio de noticias.

1.5 Aportaciones.

- Análisis de diversas aproximaciones para calcular el grado de reputación de una entidad nombrada.
- Un corpus de noticias de la república mexicana.

CAPITULO II MARCO TEORICO

2.1Entidades Nombradas

Las conferencias de entendimiento de mensajes (MUC) se crearon para fomentar el desarrollo de nuevos y mejores métodos de extracción de información. En la sexta conferencia se agregó la tarea de reconocimiento de entidades nombradas dándole la siguiente definición a una entidad nombrada:

Se conoce como una entidad nombrada: *a una o un conjunto de palabras que se utilizan para identificar a una persona, ubicación, organización, tiempo o cantidad.*

Estas aparecen en los textos comúnmente como sustantivos de las oraciones y nos permiten saber de qué o de quien se está hablando.

Por ejemplo si tenemos en siguiente texto: “El Presidente de la Republica visito a los afectados por los Huracanes Ingrid y Manuel” tendríamos que las entidades de esta línea serian: Presidente, Republica, Huracanes, Ingrid y Manuel.

La habilidad para identificar entidades nombradas ha sido establecida como una importante tarea en muchas áreas, incluyendo temas de detección y rastreo, traducciones automáticas, y recuperación de información.

La meta es la identificación de las entidades mencionadas en el texto y su etiquetado con una de muchos tipos de etiquetas, se debe notar que una entidad como George Bush puede identificarse con las siguientes entidades: “George Bush” o “Bush” y “Bush” se puede referir a múltiples entidades. La entidad Washington se pude referir tanto a una ciudad, una persona, una organización, un equipo. Por lo que se debe de hacer uso de la desambiguación.

2.1.1 reconocimiento de entidades nombradas

En las distintas áreas del Procesamiento del Lenguaje Natural (PLN), un problema común es obtener información relevante relacionada con nombres de personas, lugares u organizaciones, por lo cual se vuelve importante el poder extraer y distinguir este tipo de elementos de todo el conjunto de palabras que componen a un documento. Aun cuando algunos elementos son relativamente fáciles de identificar mediante el uso de patrones (por ejemplo fechas o datos numéricos) existen otros elementos, como personas, lugares u organizaciones, que presentan otras dificultades para ser identificados como pertenecientes a un

tipo específico, además involucra el procesamiento de documentos estructurados y no estructurados. El extraer y distinguir este tipo de elementos es el objetivo de la Extracción de Entidades Nombradas (EEN).

Por ejemplo del siguiente texto: “El Gobernador de Puebla visito la Catedral” las entidades extraídas serian: Gobernador, Puebla y Catedral.

Una de las tareas que realiza es la de delimitación de entidades nombradas y lo que realiza es determinar la longitud de la entidad nombrada, es decir sí esta consta de más de una palabra y si estas palabras vienen consecutivas o están unidas por otra palabra que les dé un sentido de pertenencia como por ejemplo un artículo, en el ejemplo anterior “El Gobernador de Puebla visito la Catedral” al incluir la función de delimitador se tendrían las siguientes entidades: “Gobernador de Puebla” y Catedral de esta manera se pueden crear entidades más específicas y hacer búsquedas más exactas cuando se buscan entidades que coincidan exactamente con los datos de búsqueda.

Para poder identificar los límites que las entidades se utilizan etiquetas al analizar el texto (B, I, O) que significan:

- B para el inicio de la entidad.
- I para la pertenencia a una entidad.
- O para cuando no pertenece a una entidad.

En el ejemplo manejado quedaría así: “El_O Gobernador_B de_I Puebla_I visito_O la_O Catedral_B”. Para poder identificar estos patrones se utilizan características léxicas, sintácticas, ortográficas, afijos y las características de las palabras que se encuentran alrededor de posible entidad. También es importante contemplar atributos importantes como la aparición de letras mayúsculas ya que un nombre propio siempre comienza con una mayúscula o las siglas de organizaciones siempre se escriben en este formato.

Cuando ya se tiene definida claramente a las entidades, lo que se debe de realizar es una categorización de ellas, estas se pueden agrupar en estas categorías: Persona, Organización, Lugar o Miscelánea colocando una etiqueta a cada una quedarían como PER, ORG, LOC, MISC dando más significado y mejor descripción a las entidades para posteriores procesos, de este modo las entidades anteriormente utilizadas quedarían como: Gobernador_de_puebla_PER y Catedral_LOC pudiendo identificar que se mencionan a una persona y a un lugar.

Para identificar la categoría de las entidades se hace uso de recursos similares que se usan para la delimitación, además del uso de diccionarios y técnicas manuales, aprendizaje automático y combinaciones de ambos.

2.1.2 Desambiguación de entidades nombradas.

La tarea de desambiguación de entidades nombradas se base en realizar una diferenciación entre entidades en un texto y enlazarlas a su entrada correspondiente en una base de conocimiento como Wikipedia. Tales desambiguaciones ayudan a fortalecer la confiabilidad y agregar semántica al texto plano. Además es un paso central para la construcción de una red de información de alta calidad de texto sin estructura.

De este modo se pueden identificar a las entidades que aparecen en el texto con su correspondiente en la vida real. Es un caso frecuente que existan consultas relacionadas a entidades nombradas lo cual constituye una significante fracción de las consultas web más populares de acuerdo a logs de motores de búsqueda.

Cuando se envían consultas tales como México o Java, a los usuarios de los motores de búsqueda se les podría mostrar además una compilación de factores y atributos específicos acerca de esas entidades, más que un conjunto de los mejores resultados que coincidan con su búsqueda [12].

Uno de los retos para crear tales alternativas de resultados de búsqueda es la ambigüedad de las consultas, como muchas instancias de la misma clase (diferentes personas), diferentes clases (tipos de serpientes, de lenguajes de programación, películas), que pueden compartir el mismo nombre en la consulta [13].

Dada una consulta el sistema seleccionara entradas de la base de datos que coincidan con los siguientes enfoques:

- El titulo coincida exactamente con lo consultado.
- Los títulos contengan completamente lo consultado.
- Las letras iniciales de lo consultado coincidan con algún acrónimo e alguna organización.
- El titulo coincida con algún alias conocido para la entidad.

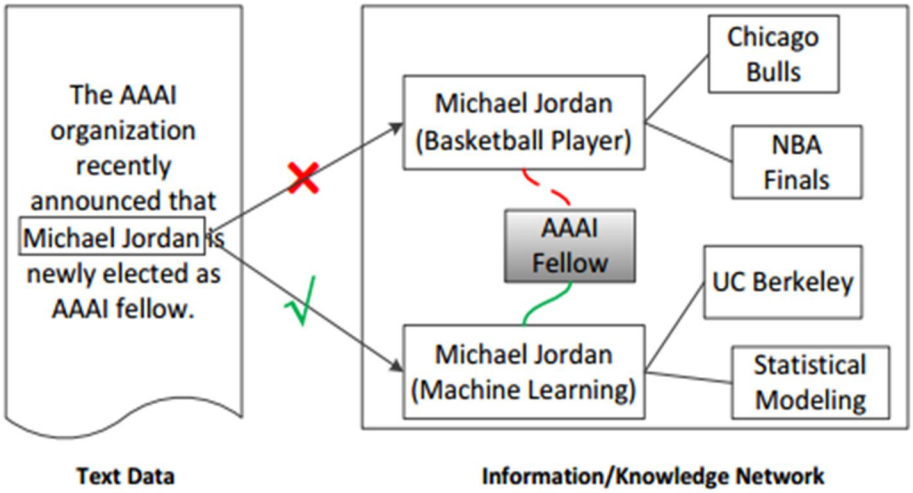

Figura 1: Ejemplo de desambiguación.

En la figura 1 tenemos un ejemplo de lo que se debe analizar en este se analiza la siguiente oración: “La organización AAAI recientemente anuncio que Michael Jordan es el nuevo elegido como seguidor AAAI” la entidad de la oración es Michael Jordan y esta se puede referir a dos entidades del mundo real las que son: Michael Jordan el ex jugador de basquetbol profesional que jugo en la NBA y fue perteneciente al equipo de los Toros de Chicago, o al miembro de la AAAI, como se puede ver si no se tuviera las palabras AAAI, organización el sistema podría regresar un resultado erróneo al elegir como respuesta al Michael Jordan que es el jugador de basquetbol, al existir estas palabras clave se deben de tomar en cuenta para poder dar un resultado correcto, así el sistema puede realizar la desambiguación de mejor manera.

2.1.3 Evaluación de entidades nombradas.

La evaluación de sistemas de Extracción de Información (EI) hace uso de las medidas de precisión, recall y de la medida F, las cuales han sido adoptadas como medidas estándar en el área de EEN.

En las evaluaciones MUC y MET, una respuesta correcta es aquella donde la etiqueta y los límites están correctos. Una respuesta está medio correcta si la etiqueta (el tipo y el atributo) está correcta pero solo 1 límite está correcto. Alternativamente, también es medio correcto si solo el tipo de la etiqueta (y no el atributo) y ambos límites están correctos [14].

Un modelo de puntuación desarrollado para las evaluaciones MUC y MET mide la precisión (P) y el recall (R) como se muestra en la Ecuación 1 y en la Ecuación 2:

$$P = \frac{\#\text{ de respuestas correctas}}{\#\text{ de respuestas}}$$

Ecuación 1. Fórmula para medir la precisión

$$R = \frac{\#\text{ de respuestas correctas}}{\#\text{ de correctas en llave}}$$

Ecuación 2. Fórmula para medir el *recall*

El término respuesta es usado para denotar "respuesta dada por el sistema", el término llave es usado para denotar "un fichero anotado que contiene las respuestas correctas". Informalmente, R mide el número de "hits" vs. El número de posibles respuestas correctas como se especifica en el fichero llave, mientras que P se define como una medida de la proporción de elementos clasificados por el sistema que en realidad son correctos [14].

Por ejemplo si se tiene que de 1000 respuestas 850 fueron correctas por el sistema y que se tenían registras 950 en un archivo llave tenemos que:

$$P = \frac{850}{1000} = 0.85 \qquad R = \frac{850}{950} = 0.89$$

Estas 2 medidas de rendimiento se combinan para formar una tercera medida, la medida F, la cual se calcula como se muestra en la Ecuación 3:

$$F_\beta = \frac{(\beta^2 + 1)RP}{(\beta^2 R) + P}$$

Ecuación 3. Fórmula para calcular la medida F

Donde β es un factor que determina la importancia que se le da a cada una de estas medidas. Típicamente tiene valor 1 quedando la ecuación 3 de la siguiente forma:

$$F = \frac{RP}{1/2(R+P)}$$

Continuando con el ejemplo tendríamos que:

$$F = \frac{(0.85 * 0.89)}{\frac{1}{2}(0.85 + 0.89)} = \frac{0.7565}{0.87} = 0.8695$$

2.2 Minería de opinión

La minería de opinión nos permite determinar la actitud de alguna persona respecto a algún tema o a la polaridad contextual de un documento. Trata de determinar una orientación ya sea positiva, negativa o neutral.

Con el crecimiento de internet y las redes sociales cada vez es más común encontrar la opinión de las personas para casi cualquier cosa, lo que ha despertado un gran interés académico e industrial.

Formalmente se define a una opinión como una característica *f* que tiene un *sentimiento* adherido comúnmente positivo o negativo, la persona que emite la opinión es llamado *titular de la opinión*, así la opinión es un quíntupla (o_j,f_{jk}, oo_{ijkl},h_i,t_i) [15] donde:

- o_j Es el objeto de la opinión.
- f_{jk} Es la característica del objeto o_j acerca del cual la opinión es expresada.
- oo_{ijkl} Es la polaridad de la opinión acerca de la característica f_{jk} del objeto o_j puede ser positiva, negativa o neutral.
- h_i Es el titular de la opinión.
- t_i es el tiempo cuando la opinión es expresada por h_i.

Existen dos categorías en las cuales se puede separar la minería de opinión el análisis manual o humano y el análisis automático. Las más notables diferencias recaen en la eficiencia del sistema y la precisión del análisis, aunque se puede utilizar una combinación de ambas categorías.

En el análisis humano, una persona es la que analiza las opiniones y mediante la aplicación de reglas y diccionarios, es como va generando resultados. Por otro lado en el análisis automático la minería de opinión puede convertirse en una tarea de clasificación de esta manera se pueden utilizar técnicas de aprendizaje de máquina, estas técnicas requieren de un corpus que contenga un amplio número de ejemplos etiquetados manualmente.

Para la tarea de clasificación, los métodos de aprendizaje de máquinas más utilizados son Nayve Bayes, Máxima entropía y máquinas de soporte vectorial, con estos métodos se han reportado resultados relativamente altos, obteniendo los porcentajes siguientes 81%, 80.4% y 82.9% respectivamente utilizando dos clases, positiva y negativa.

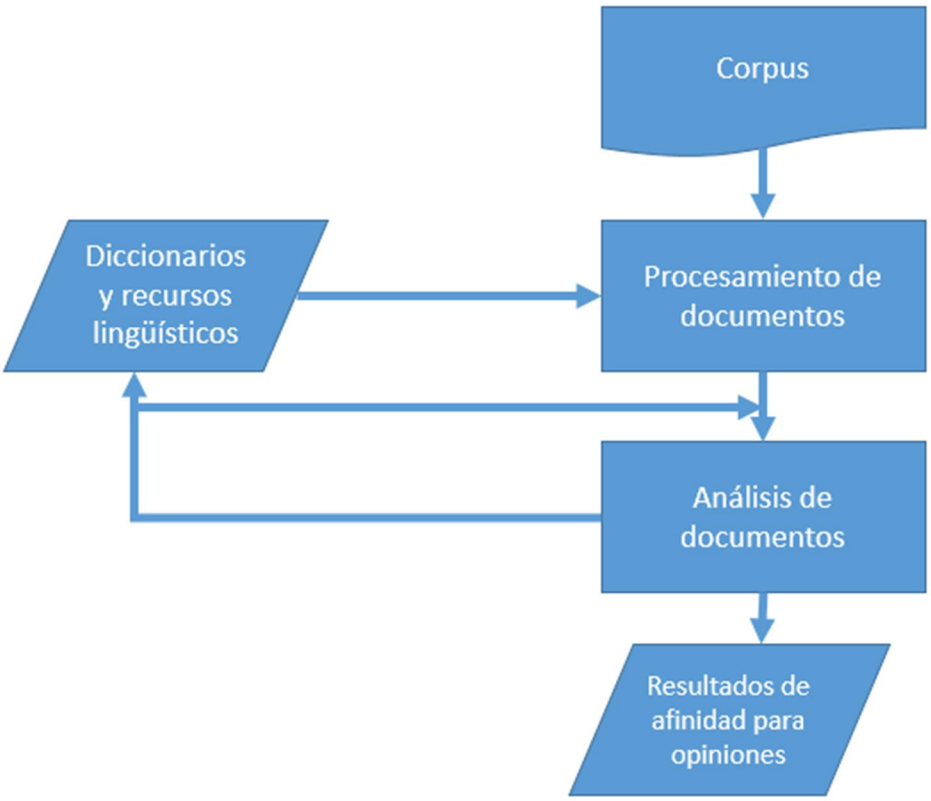

Figura 2: Proceso de la minería de opinión

La entrada del sistema es un corpus de documentos, estos documentos son convertidos a texto y pre-procesados usando una variedad de herramientas lingüísticas tales como: stemming, tokenización, etiquetado de portes de la oración, extracción de entidades y extracción de relaciones. El sistema además utiliza un conjunto de diccionarios y recursos lingüísticos.

El componente principal del sistema es el módulo de análisis de documentos, el cual los recursos lingüísticos para anotar los documentos pre-procesados anotaciones de opinión. Las anotaciones pueden ser fijadas a:

- todo el documento (para opinión basada en documentos)
- a oraciones individuales (para opinión basada oraciones)
- a aspectos específicos de entidades (para opinión basada en aspectos).

Opinión basada en documentos

Es la forma más simple de análisis de opinión y se asume que el documento contiene una opinión de un objeto en particular expresada por el autor del documento [16]. Existen dos enfoques:

- Aprendizaje supervisado.
  En este enfoque se asume que hay un conjunto finito de clases en las que puede ser clasificado, además de contar con datos de entrenamiento para cada clase, el caso más simple es que existan dos clases positiva y negativa. Se pude agregar la clase neutral o colocar una escala discreta en la cual colocar al documento. Dado el conjunto de entrenamiento se pueden usar los sistemas de clasificación más comunes como: SVM, Naïve Bayes, regresión lógica o KNN.
- Aprendizaje no supervisado.
  En este enfoque se busca determinar la orientación semántica de frases específicas dentro del documento. Si el promedio de la orientación semántica está por encima de algún umbral predefinido el documento es clasificado como positivo, de lo contrario es clasificado como negativo. Para realizar esto se puede optar por un conjunto de patrones de POS o utilizar diccionarios de palabras de opinión y frases.

Opinión basada oraciones

Un solo documento puede contener múltiples opiniones respecto a las mismas entidades. Cuando se requiere un punto de vista más fino de las diferentes opiniones expresadas en el documento acerca de las entidades se debe analizar al nivel de oraciones [17].

Se asume que se conocen la identidad de las entidades nombradas en la oración, también es asume que solo hay una opinión en la oración, esto se puede hacer dividiendo la oración en frases que solo contengan una opinión. Antes de analizar la oración se debe determinar si está es una oración subjetiva u objetiva. Solo las oraciones subjetivas se analizan ya que son más fáciles de analizar que las oraciones objetivas. Al centrarnos en las oraciones subjetivas se procede a clasificarlas en positivas o negativas para ello se ocupan métodos de aprendizaje supervisados o no supervisados.

Opinión basada en aspectos

Los dos enfoques anteriores trabajan bien cuando en el documento completo o en las oraciones se refiere a una sola entidad. Sin embargo en muchos casos las personas hablan acerca de entidades que tienen muchos aspectos (atributos) y estos tienen diferentes opiniones acerca de cada uno de estos atributos. Esto suele ocurrir en análisis acerca de productos o en foros de discusión dedicados a categorías de productos específicos (tales como autos, cámaras, teléfonos, y hasta drogas farmacéuticas) [18].

El enfoque tradicional el cual es usado por muchas compañías para la identificación de todos los aspectos en el corpus de análisis productos para extraer todos los sustantivos de las frases y quedarse solo con los que tienen una frecuencia mayor sobre la establecida. Otro enfoque para la identificación de aspectos es usar una dependencia de frases que utilicen expresiones para encontrar aspectos que son inusuales.

Adquisición de diccionarios de opinión

El diccionario de opinión es el recurso más crucial para la mayoría de algoritmos de análisis de opinión. Existen tres opciones para obtener diccionarios de opinión:

- Enfoque manual: Es el cual las personas codifican en diccionario a mano.
- Enfoque basado en diccionarios: Es en el cual un conjunto pequeño de palabras creado para el dominio especificado luego es expandido utilizando un diccionario grande como WordNet [19] buscando sinónimos y antónimos.
- Enfoque basado en corpus: es n el cual un conjunto de palabras es expandido utilizando un gran corpus de documentos de un solo dominio.

Claramente el enfoque manual en general no es recomendable ya que cada dominio requiere de su propio léxico y tal esfuerzo es prohibitivo. La principal desventaja del enfoque basado en diccionarios es que el léxico adquirido no captura las peculiaridades específicas de un dominio en específico. Si se desea crear un léxico que este especificado para el dominio apropiado se debe recurrir a algoritmos basados en corpus [20].

Estas anotaciones son la salida del sistema y estas pueden ser presentadas al usuario usando una variedad de herramientas de visualización.

Existen muchos temas abiertos en el análisis de opinión:

- Se necesita un mejor modelado de la composición de la opinión. Al nivel de oraciones significa mayor precisión de cálculo de la opinión.
- Cada producto puede tener diferentes nombres para referirse a este, incluso en el mismo documento y claramente a través de documentos.
- Cuando en un texto se encuentran varias entidades, es importante identificar el texto relevante para cada entidad.
- Aunque existen métodos para detectar el sarcasmo, a un no han sido integrados a un sistema autónomo de identificación de análisis de opinión.

- Textos ruidosos (aquellos con errores gramaticales, problemas de puntuación o falta de esta, lenguaje específico de localidades).
- Los actuales enfoques de análisis de opinión determinan la opinión de oraciones subjetivas y pasan por alto las oraciones objetivas.

Aplicaciones

La más común de las aplicaciones del análisis de opinión es en el área de análisis de productos consumibles y servicios. Existen muchos sitios web que proveen resúmenes automáticos de análisis de productos y de aspectos específicos un ejemplo de esto sería “Google Product Search”.

Twitter y Facebook son el centro de muchas aplicaciones de análisis de opinión, las más comunes monitorean la reputación de marcas específicas.

El análisis de opinión puede proveer un valor sustancial para candidatos que se encuentran en campaña, ya que se puede seguir a los votantes y conocer las opiniones que realizan acerca de las propuestas que realizan los candidatos.

Otro dominio importante para el análisis de opinión es el mercado financiero. Hay numerosos artículos, elementos, blogs y tweets acerca de compañías públicas. Un sistema de análisis de opinión puede encontrar varias fuentes para encontrar artículos que discutan y agreguen una calificación de opinión como un solo resultado que puede ser ocupado en un sistema de intercambio automático.

2.3 Etiquetado gramatical

El etiquetado gramatical es el proceso de asignar a cada una de las palabras de un texto su categoría gramatical que puede ser una de las siguientes:

- Sustantivo
- Pronombre
- Verbo
- Adjetivo
- Adverbio
- Preposición
- Conjunción
- Intersección
- Determinante

Este proceso se puede realizar de acuerdo a la definición de la palabra o el contexto en el que aparece. Uno de los usos del etiquetado tiene lugar en el contexto de la lingüística computacional, mediante el uso de algoritmos que realizan el etiquetado mediante etiquetas descriptivas predefinidas.

Aproximaciones de Aprendizaje Automático

Estas aproximaciones construyen un modelo de lenguaje utilizando métodos de aprendizaje a partir de datos. Estas aproximaciones difieren entre sí en el método de aprendizaje y en la complejidad del modelo construido. Muchos son los formalismos utilizados: Modelos de Markov o n-gramas, reglas de transformación, árboles de decisión, redes neuronales, autómatas y transductores de estados finitos, etc.

La aproximación más utilizada son los Modelos de Markov Ocultos o n-gramas.

Esta técnica consiste en construir un modelo de lenguaje estadístico, que se utiliza para obtener, a partir de una frase de entrada, la secuencia de etiquetados léxicos que tiene mayor probabilidad. Por ejemplo, si hemos etiquetado una palabra como artículo, la próxima palabra será un nombre con un 40% de probabilidad, un adjetivo con otro 40% y un número el 20% restante. Conociendo esta información, un sistema puede decidir que la palabra "vino" en la frase "el vino" es más probable que sea un nombre a que sea un verbo.

Algunos MMO más avanzados aprenden las probabilidades de pares, triples e incluso secuencias más largas. Por ejemplo, si acabamos de etiquetar un artículo y un verbo, la siguiente palabra probablemente será una preposición, un artículo o un nombre, pero difícilmente será otro verbo.

Tree Tagger

Tree Tagger es una herramienta para la anotación de texto con información del lema y de parte de la oración. Fue desarrollado por Helmut Schmid en el proyecto de cooperación técnica en el Instituto de Lingüística Computacional de la Universidad de Stuttgart.

Se ha utilizado con éxito para etiquetar Alemán, Portugués Inglés, francés, italiano, holandés, español, búlgaro, ruso, gallego, chino, swahili, eslovaco, latín, estonio y viejos textos en francés y es adaptable a otros idiomas, si un léxico y un corpus de entrenamiento etiquetado manual están disponibles.[21]

Tree Tagger es un etiquetador gramatical basado en modelos de Markov, el cual usa un árbol de decisión para obtener más fiabilidad estimada para parámetros contextuales.

Las etiquetas que para las diferentes partes de la oración que utiliza Tree Tagger son las que se muestran a continuación:

Etiqueta	Descripción
ACRNM	Acrónimos
ADJ	Adjetivos
ADV	Adverbios
ALFP	Letras plurales del alfabeto
ALFS	Letras singulares del alfabeto
ART	Artículos
BACKSLASH	Backslash (\)
CARD	Cardinales
CC	Conjunción coordinada (y, o)
CCAD	Conjunción coordinada adversativa (pero)
CCNEG	Conjunción coordinada negativa (ni)
CM	Coma (,)
CODE	Código alfanumérico
COLON	Dos puntos (:)
CQUE	Que (como conjunción)
CSUBF	Conjunción subordinada que introduce clausulas finitas (apenas)
CSUBI	Conjunción subordinada que introduce clausulas infinitas (al)
CSUBX	Conjunción subordinada sub-especificada (aunque)
DASH	Guion (-)
DM	Pronombres demostrativos (esas, ese, esta)
DOTS	Etiqueta para …
FO	Formula
FS	Fin de las marcas de puntación
INT	Pronombres interrogativos (quienes, cuantas, cuanto)
ITJN	Intersección (oh, ja)
LP	Paréntesis izquierdo (“(“,”[”)
NC	Sustantivos comunes (mesa, mesas, libro, ordenador)
NEG	Negación
NMEA	Sustantivos medición (litros, metros)
NMON	Nombre de los meses
NP	Sustantivos propios
ORD	Ordinales (primer, primera, primeras)
PAL	Palabra formada por a y el
PDEL	Palabra formada por de y el
PE	Palabra foránea
PERCT	Signo de porcentaje
PNC	Palabra sin clasificación

PPC	Pronombre personal clítico (le, les)
PPO	Pronombre posesivo (mi, su, sus)
PPX	Pronombres personales y cliticos (nos, me nosotras, te).
PREP	Preposición negativa (sin)
PREP	Proposición
PREP/DEL	Preposición compleja (después del)
QT	Símbolo comillas (“,’,)
QU	Cuantificadores (sendas, cada)
REL	Pronombres relativos (cuyas, cuyo)
RP	Paréntesis derecho (“)”,”]”)
SE	Se (como particula)
SEMICOLON	Punto y coma (;)
SLASH	Diagonal (/)
SYM	Símbolos
UMMX	Unidad de medición (MHz, Km, mA)
VCLIger	Verbo gerundio clítico
VCLIinf	Verbo infinitivo clítico
VCLIfin	Verbo finito clítico
VEadj	Verbo “estar” pasado participio
VEfin	Verbo “estar” pasado finito
VEgen	Verbo “estar” pasado Gerundio
Veinf	Verbo “estar” pasado infinitivo
VHadj	Verbo “Haber” pasado participio
VHfin	Verbo “Haber” pasado finito
VHgen	Verbo “Haber” pasado Gerundio
VHinf	Verbo “Haber” pasado infinitivo
VLadj	Verbo léxico pasado participio
VLfin	Verbo léxico pasado finito
VLgen	Verbo léxico pasado Gerundio
VLinf	Verbo léxico pasado infinitivo
VMadj	Verbo modal pasado participio
VMfin	Verbo modal pasado finito
VMgen	Verbo modal pasado Gerundio
VMinf	Verbo modal pasado infinitivo
VSadj	Verbo “ser” pasado participio
VSfin	Verbo “ser” pasado finito
VSgen	Verbo “ser” pasado Gerundio
VSinf	Verbo “ser” pasado infinitivo

Figura 3: Etiquetas de Tree Tagger.

De la siguiente oración “El Gobernador Rafael Moreno Valle inauguro el nuevo Hospital de la Ciudad de Puebla” tenemos el siguiente resultado:

PALABRA	ETIQUETA	DESCRIPCIÓN	LEMA
El	ART	Articulo	el
Gobernador	NP	Sustantivo propio	Gobernador

Rafael	NP	Sustantivo propio	Rafael
Moreno	NP	Sustantivo propio	Moreno
Valle	NP	Sustantivo propio	Valle
inauguro	VLfin	Verbo léxico pasado finito	inauguro
el	ART	Articulo	el
nuevo	ADJ	Adjetivo	nuevo
Hospital	NP	Sustantivo propio	Hospital
de	PREP	Preposición	de
la	ART	Articulo	la
Ciudad	NP	Sustantivo propio	Ciudad
de	PREP	Preposición	de
Puebla	NP	Sustantivo propio	Puebla

Figura 4: Ejemplo de la ejecución de Tree Tagger.

Obteniendo como nombres propios a las siguientes palabras: Gobernador, Rafael, Moreno, Valle, Hospital, Ciudad, Puebla que pueden ser las entidades que se mencionan en esa oración.

Freeling Tagger

El paquete Freeling consiste en una librería que provee un servicio de análisis del lenguaje. Actualmente soporta los siguientes lenguajes: español, catalán, francés, gallego, italiano, inglés, ruso, portugués, el galés y el asturiano.

Este etiquetador actúa de dos enfoques diferentes y el decide automáticamente cual utilizar:

- El primero se basa en un etiquetador clásico Markoviano de trigramas.
- El segundo es un sistema hibrido capaz de integrar conocimiento estadístico con conocimiento codificado manualmente [22].

Algunos de los servicios que ofrece son:

- Tokenización de texto.
- División de oraciones.
- Análisis morfológico.

- Etiquetado PoS.
- Detección de entidades nombradas.
- Reconocimiento de fechas, números, y las magnitudes físicas (velocidad, peso, temperatura, densidad, etc.).

Las etiquetas que utiliza Freeling para etiquetar a cada una de las partes de la oración provienen del grupo EAGLES son las siguientes:

Clasificación de Adjetivos:

ADJETIVOS			
Pos.	**Atributo**	**Valor**	**Código**
1	Categoría	Adjetivo	A
2	Tipo	Calificativo	Q
		Ordinal	O
		-	0
3	Grado	-	0
		Aumentativo	A
		Diminutivo	C
		Superlativo	S
4	Género	Masculino	M
		Femenino	F
		Común	C
5	Numero	Singular	S
		plural	P
		invariable	N
6	Función	-	0
		participio	P

Ejemplos:

FORMA	LEMA	ETIQUETA
alegres	alegre	AQ0CP0
alegre	alegre	AQ0CS0
bonita	bonito	AQ0FS0
grandazo	grande	AQAMS0

Clasificación de Adverbios:

ADVERBIOS			
Pos.	**Atributo**	**Valor**	**Código**
1	Categoría	Adverbio	R
2	Tipo	General	N
		Negativo	N

Ejemplos:

FORMA	LEMA	ETIQUETA
despacio	despacio	RG
ahora	ahora	RG
siempre	siempre	RG
hábilmente	hábilmente	RG

Clasificación de Determinantes:

DETERMINANTES			
Pos.	**Atributo**	**Valor**	**Código**
1	Categoría	Determinantes	D
2	Tipo	Demostrativo	D
		Posesivo	P
		Interrogativo	T
		Exclamativo	E
		Indefinido	I
		Articulo	A
3	Persona	Primera	1
		Segunda	2
		Tercera	3
4	Género	Masculino	M
		Femenino	F
		Común	C
		Neutro	N
5	Número	Singular	S

		Plural	P
		Invariable	N
6	Poseedor	Singular	S
		Plural	P

Ejemplos:

FORMA	LEMA	ETIQUETA
aquel	aquel	DD0MS0
aquella	aquel	DD0FS0
aquellas	aquel	DD0FP0
aquellos	aquel	DD0MP0
esa	ese	DD0FS0
esas	ese	DD0FS0

Clasificación de Nombres:

NOMBRES			
Pos.	**Atributo**	**Valor**	**Código**
1	Categoría	Nombre	N
2	Tipo	Común	C
		Propio	P
3	Género	Masculino	M
		Femenino	F
		Común	C
4	Número	Singular	S
		Plural	P
5-6	Clasificación semántica	Persona	SP
		Lugar	G0
		Organización	O0
		Otros	V0
7	Grado	Aumentativo	A
		Diminutivo	D

Ejemplos:

FORMA	LEMA	ETIQUETA
chico	chico	NCMS000
chicas	chico	NCFP000
gatito	gato	NCMS00D
oyente	oyente	NCCS000

Clasificación de Conjunciones:

CONJUNCIONES			
Atributo	**Pos.**	**Valor**	**Código**
Categoría	1	Conjunción	C
Tipo	2	Coordinada	C
		Subordinada	S

Ejemplos:

FORMA	LEMA	ETIQUETA
e	e	CC
aunque	aunque	CS

Clasificación de interjecciones.

INTERJECCIONES			
Pos.	**Atributo**	**Valor**	**Código**
1	Categoría	Interjección	I

Ejemplos:

FORMA	LEMA	ETIQUETA
ah	ah	I
eh	eh	I

Clasificación de Verbos:

VERBOS			
Pos.	**Atributo**	**Valor**	**Código**
1	Categoría	Verbo	V
2	Tipo	Principal	M
		Auxiliar	A
		Semiauxiliar	S
3	Modo	Indicativo	I
		Subjuntivo	S
		Imperativo	M
		Infinitivo	N
		Gerundio	G
		Participio	P

4	Tiempo	Presente	P
		Imperfecto	I
		Futuro	F
		Pasado	S
		Condicional	C
		-	0
5	Persona	Primera	1
		Segunda	2
		Tercera	3
6	Número	Singular	S
		Plural	P
7	Género	Masculino	M
		Femenino	F

Ejemplos:

FORMA	LEMA	ETIQUETA
cantada	cantar	VMP00SF
cantadas	cantar	VMP00PF
cantado	cantar	VMP00SM

Clasificación de preposiciones:

PREPOSICIONES			
Pos.	**Atributo**	**Valor**	**Código**
1	Categoría	Adposición	S
2	Tipo	Preposición	P
3	Forma	Simple	S
		Contraída	C
4	Género	Masculino	M
5	Número	Singular	S

Ejemplos:

FORMA	LEMA	ETIQUETA
al	al	SPCMS
del	del	SPCMS
a	a	SPS00
ante	ante	SPS00

Clasificación de pronombres:

PRONOMBRES			
Pos.	**Atributo**	**Valor**	**Código**
1	Categoría	Pronombre	P
2	Tipo	Personal	P
		Demostrativo	D
		Posesivo	X
		Indefinido	I
		Interrogativo	T
		Relativo	R
		Exclamativo	E
3	Persona	Primera	1
		Segunda	2
		Tercera	3
4	Género	Masculino	M
		Femenino	F
		Común	C
		Neutro	N
5	Número	Singular	S
		Plural	P
		Impersonal/invariable	N
6	Caso	Nominativo	N
		Acusativo	A
		Dativo	D
		Oblicuo	O
7	Poseedor	Singular	S
		Plural	P
8	Politeness	Polite	P

Ejemplos:

FORMA	LEMA	ETIQUETA
yo	yo	PP1CSN00
me	me	PP1CS000
aquello	aquello	PD0NS000
ésas	ese	PD0FP000

Clasificación de Signos de puntuación.

SIGNOS DE PUNTUACIÓN			
Pos.	**Atributo**	**Valor**	**Código**
1	Categoría	Puntuación	F

Ejemplos:

FORMA	LEMA	ETIQUETA
¡	¡	Faa
¡	¡	Fat
,	,	Fc
[	[	Fca

Clasificación de cifras:

CIFRAS			
Pos.	Atributo	Valor	Código
1	Categoría	Cifra	Z
2	Tipo	Partitivo	D
		Moneda	M
		Porcentaje	P
		Unidad	U

Ejemplos:

FORMA	LEMA	ETIQUETA
239	239	Z
docientos_veinte	220	Z
un_millón	1000000	Zd
una_docena	12	Zd

Clasificación de fechas y horas:

FECHAS Y HORAS			
Pos.	**Atributo**	**Valor**	**Código**
1	Categoría	Fecha / Hora	W

Ejemplos:

FORMA	LEMA	ETIQUETA
Viernes_26_de_septiembre_de_1992	[V:26:09:1992:??.??]	W
Las tres de la tarde del 26 de septiembre de 1992	[??:26:09:1992:03:00:pm]	W

De la siguiente oración: “El Gobernador Rafael Moreno Valle inauguro el nuevo Hospital de la Ciudad de Puebla” tenemos los siguientes resultados.

Freeling

TEXTO	LEMA	ETIQUETA
El	el	DA0MS0
Gobernador_Rafael_Moreno_Valle	gobernador_rafael_moreno_valle	NP00000
inauguro	Inaugurar	VMIP1S0
el	el	DA0MS0
nuevo	nuevo	AQ0MS0
Hospital_de_la_Ciudad_de_Puebla	hospital_de_la_ciudad_de_puebla	NP00000

Obteniendo como nombre propios a: *gobernador_rafael_moreno_valle* y *hospital_de_la_ciudad_de_puebla.* Cabe resaltar que Freeling ya realiza un delimitado de entidades, obteniendo entidades más específicas.

2.4 Sistemas de recuperación de información

La Recuperación de Información (IR, Information Retrieval) es el área de la ciencia y la tecnología que trata de la adquisición, representación, almacenamiento, organización y acceso a elementos de información. Desde un punto de vista práctico, dada una necesidad de información del usuario, un proceso de IR produce como salida un conjunto de documentos cuyo contenido satisface potencialmente dicha necesidad.

Los sistemas de recuperación de información son sistemas que tratan con bases de datos compuestas por documentos y procesan las consultas de los usuarios permitiéndoles acceder a la información relevante en un tiempo de intervalo apropiado. Estas consultas son sentencias formales mediante las cual se el usuario expresa sus necesidades de información, formuladas en un lenguaje de consulta.

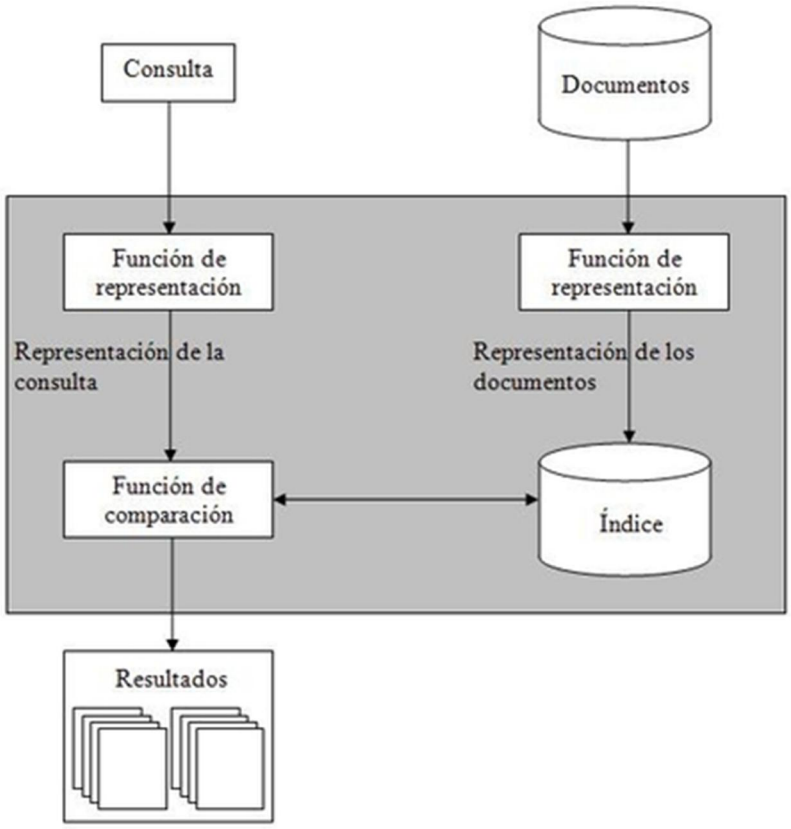

Figura 5: Arquitectura de un sistema de información.

A la hora de diseñar un sistema de recuperacion de informacion es precisoestrablecer previamente como representar los documentos y las necesidades de informacion del usuario, y como comparar ambas representaciones. Es preciso definir el modelo de recuperacion sobre el que ha de desarrollarse el sistema.

Se define formalmente el concepto de modelo de recuperación como una cuádrupla [D, Q, *F*, $R(q_i, d_j)$] donde:

- D es el conjunto de representaciones de los documentos de la colección.
- Q es el conjunto de representaciones de las necesidades de información del usuario, representaciones denominadas consultas.
- *F* es el marco formal dentro del cual modelar las representaciones de documentos, consultas y las relaciones entre ambos.
- $R(q_i, d_j)$ es una función de ordenación que asocia un número real a los diferentes pares consulta $q_i \in Q$ – representación de documento $d_j \in D$.

Dicha ordenación define una relación de orden entre los documentos de la colección respecto a la consulta q_i.

Entre los modelos más representativos están: el booleano, vectorial y el probabilístico.

Un sistema de recuperación de información lleva a cabo las siguientes tareas para responder a las consultas del usuario:

1. Indexación de la colección de documentos: en esta fase, mediante la aplicación de técnicas de NLP, se genera un índice que contiene las descripciones de los documentos. Normalmente, cada documento es descrito mediante el conjunto de términos que, hipotéticamente, mejor representa su contenido.
2. Cuando un usuario formula una consulta el sistema la analiza, y si es necesario la transforma, con el fin de representar la necesidad de información del usuario del mismo modo que el contenido de los documentos.
3. El sistema compara la descripción de cada documento con la descripción de la consulta, y presenta al usuario aquellos documentos cuyas descripciones más se asemejan a la descripción de su consulta.
4. Los resultados suelen ser mostrados en función de su relevancia, es decir, ordenados en función del grado de similitud entre las descripciones de los documentos y de la consulta.

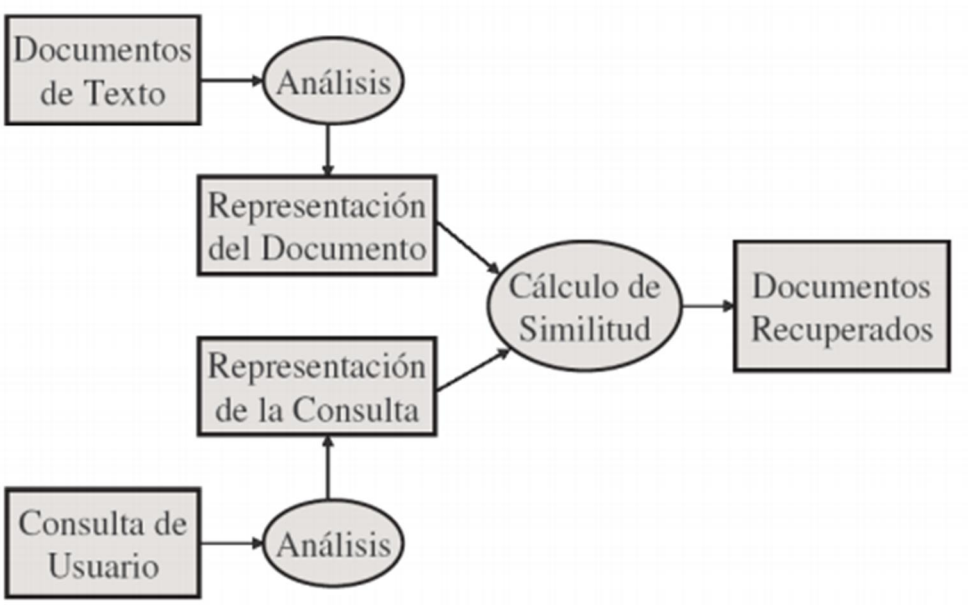

Figura 6: Operaciones para la recuperación de documentos.

Los Componentes de los Sistemas de Recuperación de Información son tres:

- En primer lugar se encuentra la base de datos documental, refiriéndose al conjunto de datos representados en forma de texto, graficas, videos animados fotográficos, ilustraciones, audio, etc.
- El segundo componente es el subsistema de consultas compuesto por la interfaz permitiendo al usuario formular sus consultas utilizando un analizador sintáctico que toma la consulta que el usuario ha escrito, donde la consulta es desglosada en sus partes integrantes, realizando esta tarea con un lenguaje de consulta con reglas para generar consultas apropiadas y la interfaz mostrará al usuario el resultado de su búsqueda, una vez procesada su consulta.
- El tercer componente llamado Subsistema de Evaluación que calcula el grado RSV (Retrieval Status Value), donde las representaciones de los documentos satisfacen las exigencias de la consulta, recuperando los documentos más relevantes.

CAPITULO III SISTEMA DE RECOLECCIÓN DE NOTICIAS

Como se mencionó anteriormente la extracción de las entidades nombradas será de textos provenientes de noticias que se publican en internet por distintos periódicos de la mayoría de los estados de la República Mexicana. En este capítulo se describe la realización de este sistema y de la creación del corpus para la extracción de entidades nombradas.

Como fuente de las noticias se tomaran aquellos periódicos que tienen un sistema de RSS (Really Simple Syndication) la lista de los periódicos tomados como fuente son los siguientes:

- Periódico Reforma
- Periódico el universal
- Red política
- El grafico
- Milenio
- Crónica
- Razón
- Impacto
- El sol de México
- Excelsior
- Mas por mas
- Publimetro
- DiarioDF
- Record
- ESTO
- Economista
- Mundo ejecutivo
- El periódico de México
- El Heraldo
- Aguas digital
- Expreso Chiapas
- Periódico el orbe
- Carteles de Comitán
- El informador de Chiapas
- Periódico el zócalo
- Periódico Vanguardia
- El heraldo de Saltillo
- El siglo de Torreón
- Periódico León

- El heraldo del Bajío
- Periódico Correo
- Plaza Juárez
- Crónica Hidalgo
- El Informante de México
- Edomex al día
- Diario Fuerza
- 8 Columnas
- Diario Portal
- La Unión
- El Regional
- El Norte
- El Porvenir
- Red Crucero
- Periódico ABC
- El Heraldo de Puebla
- EL Popular
- Diario Como
- Sipse
- Milenio Yucatán
- Novedades de Quintana Roo
- Novedades de Chetumal
- México Eclipse
- Diario Imagen de Quintana Roo
- El quintanarroense
- Noroeste
- Tabasco Hoy
- Diario Presente
- El correo de Tabasco
- Periódico Yucatán
- Milenio Yucatán
- Periódico Frontera
- La Crónica
- Novedades de Campeche
- Expreso Campeche
- Diario
- Diario de Colima
- Contexto de Durango
- Novedades de Acapulco
- Voz Zihuatanejo

- El Informador
- Mural
- La Jornada de Jalisco
- NNC
- El Faro MX
- Voz de Michoacán
- El Diario Visión
- La Extra
- El Independiente de Zamora
- Imparcial
- Despertar de Oaxaca
- El Regional de Sonora
- Diario Sonora
- Nuevo día
- El Mañana
- La Prensa
- El Eco delmante
- Respuesta en línea
- Imagen Zacatecas
- Página 24 Zacatecas
- Oye Veracruz
- Noroeste
- Diario la Info
- El mundo de Poza Rica
- Noticias Perfil

De estos 94 periódicos y de sus secciones se tomó la información necesaria que se lista a continuación:

- Título de la noticia.
- La fecha de publicación.
- El enlace a la página de la noticia.
- La descripción de la noticia.

El formato como se encuentra esa información es el siguiente:

```
<item>
<title>"Caen" 10 toneladas de droga en Mariposa</title>
<link>
http://www.nuevodia.com.mx/local/caen-10-toneladas-de-droga-en-
mariposa/
```

```
</link>
<pubDate>Tue, 19 Nov 2013 12:03:23 +0000</pubDate>
<description>
<![CDATA[
Un cargamento de 20 mil libras de marihuana fue asegurado por los
agentes de la Aduana Americana en la garita comercial Mariposa; el
decomiso estableció un nuevo record superando las catorce mil 121
libras de droga detectadas en enero de este año.
]]>
</description>
</item>
```

Cada noticia almacenada en los archivos RSS representa una etiqueta <ítem> dentro de esta se encuentran más etiquetas como: <title> para el título, <PubDate> para la fecha, <link> para el enlace a la página de la noticia y <description> para la descripción de la noticia, algunos archivos contienen más información como el autor de la noticia, comentarios, etc. Pero al no ser constante en todos o la mayoría de los archivos no se tomaran en cuenta, además de que no se sigue un estándar al momento de codificar estos archivos por lo que algunos no incluyen etiquetas para la fecha o el enlace y en otros casos las fechas no siguen el mismo formato.

3.1 Diseño

Conociendo el formato en el que se encuentran almacenadas las noticias se procedió a realizar un algoritmo que recolecte estas noticias y las almacene en una base de datos.

La base de datos se llamara Noticias y contendrá una tabla llamada detalles_noticias la cual contendrá las siguientes columnas:

COLUMNA	DETALLES
TITULO	Varchar, Primary key
FECHA	Date
LINK	Varchar
DESCRIPCIÓN	Varchar

Figura 7: Información de la Tabla Detalles_noticia.

Los pasos que realizara el algoritmo para la obtención de la información de las noticias se muestran en el siguiente diagrama de casos de uso:

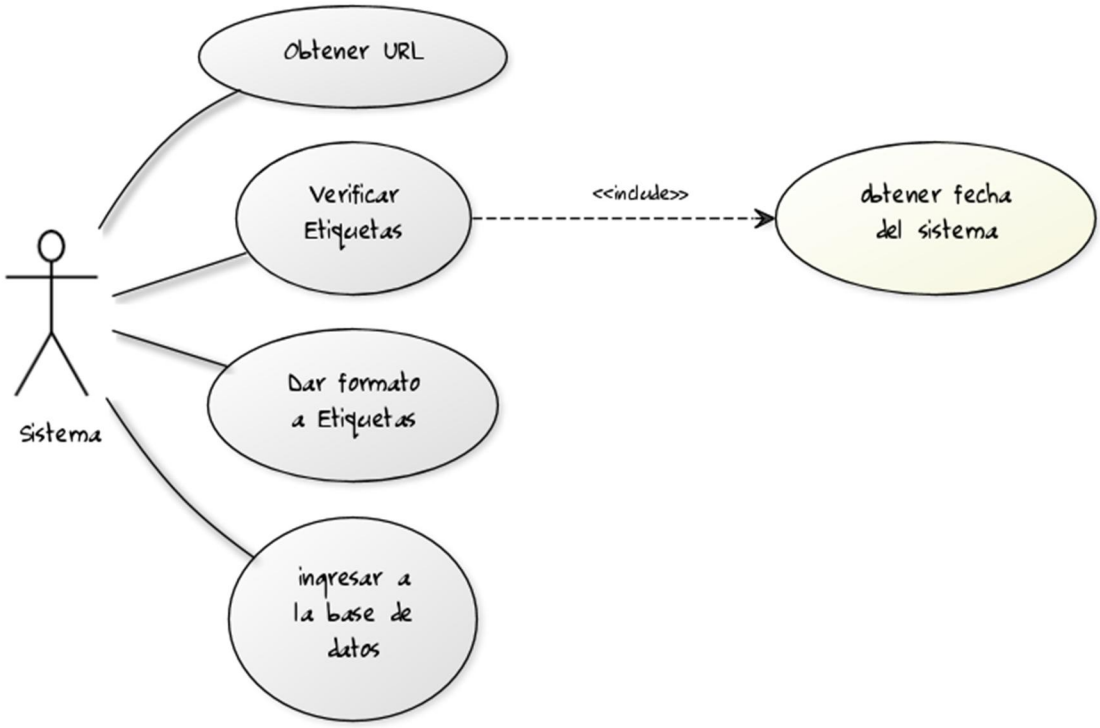

Figura 8: Diagrama de casos de uso.

3.1.1 Especificación de casos de uso

Caso de uso: Obtener URL

- Descripción: Se obtiene un archivo con extensión XML o RSS, se crea internamente un objeto de tipo SimpleXmlElement el cual tendrá toda estructura del archivo.
- Actor: Sistema
- Flujo Básico:
  - El sistema lee la dirección URL del archivo a procesar.
  - Crea un objeto SimpleXmlElement con la información del archivo
- Excepciones: Si la dirección no es válida se muestra un mensaje de falla.

Caso de uso: Verificar Etiquetas

- Descripción: Se debe verificar que las etiquetas pubDate y link existan para que la información sea lo más completa posible.

- Actor: Sistema
- Flujo Básico:
    - Se verifica que la etiqueta pubDate exista si no existe se obtiene la fecha del sistema.
    - Verifica que exista la etiqueta link, si no existe en su lugar se mandara la cadena “no link”.
- Excepciones: ninguna.

Caso de uso: Dar formato a etiquetas.

- Descripción: En este caso de uso se busca limpiar la información a ser recabado de etiquetas HTML, saltos de línea, espacios consecutivos, cambiar el formato a UTF-8.
- Actor: Sistema
- Flujo Básico:
    - El sistema obtiene el contenido de las etiquetas título y descripción.
    - Se eliminan las etiquetas HTML.
    - Se eliminan los saltos de line, retorno de carro y espacios consecutivos.
    - Se transforma el texto a UTF-8.
- Excepciones: ninguna.

Caso de uso: ingresar a la base de datos

- Descripción: Se ingresa la información a la base de datos.
- Actor: Sistema
- Flujo Básico:
    - Se obtiene la información ya limpiada y formateada para ser almacenada.
    - Se realiza la consulta de ingreso a la tabla de la base de datos.
- Excepciones: Si ya existe ese registro en la base de datos se ignora.

3.2 Implementación

Para la implementación se utilizó el lenguaje de programación PHP y para la base de datos el gestor de bases de datos MySQL. Además se agregó a este programa a las tareas programadas de Windows para su ejecución constante.

El programa principal se compone de un método llamado getFeed que realiza las operaciones para una única dirección, una función de apoyo que realiza las operaciones de formato de las etiquetas y la página principal que se encargara

de alimentar al método anterior con cada una de las direcciones que se tienen capturadas.

El método getFeed obtiene una cadena de texto que representa una dirección hacia un archivo RSS o XML que obtiene como parámetro de la página principal, esta dirección es asignada a un objeto SimpleXmlElement el cual tendrá la estructura completa de etiquetas del archivo de la dirección ingresada.

```
 #obtiene el contenido de la URL
@$content = file_get_contents($feed_url);
     #obtiene el contenido de la URL en formato XML
@$x = new SimpleXmlElement($content);
```

Figura 9: objeto SimpleXmlElement

Una vez que se tiene el contenido de la noticia, se procede a verificar que la etiqueta que representa a la fecha de la publicación de la noticia ya que esta es de vital importancia en posteriores procesos, si la etiqueta no está incluida se procede a obtener la fecha del sistema para poder agregarla, también se verifica que exista la etiqueta link, aunque tiene importancia podemos prescindir de ella por lo que si no llegara a estar simplemente colocamos la cadena 'no link', una vez realizado esto, se procede a dar formato a el contenido de las etiquetas título y descripción.

La función conversiones realiza la sustitución de las etiquetas HTML que se encuentren en las etiquetas de título y descripción, además de eliminar los saltos de línea, retorno de carro, espacios consecutivos además de cambiar el formato del texto a UTF-8 para que no se pierdan los símbolos especiales del idioma Español para ellos se utilizan las funciones strip_tags, preg_replace, iconv.

```
function conversiones($cadena){
     #Elimina las etiquetas de la acdena
    $quitar_etiquetas=strip_tags($cadena);
    #Convierte los codigos html a su correspondiente en UTF-8
    $formatoa = html_entity_decode($quitar_etiquetas,ENT_COMPAT,"UTF-8");
    setlocale(LC_CTYPE, 'es_ES.UTF-8');
    $formatoa = iconv("UTF-8","UTF-8//TRANSLIT",$formatoa);
    #Elimina los saltos de linea, retorno de carro y los espacios seguidos
    $formatoa=preg_replace("[\n|\r|\n\r|\s\s+]", ' ', $formatoa);
    return $formatoa;
```

Figura 10: función conversiones.

Una vez que se han realizado los pasos anteriores se procede a realizar la consulta que permita ingresar un nuevo elemento a la tabla detalles_noticias.

La página principal se encarga de proporcionar a la función getFeed las direcciones donde se encuentran los archivos RSS o XML. Estas direcciones se encuentran en un archivo de texto llamado l_noticias.txt este texto contiene alrededor de 441 enlaces a archivos RSS, mediante un ciclo foreach se recorren las URL's además en cada iteración se resetea el tiempo del script ya que este solo consta de 30 segundos lo que sería completamente insuficiente para terminar la ejecución de todas las iteraciones.

```
#obtniene las URL del archivo l_noticias.txt
$lineas = file('C:\wamp\www\noticias\l_noticias.txt',FILE_IGNORE_NEW_LINES);
foreach ($lineas as $num_linea => $linea) {
  #para cada linea realiza la operacion.
    set_time_limit(0); #resetea el tiempo de ejecucion del script
    $linea = rtrim($linea);#elimina los saltos de linea
    try{
        echo '<br>'.$num_linea.'   ';
        getFeed($linea);
    }catch (Exception $e) {
        echo 'Excepci&oacute;n capturada: ',  $e->getMessage(), "\n";
    }
}
```

Figura 11: Código de la página principal

El sistema RSS tiene la cualidad de que se actualiza constantemente por lo que si realizáramos una segunda ejecución inmediatamente se encontrarían nuevas noticias en al menos la mitad de los enlaces, por ello se activó una tarea programada que sea ejecutada en segundo plano cada 10 minutos, se eligió esta cantidad de tiempo ya que es un tiempo recomendable para que se acumulen las noticias.

CAPITULO IV CALCULO DE REPUTACION DE ENTIDADES NOMBRADAS

4.1 Extracción de entidades

Una vez obtenido la fuente de donde obtendremos las entidades procedemos a realizar la aplicación que extraerá las entidades del texto.

Para ello utilizaremos el lenguaje de programación JAVA junto al IDE Netbeans, para comenzar necesitaremos obtener el título y la descripción que anteriormente guardamos en la base de datos Noticias, una vez obtenidos realizaremos el llamado a dos de los programas que nos ayudaran, estos son Tree Tagger y Freeling a cada uno de ellos le pasaremos como parámetro de entrada el título y la descripción de las noticias primero lo haremos con Tree Tagger.

```
//--------------llamada a TreeTagger-------------------------------
Runtime r=Runtime.getRuntime();
Process p;
p = r.exec("script.cmd");//contiene la llamada a TreeTagger con la descripción de la noticia
BufferedReader br=new  BufferedReader(new InputStreamReader(p.getInputStream()));
String resultado;
br.readLine();//se descarta el encabezado del script
br.readLine();
String entidadT = "",linea[];
ent=false;
//se lee la salida del script
while((resultado = br.readLine())!= null){
    linea = resultado.split("\t");
    //linea[0] palabra      linea[1] tag
    entidadT = obtener_entidades(linea, entidadT);
}
```

Figura 12: llamado a Tree Tagger.

Además llamaremos a un método llamado obtener_entidades el cual lo que hará es unir las entidades ya que Tree Tagger identifica palabra por palabra pero no agrupa palabras identificadas como nombre propios por ejemplo: si tenemos como entrada la siguiente frase “El Gobernador Rafael Moreno Valle inauguro el Hospital de la Ciudad de Puebla” nos mostrara la siguiente salida:

Tree Tagger

PALABRA	ETIQUETA	LEMA
El	ART	el
Gobernador	NP	Gobernador
Rafael	NP	Rafael
Moreno	NP	Moreno

Valle	NP	Valle
inauguro	VLfin	inauguro
el	ART	el
nuevo	ADJ	nuevo
Hospital	NP	Hospital
de	PREC	de
la	ART	la
Ciudad	NP	Ciudad
de	PREC	de
Puebla	NP	Puebla

Figura 13: Salida de Tree Tagger.

La manera en la que unirá las palabras es por la clave que maneja Tree Tagger para los nombres propios utiliza una clave NP y NC para los nombres comunes de manera que para cada vez que aparezca la palabra NP en la segunda columna verificara si la palabra a continuación también tiene cualquiera de las dos claves, forman parte de la misma entidad. De esa manera obtendremos las siguientes entidades: Rafael Moreno Valle, Hospital Ciudad y Puebla.

Por parte de Freeling también se realizara una llamada a un script que se le pasara como parámetro de entrada, la salida del análisis se guardara en un archivo para que posteriormente sea leído, también se llamara a un método llamado obtener_entidadesFre el cual hará un procedimiento similar a la función para Tree Tagger detectara si aparece la clave que contenga NP o NC, además dado que en la salida Freeling al unir las palabras que considera como nombres propios sustituye los espacios con guiones bajos, del ejemplo anterior tendríamos la siguiente salida:

Freeling

TEXTO	**LEMA**	**ETIQUETA**
El	el	DA0MS0
Gobernador_Rafael_Moreno_Valle	gobernador_rafael_moreno_valle	NP00000
inauguro	Inaugurar	VMIP1S0
el	el	DA0MS0
nuevo	nuevo	AQ0MS0
Hospital_de_la_Ciudad_de_Puebla	hospital_de_la_ciudad_de_puebla	NP00000

Figura 14: Salida de Freeling.

Tendríamos como entidades: Gobernador Rafael Moreno Valle y Hospital de la Ciudad de Puebla, como podemos observar las entidades que obtenemos con Freeling son más específicas que las de Tree Tagger.

```
Runtime r1=Runtime.getRuntime();
Process p1;
p1= r1.exec("script1.cmd");
//contiene la llamada a Freeling con la descripción de la noticia
while (p1.waitFor()!=0);
BufferedReader bf1 = new BufferedReader(new FileReader("salidatag.txt"));
//salidatag es el archivo que contiene
String entidadF = "";//la salida: tags de Freeling
try{
while((resultado = bf1.readLine())!= null){
    if(!resultado.equals("")){
    linea = resultado.split(" ");
    //linea[0] palabra      linea[2] tag
    //mientras va lleyendo las lineas del archivo de salida del
    //script se van obteniendo las entidades
    entidadF = obtener_entidadesFre(linea, entidadF);
    }
}
```

Figura 15: llamado a Freeling

Al aplicar estos Scripts a cada una de las noticias obtenemos las entidades nombradas de cada una de ellas, así al estar guardadas en la base de datos las podemos utilizar más adelante, también las guardaremos en un archivo de texto llamado entidades.txt para procesarlas en el conteo de entidades.

4.2 Conteo de entidades nombradas

Una vez obtenidas las entidades nombradas debemos contabilizarlas para poder generar una línea de tendencia y así conocer el comportamiento de la entidad nombrada con el tiempo además de pre-procesar la información para cuando se realice el cálculo de la reputación.

Para comenzar el análisis debemos procesar la información de las noticias que se obtuvieron anteriormente y que se encuentran almacenadas en la base de datos para ello colocaremos la información en un archivo de texto llamado noticias.txt con el siguiente formato:

1 | Fecha | título y descripción
2 | Fecha | título y descripción
.
.
.
n | Fecha | título y descripción

Un ejemplo seria el siguiente:

```
1 | 22 de febrero de 2013 | Inicia sesión solemne en Senado por centenario del Ejército El
presidente de la Mesa Directiva, Ernesto Cordero, decretó un receso para recibir a los
invitado, entre los que destacan el presidente Enrique Peña Nieto y el titular de la
Sedena, Salvador Cienfuegos.
2 | 22 de febrero de 2013 | Informará procuradora de Guerrero a diputados sobre caso de
españolas Martha Elva Garzón Bernal recibirá en la PGJE a la Comisión Especial del Congreso
local para atender inquietudes de los legisladores sobre las turistas violadas.
3 | 22 de febrero de 2013 | Desarticulan autogobierno al interior del penal en Oaxaca
Policías federales y estatales ingresaron esta mañana a la cárcel de alta seguridad de
Miahuatlán de Porfirio Díaz, en la Sierra Sur del estado y aseguraron armas blancas,
pequeñas dosis de droga entre otros objetos.
4 | 22 de febrero de 2013 | Piden juristas y especialistas garantizar libertad religiosa
Especialistas del Derecho y la Filosofía Política de instituciones públicas y privadas
pidieron promover la tolerancia como principio rector de la democracia en México.
5 | 22 de febrero de 2013 | La Fepade, institución de confianza para los jóvenes, según
encuesta De acuerdo con la encuesta "La Cultura Política de los Jóvenes en México", el 40
por ciento de los jóvenes considera que el IFE está influido por el gobierno.
6 | 22 de febrero de 2013 | Proponen premiar a policías por detener al 'Niño Verde' La
asambleísta del PRD, Dione Anguiano, consideró que la actitud del agente de la SSP DF,
Antonio Caracho, se vuelve un candidato a que se le entregue la Medalla al Mérito Policial.
7 | 22 de febrero de 2013 | Asesinan a dos en emboscada en Guerrero El ataque se registró
esta mañana cuando tres personas regresaban de cortar leña en la Montaña.
8 | 22 de febrero de 2013 | Protesta "Chucho el Roto" ante la CFE por tarifas, en Veracruz
La Secretaría de Gobierno estatal se comprometió a buscar un acuerdo para que la
paraestatal haga una revisión minuciosa de las tarifas eléctricas.
```

Utilizaremos el archivo de texto entidade.txt don están guardadas todas las entidades nombradas extraídas. Al realizar esto crearemos un algoritmo que nos identifique en que días aparecen cada una de las entidades y con el índice de la noticia. Esto se realizara en el lenguaje de programación AWK y tendrá como parámetros de entrada los dos archivos de texto que se mencionaron anteriormente:

Ejecutando el programa de la siguiente forma:

./cuentaentidades.awk entidades.txt noticias.txt > resultado.txt

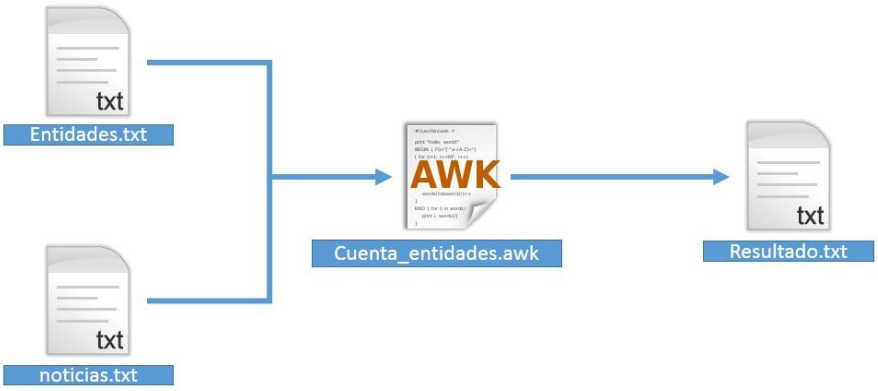

Figura 16: diagrama de la ejecución del archivo cuenta_entidades.awk

Un ejemplo de las líneas del archivo resultado.txt es la siguiente:

mario riestra piña | 21 de mayo de 2013 | 21-05-2013 | 2 | 2 | ,215751 ,216692

A continuación se explican cada uno de los campos que están divididos por el símbolo '|':

CAMPO	VALOR	DESCRIPCIÓN
Entidad nombrada	Mario Riestra Piña	La entidad que fue contada.
Fecha larga	21 de mayo de 2013	Fecha en formato largo.
Fecha corta	21-05-2013	Fecha en formato corto.
Conteo total	2	Número de veces en la que aparece la entidad en la fecha seleccionada.
Conteo sin repeticiones	2	Número de veces en las que aparece la entidad sin contar repeticiones en la misma noticia. Este valor debe coincidir con el número de índices de las noticias.
Listado de índices de las noticias	215751, 216692	Se muestran los índices de las noticias donde aparece la entidad para la fecha seleccionada.

Ahora que ya tenemos las entidades por fecha, debemos de agrupar los datos por entidad ya que por el momento aparecen las entidades repetidas por el número de días que aparecen haciendo el archivo muy extenso y pesado para trabajar con él, para ello debemos de agrupar las fechas en las que aparecen las entidades del siguiente modo:

Entidad: Fecha * frecuencia * indice|…FechaN * frecuenciaN * indicen |

Por ejemplo una línea del nuevo archivo seria:

alejandro benítez: 26-05-2013 * 1 * ,241585 | 06-06-2013 * 1 * ,286041 | 22-05-2013 * 1 * ,221822 | 24-05-2013 * 1 * ,236017

CAMPO	VALOR	DESCRIPCIÓN
Entidad	Alejandro Benítez	Nombre de la entidad nombrada
Datos de la entidad	26-05-2013 * 1 * ,241585	Fecha corta en la que la entidad fue nombrada, frecuencia sin repetición, índices de las noticias.

Cada dato de la entidad fue dividido por el símbolo '|' de esta manera se puede contabilizar la veces en las que la entidad fue nombrada, a su vez los datos específicos de cada fecha fueron divididos con el símbolo '*' para un acceso fácil de estos datos.

Como se explicó anteriormente la entidad nombrada Alejandro Benítez apareció en cuatro fechas distintas en 1 ocasión y en los índices 241585, 286041, 221822 y 236017.

De esta manera el archivo reduce su tamaño y se tiene una mejor organización de los datos.

4.3 Calculo de la reputación de las entidades nombradas

Para llevar a cabo esta tarea se optó por utilizar un diccionario que contiene palabras asociadas a un valor de afectividad [23] a las siguientes emociones: alegría, furia, miedo, tristeza, sorpresa y disgusto. Las cuales fueron separadas en dos categorías positiva (alegría y sorpresa) y negativa (furia, miedo y disgusto). Una muestra de los datos en el diccionario [23] es la siguiente:

palabra	Valor de afinidad	Tipo	palabra	Valor de afinidad	Tipo
abundancia	0.83	Alegría	aborrecimiento	0.966	Repulsión
acabalar	0.396	Alegría	abusar	0.697	Repulsión
acallar	0.198	Alegría	adversión	0.497	Repulsión
acatar	0.198	Alegría	alergia	0.397	Repulsión
acción	0.397	Alegría	ampolla	0.497	Repulsión
aceptable	0.594	Alegría	animadversión	0.664	Repulsión
aceptación	0.696	Alegría	animosidad	0.297	Repulsión
acicate	0.429	Alegría	antipatía	0.662	Repulsión
aclamación	0.799	Alegría	antipático	0.728	Repulsión
aclamar	0.799	Alegría	apestar	0.831	Repulsión
acogedor	0.83	Alegría	apestoso	0.899	Repulsión
acoger	0.729	Alegría	aprensión	0.529	Repulsión
acomodadamente	0.729	Alegría	araña	0.429	Repulsión
acuciar	0.264	Alegría	asco	0.966	Repulsión
acucioso	0.232	Alegría	asqueadamente	0.932	Repulsión
adecuar	0.331	Alegría	asquear	0.932	Repulsión
adicto	0.429	Alegría	asquerosamente	0.966	Repulsión
admirable	0.764	Alegría	asqueroso	1	Repulsión
admirablemente	0.765	Alegría	astroso	0.629	Repulsión

admiración	0.765	Alegría	aversión	0.629	Repulsión
admirar	0.731	Alegría	basca	0.865	Repulsión
admitir	0.53	Alegría	bascosidad	0.798	Repulsión
adorable	0.898	Alegría	bascoso	0.664	Repulsión
adorablemente	0.865	Alegría	basura	0.796	Repulsión
adoración	0.765	Alegría	birria	0.331	Repulsión
adorador	0.664	Alegría	borracho	0.563	Repulsión
adorar	0.764	Alegría	caca	0.898	Repulsión
afable	0.696	Alegría	carroña	0.864	Repulsión
afán	0.764	Alegría	cicatriz	0.396	Repulsión
afectivo	0.864	Alegría	coágulo	0.397	Repulsión
afecto	0.899	Alegría	cochino	0.562	Repulsión
afectuosamente	0.966	Alegría	cocolía	0.631	Repulsión
afectuosidad	0.932	Alegría	contaminar	0.63	Repulsión
afectuoso	0.898	Alegría	corromper	0.596	Repulsión
afervorizar	0.731	Alegría	cortar	0.165	Repulsión
afición	0.729	Alegría	desagradable	0.898	Repulsión
aficionar	0.595	Alegría	desagradablemente	0.831	Repulsión
afortunadamente	0.831	Alegría	desagrado	0.798	Repulsión
afortunado	0.932	Alegría	desasosegar	0.298	Repulsión
agradable	0.899	Alegría	descomposicón	0.763	Repulsión
agradar	0.798	Alegría	deshonrar	0.63	Repulsión
agradecer	0.763	Alegría	desperdicio	0.562	Repulsión
agradecido	0.764	Alegría	detestable	0.966	Repulsión
agudeza	0.331	Alegría	detestablemente	0.932	Repulsión
aguzar	0.199	Alegría	detestación	0.932	Repulsión
ahínco	0.73	Alegría	diarrea	0.697	Repulsión
ahíto	0.564	Alegría	disgustar	0.595	Repulsión
airoso	0.565	Alegría	disgusto	0.595	Repulsión
alabar	0.53	Alegría	distanciamiento	0.363	Repulsión
alborozar	0.899	Alegría	empalagamiento	0.33	Repulsión
alborozo	0.899	Alegría	empalagar	0.33	Repulsión
alegrar	0.966	Alegría	empalagoso	0.363	Repulsión
alegre	1	Alegría	emporcar	0.595	Repulsión
alegremente	1	Alegría	enfermedad	0.463	Repulsión
alegría	1	Alegría	enfermizo	0.496	Repulsión
alentar	0.73	Alegría	enfermo	0.564	Repulsión
aliviar	0.763	Alegría	ensuciar	0.662	Repulsión
alivio	0.763	Alegría	envilecer	0.798	Repulsión
allegar	0.364	Alegría	escrúpulo	0.33	Repulsión
alma	0.397	Alegría	espanto	0.264	Repulsión

altivo	0.132	Alegría	estiércol	0.73	Repulsión
altruismo	0.529	Alegría	excremento	0.798	Repulsión
alzar	0.298	Alegría	execrable	0.763	Repulsión
amabilidad	0.728	Alegría	execrablemente	0.763	Repulsión
amable	0.762	Alegría	fastidiar	0.53	Repulsión
amablemente	0.762	Alegría	fastidio	0.597	Repulsión
amante	0.731	Alegría	fétido	0.898	Repulsión
amar	0.899	Alegría	fila	0.366	Repulsión
amartelado	0.966	Alegría	fobia	0.663	Repulsión
amartelar	0.731	Alegría	gases	0.663	Repulsión
amativo	0.764	Alegría	gelatinoso	0.231	Repulsión
amatorio	0.832	Alegría	grima	0.629	Repulsión
ambición	0.429	Alegría	guácala	0.966	Repulsión
ambicioso	0.363	Alegría	gusano	0.462	Repulsión
amigabilidad	0.898	Alegría	harto	0.497	Repulsión
amigable	0.898	Alegría	hastiar	0.596	Repulsión
amigablemente	0.864	Alegría	hastío	0.563	Repulsión
amigo	0.898	Alegría	hediondo	0.864	Repulsión
amistad	0.966	Alegría	herida	0.33	Repulsión
amistosamente	0.898	Alegría	hincha	0.664	Repulsión
amistoso	0.966	Alegría	hinchado	0.565	Repulsión
amor	0.966	Alegría	hongos	0.396	Repulsión
amorosamente	0.932	Alegría	horrendo	0.731	Repulsión
amorosidad	0.898	Alegría	horrible	0.764	Repulsión
amoroso	0.898	Alegría	horripilación	0.697	Repulsión
amuleto	0.398	Alegría	horror	0.764	Repulsión
anhelo	0.297	Alegría	horroroso	0.764	Repulsión
animación	0.562	Alegría	hostil	0.498	Repulsión
animado	0.865	Alegría	impúdico	0.496	Repulsión
animar	0.797	Alegría	impuro	0.597	Repulsión
ánimo	0.864	Alegría	incasto	0.43	Repulsión
animoso	0.83	Alegría	indecente	0.63	Repulsión
anticipar	0.231	Alegría	indecoroso	0.563	Repulsión
apaciguar	0.298	Alegría	indisponer	0.364	Repulsión
apasionadamente	0.663	Alegría	infame	0.63	Repulsión
apasionamiento	0.696	Alegría	infección	0.531	Repulsión
apasionante	0.798	Alegría	infecto	0.664	Repulsión
apego	0.597	Alegría	infrahumano	0.431	Repulsión
apetecer	0.363	Alegría	abierto	0.298	Sorpresa
aplaudir	0.797	Alegría	abobamiento	0.199	Sorpresa
aplauso	0.763	Alegría	abobar	0.298	Sorpresa

apreciar	0.629	Alegría	acertijo	0.531	Sorpresa
aprobación	0.596	Alegría	acojonante	0.73	Sorpresa
abominable	0.53	Enojo	adivinanza	0.498	Sorpresa
abominación	0.464	Enojo	admirable	0.73	Sorpresa
abominar	0.696	Enojo	admirablemente	0.663	Sorpresa
aborrecer	0.798	Enojo	admiracion	0.664	Sorpresa
aborrecible	0.729	Enojo	admiración	0.764	Sorpresa
aborreciblemente	0.629	Enojo	admirar	0.73	Sorpresa
aborrecimiento	0.663	Enojo	agenciar	0.198	Sorpresa
abusar	0.463	Enojo	alcanzar	0.265	Sorpresa
acometedor	0.463	Enojo	alelamiento	0.198	Sorpresa
acometer	0.497	Enojo	alelar	0.132	Sorpresa
acometiente	0.464	Enojo	anonadamiento	0.765	Sorpresa
acometimiento	0.431	Enojo	anonadar	0.731	Sorpresa
acometividad	0.397	Enojo	apoteósico	0.598	Sorpresa
acosar	0.465	Enojo	arrocinar	0.165	Sorpresa
acribillar	0.666	Enojo	asombramiento	0.932	Sorpresa
acrimonia	0.363	Enojo	asombrar	0.966	Sorpresa
acritud	0.33	Enojo	asombro	0.966	Sorpresa
adversario	0.363	Enojo	asombrosamente	0.966	Sorpresa
aferrar	0.265	Enojo	asombroso	0.966	Sorpresa
afrenta	0.53	Enojo	atarantar	0.132	Sorpresa
agarrar	0.231	Enojo	atolondrar	0.199	Sorpresa
agobiar	0.297	Enojo	atónito	0.799	Sorpresa
agobio	0.231	Enojo	atontamiento	0.232	Sorpresa
agravamiento	0.397	Enojo	atontar	0.232	Sorpresa
agravante	0.463	Enojo	atrapar	0.463	Sorpresa
agravar	0.43	Enojo	atronar	0.265	Sorpresa
agraviar	0.597	Enojo	aturdimiento	0.165	Sorpresa
agravio	0.43	Enojo	aturdir	0.165	Sorpresa
agresividad	0.831	Enojo	aturrullar	0.165	Sorpresa
agresivo	0.865	Enojo	aturullar	0.165	Sorpresa
agresor	0.83	Enojo	batir	0.066	Sorpresa
agriado	0.497	Enojo	bizco	0.265	Sorpresa
airadamente	0.596	Enojo	bobalicón	0.066	Sorpresa
airamiento	0.563	Enojo	bobo	0.099	Sorpresa
airar	0.663	Enojo	bombo	0.331	Sorpresa
alevosía	0.396	Enojo	boquiabierto	0.799	Sorpresa
altivez	0.264	Enojo	bruto	0.165	Sorpresa
altivo	0.33	Enojo	chasco	0.399	Sorpresa
amargar	0.363	Enojo	coger	0.429	Sorpresa

amargo	0.165	Enojo	confundir	0.264	Sorpresa
amargor	0.132	Enojo	confusión	0.264	Sorpresa
amohinar	0.599	Enojo	conmoción	0.631	Sorpresa
amoscamiento	0.397	Enojo	conmocionar	0.663	Sorpresa
animadversión	0.497	Enojo	conseguir	0.33	Sorpresa
ánimo	0.099	Enojo	consternación	0.463	Sorpresa
animosidad	0.132	Enojo	consternar	0.43	Sorpresa
antagónico	0.199	Enojo	conturbar	0.632	Sorpresa
antagonismo	0.364	Enojo	descomunalmente	0.798	Sorpresa
antipatía	0.662	Enojo	desconcertar	0.664	Sorpresa
antipático	0.695	Enojo	desconcierto	0.631	Sorpresa
apesadumbrar	0.495	Enojo	descuidar	0.363	Sorpresa
arrebatamiento	0.63	Enojo	deslumbramiento	0.597	Sorpresa
arrebato	0.497	Enojo	deslumbrantemente	0.664	Sorpresa
arrojar	0.496	Enojo	deslumbrar	0.664	Sorpresa
asar	0.165	Enojo	desmayar	0.53	Sorpresa
asediar	0.364	Enojo	desmayo	0.53	Sorpresa
aspereza	0.298	Enojo	despampanante	0.833	Sorpresa
asurar	0.398	Enojo	despistar	0.297	Sorpresa
atacar	0.832	Enojo	desprevenido	0.661	Sorpresa
ataque	0.899	Enojo	distraído	0.463	Sorpresa
atizar	0.465	Enojo	embobamiento	0.132	Sorpresa
atormentar	0.664	Enojo	embobar	0.099	Sorpresa
atosigar	0.563	Enojo	embobecer	0.165	Sorpresa
atrabiliario	0.664	Enojo	embrutecer	0.165	Sorpresa
atrocidad	0.331	Enojo	encandilar	0.597	Sorpresa
atropellar	0.398	Enojo	enigma	0.565	Sorpresa
atropello	0.397	Enojo	enmudecer	0.664	Sorpresa
atroz	0.697	Enojo	entontecer	0.198	Sorpresa
atufar	0.398	Enojo	entontecimiento	0.099	Sorpresa
azuzar	0.264	Enojo	espantar	0.831	Sorpresa
baquetear	0.365	Enojo	espanto	0.831	Sorpresa
barbaridad	0.364	Enojo	estólido	0.099	Sorpresa
batallar	0.463	Enojo	estrellas	0.232	Sorpresa
belicosidad	0.53	Enojo	estruendo	0.497	Sorpresa
belicoso	0.53	Enojo	estupefacción	0.798	Sorpresa
beligerante	0.396	Enojo	estupefacto	0.899	Sorpresa
berrinche	0.629	Enojo	estupendamente	0.664	Sorpresa
bilis	0.532	Enojo	estupendo	0.698	Sorpresa
bochinche	0.363	Enojo	estúpido	0.099	Sorpresa
bravo	0.563	Enojo	estupor	0.765	Sorpresa

bufar	0.532	Enojo	exclamación	0.497	Sorpresa
cabrear	0.597	Enojo	extrañar	0.198	Sorpresa
cargoso	0.363	Enojo	extrañeza	0.464	Sorpresa
celos	0.63	Enojo	extraordinario	0.665	Sorpresa
chalar	0.232	Enojo	fabuloso	0.597	Sorpresa
chinche	0.396	Enojo	fantásticamente	0.696	Sorpresa
cinismo	0.298	Enojo	fantástico	0.763	Sorpresa
cizañar	0.463	Enojo	fascinación	0.663	Sorpresa
cocolía	0.53	Enojo	fascinante	0.73	Sorpresa
cólera	0.73	Enojo	frustrar	0.165	Sorpresa
coléricamente	0.83	Enojo	golpe	0.265	Sorpresa
colérico	0.762	Enojo	grandioso	0.63	Sorpresa
concitar	0.363	Enojo	gritar	0.528	Sorpresa
contrariar	0.463	Enojo	grito	0.528	Sorpresa
contrariedad	0.331	Enojo	horrible	0.429	Sorpresa
convulsión	0.198	Enojo	idiotizar	0.099	Sorpresa
coraje	0.899	Enojo	abajamiento	0.497	Tristeza
corajudo	0.831	Enojo	abajo	0.364	Tristeza
corroer	0.297	Enojo	abandonamiento	0.865	Tristeza
crispar	0.43	Enojo	abandonar	0.898	Tristeza
crucificar	0.198	Enojo	abatidamente	0.832	Tristeza
abominable	0.797	Miedo	abatido	0.865	Tristeza
accidente	0.696	Miedo	abatimiento	0.798	Tristeza
acobardar	0.865	Miedo	abatir	0.765	Tristeza
acomplejado	0.597	Miedo	abochornar	0.264	Tristeza
acoquinamiento	0.598	Miedo	abrumar	0.43	Tristeza
acoquinar	0.831	Miedo	aburrir	0.264	Tristeza
agüero	0.297	Miedo	aciago	0.563	Tristeza
ahuyentar	0.696	Miedo	acongojar	0.796	Tristeza
alarma	0.596	Miedo	acuitadamente	0.463	Tristeza
alarmado	0.695	Miedo	acuitar	0.562	Tristeza
alarmante	0.595	Miedo	adolecer	0.563	Tristeza
alarmar	0.63	Miedo	adverso	0.464	Tristeza
alertar	0.53	Miedo	aflicción	0.764	Tristeza
alma	0.066	Miedo	afligidamente	0.83	Tristeza
amedentrar	0.898	Miedo	afligido	0.83	Tristeza
amedrentador	0.831	Miedo	afligir	0.796	Tristeza
amedrentamiento	0.831	Miedo	agobiado	0.764	Tristeza
amenazar	0.865	Miedo	agobiante	0.73	Tristeza
amilanar	0.831	Miedo	agobiantemente	0.73	Tristeza
angustia	0.797	Miedo	agobiar	0.697	Tristeza

angustiado	0.831	Miedo	agraviar	0.331	Tristeza
angustiosamente	0.831	Miedo	aguantar	0.364	Tristeza
ansioso	0.264	Miedo	aherrojar	0.198	Tristeza
aprensión	0.397	Miedo	ahogo	0.463	Tristeza
aprensivo	0.364	Miedo	alicaído	0.464	Tristeza
araña	0.463	Miedo	aliquebrado	0.497	Tristeza
arredrar	0.463	Miedo	alma	0.165	Tristeza
asesinar	0.73	Miedo	amargo	0.264	Tristeza
asustadizo	0.864	Miedo	amargura	0.528	Tristeza
asustar	0.864	Miedo	amurriar	0.697	Tristeza
atemorizado	0.932	Miedo	angustia	0.763	Tristeza
atemorizante	0.932	Miedo	angustiadamente	0.697	Tristeza
atemorizar	0.932	Miedo	angustiar	0.729	Tristeza
atento	0.099	Miedo	angustioso	0.763	Tristeza
aterrado	0.966	Miedo	aniquilar	0.463	Tristeza
aterrador	1	Miedo	añicos	0.463	Tristeza
aterradoramente	1	Miedo	apabullar	0.429	Tristeza
aterrar	1	Miedo	apenar	0.397	Tristeza
aterrorizador	0.966	Miedo	apiadarse	0.464	Tristeza
aterrorizar	0.966	Miedo	aplanar	0.166	Tristeza
atroz	0.696	Miedo	aplastar	0.232	Tristeza
avergonzado	0.297	Miedo	aquejar	0.497	Tristeza
avergonzar	0.33	Miedo	arrastradamente	0.297	Tristeza
azoramiento	0.529	Miedo	arrastrar	0.198	Tristeza
azorar	0.696	Miedo	arrepentimiento	0.462	Tristeza
barbarie	0.463	Miedo	arrepentir	0.462	Tristeza
bruja	0.495	Miedo	asolar	0.53	Tristeza
brujo	0.495	Miedo	atormentadamente	0.496	Tristeza
cadaver	0.697	Miedo	atormentar	0.53	Tristeza
calaca	0.462	Miedo	atribular	0.529	Tristeza
calamidad	0.396	Miedo	atrición	0.495	Tristeza
calamitoso	0.363	Miedo	avergonzar	0.297	Tristeza
calofrío	0.463	Miedo	bajo	0.264	Tristeza
cohibición	0.463	Miedo	cabizbajo	0.898	Tristeza
cohibidamente	0.43	Miedo	cabizcaído	0.898	Tristeza
cohibir	0.496	Miedo	caer	0.528	Tristeza
confuso	0.099	Miedo	caerse	0.462	Tristeza
conminar	0.529	Miedo	caída	0.529	Tristeza
convulsión	0.33	Miedo	caído	0.528	Tristeza
cortar	0.297	Miedo	calvario	0.763	Tristeza
cortedad	0.165	Miedo	cancamurria	0.799	Tristeza

cruento	0.33	Miedo	carcel	0.596	Tristeza
demonio	0.63	Miedo	carecer	0.529	Tristeza
desesperado	0.33	Miedo	cargar	0.33	Tristeza
desfallecer	0.364	Miedo	castigo	0.595	Tristeza
desfallecimiento	0.43	Miedo	cetrino	0.165	Tristeza
desgracia	0.629	Miedo	compadecer	0.564	Tristeza
desmayar	0.363	Miedo	compasión	0.63	Tristeza
desorientado	0.132	Miedo	compasivo	0.63	Tristeza
despavorir	0.864	Miedo	compunción	0.464	Tristeza
despeluznante	0.831	Miedo	compungidamente	0.331	Tristeza
despeluznar	0.696	Miedo	compungir	0.397	Tristeza
despiadadamente	0.663	Miedo	condolencia	0.796	Tristeza
despiadado	0.663	Miedo	condoler	0.796	Tristeza
diablo	0.664	Miedo	confuso	0.231	Tristeza
encoger	0.399	Miedo	congoja	0.496	Tristeza
enfriar	0.066	Miedo	congojoso	0.529	Tristeza
escalofrío	0.73	Miedo	consternación	0.595	Tristeza
espantable	0.831	Miedo	contrición	0.429	Tristeza
espantadizo	0.899	Miedo	contristar	0.663	Tristeza
espantar	0.899	Miedo	contrito	0.63	Tristeza
espanto	0.899	Miedo	crisis	0.696	Tristeza
espantoso	0.899	Miedo	cuita	0.73	Tristeza
espeluznante	0.932	Miedo	cuitado	0.597	Tristeza
espeluznar	0.932	Miedo	culpa	0.598	Tristeza
esperpento	0.529	Miedo	culpabilidad	0.631	Tristeza
espiritu	0.231	Miedo	culpable	0.565	Tristeza
estremecer	0.628	Miedo	culpado	0.631	Tristeza
execrable	0.429	Miedo	culposo	0.631	Tristeza
exorcismo	0.629	Miedo	dañino	0.562	Tristeza
fantasma	0.596	Miedo	debilitar	0.33	Tristeza
fealdad	0.396	Miedo	decaer	0.729	Tristeza
feo	0.297	Miedo	decaído	0.831	Tristeza
fobia	0.932	Miedo	decaimiento	0.797	Tristeza
frío	0.099	Miedo	decepción	0.898	Tristeza
friolero	0.099	Miedo	decepcionar	0.898	Tristeza
gallina	0.265	Miedo	defectuoso	0.33	Tristeza
hechizo	0.297	Miedo	deficiente	0.363	Tristeza
helado	0.099	Miedo			
helar	0.033	Miedo			
histeria	0.562	Miedo			
abominablemente	0.73	Repulsión			

abominar	**0.797**	**Repulsión**			
aborrecer	**0.966**	**Repulsión**			
aborrecible	**0.966**	**Repulsión**			

Una vez que se obtuvo el diccionario lo siguiente es obtener las oraciones donde estén incluidas las entidades nombradas se tomaron varias muestras con respecto a la longitud que se deberían de tomar con respecto a la posición de la entidad nombrada dentro del texto, las opciones fueron de 10, 20 y 50 palabras, siendo la ultima la que dio mejores resultados además de eliminar de las oraciones las palabras cerradas que son aquellas que no tienen valor de afectividad para la oración.

A continuación se muestran algunas de las palabras cerradas que se eliminaron del texto:

PALABRA	PALABRA	PALABRA
asimismo	**dará**	**deba**
debiendo	**debieses**	**debíamos**
dejando	**dejarás**	**dejaste**
dicen	**diese**	**dijera**
diremos	**diéramos**	**estar**
dar	**darán**	**debamos**
debiera	**debimos**	**debían**
dejara	**dejaré**	**deje**
dices	**diesen**	**dijeran**
dirá	**diéremos**	**estaba**
da	**darás**	**deban**
debieran	**debiste**	**debías**
dejaran	**dejaría**	**dejemos**
diciendo	**dieses**	**dijeras**
dirán	**diésemos**	**estaban**
daba	**daré**	**debas**
debieras	**debiéramos**	**decir**
dejaras	**dejaríamos**	**dejen**
diera	**diga**	**dijere**
dirás	**doy**	**estabas**
daban	**daría**	**debe**
debiere	**debiéremos**	**decimos**
dejare	**dejarían**	**dejes**
dieran	**digamos**	**dijeren**
diré	**dábamos**	**estamos**
dabas	**daríamos**	**debemos**

debieren	**debiésemos**	**decía**
dejaremos	**dejarías**	**dejo**
dieras	**digan**	**dijeres**
diría	**dé**	**estando**
damos	**darían**	**deben**
debieres	**debió**	**decíamos**
dejares	**dejas**	**dejábamos**
diere	**digas**	**dijese**
diríamos	**ser**	**estaremos**
dan	**darías**	**deberemos**
debieron	**debo**	**decían**
dejaron	**dejase**	**dejáramos**
dieren	**digo**	**dijesen**
dirían	**eran**	**estará**
dando	**das**	**deberá**
debiese	**debí**	**decías**
dejará	**dejasen**	**dejáremos**
dieres	**dije**	**dijeses**
dirías	**eres**	**estarán**
daremos	**deber**	**deberán**
debiesen	**debía**	**dejar**
dejarán	**dejases**	**dejásemos**
dieron	**dijeron**	**dijimos**
diste	**deberás**	**estarás**

Al aplicar una comparación con las palabras del diccionario, sumando o restando el valor de afectividad de las mismas a un contador, de esta forma se pueda tener al final un valor total y verificar que:

- si es mayor a 0 entonces en esa oración se habla de manera positiva de la entidad.
- Si el valor es 0 se dice que la reputación es neutral.
- Si el valor está por debajo de 0 entonces se habla de forma negativa de la entidad.

A continuación se muestran algunos de los resultados del cálculo de la reputación de las entidades nombradas.

Reputación Negativa:

-1.429 | Negativo |todo lo que ésta representa dejando a nuestro mundo en un estado de crisis con una cuenta personal que saldar el capitán kirk dirigirá una cacería humana en un mundo en guerra para capturar a un hombre que es en realidad un arma de destrucción masiva mientras nuestros héroes son arrojados.

Palabra	Afinidad	Valor de afinidad
crisis	Negativa	**-0.696**
guerra	Negativa	**-0.832**
capturar	Positiva	**0.099**
		-1.429

Reputación Positiva:

1.393 | Positivo |recomienda secretaría de salud realizarse la vasectomía la secretaría de salud de baja california invita a todos los hombres que tengan su paternidad satisfecha a que se realicen la vasectomía sin bisturí un método rápido sencillo y gratuito de planificación familiar el secretario de salud,

Palabra	Afinidad	Valor de afinidad
satisfecha	Positiva	**0.864**
familiar	Positiva	**0.529**
		1.393

Reputación Neutral:

0 | Neutral |con motivo de la intención que tiene un grupo de empresarios de la educación para poner en operación una unidad de la universidad autónoma de Durango en esta ciudad específicamente en la colonia torreón jardín se trata de un proyecto que abriría un espacio para más de 1400 alumnos y además con.

De las 2, 238,288 oraciones analizadas tenemos los siguientes resultados:

Reputación	Cantidad	Porcentaje
Positiva	**562,278**	**25.12%**
Negativa	**370,289**	**16.54%**
Neutral	**1,305,721**	**58.33%**
Total	**2,238,288**	**100%**

Figura 17: Resultados Finales.

CAPÍTULO V VISUALIZACIÓN DE RESULTADOS

Una vez realizados los procesos anteriores se procede a mostrar los resultados obtenidos, de manera que se puedan utilizar para analizar las tendencias y la reputación de las entidades extraídas, para ello se creara una página web donde se puedan consultar los datos de las entidades.

Para comenzar los datos que se generaron a lo largo de este trabajo fueron almacenados en la base de datos Noticias que se creó en el capítulo III de esta forma la consulta es muy rápida y eficiente.

A continuación se muestra la estructura de la página:

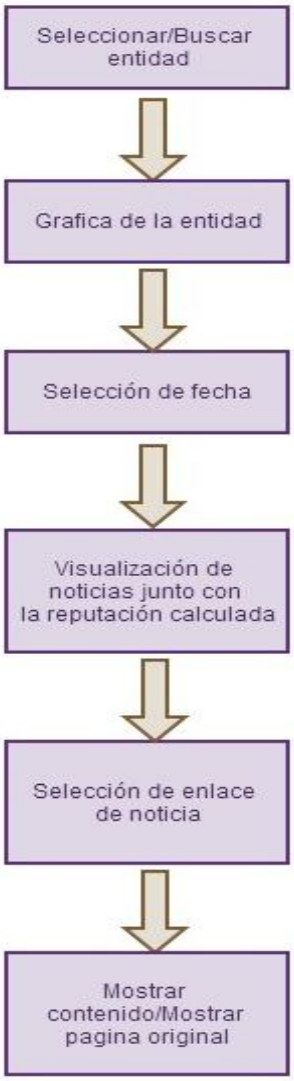

Figura 18: Estructura de la página web.

Los detalles de cada una de las secciones de la página se describen a continuación.

- Seleccionar / Buscar Entidad

  Al cargar la página se mostrara una lista de 2000 entidades, mediante el Scroll del lado derecho de la lista puede desplazarse, si desea cargar otras 2000 entidades lo puede hacer pulsando el botón siguiente, de este modo puede recorrer todas las entidades almacenadas en la base de datos, si desea visualizar las entidades anteriores lo puede hacer pulsando el botón de Anterior, así la página mantiene fluidez y no se ve entorpecida al cargar y almacenar todas las entidades en la lista.

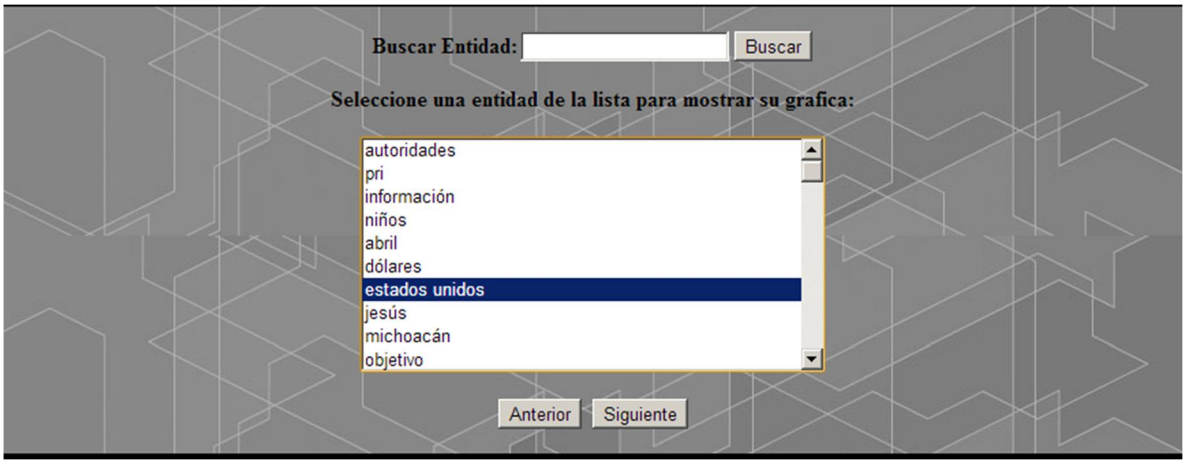

Figura 19: selección de entidades.

Si desea buscar una entidad en particular lo puede hacer escribiendo en la caja de texto y pulsando en el botón buscar que se encuentra arriba de la lista de entidades, le mostrara todas las entidades que comiencen con la palabra o palabras que se escribieron en la caja de texto.

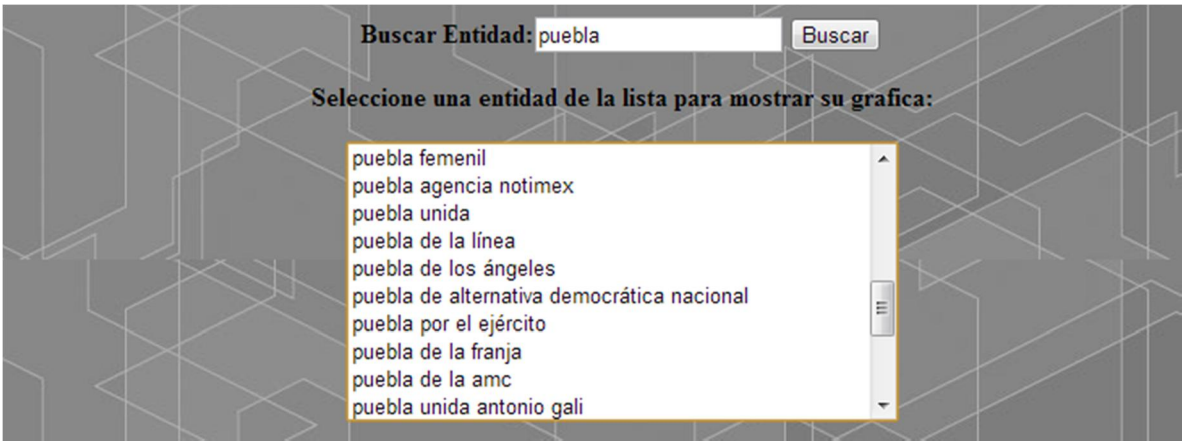

Figura 20: Búsqueda de una entidad

- Grafica de la entidad

  Una vez seleccionada una entidad de la lista, se procederá a generar una gráfica que mostrara la evolución de la entidad a través de las fechas en las que fue mencionada, es decir se mostrara la fecha en la que fue mencionada y las veces que lo hicieron, de modo que se podrá conocer entre que fechas fue creciendo el número de ocurrencias de esta. La grafica es generada mediante la implementación de la librería PHPlot y es generada al momento con los datos consultados de la base de datos.

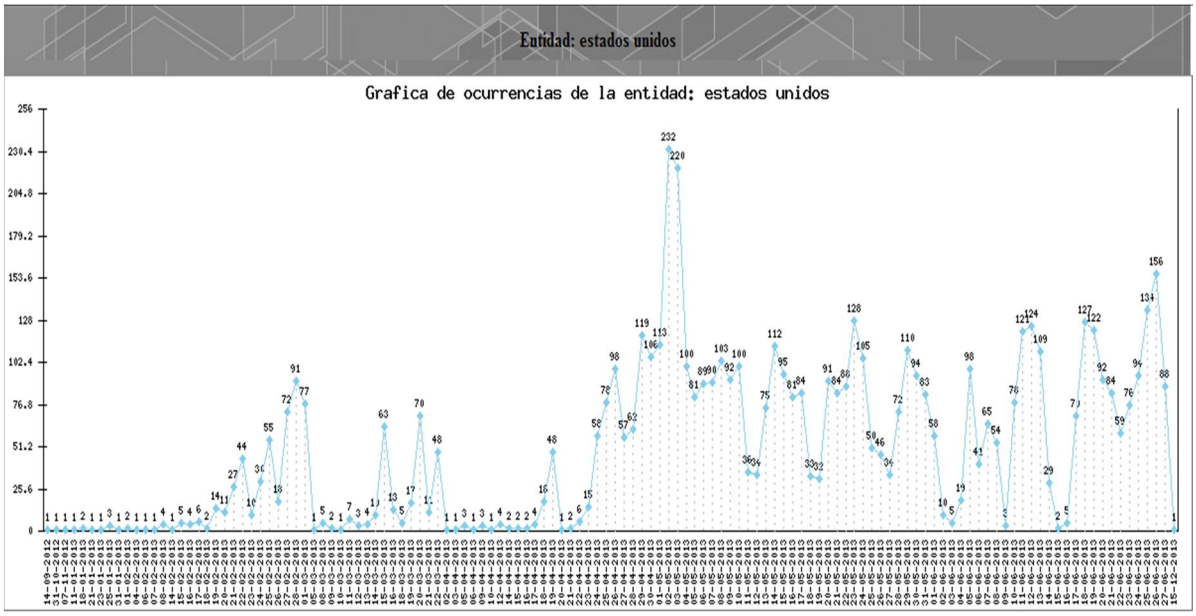

Figura 21: Grafica de la entidad seleccionada

- Selección de fecha

  En esta parte de la página se puede recorrer las fechas que se muestran en la gráfica, se incluyen las fechas ordenadas cronológicamente.

Figura 22: Lista de fechas donde aparece la entidad seleccionada

- Visualización de noticias junto a su reputación calculada

  En esta sección se muestran los títulos de las noticias de la fecha seleccionada, además junto a ellos se muestra la reputación de la noticia con un led color verde una reputación positiva, con un led color azul una reputación neutral y con un led color rojo una reputación negativa, también se incluyen las palabras clave, con las que se basó el cálculo de la reputación. Al igual que las secciones anteriores el cálculo de la reputación se hace en el momento en que se selecciona la fecha para mostrar.

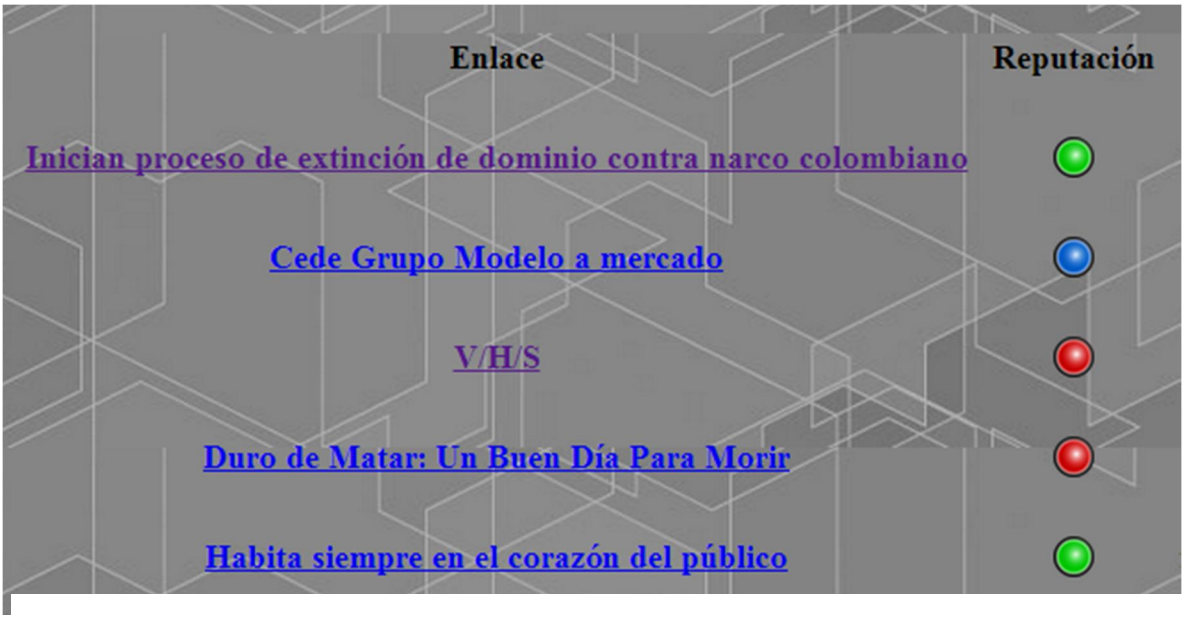

Figura 23: Visualización de noticias junto a su reputación.

- Selección de enlace de noticia

  Al colocar el cursor del ratón sobre el enlace de la noticia se mostrara el texto con el cual fue calculada la reputación, esto es importante ya que algunas noticias no tienen enlace a la noticia original o este ya no es válido por ser eliminado del servidor del periódico.

Figura 24: Visualización del texto procesado.

- Mostrar página original

  Por último, si se desea visitar el sitio original de periódico donde se encuentra alojada la noticia completa se puede hacer haciendo clic en el enlace de la noticia, esto hará que se habrá una nueva pestaña o ventana según su navegador y podrá visualizar la noticia, puede que ocurra un error que no se puede mostrar el sitio, pero eso es por parte del servidor del periódico ya que el enlace ha sido borrado del mismo, aunque la mayoría de los periódicos mantienen por un largo periodo de tiempo las noticias en su servidor.

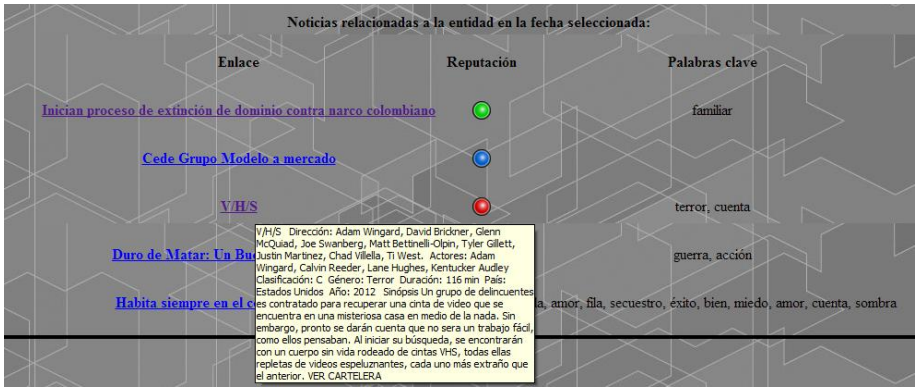

Figura 25: Enlaces a las noticias originales

Al hacer clic en el enlace a la noticia V/H/S por ejemplo, nos llevara a la página donde la noticia fue publicada originalmente por el periódico.

Figura 26: Página original de la noticia seleccionada.

Conclusiones

Con el trabajo realizado se realizó un primer acercamiento al cálculo de reputación de las entidades nombradas con lo cual se abre un gran abanico de posibilidades para mejorar el trabajo realizado ya que existen muchos factores que se deben tomar en cuenta en este análisis para que el sistema tenga una mayor efectividad.

Al momento de recabar la información que será utilizada para crear el corpus, se debe de realizar una limpieza de estos datos, al ser extraídos de páginas de internet se debe de tener cuidado de que el texto no contenga etiquetas html, caracteres especiales que no sean mostrados correctamente como es el caso de algunos tipos de comillas, saltos de línea, retorno de carro, además de que al manejar el idioma español se debe considerar la letra ñ ya que la mayoría de la codificación continua siendo en formato ascii.

Se debe procurar eliminar algunas palabras que los periódicos colocan al comienzo de cada una de sus publicaciones como lo son el lugar por ejemplo la palabra México aparece constantemente al inicio de las noticias lo que puede provocar resultados falsos al momento de hacer un conteo de frecuencia de esa entidad ya que aparecería pero la noticia no tiene relación con esa entidad es decir, no se habla de la entidad en sí. Otras de las palabras que tienen el mismo error son el nombre de los periódicos, el autor de la nota.

También se debe procurar que el tamaño del texto sea lo suficientemente grande para que exista un verdadero análisis, ya que el sistema de RSS es comúnmente utilizado para dar una breve descripción de la noticia y no se muestra completa o como se pudo verificar algunos sistemas no incluyen contenido de la noticia y solo muestran el título y un enlace a la página de la noticia.

El uso de los etiquetadores gramaticales es de gran ayuda para la identificación de entidades ya que al detectar a los sustantivos propios y a los comunes, al ser llamados externamente se incrementa el tiempo de análisis de las entidades por lo que se puede tratar de que estos se ejecuten de forma paralela y así ahorrar tiempo de ejecución, por ejemplo Freeling incluye las clases hechas en java que pueden ser incluidas en el proyecto principal tal vez de esta manera se ejecuten más rápido las consultas a este programa que si se ejecuta de manera externa como se estuvo haciendo.

Al usar tree tagger se tuvo que unir a las entidades que consistían de más de una palabra ya que este programa analiza palabra por palabra y no hace delimitaciones de sustantivos propios aun si estos aparecen de manera continua y menos si están unidos por algún artículo, por otro lado Freeling si delimita los nombres propios, pero pueden ocurrir casos en los que al globalizar la entidad a lo

máximo que se permita, puede ser que se abarque a dos o más entidades en una sola lo que haría que esa entidad solo aparezca una sola vez en todo el corpus y no se contabilicen las entidades que se están englobando en sus respectivas frecuencias. Con esto se pueden generar entidades nombradas que no son relevantes y que sean consideradas como erróneas. También se pude optar por utilizar el identificador de entidades nombradas que contiene la herramienta de Freeling si esta toma en cuenta otros factores para su extracción de entidades y no solo ocupa el etiquetador gramatical se pueden obtener mejores entidades.

Otra mejora que se puede implementar es dividir las noticias en dominios o secciones como en los periódicos para poder interpretar las palabras según sea el dominio ya que el léxico de las secciones de un periódico pueden ser muy diferentes por ejemplo si se encuentra la palabra “pelea” en una noticia si la noticia tiene que ver con una pelea callejera o entre organizaciones se tomaría como algo negativo mientras que si hace alusión a una pelea de box se tome como algo neutral de modo que esto puede afectar la reputación de la entidad, convirtiendo algo positivo o neutral en negativo.

Se pudo notar como es la tendencia de las entidades que van creciendo y disminuyendo con el pasar del tiempo como es el caso de entidades relacionadas a celebraciones como por ejemplo la Ciudad de Puebla incrementa su frecuencia de nombramientos cuando se acerca el mes de mayo y en especial el 5 de mayo es un día en el cual esta entidad es una de las más nombradas. Algunas otras como las entidades relacionadas con la palabra Presidente que es común ser encontrada a lo largo del tiempo, con esto podemos notar que algunas entidades aparecen y desaparecen con el tiempo mientras que otras son constantes.

El incrementar las palabras que aparecen en el diccionario ampliaría el margen de exactitud al analizar las reputaciones y poder clasificar de mejor manera ya que se podrá tener un mayor número de palabras que puedan hacer que las noticias caigan en una de las dos clases que se tienen positiva y negativa, ya que los resultados que se obtuvieron indican que solo el 25% de las noticias fueron resultados positivos, mientras que el 16% fueron resultado negativos y el 58% fueron resultados neutrales lo que indica que muchas palabras que pueden ser positivas o negativas no están incluidas en el diccionario que se utilizó para el análisis ya que la gran mayoría de los resultados neutrales son por este problema y no fueron por que se hayan contrarrestado las palabras positivas con las negativas.

También se puede optar por incluir otro diccionario que contenga más palabras aunque existen algunos que solo contienen las palabras pero carecen de un valor de afinidad, lo que se podría realizar seria dar un valor de 1 a todas las palabras y así llevar un conteo de la reputación.

Con la integración de una base de datos que almacene toda la información recabada se pudo optimizar la página web haciéndola más fluida y los tiempos de procesamiento ya sea de mostrar las entidades, la generación de la gráfica de las ocurrencias de las entidades, el análisis de la reputación y la muestra de los resultados fueron reducidos.

Al mostrar la gráfica de frecuencia de las entidades se pudo notar que si las entidades son muy populares su grafica puede crecer a un tamaño en el cual cada una de las fechas queden muy juntas hasta el punto de ser ilegibles, lo que ocasiona que la gráfica pierda su utilidad, al no poder visualizar correctamente los datos, además esto empeora si el usuario está visualizando la página con un monitor pequeño, lo que se podría realizar seria contar el número de ocurrencias que tiene en total la entidad y dividir la gráfica en dos o más partes y poder visualizar parte por parte la gráfica para mejorar la visión, también se puede asignar el ancho de la imagen en base a el tamaño total de la frecuencia para que de ese modo colocar un enlace para que el usuario pueda visualizar la imagen completa del grafico en una ventana o pestaña de su navegador. Además se puede agregar un verificador de enlaces para que se pueda mostrar un indicador de que el enlace a la noticia aun continua vigente, de ese modo el usuario no tiene que estar probando cada uno de los enlaces y no llevarse una sorpresa al ver que el navegador le muestra que la pagina ya no existe.

Bibliografía

[1] Tengfei Ma, Xiao Jun Wan *Opinion Target Extraction in Chinese News Comments.*

[2] Yuanbin Wu, Qi Zhang, Xuanjing Huang, Lide Wu *Structural Opinion Mining for Graph-based Sentiment Representation.*

[3] Kang Liu, Liheng Xu, Jun Zhao Opinion *Target Extraction Using Word-Based Translation Model.*

[4] Andrea Esuli, Fabrizio Sebastiani *Determining Term Subjectivity and Term Orientation for Opinion Mining.*

[5] Mahesh Joshi, Carolyn Penstein-Rosé Generalizing *Dependency Features for Opinion Mining*

[6] Niklas Jakob, Iryna Gurevych *Using Anaphora Resolution to Improve Opinion Target Identification in Movie Reviews*

[7] Alexandra Balahur, Andrés Montoyo *OpAL: Applying Opinion Mining Techniques for the Disambiguation of Sentiment Ambiguous Adjectives in SemEval-2 Task 18*

[8] Shusen Zhou, Qingcai Chen and Xiaolong Wang *Active Deep Networks for Semi-Supervised Sentiment Classification*

[9] Deepak Agarwal, Bee-Chung, Chen Bo Pang *Personalized Recommendation of User Comments via Factor Models*

[10] Y ejin Choi and Claire Cardie, Ellen Riloff and Siddharth Patwardhan *Identifying Sources of Opinions with Conditional Random Fields andExtraction Patterns*

[11] Stephan Greene, Philip Resnik *More than Words: Syntactic Packaging and Implicit Sentiment*

[12] Razvan Bunescu,Marius Pasca. Using Encyclopedic Knowledge for Named Entity Disambiguation.

[13] Mark Dredze, Paul McNamee, Delip Rao, Adam Gerber,Tim Finin.Entity Disambiguation for Knowledge Base Population.

[14] Bikel, D.M., et al., *Nymble: a High-Performance Learning Name-finder.*

[15] Liu, B.: *Sentiment Analysis and Subjectivity. Handbook of Natural Language Processing.*

[16] Pang, B., Lee, L. and Vaithyanathan, S. *Thumbs up? Sentiment Classification using machine learning techniques.*

[17] Yu, H. and Hatzivassiloglou, V. Towards *answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences*.

[18] Popescu, A.-M. and Etzioni, O. *Extracting product features and opinions from reviews.*

[19] Fellbaum, C.D. *Wordnet: An Electronic Lexical Database*.

[20] Hatzivassiloglou, V. and K. McKeown, *Predicting the semantic orientation of adjectives.*

[21] Helmut Schmid (1995): *Improvements in Part-of-Speech Tagging with an Application to German.*

[22] Lluís Padró. *A Hybrid Environment for Syntax {Semantic Tagging}.*

[23] Grigori Sidorov, Sabino M.J, Francisco V.J, Alexander Gelbukh, Noé Castro S, Francisco Velásquez, Ismael Díaz, Sergio Suárez, Alejandro Treviño, and Juan Gordon, *Empirical Study of Machine Learning Based Approach for Opinion Mining in Tweets.*