PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

FACULTAD DE CIENCIAS E INGENIERÍA

Implementación de una aplicación para el análisis y visualización de eventos en español usando extracción automática de ontologías

TESIS PARA OPTAR POR EL TÍTULO PROFESIONAL EN INGENIERÍA INFORMÁTICA

AUTOR:

Enrique Valeriano Loli

ASESOR:

Mag. Félix Arturo Oncevay Marcos

Lima, Enero del 2019

A mis padres Enrique y Carmen por siempre apoyarme en todas mis actividades.
A mis abuelos Antonio y Teresa por siempre cuidar de mí y educarme.
A toda mi familia por brindarme su cariño incondicional y estar conmigo en todo momento.

AGRADECIMIENTOS

Al Mag. Arturo Oncevay, por su total apoyo durante este proyecto de fin de carrera.

RESUMEN

La globalización y la aparición de diferentes tecnologías como las redes sociales han ocasionado que la información relacionada a noticias y eventos se propague de una manera más rápida. Ahora las empresas deben estar siempre pendientes a los datos generados para así ser más eficaces en su relación con ellos. Sin embargo, esta es una tarea difícil debido a la gran cantidad de datos y a la falta de procesos automáticos para analizar estos, sobre todo en el idioma español.

Como objetivo de este proyecto, se busca brindar una aplicación la cual de manera automática pueda realizar análisis de datos de eventos en español y permitan visualizar los aspectos más importantes relacionados a estos datos. Para esto se implementarán algoritmos de Análisis de Formal de Conceptos y Análisis de Patrones Léxico-Sintácticos. Además, se usarán ontologías para poder estructurar la información recolectada a partir de los algoritmos.

Se concluye que los algoritmos desarrollados permiten obtener las entidades y relaciones más relevantes encontradas en los datos con porcentajes relativamente altos de precisión y exhaustividad sobre todo usando datos limpios. Además, es posible mostrar la información recolectada de manera adecuada debido a la flexibilidad de las ontologías.

Tabla de contenido

Índice de Figuras

Índice de Tablas

CAPÍTULO 1: INTRODUCCIÓN

En el siguiente apartado, se presenta la problemática, los objetivos, los resultados esperados, las herramientas, el alcance, la justificación y el análisis de viabilidad del presente proyecto.

1. Problemática

La globalización y la aparición de diferentes tecnologías como las redes sociales han ocasionado que la información relacionada a noticias y eventos se propague de una manera más rápida [LOBZHANIDZE et al, 2013]. Ahora las empresas deben estar pendientes a los datos generados por sus clientes para así ser más eficaces en su relación con ellos [REDSPIRE 2016]. Es en este contexto en el que el análisis de eventos obtiene mayor relevancia, ya que es posible contar con una mayor cantidad de datos relacionados a distintos eventos respecto a lo que se tenía en años anteriores. [VALKANAS, GUNOPULOS 2013]

Para realizar análisis de eventos, se debe contar con 2 capacidades: Adquisición y procesamiento de datos de eventos [CUELOGIC 2016]. En primer lugar, adquisición se refiere a la capacidad de obtener data en grandes volúmenes para su procesamiento posterior, esto suele ir de la mano con técnicas de transmisión de datos en tiempo real y *Big Data* [CUELOGIC 2016]. Dicho esto, la segunda capacidad mencionada es la más relevante ya que, según OpsClarity [1] [OPSCLARITY 2016] en un estudio del estado de la data veloz y de las aplicaciones de flujos de datos, más del 90% de organizaciones planea incrementar su inversión en procesamiento de datos este año. Esto indica que la tendencia en las empresas es usar procesamiento de datos para obtener información valiosa. En este estudio, se detalla también que alrededor del 68% de entrevistados citan la falta de conocimiento del dominio a analizar y la complejidad de las aplicaciones como principales barreras de entrada para realizar procesamiento de datos [OPSCLARITY 2016]. Este problema detallado en el estudio de OpsClarity junto con la falta de herramientas para análisis de eventos en español observada en la revisión del estado del arte es el que se buscará resolver en el presente trabajo de fin de carrera.

Los eventos que se van a analizar se definen como una ocurrencia del mundo real que se lleva a cabo en un espacio y en un tiempo específico [ATEFEH, KHREICH 2013]. Es por ello que en esta propuesta se buscará dar un enfoque mayor a la característica temporal de los eventos, ya que es una propiedad que no se explota en gran parte de las investigaciones y es una forma de comparar momentos de tiempo y apreciar las distintas actividades desarrolladas en cada día de los eventos.

Los datos propios de eventos de cualquier tipo presentan un reto, debido a que los eventos pueden estar relacionados a distintos dominios tales como: Deporte, Artístico, Político, entre

[1] Empresa enfocada a brindar soluciones de monitoreo de rendimiento para aplicaciones modernas e infraestructuras de datos.

otros; y se buscará estructura estos de una manera homogénea. Además, el conocimiento que se extraerá respecto a los eventos debe ser compatible con los distintos tipos de eventos que existan Es decir, se debe buscar una estructura que sea lo suficientemente flexible y escalable para poder almacenar el conocimiento que se buscará obtener de los eventos.

Es por esto que se plantea usar ontologías como base para estructurar los eventos con los cuales se trabajarán, ya que estas permitirán tener una estructura común para todos los eventos, definir entidades propias de estos relacionadas a un dominio en particular y finalmente relacionar estas entidades entre sí [WU et al, 2003]. Se define una ontología como una definición formal de conceptos y relaciones entre dichos conceptos, usados para describir un área de conocimiento [W3C 2016].

De esta forma, para la construcción de una ontología de eventos, la cual es una ontología en la cual cada concepto representa una entidad la cual puede estar definida por características y cada relación una actividad en la que participan ambas entidades [RAIMOND, ABDALLAH 2007], se definen dos tipos de conceptos principales: instancias de eventos y clases de eventos. Las instancias representan la ocurrencia de un evento mientras que las clases son un conjunto de instancias con una tipología común o un dominio similar. Las principales características que se analizan de las instancias son el actor, objeto o receptor de la acción, tiempo y lugar en el que se realiza. Asimismo, también se analizarán las relaciones entre instancias, las cuales son de tipo causal y de orden, mientras que las relaciones entre clases son de tipo parte-todo. [KANEIWA et al, 2007]

Dentro de la construcción de ontologías se presentan distintas metodologías. Existen metodologías manuales (realizadas íntegramente por expertos), semiautomáticas (realizan parte de la labor de manera automática pero necesitan una elevada interacción con el usuario) y automáticas (necesitan poca o ninguna interacción del usuario) [CIMIANO 2006]. Esta última metodología para la extracción de ontologías es muy beneficiosa debido a que se evita la necesidad de contar con expertos en construcción de ontologías o en el área de conocimiento que se quiera analizar, lo cual implica un menor costo [CIMIANO 2006]. Sin embargo, esta es una tarea muy difícil de conseguir debido a que en una amplia cantidad de casos hay un cuello de botella en la adquisición del dominio, es decir, en el modelado del conocimiento relevante al dominio a analizar [MARYAM et al, 2011]. Es debido a esto que existen diversos estudios e investigaciones que se enfocan en postular procedimientos para el aprendizaje o extracción automática de las ontologías a partir de textos, dependiendo de la estructura y el dominio de la ontología que se desee construir. [CIMIANO 2006]

Resumiendo lo previamente mencionado, se observa que las ontologías son estructuras que permiten representar todo tipo de datos homogeneizando así los datos recolectados, y la construcción automática de estas permite ahorrar costos en recursos valiosos y evitar la necesidad de contar con expertos en el dominio a analizar o en el manejo de la herramienta a usar.

De este modo, este trabajo de fin de carrera plantea el desarrollo de una aplicación para el análisis y visualización de eventos usando ontologías. Esto permitirá a cualquier persona o

grupo de investigación que desee obtener información relevante de los eventos que obtenga de cualquier fuente, generar una ontología para los eventos que se van a analizar y usarla para estructurar la información recopilada, facilitando así el proceso de recolección y análisis de los eventos, lo cual luego se podrá integrar a un sistema de análisis de eventos en tiempo real.

Para cumplir con este objetivo, el proyecto se centrará en las publicaciones en español, adaptando modelos ontológicos para eventos definidos en estudios previos a la nueva información que se pueda recolectar a partir de los eventos y a las características particulares del idioma español. La aplicación que se desarrollará en este trabajo podrá ser usada con cualquier dominio de eventos en español. Sin embargo, para el motivo de pruebas se usarán eventos recopilados de Twitter, ya que estos datos son fácilmente accesibles y existen conjuntos de publicaciones anotadas para facilitar las pruebas.

En conclusión, el presente proyecto de investigación se enfocará en dar solución a la falta de una forma o aplicación para analizar y visualizar eventos en español automáticamente evidenciada en el estudio de OpsClarity mencionado previamente [OPSCLARITY 2016], buscando responder la pregunta: ¿Cómo se podría estructurar y visualizar datos de eventos haciendo uso de ontologías?

2. Objetivos

En esta sección se describen los objetivos que se buscan alcanzar en el presente proyecto de investigación.

2.1. Objetivo General

Implementar una aplicación para el análisis y visualización de eventos en español usando extracción automática de ontologías

2.2. Objetivos Específicos

1) Desarrollar un componente para la adquisición automática y almacenamiento de publicaciones de eventos en Twitter.
2) Desarrollar un componente para la extracción automática de conceptos y relaciones entre estos, usados para representar los eventos.
3) Implementar un módulo de *software* que integre los componentes desarrollados y permita la visualización de los eventos.

3. Resultados esperados

En esta sección se detallan los resultados esperados relacionados a los objetivos específicos previamente planteados.

- Para el objetivo específico 1:

Resultados Esperados	**Indicadores**

1) Componente de *software* para adquisición automática de publicaciones de eventos en Twitter.	- Pseudocódigo de las consultas a Twitter para la obtención de publicaciones. - El componente funcionando dentro de la aplicación integrada.
2) Estructuras de datos para almacenar las publicaciones adquiridas.	- Ejemplo de una publicación almacenada usando las estructuras desarrolladas.
3) Conjuntos de publicaciones de un evento en particular a escogerse, almacenadas las estructuras de datos previamente desarrolladas.	- Repositorio con las publicaciones de eventos extraídas.
4) Estructuras de datos base que serán usadas almacenar las entidades y relaciones a usarse para representar los eventos.	- Ejemplo de entidades y relaciones usando las estructuras desarrolladas

Tabla 1.1 Tabla de relación entre Resultados esperados del objetivo 1 y sus indicadores

- Para el objetivo específico 2:

Resultados Esperados	**Indicadores**
5) Componente de *software* para la extracción automática de conceptos de las ontologías de eventos usado un algoritmo basado en métodos de agrupamiento y Análisis Formal de Conceptos (FCA).	- Pseudocódigo o diagrama de flujo del algoritmo - El componente funcionando dentro de la aplicación integrada
6) Componente de *software* para la extracción automática de relaciones entre entidades usando un algoritmo basado en patrones léxico-sintácticos.	- Pseudocódigo o diagrama de flujo del algoritmo - El componente funcionando dentro de la aplicación integrada
7) Ontologías relacionadas al evento a analizarse obtenido en el resultado 3).	- Repositorio con las ontologías extraídas almacenadas en formato RDF

Tabla 1.2 Tabla de Relación entre Resultados esperados del objetivo 2 y sus indicadores

- Para el objetivo específico 3:

Resultados Esperados	**Indicadores**
8) Componente de *software* para visualizar eventos usando las ontologías extraídas en el resultado 7).	- Mockups de visualización de las ontologías. - El componente funcionando dentro de la aplicación integrada
9) Módulo de *software* que integra todos los componentes previamente obtenidos.	- Aplicación de *software* la cual permita el uso de todos los componentes previamente desarrollados

10) Análisis cuantitativo de métricas relacionadas a la precisión y exhaustividad de las ontologías extraídas para representar el evento a analizarse, comparadas con una Línea Base de la ontología del evento.	- Reporte con las métricas de precisión y exhaustividad para cada evento analizado

Tabla 1.3 Tabla de Relación entre Resultados esperados del objetivo 3 y sus indicadores

4. Herramientas, metodologías y procedimientos

A continuación se introducirán las herramientas, métodos y procedimientos que se utilizarán para el desarrollo del trabajo de fin de carrera.

4.1. Relación de resultados esperados

Se presenta un listado de los resultados esperados junto con las herramientas que se utilizarán para la realización de cada uno.

Resultados Esperados	**Herramientas, métodos y procedimientos**
Resultado Esperado 1: Componente de *software* para adquisición automática de publicaciones de eventos en Twitter	- Python - API de Twitter
Resultado Esperado 2: Estructuras de datos para almacenar las publicaciones adquiridas	- Python
Resultado Esperado 3: Un conjunto de publicaciones de un evento en particular a escogerse	- Python - API de Twitter
Resultado Esperado 4: Estructuras de datos base que serán usadas almacenar las entidades y relaciones a usarse para representar los eventos.	- Python
Resultado Esperado 5: Componente de *software* para la extracción automática de conceptos de las ontologías de eventos usado un algoritmo basado en métodos de agrupamiento y Análisis Formal de Conceptos (FCA)	- Python - NLTK - KDD - FCA - UDPipe - NER de Stanford
Resultado Esperado 6: Componente de *software* para la extracción automática de relaciones entre eventos usando un algoritmo basado en patrones léxico-sintácticos	- Python - NLTK - KDD - Análisis de patrones léxico-sintácticos - UDPipe
Resultado Esperado 7: Ontologías relacionadas al evento a analizarse obtenido en el resultado 3)	- Python - NLTK - KDD

Resultado Esperado 8: Componente para visualizar eventos usando las ontologías extraídas en el resultado 7)	- Python
Resultado Esperado 9: Módulo de *software* que integra todos los componentes previamente obtenidos	- Python
Resultado Esperado 10: Reporte de métricas relacionadas a la precisión y exhaustividad de las ontologías extraídas para representar el evento a analizarse	- Python - KDD - Métricas de evaluación

Tabla 1.4 Tabla de Relación entre Herramientas y Resultados Esperados

4.2. Python

Python es un lenguaje de programación simple pero potente, el cual cuenta con excelente funcionalidad para el procesamiento de datos lingüísticos. Además, Python cuenta con cualidades como: [BIRD et al, 2009]

- Baja curva de aprendizaje.
- Permite la encapsulación de datos y métodos para su reutilización.
- Librería estándar extensa, la cual incluye componentes para programación gráfica, procesamiento numérico, conectividad web, entre otras cosas.
- Amplia cantidad de librerías externas sobre todo enfocadas al procesamiento lingüístico.

Se usará Python para todas las actividades de programación, es decir, para desarrollar todos los componentes de *software* que se usarán en la aplicación. Esto facilitará las labores de procesamiento de texto y permitirá enfocarse en los algoritmos a desarrollar más que en el pre-procesamiento de textos.

4.3. NLTK

NLTK es una librera externa de Python la cual define una infraestructura para ser usada en la construcción de programas de procesamiento de lenguaje natural en Python. En esta infraestructura se cuenta con clases básicas para representar datos relevantes al procesamiento de lenguaje natural, así como interfaces estandarizadas para realizar tareas de procesamiento. [BIRD et al, 2009]

Se hará uso de esta librería para aprovechar la funcionalidad que provee en procesamiento de texto y facilitar esta labor.

4.4. Interfaz de Programación de Aplicaciones (API) de Twitter

La API de Twitter cuenta con una colección de métodos para la lectura y escritura de publicaciones y, en general, de los metadatos con los que cuenta Twitter. Estos métodos se dividen en 2 colecciones: [TWITTER 2016]

- REST: Provee métodos de acceso a datos de Twitter pasados, es decir, datos ya publicados.
- API de flujos (*Streaming*): Provee métodos de acceso a datos de Twitter en línea, es decir, datos en tiempo real.

En este trabajo usaremos el API REST para brindar a los usuarios la capacidad de obtener datos de eventos transcurridos en los últimos días.

4.5. UDPipe

UDPipe es una librería la cual permite entrenar un modelo para tokenizar, etiquetar, lematizar y obtener las dependencias sintácticas de textos en distintos idiomas. [UDPIPE 2017]

En este trabajo se usará UDPipe para realizar el etiquetado gramatical en los datos a usarse, así como obtener las relaciones de dependencia en estos. Esto ayudará a obtener las entidades y las relaciones entre estas para los eventos.

4.6. NER (Name Entity Recognition) de Stanford

El NER de Stanford permite etiquetar las palabras en textos en base a 3 categorías principales: [STANFORD 2017]

- Persona
- Organización
- Lugar

En este trabajo se usará el NER de Stanford para tratar de obtener las entidades de los eventos-

4.7. Métodos y procedimientos

4.5.1. Descubrimiento de Conocimiento en Bases de Datos (KDD)

La metodología KDD se refiere al proceso de descubrir información relevante a partir de datos de entrada que provienen de una base dados, siguiendo un conjunto de pasos claramente definidos. [FAYYAD et al, 1996]

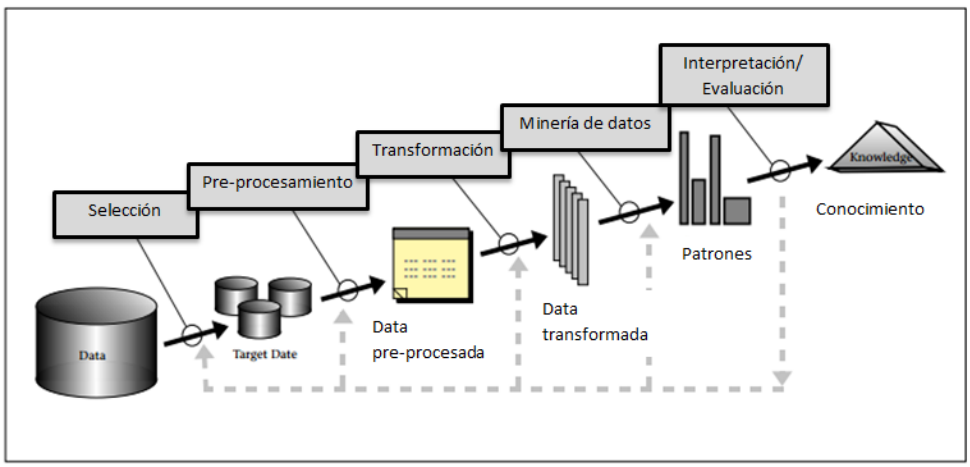

Figura 1.1 Proceso KDD[FAYYAD et al, 1996]

En este caso, se seguirá el proceso definido en la Figura 1.1 para realizar el análisis de los eventos y la adquisición de ontologías relacionadas a estos.

4.5.2. Métricas de evaluación

Para verificar la eficacia del algoritmo que se desarrollará, se definen las siguientes métricas a tomar en consideración [CIMIANO 2006]:

- Precisión: Indica la cantidad de ítems encontrados que fueron relevantes, decir, se define por la fórmula:

$$\frac{\text{Ítems relevantes encontrados}}{\text{Total ítems encontrados}}$$

- Exhaustividad (*recall*): Indica la cantidad de ítems relevantes que fueron encontrados. Se define por la fórmula:

$$\frac{\text{Ítems relevantes encontrados}}{\text{Total ítems relevantes}}$$

Estas métricas se usarán para evaluar las siguientes características:

- Entidades: Se analizarán las entidades extraídas automáticamente.
- Relaciones: Se analizarán las relaciones extraídas automáticamente.

Se hallará el valor de estas métricas usando el algoritmo desarrollado usando los conjuntos de datos extraídos y luego se comparará con una línea base la cual se definirá manualmente a través de la búsqueda de información de los eventos.

4.5.3. Análisis Formal de Conceptos (FCA)

El Análisis Formal de Conceptos se define por el proceso mostrado en la Figura 1.2. A través de este proceso se busca obtener los conceptos más relevantes para la ontología a partir del conjunto de datos que se está analizando. Estos conceptos serán la estructura base de la ontología que será después usada junto a otros algoritmos para obtener características y/o relaciones. [CIMIANO 2006]

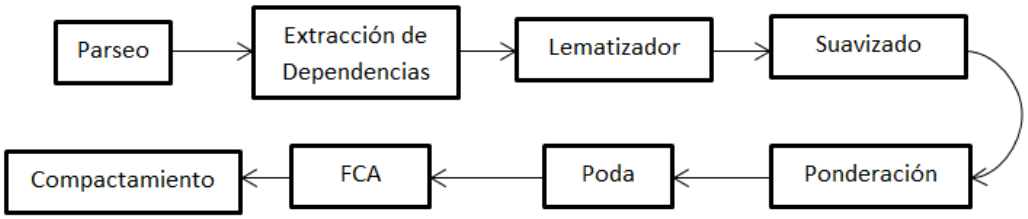

Figura 1.2 Proceso general de inducción de conceptos usando FCA [CIMIANO 2006]

4.5.4. Análisis de patrones léxico-sintácticos

Los patrones léxico-sintácticos pueden ser aplicados en el contexto de descubrimiento de relaciones, específicamente, en la tarea de aprender relaciones de tipos tales como: [CIMIANO 2006]

- Parte-todo
- Causa
- Motivo
- Sujeto-Acción
- Etc.

Se usará este análisis para obtener relaciones entre conceptos de la ontología, independientemente del tipo de evento que se esté analizando enfocándose sobre todo a las relaciones de tipo Sujeto-Acción.

5. Alcance

La aplicación que se desarrollará en el presente trabajo se basará en 2 algoritmos relacionados a extracción automática de ontologías: Análisis Formal de Conceptos para extracción de conceptos y Análisis de Patrones Léxico-sintácticos para la extracción de relaciones.

Esta aplicación podrá trabajar con datos de entrada de cualquier fuente mientras sean datos textuales de eventos en español y tengan información acerca de la fecha en la que se generaron o publicaron los datos.

El componente de visualización que se desarrollará permitirá observar descripciones del evento que está siendo analizado así como una línea de tiempo en la que se verá reflejada la evolución del evento en el tiempo, es decir, la nueva información del evento que surge en el tiempo, según los datos de entrada brindados.

Además, también se brindará un módulo de recopilación de datos de eventos a partir de Twitter. Este modelo podrá ser usado para recopilar datos en caso el usuario final desee utilizar la aplicación basándose en Twitter. Sin embargo, como se mencionó previamente, también se pueden usar datos de otras fuentes recopilados de manera propia pero no está contemplado en el alcance proporcionar un módulo para su recolección. Se eligió realizar un módulo para recolección a partir de Twitter ya que las publicaciones son de acceso público, hay una amplia cantidad de estas y un alto porcentaje tienen carácter social que es lo que se busca para los eventos. Además, la existencia de conjuntos de datos anotados de publicaciones en Twitter facilita anotar y validar los datos que se usen en el proyecto.

5.1. Limitaciones

Entre las limitaciones del proyecto se tiene:

- El proyecto desarrollado sólo permitirá el uso de datos de entrada en español.

- No existen trabajos similares con datos públicos para realizar una comparación directa o experimentación numérica.

5.2. Riesgos

A continuación se presentan los riesgos identificados para el proyecto:

Riesgo	Impacto en el proyecto	Resultados esperados involucrados	Tipo de tratamiento de riesgo	Medidas correctivas
No se presenta un evento con suficiente cantidad de datos para realizar las pruebas en el periodo de realización de la aplicación.	No se tendría datos actuales de eventos para realizar pruebas.	Resultado Esperado 3	Mitigar	Usar datos de eventos pasados brindados por el grupo de investigación.
Los datos obtenidos presentan ruido o mucho contenido que no aporta valor.	El uso de datos con ruido puede sesgar los resultados finales obtenidos.	Resultado esperado 7	Mitigar	Descartar publicaciones con ruido mayor a cierto límite.

Tabla 1.5 Tabla de Riesgos del Proyecto

6. Justificación

Como se mencionó en la problemática, el principal problema planteado es la necesidad de una aplicación para el análisis y visualización de eventos en español automáticamente.

Para la solución de este problema se optó por utilizar algoritmos de Análisis Formal de Conceptos para extracción de conceptos y Análisis de Patrones Léxico-sintácticos ya que, entre los distintos algoritmos de extracción de conceptos y relaciones que se encontraron en el estado del arte, se pudo observar que estos algoritmos son los que mejor se acomodan a la extracción automática de cualquier tipo de eventos basándose en patrones propios del idioma español.

La aplicación que será desarrollada podrá ser usada por distintos grupos beneficiarios, entre los cuales se identifican como principales los siguientes:

- Organizaciones no gubernamentales u organizaciones del estado: Estas organizaciones tienen un contacto más directo con actividades de carácter social y es por esto que el análisis de eventos sociales les brindaría información valiosa en sus actividades.
  Por ejemplo, cuando una de estas organizaciones este a cargo de un evento social o desee saber la respuesta pública frente a una nueva iniciativa o servicio, pueden usar la aplicación para conocer información de esto a detalle.

- Grupos de investigación: Algunos grupos de investigación requieren hacer estudios sobre eventos sociales y podrían usar esta aplicación para automatizar esta etapa y facilitar su trabajo.
  Por ejemplo, cuando se realiza un trabajo investigativo para analizar el comportamiento en eventos como desastres, se puede usar esta aplicación para consolidar la información y conocer zonas o situaciones importantes en el desarrollo del evento.

- Organizaciones que trabajen con noticias (noticieros, canales de televisión, etc.): Estas organizaciones deben estar al tanto de distintas noticias y podrían usar la aplicación para organizar la información relacionada a algún evento en particular.
  Por ejemplo, si además de los medios de información tradicionales se quiere tener otro medio de detección de posibles eventos interesantes para reportar, se puede usar la aplicación para rápidamente observar la información más relevante de distintos eventos y estar pendiente de nuevas posibles noticias.

CAPÍTULO 2: ESTADO DEL ARTE

En este capítulo se describe el estado del arte de las herramientas e investigaciones relacionadas con el problema de estructuración de eventos y generación automática de ontologías a partir de datos textuales.

1. Objetivo

El objetivo del estado del arte es conocer el panorama de las distintas soluciones a problemas similares que existan en la actualidad. A través de esto se puede conocer el estado de los trabajos y herramientas, los problemas principales que afrontaron y las herramientas que pueden ser de ayuda para desarrollar el presente trabajo de fin de carrera.

A partir de la problemática planteada, se plantearon las siguientes preguntas de investigación, que han buscado ser respondidas en la revisión del estado del arte:

- ¿Cómo se estructura un evento en una ontología?
- ¿Qué herramientas o trabajos existen para detección y análisis de eventos usando ontologías?
- ¿Qué herramientas o trabajos existen para la generación automática de ontologías a partir de publicaciones?

Estas 3 preguntas permitirán conocer el panorama relacionado a la problemática previamente planteada tanto en investigaciones de ontologías como de análisis de eventos.

Para resolver estas preguntas se hizo una revisión de artículos y herramientas en librerías digitales como Scopus, IEEE, ACM y Springer. También se hizo uso de una librería digital elaborada por Open Semantic Framework, en la cual se listan distintas herramientas relacionadas a ontologías desarrolladas antes del 2010

Para realizar la búsqueda, se usaron cadenas cómo:

- *event AND (“ontology learning” or “ontology extraction”)*
- *event structure AND ontology*
- *ontology AND (“event detection” or “topic detection”)*

En las siguientes secciones se detallan los trabajos relacionados y el software existente encontrado que ayudaron a responder las preguntas previamente planteadas.

2. Trabajos Relacionados

a) Ontologías propuestas para estructurar eventos:

Respondiendo a la primera pregunta planteada, en este apartado se detallan los trabajos en los que se proponen modelos para definir una ontología orientada a eventos, donde se proponen tanto las entidades como las relaciones que se usan en una ontología de este tipo.

- **Extracción de Eventos de Dominio y Representación como una Ontología de Dominio [WU et al, 2003]**

  En este trabajo se introduce un método para identificar estructuras de eventos usando una ontología de dominio como recurso. Se utiliza el formato InfoMap que consiste de conceptos de dominio y subconceptos relacionados como categorías, atributos y acciones e indexación de documentos basándose en la estructura de un evento.

  Este formato se puede combinar con bases de datos contemporáneas para lograr extraer estructuras del tipo sujeto-verbo-objeto o sujeto-verbo-objeto-modificador. Con estas estructuras se pueden tener oraciones que definan información importante además de las entidades participantes en un evento.

  Se utilizó este formato para categorizar noticias que se recopilaron de la Agencia de Noticias China (CNA) en doce categorías previamente definidas, eligiendo 200 noticias para entrenar la aplicación y otras 200 noticias para realizar las pruebas.

- **Una Ontología Superior para Clasificación de Eventos y Relaciones [KANEIWA et al, 2007]**

  En este estudio se introducen posibles formas de clasificar los eventos: Acciones de agentes a otros objetos o agentes, acciones de un único agente, acciones con varios agentes. Se detalla el estado y cambio de estado por el que pueden pasar los eventos y los cambios que pueden sufrir tanto en espacio como en tiempo.

  Además, se discuten las relaciones ontológicas que se pueden definir, tales como relaciones entre instancias de eventos y relaciones entre clases. En las relaciones entre instancias se definen relaciones causa-efecto, de orden y parte-todo. Mientras que en las relaciones entre clases se definen relaciones disjuntas, de subclase y causa-efecto. En el tiempo estas relaciones pueden ser disjuntas, superpuestas, continuas o parte-todo.

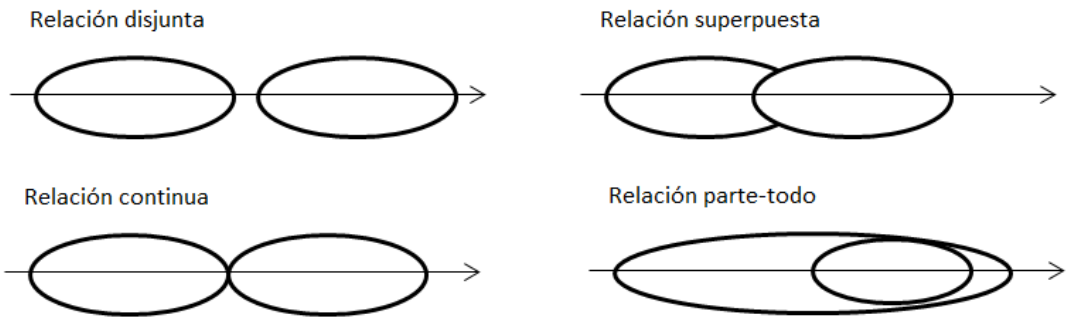

Figura 2.1 Relaciones en el tiempo [KANEIWA et al, 2007]

- **La Ontología de Eventos [RAIMOND, ABDALLAH 2007]**

  En esta investigación se describe una ontología para eventos en donde se define como principales características: lugar, momento, agentes activos, factores y productos.

  En este modelo ontológico cada evento se relaciona a distintos atributos que pueden ser agentes, factores, lugares, productos y tiempo. Además, se usan relaciones de tipo evento-subevento en la construcción de la ontología.

  Esta ontología fue desarrollada por el Centro para Música Digital y fue usado para estructurar charlas en conferencias, descripciones de conciertos, etc.

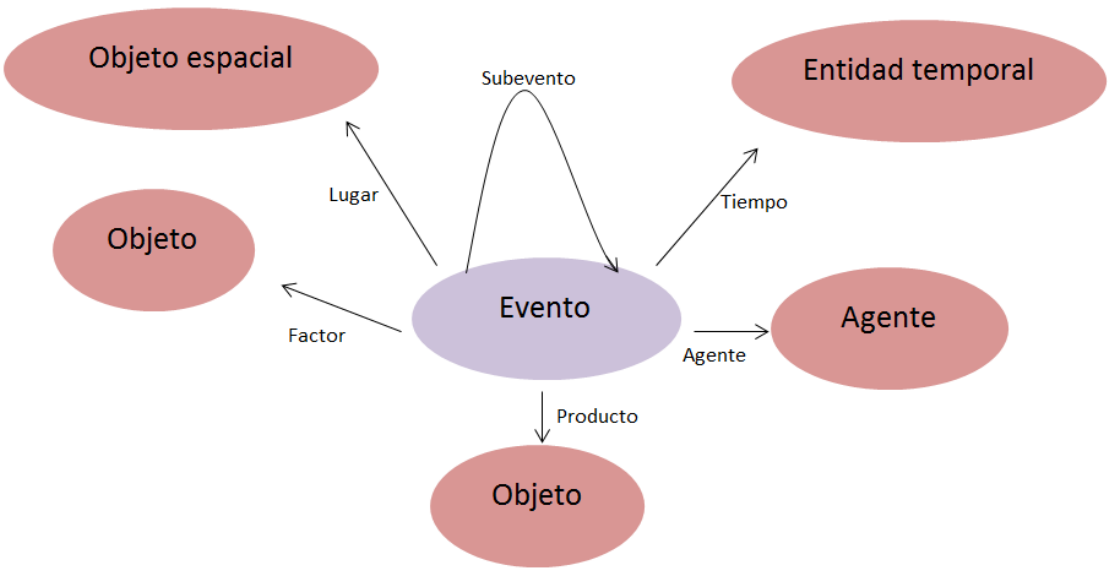

Figura 2.2 Modelo de ontología de eventos [RAIMOND, ABDALLAH 2007]

- **Modelo Entidad-Evento de Ontologías de Registros de Vida (EELOM) para la Definición del Esquema Ontológico de Registros de Vida [LEE et al, 2010]**

  En este artículo se propone un modelo de data flexible para integrar distintos tipos de registros de eventos que representan la actividad de un individuo o su interacción con

el mundo. El modelo es capaz de representar relaciones semánticas entre registros y tomar ventaja de ellas. Además, es lo suficientemente flexible para cubrir una amplia variedad de fuentes de información.

Este modelo está compuesto por un dominio, entidades y eventos. Un dominio es un grupo de entidades y eventos, el cual tiene un nombre representativo. Una entidad es representada como un conjunto de atributos, estas pueden ser compartidas a través de dominios. Finalmente, un evento es una actividad o interacción que puede ser únicamente identificada, el cual es representado como un conjunto de entidades en las que cada entidad tiene un rol.

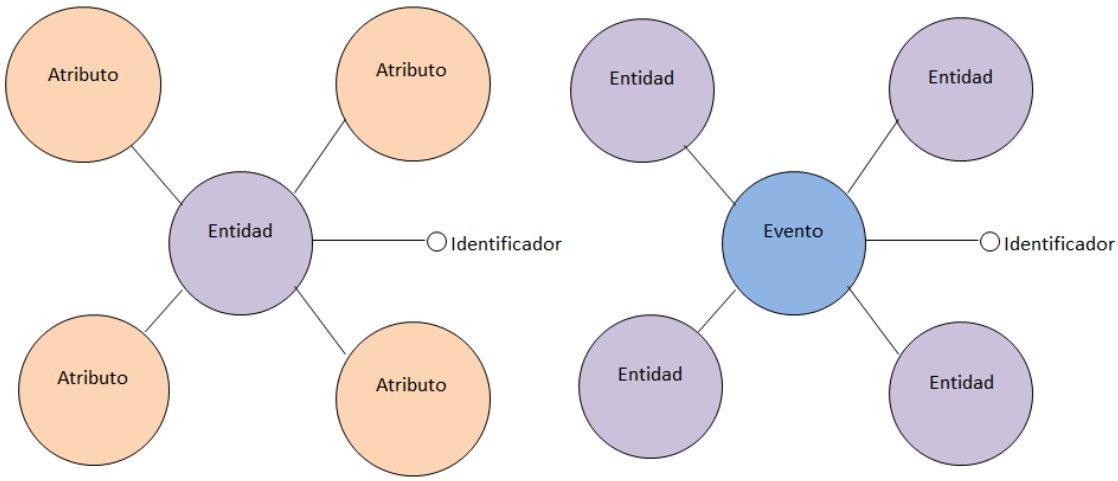

Figura 2.3 Modelo Entidad-Evento de Ontologías de Registros de Vida [LEE et al, 2010]

b) Uso de ontologías para el análisis de eventos:

Respondiendo a la segunda pregunta planteada, en este apartado se detallan los trabajos que usan ontologías para analizar publicaciones de eventos, con el propósito de extraer tópicos, temas de eventos, características de eventos, etc.

- **Aprender Relaciones Semánticas entre Entidades en Twitter [CELIK et al, 2011]**

  En este trabajo se investigan las relaciones semánticas entre entidades y su aprendizaje a través de publicaciones de *microblog*. Se desarrolla un marco de trabajo para la identificación de relaciones, este marco puede ser aplicado en la construcción de ontologías y complementar bases de conocimiento ya existentes como DBpedia.

  En el proceso de detección de entidades se obtiene como resultado un grafo que relaciona los recursos de data con entidades. Este grafo luego es usado para detectar pares de entidades que tienen cierta relación en un periodo de tiempo dado.

- **Aprender a Crear un Modelo de Ontologías de Eventos Extensible para Flujos de Medios de Comunicación Social [LEE et al, 2013]**

  En este estudio se propone un diseño el cual pueda soportar la construcción de entidades para ontologías de eventos de desastres de alto impacto. Se desarrolla un sistema de detección de eventos el cual es usado en conjunto con un módulo de construcción de ontologías de eventos para predecir la posible evolución e impacto de eventos futuros.

  Además, se extraen características temporales a partir de agrupaciones de eventos, se define el momento inicial como la primera fuente de información recibida y el momento final como la última fuente de información recibida. Luego de esto, se extraen características espaciales eligiendo los términos con mayor valor según una función de relevancia de términos.

- **Extrayendo Entidades de Eventos Emergentes a partir de Flujos Sociales Basadas en un Enfoque de División de Clúster de Datos para Ingeniería de Ontologías [LEE, WU 2015]**

  En este investigación se describe un modelo y método de extracción de entidades clave a partir de mensajes de redes sociales, el cual está enfocado a eventos emergentes para mejorar la ingeniería de ontologías, permitiendo obtener una solución para la prevención de desastres similares.

  Para esto, se definen como características a analizar de eventos las siguientes: línea de tiempo, centro geográfico y agrupación de energía. Además, se denota que es posible detectar eventos tempranamente gracias al monitoreo de la evolución de la energía de un evento.

- **Enfoque Basado en Conocimiento para Extracción de Eventos a partir de *Tweets* Árabes [AL-SMADI, QAWASMEH 2016]**

  En este artículo se discute un enfoque no supervisado para la extracción de eventos a partir de *tweets* arábicos. Se unen menciones de entidades en eventos a entidades correspondientes en Wikipedia[2] y DBpedia[3] a través de una base de conocimiento basada en ontologías diseñada para representar entidades de eventos. También, se establecen reglas para hacer uso de la composición natural del idioma árabe en la extracción de entidades y finalmente introducirlos en la ontología.

  Además, en este trabajo se definen las expresiones de eventos con los siguientes argumentos: agente del evento, lugar del evento, foco del evento, disparador del

[2] https://www.wikipedia.org/
[3] http://wiki.dbpedia.org/

evento, producto del evento y momento del evento. Estos argumentos son los que se proceden a extraer de los *tweets*.

c) Generación automática de ontologías desde publicaciones de texto:

Respondiendo a la última pregunta planteada, en este apartado se detallan las técnicas y modelos propuestos para la generación automática de ontologías de cualquier tipo.

- **Extracción de Ontologías usando Redes Sociales [HAMASAKI et al, 2007]**

  En este trabajo se propone la integración de una red social con un modelo tripartito de ontologías. Este modelo se basa en tres dimensiones: actores, conceptos e instancias y usa relaciones de tipo actor-concepto, concepto-instancia y actor-actor.

  En primer lugar se construyen grafos que representen los tipos de relaciones previamente mencionados. Luego, se agrupan 2 conceptos si comparten más de una cierta cantidad de actores e instancias para finalmente reducir la cantidad de conceptos repetidos.

- **Enfoque basado en Ontologías para Detección de Eventos en Flujos de Data de Twitter [RAMACHANDRAN et al, 2015]**

  En este investigación se propone un modelo ontológico para eventos en el que las entidades se extraen haciendo uso del analizador de *tweets* de CMU[4] (Carnegie Mellon University) [GIMPEL et al, 2011] en el que las relaciones son inferidas de Wikipedia, DBpedia y documentos en la red de las entidades extraídas.

  Para esto, se usa una técnica conocida como etiquetado gramatical, la cual identifica automáticamente la categoría de las palabras en una oración, para la extracción de un conjunto de entidades requeridas. Una vez terminado este proceso, se infieren las relaciones entre las entidades y se pueden hacer consultas para recuperar documentos relacionados al dominio de la ontología.

3. Software existente

- **Protége + OntoLT**

  Protégé[5][MUSEN 2015] es un editor de ontologías gratuito y de fuente abierta. Además, sirve también como marco de referencia para construir sistemas inteligentes.

[4] http://www.cs.cmu.edu/~ark/TweetNLP/
[5] http://protege.stanford.edu/

Protégé cuenta con una arquitectura basada en complementos (funcionalidades añadidas) las cual puede ser adaptada para construir aplicaciones basadas en ontologías ya sean simples o complejas. Algunas características con las que cuenta son:

- Una comunidad activa que contribuye en documentación y desarrollo de nuevos complementos.
- Soporte para los estándares definidos por W3C, específicamente para el Lenguaje de Ontologías Web y el marco de trabajo de descripción de recursos.

OntoLT [6] [BUITELAAR et al, 2003] es un complemento de Protégé en el que conceptos y relaciones pueden ser extraídos automáticamente. Este complemento apunta a tener una conexión más directa entre ingeniería de ontologías y análisis lingüístico.

Para esto, OntoLT provee reglas de asignación definidas por un lenguaje de precondiciones que permite una asignación entre entidades lingüísticas y candidatas a clases en Protégé.

- **Open Calais**

Open Calais [7] permite etiquetar semánticamente contenido basado en entidades, hechos, eventos o tópicos. Los cuales luego pueden ser utilizados para obtener información más específica acerca del documento que se esté analizando. Se usa sobre todo en el ámbito de información de empresas, para poder detectar áreas de labor de una empresa o entidades relacionadas a esta.

- **Text2Onto [CIMIANO, VÖLKER 2005]**

Text2Onto[8] es un *software* el cual ayuda a construir ontologías. Esta herramienta ha sido desarrollada para soportar la creación de ontologías a partir de documentos textuales. Usa algoritmos como herencia de conceptos, así como similitud vectorial y de conceptos para ayudar a definir las ontologías.

4. Conclusiones

Dentro de las publicaciones relacionadas a ontologías para eventos, se observa que hay distintos modelos para estructurar eventos, los cuales varían ya sea en entidades que consideran, tipos de relaciones entre estas entidades o ambas.

[6] http://olp.dfki.de/OntoLT/OntoLT.htm
[7] http://www.opencalais.com/
[8] http://neon-toolkit.org/wiki/1.x/Text2Onto.html

Luego, en el ámbito de las investigaciones relacionadas al uso de ontologías de eventos, se observa que la mayor parte de estas se enfoca en la detección de entidades o en proponer marcos o diseños de construcción de ontologías. En Al-Smadi y Qawasmeh (2016) es donde se usan técnicas de extracción relacionadas al idioma árabe para extraer tanto entidades como relaciones y formar una ontología con una estructura previamente definida.

Finalmente, en los trabajos relacionados a la generación automática de ontologías, se puede apreciar que la mayor parte de estas se enfoca a extraer entidades para ayudar en la detección en tiempo real de eventos de desastres de gran escala y la descripción de enfoques de extracción automática de ontologías sin tomar en cuenta algún dominio en particular. Además, no se encuentran trabajos relacionados a extracción automática a partir de datos en español.

En conclusión, se determina que la solución planteada, a diferencia de las soluciones anteriores, está enfocada en la descripción de eventos considerando tanto entidades y relaciones, como atributos temporales y espaciales. Además, se desarrollará una ontología la cual nos permita describir eventos de interés social en español y no se enfocará en un tipo específico de evento.

CAPÍTULO 3: MARCO TEÓRICO

1. Marco Conceptual

En esta sección se explicarán conceptos relacionados al problema planteado así como a la solución planteada.

1.1. Evento

Un evento es una ocurrencia del mundo real que se lleva a cabo en un espacio y en un tiempo específico, donde se pueden identificar dos elementos importantes: entidades y actividades [ATEFEH, KHREICH 2013]. Los eventos son las unidades de estudio en este proyecto y se buscará extraer e interpretar información valiosa de estos.

Definimos una entidad como el sujeto que realiza una actividad en un evento. Este sujeto puede ser una persona, una máquina, un fenómeno natural, etc [ATEFEH, KHREICH 2013]

Además, una actividad es una acción que relaciona distintas características de un evento en un dominio dado, entre las cuales tenemos [ATEFEH, KHREICH 2013]:

- Conjunto de entidades participantes
- Tiempo
- Lugar
- Acción

Se buscará usar tanto entidades como actividades y relaciones entre entidades para definir los eventos y poder trabajar fácilmente con estos.

1.2. Procesamiento de Lenguaje Natural

El Procesamiento de Lenguaje Natural es un conjunto de técnicas computacionales para analizar y representar textos que ocurren de manera natural en uno o más niveles de análisis lingüístico, con el propósito de conseguir un procesamiento similar al realizado por una persona [LIDDY 2001]. En este trabajo nos enfocaremos en el análisis morfológico, sintáctico y semántico para poder obtener información relevante de los datos de eventos.

a) Análisis Morfológico

El análisis morfológico es un campo de la lingüística el cual estudia la formación y la estructura interna de las palabras. Las unidades de estudio son los morfemas, los cuales son definidos como la unidad mínima de lenguaje. Se definen tres áreas de estudio dentro del análisis morfológico [BENDER 2013]:

- Morfotáctica: Reglas de combinación de morfemas para formar palabras.

- Morfofonología: Cómo la forma de los morfemas está condicionada por los otros mo/rfemas con los que se combinan.
- Morfosintaxis: Cómo los morfemas en una palabra afectan sus combinaciones posibles.

b) Análisis Sintáctico

La sintaxis es el conjunto de reglas para construir posibles oraciones a partir de palabras, determinando su forma y significado. El análisis sintáctico hace uso de estas reglas para definir categorías sintácticas y relaciones entre dichas categorías. [BENDER 2013]

c) Análisis Semántico

El análisis semántico es el proceso de determinar la similitud semántica de distintas palabras y encontrar el sentido o la correcta definición de una palabra según el contexto en el que se emplea. Esto encuentra aplicaciones en procesos como el de desambiguación de palabras, por lo cual es muy importante cuando se desea analizar datos de lenguaje natural. [CIMIANO 2006]

1.3. Ontologías

Las ontologías son un vocabulario de términos junto a una especificación de su significado. Esto incluye definiciones e una indicación de como los conceptos están interrelacionados.

Se usarán las ontologías para estructurar los eventos obtenidos y poder contar con la información almacenada de esta manera.

a) Concepto

Los conceptos representan un conjunto de propiedades pertenecientes a una clase en particular.

Las ontologías pueden contar con atributos, que son características que definen a un concepto en particular. Los conceptos se dividen en dos tipos [BECHHOFER et al, 2002]:

i. **Primitivos**

Los conceptos primitivos son los que tienen las condiciones necesarias para pertenecer a una clase, es decir, cuentan con las propiedades que involucra la clase. [BECHHOFER et al, 2002]

ii. **Definidos**

Los conceptos definidos son conceptos cuya descripción es suficiente y necesaria para que un objeto pertenezca a una clase, es decir, definen las propiedades de la clase a la que pertenecen. [BECHHOFER et al, 2002]

Por ejemplo, para la clase “Persona” un concepto primitivo podría contar con atributos de “Nombre” y “Género”. Mientras que un concepto definido podría contar con atributos de “Nombre”, “Género”, “Fecha de Nacimiento”, “Procedencia”, “Talla”, “Peso”, etc.

b) Relación

Las relaciones describen interacciones entre entidades y propiedades de las entidades. Las relaciones se dividen en dos grupos [BECHHOFER et al, 2002]:

i. **Taxonómicas**

Relaciones que organizan conceptos en estructuras de tipo sub-super entidad. Los tipos más comunes de relaciones taxonómicas son [BECHHOFER et al, 2002]:

- Parte-todo: Describen pertenencia y subdivisión de conceptos.
- “Es un tipo de”: Describen especializaciones y clases de conceptos.

ii. **Asociativas**

Relaciones que describen interacciones entre conceptos o propiedades de los conceptos. Los tipos más comunes de relaciones asociativas son [BECHHOFER et al, 2002]:

- Nominales: Describen conceptos y sus propiedades.
- De locación: Describen locaciones de conceptos respecto a otras.
- Asociativas: Representan funciones, procesos o actividades en las que un concepto está involucrado.

c) Extracción automática de ontologías

Se define la extracción automática de ontologías como la adquisición de un modelo de dominio a partir de un conjunto de datos [CIMIANO 2006]. En este trabajo nos enfocaremos en la extracción a partir de datos textuales.
Se describen tres pasos para la extracción automática de ontologías [CIMIANO 2006]:

- Inducción de jerarquía de conceptos: Se induce la jerarquía de conceptos, es decir, se estructura la información en categorías tal que facilite la búsqueda, reutilización y entendimiento de esta.

- Aprendizaje de atributos y relaciones: Se deben recuperar los atributos de los conceptos y las relaciones entre estos. Estas relaciones deben ser axiomáticamente definidas y combinadas con otras relaciones o conceptos formando reglas.

- Popular datos: Finalmente, se debe popular la ontología basándose en documentos de texto. Esto involucra encontrar instancias tanto de relaciones cómo de conceptos.

1.4. Enfoques para extraer ontologías automáticamente

A continuación se detallan los enfoques para extraer ontologías automáticamente, debido a que un punto importante para resolver el problema planteado es brindar una aplicación que trabaje de manera automática.

a) Aprendizaje de Máquina

El Aprendizaje de Máquina es el proceso de reconocer y detectar automáticamente ciertos patrones y regularidades en datos. Este proceso es basado en inducción, es decir, realizando inferencias o generalizaciones a partir de los datos de entrenamiento que se tienen. [BISHOP 2006]

Se distinguen dos tipos de aprendizaje inductivo:

- Aprendizaje supervisado: El fin de este tipo de aprendizaje es aprender una función matemática a partir de los datos de entrenamiento para luego realizar labores de clasificación o regresión. [BISHOP 2006]
- Aprendizaje no supervisado: Este tipo de aprendizaje se usa para explorar estructuras frecuentes en los datos que se están procesando y no necesita datos de entrenamiento previos. [BISHOP 2006]

b) Enfoque Estadístico

El enfoque estadístico comienza a construir una ontología usando palabras clave relacionadas a los conceptos del tipo de ontología a extraerse. Luego se realizan búsquedas de las palabras clave en algún motor de búsqueda y se procesan los documentos de texto recopilados. [MARYAM et al, 2011]

Se seleccionan los conceptos representativos de la ontología de acuerdo a las siguientes métricas [MARYAM et al, 2011]:

- Número total de apariciones
- Número de documentos distintos que contienen el concepto
- Número estimado de resultados que retorna el motor de búsqueda buscando sólo el concepto evaluado.
- Número estimado de resultados que retorna el motor de búsqueda buscando el concepto evaluado y la palabra clave inicial.
- Razón entre las 2 últimas métricas mencionadas.

1.5. Explotación de datos

La explotación de datos se refiere a obtener información valiosa para la persona o empresa que lo requiera a partir de un flujo de datos. Se buscará combinar este punto con el de Procesamiento de Lenguaje Natural para lograr obtener la mayor cantidad de información relevante posible de los eventos.

a) Recopilación de datos

La recopilación de datos es el proceso de recopilar y medir la información en variables de interés, de manera sistemática, lo cual permite investigar preguntas, probar hipótesis y evaluar resultados. [RCR 2016]

b) Análisis de datos

El análisis de datos es el proceso de aplicar sistemáticamente técnicas estadísticas o lógicas para describir, ilustrar, condensar, recapitular o evaluar datos. [RCR 2016]

En las investigaciones el análisis de datos suele contener procedimientos estadísticos y muchas veces se vuelve un proceso iterativo continuo en el cual los datos son continuamente recopilados y analizados. [RCR 2016]

c) Visualización de datos

Se define la visualización de datos como la comunicación de datos usando representaciones gráficas. [WARD et al, 2010]

La visualización de datos es importante ya que, por un lado, los humanos somos seres que usamos la visión como uno de nuestros principales sentidos. También se usa la visualización para proveer vistas alternativas de los datos y ayudar a describir estructuras, patrones o anomalías en estos. [WARD et al, 2010]

CAPÍTULO 4: Desarrollo del componente para extracción automática de publicaciones de eventos en Twitter

1. Componente de *software* para adquisición automática de publicaciones de eventos en Twitter

Este componente será usado para poder obtener un conjunto de textos sobre distintos eventos para realizar pruebas. Además, los usuarios finales podrán hacer uso de este para obtener datos de Twitter si desean usar esta plataforma como fuente para el análisis.

Para el desarrollo del componente se hizo uso de la librería Tweepy, la cual permite usar el REST API propio de Twitter para extraer publicaciones recientes. Las publicaciones adquiridas se filtran para obtener solo el idioma español y evitar usar *retweets* en los datos.

1.1. Creación de aplicación en Twitter

El primer paso fue la creación de la aplicación dentro de Twitter para obtener todos los permisos necesarios por el REST API. Los usuarios finales deberán crear su propia aplicación en Twitter para obtener los permisos necesarios.

Se tienen 4 archivos con las llaves y contraseñas provistas por la aplicación, las cuales serán usadas por Tweepy para obtener los accesos necesarios.

Una vez configurada la aplicación se puede comenzar la extracción de *tweets*. Para esto se hacen uso de palabras claves proporcionadas por el usuario las cuales formarán parte de la consulta a Twitter.

```
consulta = '';
para palabra en palabras_clave hacer
  | consulta += 'OR '+palabra
fin
mientras se obtengan nuevos tweets en consulta hacer
  | para tweet en nuevos_tweets hacer
  |   | si tweet.retweet es falso Y 'RT @' no aparece en texto entonces
  |   |   | a = 'text':texto,'timestamp':tiempo;
  |   |   | añadir a al archivo según su tiempo;
  |   | en otro caso
  |   |   | procesar siguiente tweet;
  |   | fin
  | fin
fin
```

Figura 4.1 Pseudocódigo del algoritmo para recopilar *tweets*

2. Estructuras de datos para almacenar las publicaciones adquiridas

En segundo lugar, se definieron las estructuras de datos para almacenar las publicaciones. Se optó por usar una estructura de datos que pueda ser adaptable a cualquier fuente, es decir, no tenga características difíciles de encontrar fuera de Twitter.

Es por esto que se decidió almacenar los datos en archivos en los cuales el nombre del archivo indica la fecha a la cual corresponde los textos dentro de este. Además, dentro del archivo cada línea representa un registro en formato JSON en el cual la llave "*text*" representa un texto, frase u oración a ser considerada por la aplicación.

Figura 4.2 Contenido en archivo "17_01_2017.txt" para uno de los casos de prueba tratados

3. Conjuntos de publicaciones de un evento en particular a escogerse, almacenadas en las estructuras de datos previamente definidas

Finalmente, se usó el componente previamente mencionado para extraer eventos de distintos tipos para realizar las pruebas. Estos son:

- Abierto de Australia [AUS]: En este caso se usaron las frases de búsqueda "Australian Open" y "Abierto de Australia". Se recolectaron 2217 publicaciones desde 21/01/2017 hasta 30/01/2017
- Marcha contra la corrupción [CORR]: Se utilizó la frase de búsqueda "Marcha contra la corrupción". Se recolectaron 1493 publicaciones desde 11/02/2017 hasta 20/02/2017.
- Reclamos por el peaje de Puente Piedra [PIED]: Se usó las frases de búsqueda "PuentePiedra" y "Peaje Puente Piedra". Se recolectaron 3882 publicaciones desde 08/01/2017 hasta 18/01/2017.

4. Estructuras de datos base para almacenar las entidades de los eventos y las relaciones entre estas

El modelo ontológico que se usará se define de la siguiente manera:

Conceptos: Los conceptos que compondrán la ontología serán sujetos como personas, objetos o lugares. Estos tendrán las acciones en las que participa cada sujeto en base a los datos del evento proporcionados y las relaciones que guardan con otros conceptos.

Relaciones: Las relaciones que se buscarán extraer son las siguientes:

- Sujeto-acción: Esta relación indica actividades en las que participó un sujeto por cuenta propia.
- Sujeto-acción-objeto: Esta relación indica actividades en las que participó un sujeto interactuando con otro objeto el cual, por ejemplo, podría tratarse de otro sujeto dentro de la ontología.

A continuación se muestra el esquema para las relaciones sujeto-acción-objeto, el caso de sujeto-acción sería una simplificación de esta ya que el agente contendría tanto el sujeto como la acción.

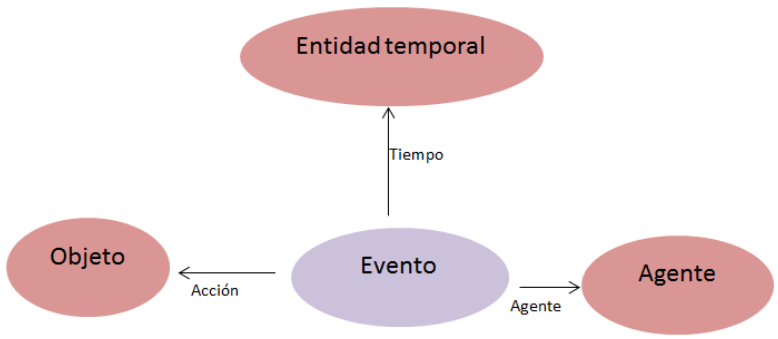

Figura 4.3 Esquema de estructura base para un nodo en la ontología

En relación a esto, se muestra un ejemplo de cómo se verían dos nodos y su relación en la ontología para el caso del Abierto de Australia.

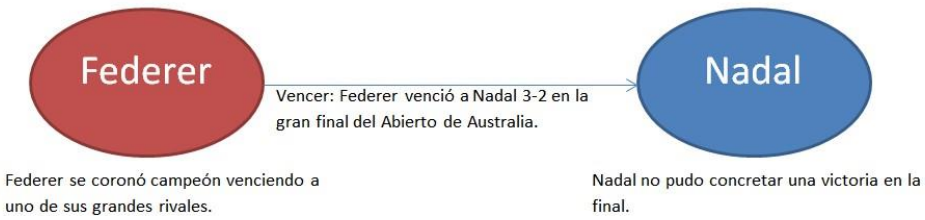

Figura 4.4 Ejemplo de parte de la ontología del Abierto de Australia

CAPÍTULO 5: Desarrollo del componente para la extracción automática de ontologías de eventos

1. Componente de *software* para la extracción automática de conceptos de las ontologías de eventos usado un algoritmo basado en métodos de agrupamiento y Análisis Formal de Conceptos (FCA)

A continuación se explicarán los pasos que sigue el componente para obtener los conceptos relevantes a partir de los datos proporcionados. El flujo que se sigue de manera general se puede observar en la Figura 6.1.

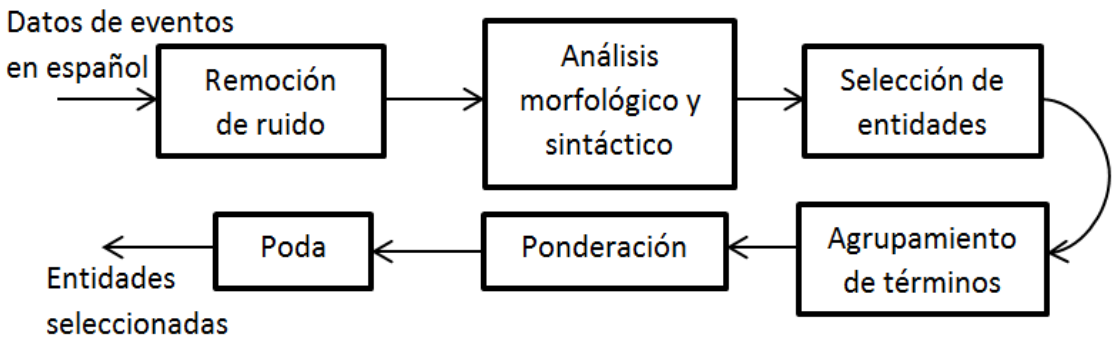

Figura 5.1 Flujo del procesamiento de datos

1.1. Remoción de ruido

En esta etapa se busca reducir el ruido removiendo algunos caracteres o frases las cuales dificultan el procesamiento de los datos.

En particular, se busca eliminar algunos caracteres Unicode que no son propios del idioma español, signos de puntuación o símbolos que dificultan el procesamiento de texto; y links de páginas para evitar identificarlas como conceptos dentro de la ontología.

1.2. Análisis morfológico y sintáctico

Una vez se tienen los datos limpiados, se procede a realizar el análisis morfológico y sintáctico para encontrar las palabras o entidades iniciales que serán procesadas.

Para esto se hace uso de la librería UDPIPE, la cual permite obtener las categorías gramaticales y el árbol de dependencia de las oraciones a ser analizadas. Una vez se tiene esto se procede a extraer pares sujeto-verbo, los cuales definen el contexto formal. La forma de extracción de estos pares será explicad con más detalle en el siguiente punto.

Además, en esta etapa se crean y almacenan algunas listas que serán usadas posteriormente para el agrupamiento de términos. El contenido de estas listas son valores

como la frecuencia de aparición de las entidades o acciones, así como las combinaciones de estas.

1.3. Selección de entidades

Una vez se cuenta con las categorías gramaticales de las palabras dentro de una oración así como el árbol de dependencia de este, se procede a seleccionar las entidades que inicialmente se considerarán para luego ser finalmente agrupadas y filtradas.

Para esto se hacen uso de las siguientes categorías gramaticales:

- 'PROPN': Esta categoría indica las entidades que serán consideradas, pueden ser nombres propios, entidades, lugares, entre otros.
- 'VERB': Las palabras categorizadas con 'PROPN' solo serán consideradas si van acompañadas de la categoría 'VERB', es decir, si están realizando alguna acción dentro de la oración analizada.

A partir de estas 2 categorías se consiguen pares (entidad, acción) las cuales en conjunto forman el contexto formal que será procesado en los puntos posteriores.

Además de esto, se experimentó con el NER de Stanford para reconocimiento de entidades. Sin embargo, debido a que esta herramienta no funcionaba bien con datos poco formales los cuales abundaban en los datos de Twitter, no se consideró en la aplicación final.

1.4. Agrupamiento de términos

Se desarrolló un algoritmo de agrupamiento para encontrar términos similares dentro de los datos y agruparlos como el mismo sujeto. El pseudocódigo se muestra en la Figura 6.1 a continuación:

```
Entrada: frecuencia: lista de frecuencias entre pares de términos
distancia_promedio: lista de distancia promedio entre pares de términos
frecuencia_maxima: valor máximo de frecuencia
valor_corte: valor real entre 0 y 1 que define el punto de corte de
frecuencia
similitud_mínima: valor que define la distancia mínima promedio entre 2
términos para considerarlos similares

para término en términos_iniciales hacer
    para segundo_término en términos_iniciales hacer
        si término != segundo_termino Y
          distancia[(término,segundo_término)]_a < similitud_mínima_b
          < Y frecuencia_máxima[(término,segundo_término)]_a >
          valor_corte*frecuencia_maxima_b  entonces
            agrupar(término,segundo_término)
        fin
    fin
fin
```

Figura 5.2 Pseudocódigo del algoritmo de agrupación

Los parámetros de entrada que se usan para el algoritmo se obtienen en la etapa de análisis morfológico y sintáctico. A continuación se tiene un ejemplo de los grupos encontrados para el caso Abierto de Australia.

```
andy:andy,murray,
mischa:zverev,mischazverev,mischa,
serena:wlliams,serena,williams,serenawilliams,
australian:open,australianopen,australian,ãustralianopen,australia,
federer:federer,rogerfederer,roger,federe,
grand:grand,slam,
gael:monfils,gael,monflis,
c0nvey:cnnee,via,c0nvey,
nadal:nadal,rafaelnadal,rafael,
```

Figura 5.3 Términos agrupados para el caso Abierto de Australia

1.5. Ponderación

Respecto a la ponderación, se consideraron las siguientes métricas [CIMIANO 2006]:

- Condicional:

$$Condicional(n,v) = P(n|v) = \frac{f(n,v)}{f(v)}$$

- PMI:

$$PMI(n,v) = log_2 \frac{P(n|v)}{P(n)}$$

- Resnik:

$$Resnik(n,v) = SR(v) * P(n|v)$$

Dónde:

$$f(n,v) => Frecuencia\ de\ aparición\ de\ la\ entidad\ n\ con\ la\ acción\ v$$

$$f(v) => Frecuencia\ de\ aparicación\ de\ la\ acción\ v\ con\ cualquier\ entidad$$

$$SR(v) = \sum_n P(n|v) * log_2 \frac{P(n|v)}{P(n)}$$

El uso de estas 3 métricas en conjunto se debe a que en la literatura se denota que esto arroja mejores resultados que el uso de cualquier sub-grupo de estos y de otras métricas con las cuales se experimentó.

Una vez obtenidos, se normalizan los valores de las 3 métricas para tenerlos en valores reales entre 0 y 1.

1.6. Poda de conceptos

En esta etapa se remueven los conceptos menos relevantes según las métricas previamente obtenidas. Para esto se define una variable límite a partir de la cual se comparan las métricas.

El cálculo de esta variable se realizó un proceso de experimentación en el que se ajustó su valor observando los resultados obtenidos para los 3 eventos recolectados en el capítulo 4. Finalmente, se escogió el valor de **0.1**, debido a que es el valor sugerido en el estado del arte y además se observó que con este valor se eliminaban las entidades menos relevantes de los eventos.

Los pares (entidad, acción) para los cuales sus 3 métricas por separado superen el valor límite serán considerados como los conceptos finales con los cuales se formará la ontología.

2. Componente de *software* para la extracción automática de relaciones entre eventos usando un algoritmo basado en patrones léxico-sintácticos

Una vez obtenidos los conceptos se procede a extraer las relaciones entre conceptos, así como los textos relevantes para cada concepto, haciendo uso de las relaciones que se definieron en el capítulo 5.

Para esto usamos nuevamente la librería UDPipe con la cual obtenemos las relaciones de dependencia sintáctica entre las categorías, centrándose en las siguientes:

- Objeto directo ‘dobj’: Se buscará obtener el verbo raíz al cual está ligado el objeto directo y luego unirlo a la entidad que se relacione con este verbo.
- Objeto indirecto ‘iobj’: De manera similar, se usará el objeto indirecto para relacionarlo tanto con el verbo raíz así como la entidad que realiza la acción.

Usando estos nuevos términos se podrá establecer las relaciones entre entidades que ejecutan acciones con las entidades que las reciben, es decir, los objetos directos e indirectos.

3. Ontologías relacionadas a los eventos a analizarse

Usando los componentes previamente mencionados, se procede a realizar la extracción de las ontologías para los eventos recopilados en el Capítulo 4.

Para su almacenamiento se usa el formato RDF (*Resource Description Framework*), haciendo uso de la librería rdflib. Además, se trabaja bajo el esquema FOAF (*Friend of a Friend*), la cual es una descripción de una persona a través de una ontología con la intención de crear una red social semántica [CASTRO 2008].

```
<?xml version="1.0" encoding="UTF-8"?>
<rdf:RDF
   xmlns:ns1="http://xmlns.com/foaf/0.1/"
   xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
>
  <rdf:Description rdf:nodeID="Nbbf768d7b87e456784a0f4f401ad7ea9">
    <ns1:name>australianopen</ns1:name>
    <ns1:knows rdf:nodeID="N4fbdc855dd544eb182ea76449fef837b"/>
    <ns1:Type rdf:resource="http://xmlns.com/foaf/0.1/Agent"/>
    <ns1:knows rdf:nodeID="N6a22b9ae0ad146ee99044ed710051bdd"/>
    <ns1:knows rdf:nodeID="Nf04e01ce295f444588ce90032fc38124"/>
    <ns1:name>ãustralianopen</ns1:name>
    <ns1:name>australian</ns1:name>
    <ns1:knows rdf:nodeID="Nccd047c6f63647c2b43251b22731a631"/>
    <ns1:name>australia</ns1:name>
    <ns1:knows rdf:nodeID="N23181af788ad468cb0e364a47dac00de"/>
    <ns1:name>open</ns1:name>
    <ns1:knows rdf:nodeID="N91343cd758c64bbfb813da5d7a7cad8b"/>
    <ns1:knows rdf:nodeID="Nab1aa4cae8b8439497b5f7e21b692412"/>
  </rdf:Description>
  <rdf:Description rdf:nodeID="N4fbdc855dd544eb182ea76449fef837b">
    <ns1:knows rdf:nodeID="N91343cd758c64bbfb813da5d7a7cad8b"/>
    <ns1:knows rdf:nodeID="Nbbf768d7b87e456784a0f4f401ad7ea9"/>
    <ns1:knows rdf:nodeID="N6a22b9ae0ad146ee99044ed710051bdd"/>
    <ns1:name>serenawilliams</ns1:name>
    <ns1:name>williams</ns1:name>
    <ns1:knows rdf:nodeID="Nf04e01ce295f444588ce90032fc38124"/>
    <ns1:Type rdf:resource="http://xmlns.com/foaf/0.1/Agent"/>
    <ns1:name>wlliams</ns1:name>
    <ns1:name>serena</ns1:name>
    <ns1:knows rdf:nodeID="Nccd047c6f63647c2b43251b22731a631"/>
    <ns1:knows rdf:nodeID="N23181af788ad468cb0e364a47dac00de"/>
    <ns1:knows rdf:nodeID="Nab1aa4cae8b8439497b5f7e21b692412"/>
  </rdf:Description>
  <rdf:Description rdf:nodeID="N6a22b9ae0ad146ee99044ed710051bdd">
    <ns1:knows rdf:nodeID="N91343cd758c64bbfb813da5d7a7cad8b"/>
    <ns1:knows rdf:nodeID="Nab1aa4cae8b8439497b5f7e21b692412"/>
    <ns1:knows rdf:nodeID="Nccd047c6f63647c2b43251b22731a631"/>
    <ns1:Type rdf:resource="http://xmlns.com/foaf/0.1/Agent"/>
    <ns1:knows rdf:nodeID="Nbbf768d7b87e456784a0f4f401ad7ea9"/>
    <ns1:knows rdf:nodeID="N4fbdc855dd544eb182ea76449fef837b"/>
    <ns1:knows rdf:nodeID="N23181af788ad468cb0e364a47dac00de"/>
    <ns1:name>ausopen</ns1:name>
    <ns1:knows rdf:nodeID="Nf04e01ce295f444588ce90032fc38124"/>
  </rdf:Description>
  <rdf:Description rdf:nodeID="N91343cd758c64bbfb813da5d7a7cad8b">
    <ns1:Type rdf:resource="http://xmlns.com/foaf/0.1/Agent"/>
    <ns1:knows rdf:nodeID="Nf04e01ce295f444588ce90032fc38124"/>
    <ns1:name>tenis</ns1:name>
    <ns1:knows rdf:nodeID="Nbbf768d7b87e456784a0f4f401ad7ea9"/>
    <ns1:knows rdf:nodeID="Nab1aa4cae8b8439497b5f7e21b692412"/>
    <ns1:knows rdf:nodeID="N4fbdc855dd544eb182ea76449fef837b"/>
    <ns1:knows rdf:nodeID="N6a22b9ae0ad146ee99044ed710051bdd"/>
    <ns1:knows rdf:nodeID="Nccd047c6f63647c2b43251b22731a631"/>
    <ns1:knows rdf:nodeID="N23181af788ad468cb0e364a47dac00de"/>
  </rdf:Description>
```

Figura 5.4 Ejemplo de ontología en formato RDF

Una vez se tienen las ontologías guardadas, se procede a almacenar los datos con los cuales se relacionan tanto las entidades como las relaciones extraídas. Estas se guardan en otros archivos y junto con la ontología, podrán ser cargados luego en el proceso de visualización.

CAPÍTULO 6: Desarrollo del componente de visualización e integración de la aplicación

1. Componente de visualización

A continuación se muestran las pantallas desarrolladas para la interfaz de la aplicación..

En primer lugar se muestra la pantalla de inicio en la Figura 6.1, esta pantalla será usada como bienvenida.

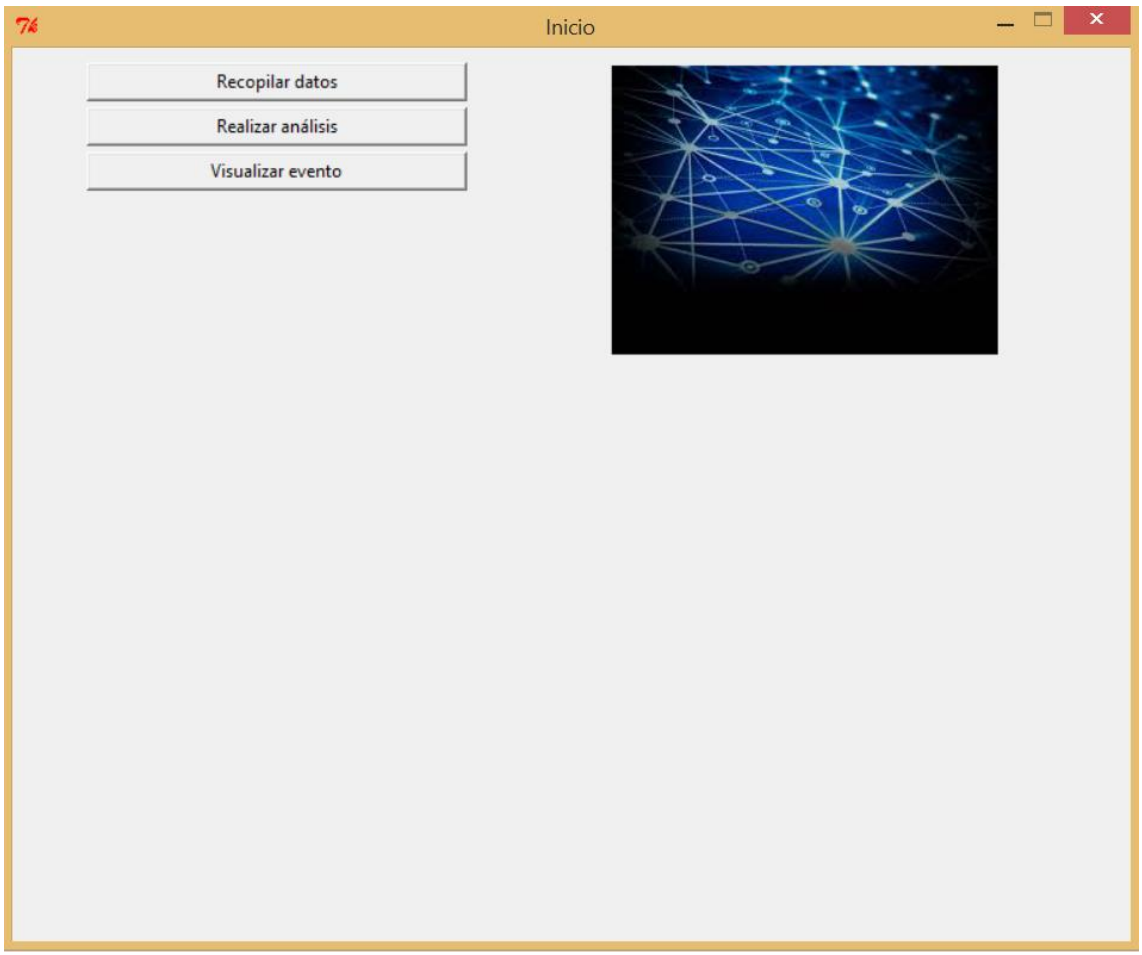

Figura 6.1 Pantalla de inicio

Luego, el usuario tendrá la opción de seleccionar recopilar datos para hacer uso del módulo de extracción de publicaciones de Twitter. En caso ya cuente con los datos necesarios, podrá hacer clic en “Realizar análisis” para procesar los datos o “Visualizar evento” en caso ya se tengan los datos procesados.

En caso seleccione “Recopilar datos”, se procederá a la pantalla mostrada en la Figura 6.2.

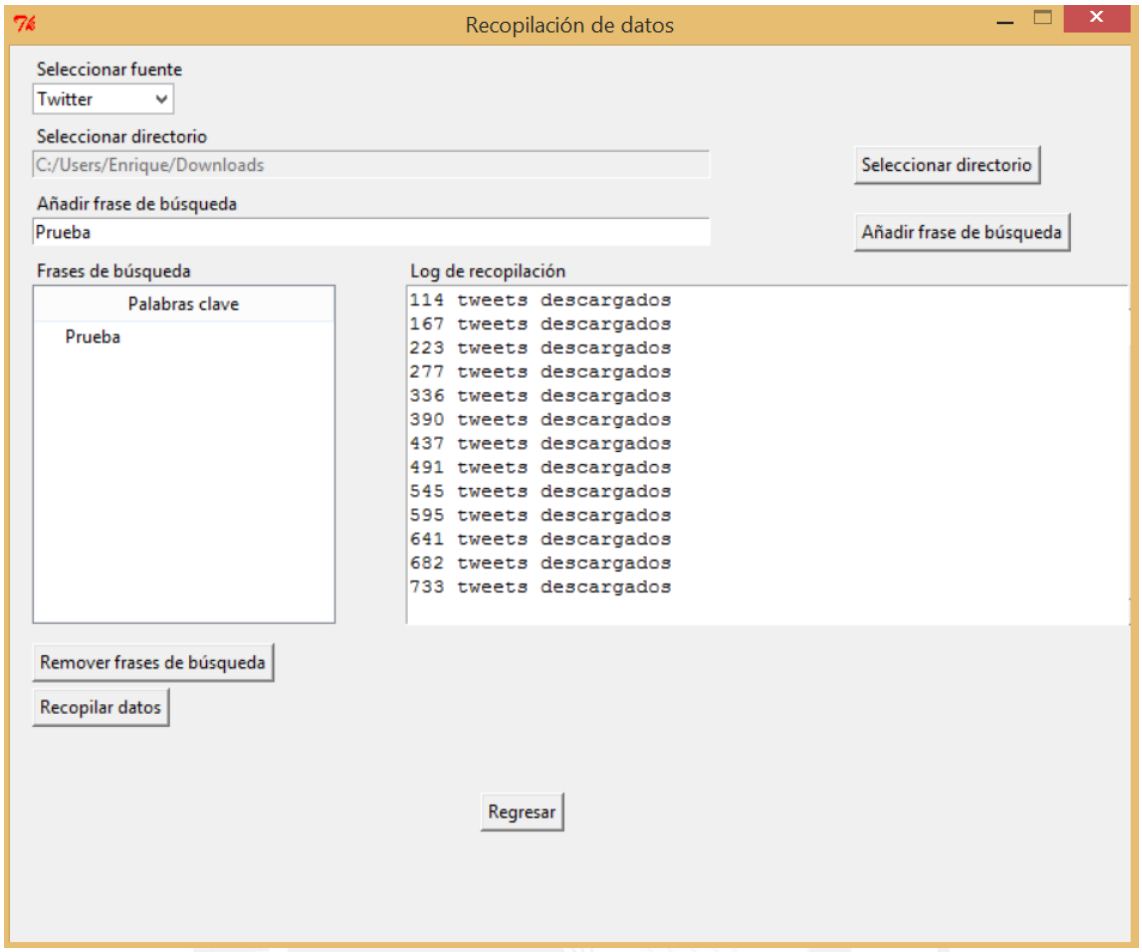

Figura 6.2 Pantalla de recopilación de datos

Para realizar la recopilación de datos, el usuario deberá seleccionar la fuente, cabe mencionar que en caso de este proyecto solo se ofrecerá la opción de Twitter, también deberá seleccionar un directorio en el cual se guardarán los archivos con los datos del evento que va a analizar y definir algunas palabras o frases clave las cuales serán usadas para realizar la consulta a Twitter.

Se podrán remover frases de búsqueda seleccionándolas y haciendo clic en el botón “Remover frases de búsqueda” y en el “Log de recopilación” se mostrará el estado de la recopilación así como cualquier dificultad encontrada durante el proceso.

Una vez se cuenten con los datos necesarios ya sea usando el módulo previamente mencionado o recolectándolos por cuenta propia, se podrá comenzar el análisis de los datos. Para esto desde la Figura 8.1 se deberá hacer clic en “Realizar análisis” en la “Pantalla de Inicio” y se procederá a la pantalla mostrada en la Figura 6.3.

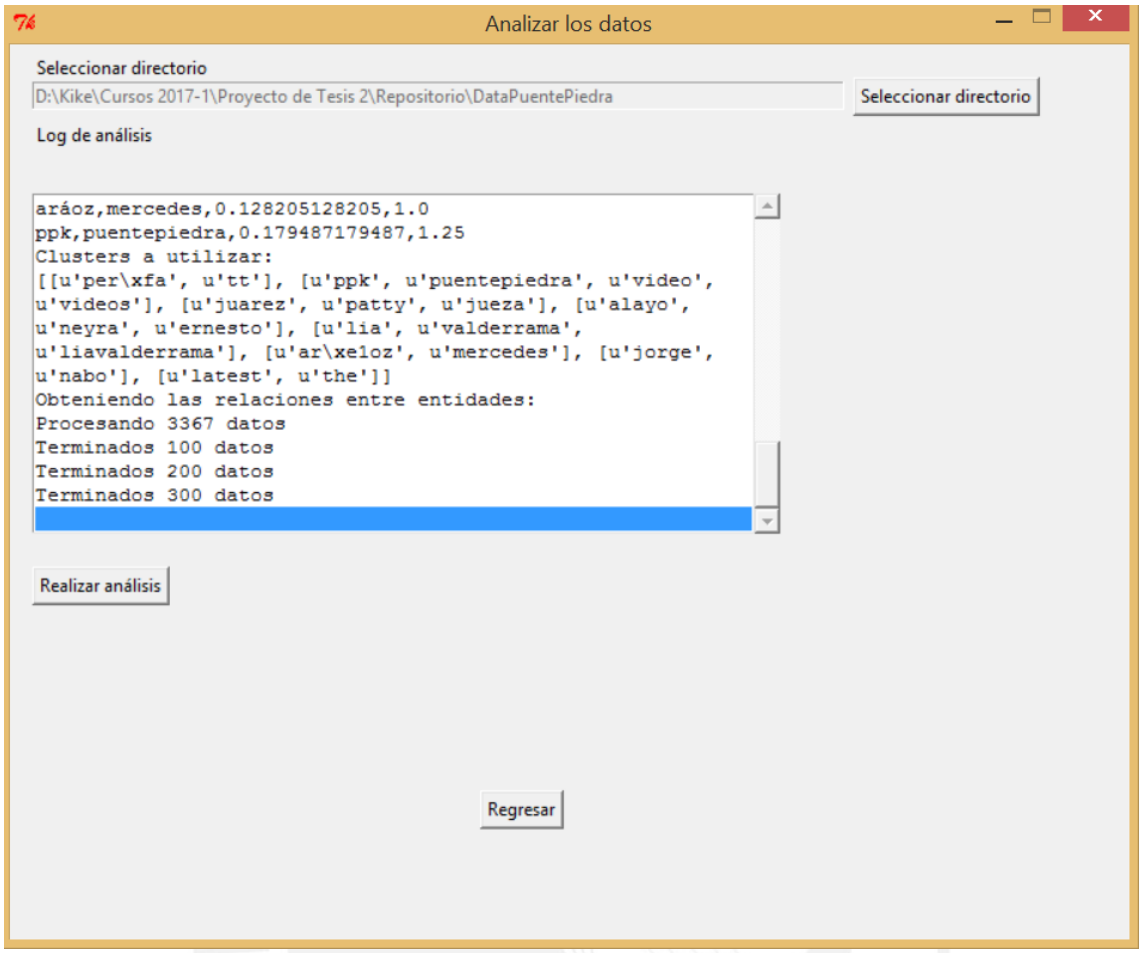

Figura 6.3 Pantalla de análisis de datos

En esta pantalla se procederá a seleccionar el directorio donde se tienen los archivos con los datos de entrada y se hará clic en “Realizar análisis”.

Mientras se procesan los datos se mostrará un log para mantener al usuario informado del progreso y mostrar las alertas u observaciones que sean necesarias.

Una vez finalizado este proceso, se guardarán los resultados del análisis en algunos archivos para poder cargarlos posteriormente y evitar realizar un nuevo análisis.

Posteriormente, se podrá hacer clic en el botón “Visualizar evento” para mostrar la información obtenida a partir del análisis.

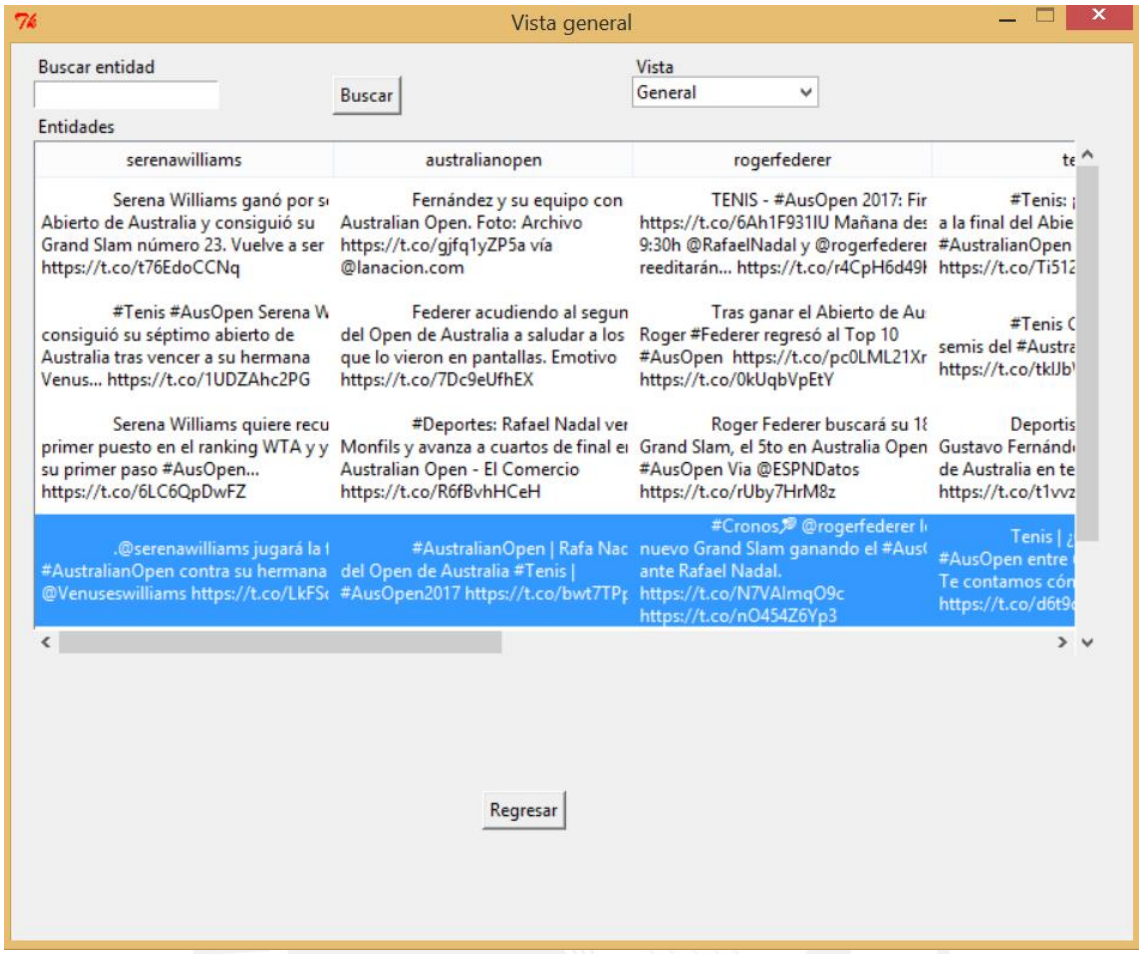

Figura 6.4 Pantalla de la Vista general

En esta pantalla se muestran todas las entidades consideradas relevantes dentro del evento así como algunos de los datos relacionados a ellos más importantes. A partir de aquí se puede pasar a las vistas de “Grafo” y “Línea de Tiempo”, así como realizar búsquedas para alguna entidad en específico.

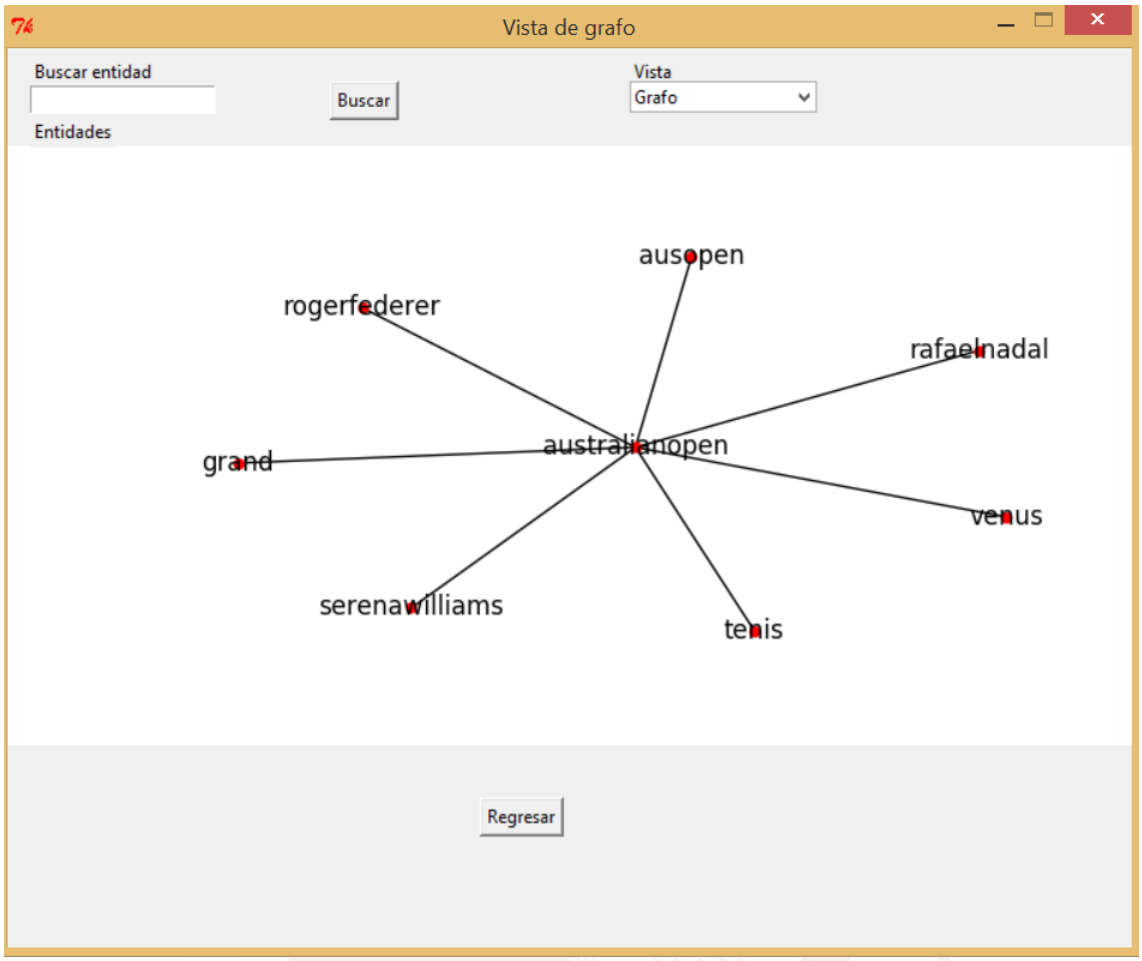

Figura 6.5 Pantalla de la vista del grafo obtenido

En esta pantalla se podrá apreciar de manera visual las relaciones entre entidades participantes del evento. Si se hace clic al nodo central del grafo se podrá ver el detalle de la entidad correspondiente mientras que si se hace clic a alguno de sus nodos adyacentes se podrá explorar el grafo cogiendo el nodo seleccionado como central.

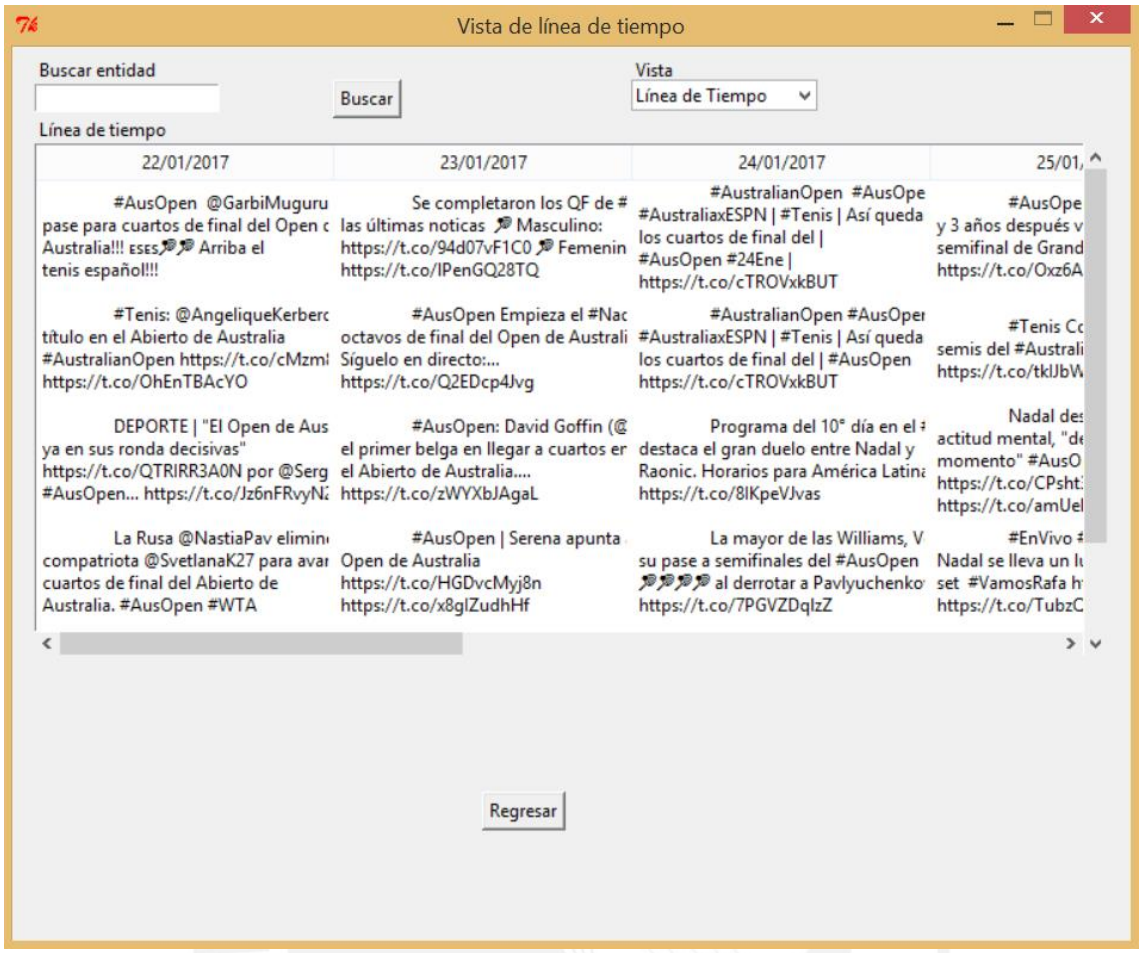

Figura 6.6 Pantalla de la vista de línea de tiempo

En esta pantalla se podrá ver el tiempo de duración del evento y se mostrarán algunos de los datos más relevantes por cada día del evento.

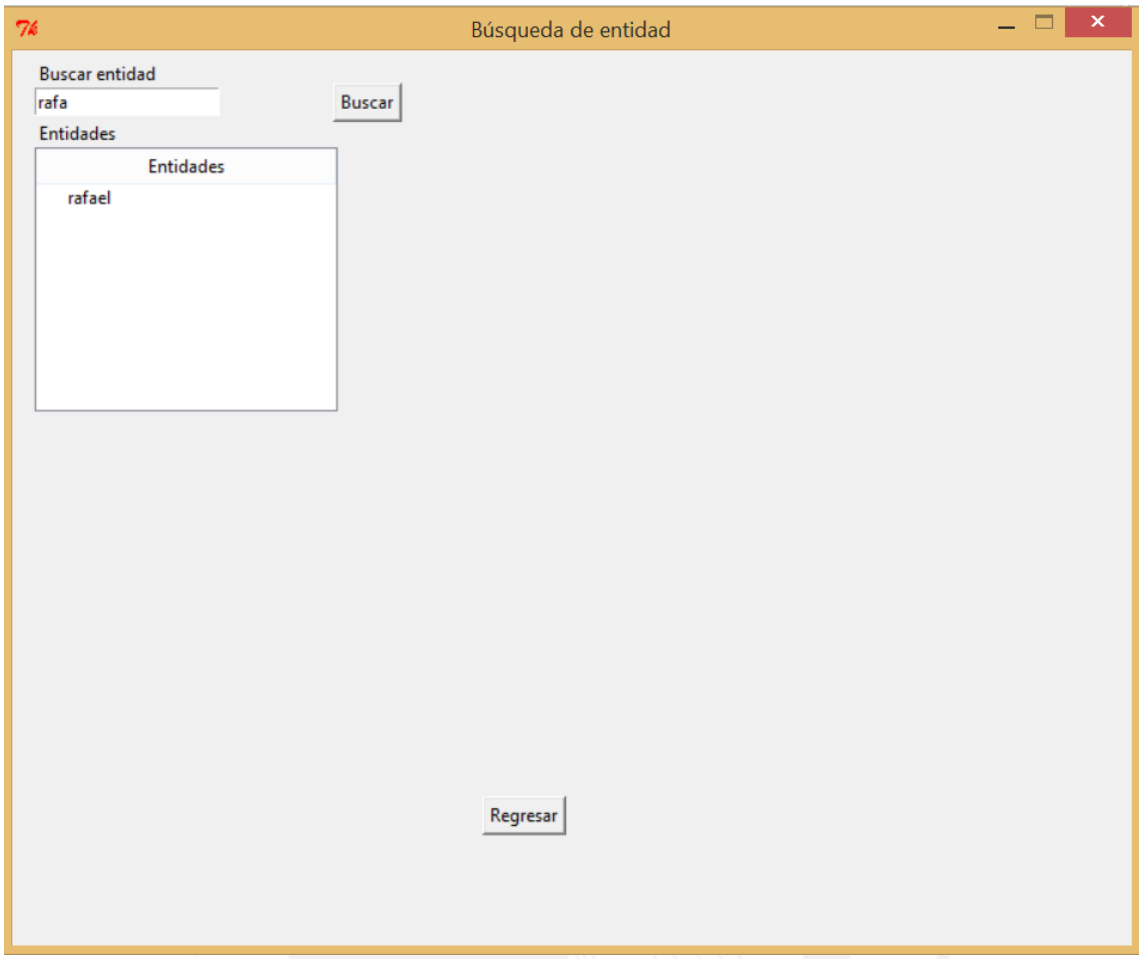

Figura 6.7 Pantalla de búsqueda de entidades

Esta pantalla se mostrará cuando se realice la búsqueda de una entidad desde alguna de las 3 vistas principales. Se mostrará la lista de entidades que correspondan a la búsqueda ingresada y se podrá hacer doble clic sobre las entradas de la tabla para poder mostrar el detalle de alguna de estas.

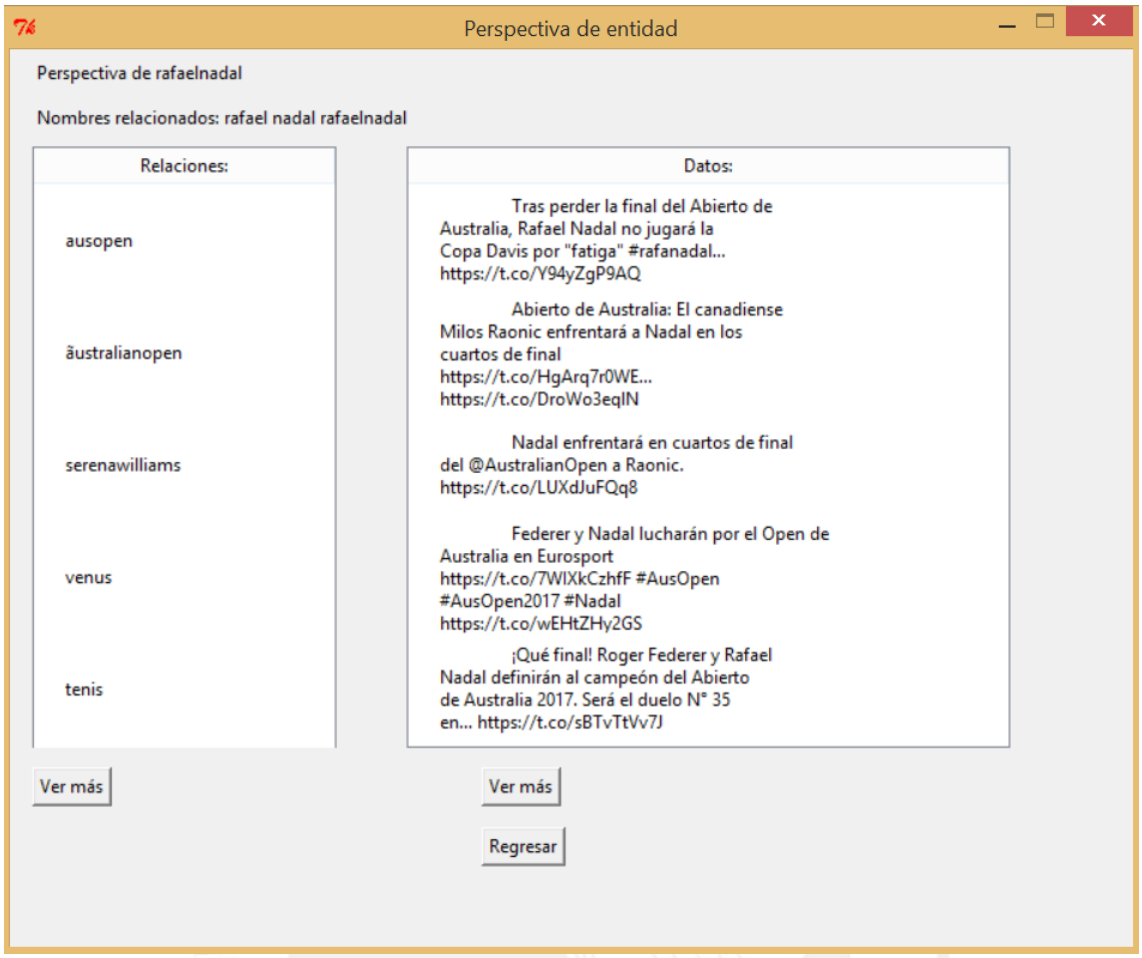

Figura 6.8 Pantalla de perspectiva de una entidad

En esta pantalla se podrán apreciar los datos más relevantes para una entidad en particular, así como algunas de las otras entidades con las que se relacione, si se quiere ver la totalidad de los datos relevantes obtenidos así como las relaciones extraídas, se puede hacer clic en “Ver más” para cualquiera de las 2 tablas.

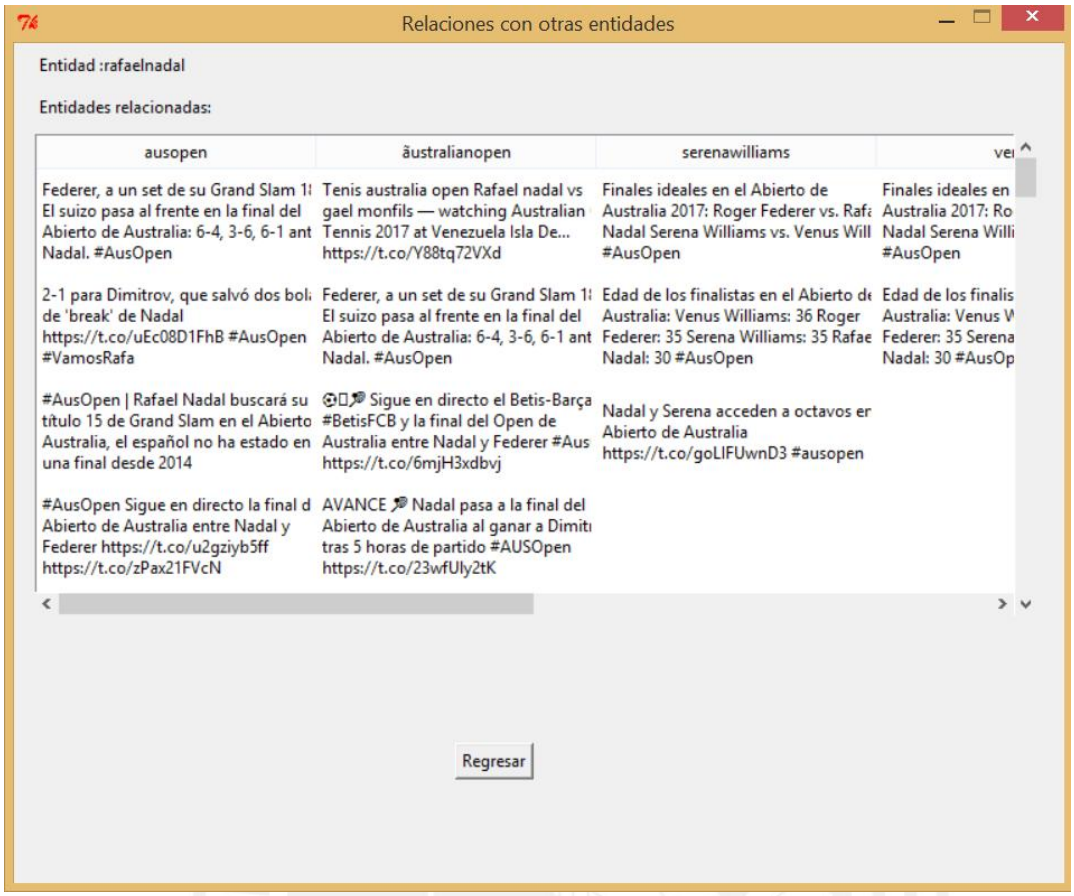

Figura 6.9 Pantalla de relaciones entre entidades

En esta pantalla se puede apreciar todas las entidades con las que se relaciona la entidad inicial así como el detalle de los datos a partir de los cuales se extrapolan estas relaciones.

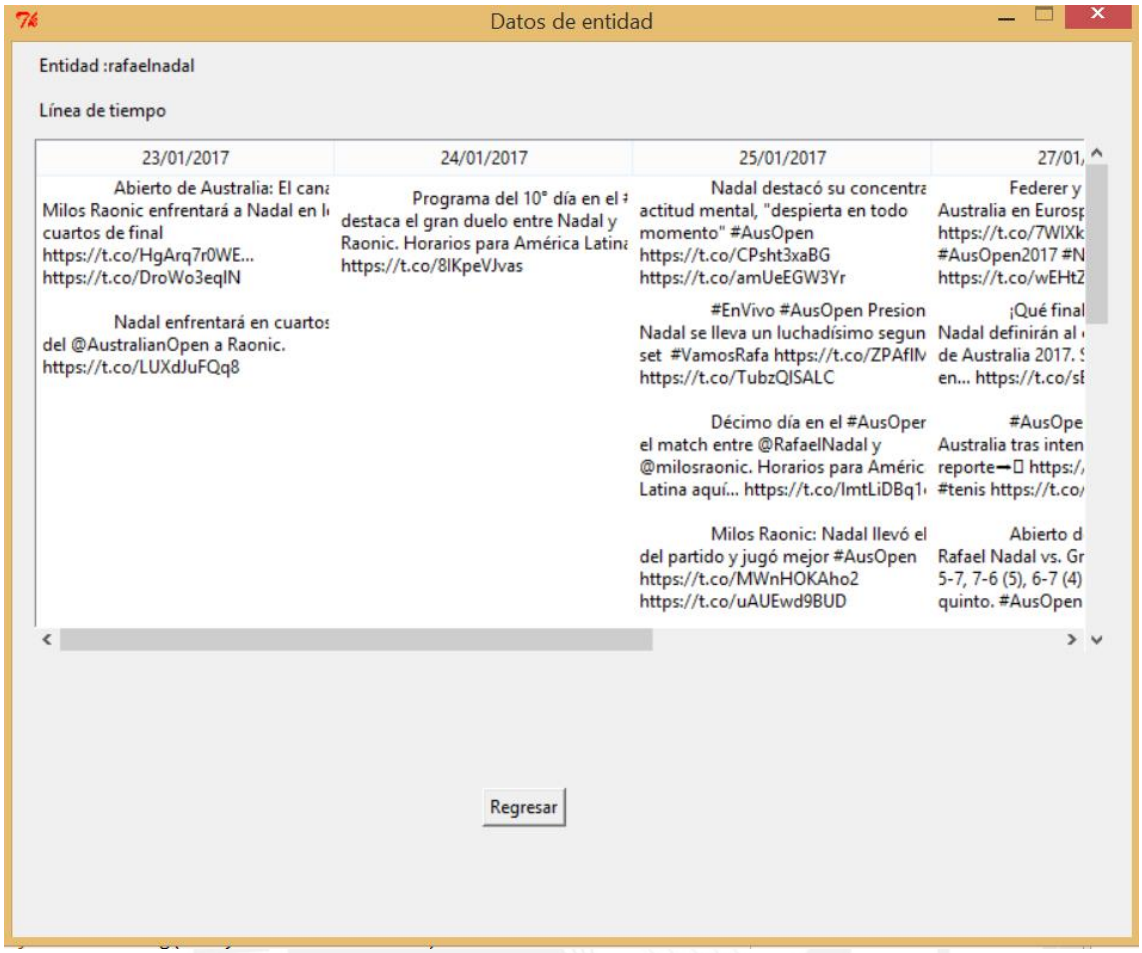

Figura 6.10 Pantalla de datos de una entidad

En esta pantalla se muestran los datos que definen a alguna entidad. Esto se realiza haciendo uso de una tabla la cual trabaja como una línea de tiempo, en la que cada día es acompañado por los datos relacionados a la entidad en ese día en particular.

CAPÍTULO 7: Resultados y Discusión

1. Experimentación

1.1.Selección de entidades

En la elaboración del componente de extracción automática de entidades para los eventos se experimentó con 2 enfoques: usando el NER de Stanford y usando la aplicación de anotado gramatical de UDipe.

De los resultados obtenidos se notó que el NER funcionaba mejor en casos donde los datos correspondían a textos formales, como pueden ser noticias o artículos en la Web. Sin embargo, al momento de trabajar con texto menos formal como son los datos de Twitter que se usaron para las pruebas, los resultados empeoraban drásticamente. Esto es debido a que cambios en alguna letra o la falta de mayúsculas en el nombre de la entidad afectaba el resultado obtenido.

Finalmente se optó por usar UDPipe como base para obtener las entidades iniciales que fueron utilizadas en pasos posteriores del algoritmo debido a que se considera importante permitir trabajar con distintas fuentes y no sólo fuentes de tipo formal, es decir, que no se asume que los nombres propios siempre van a estar escritos adecuadamente ya que esto depende de los datos que se usen para el análisis y en el caso de las pruebas realizadas en Twittter los nombres suelen no estar escritos correctamente.

Por ejemplo, cuando se ingresa a ambas herramientas la oración “rogerfederer venció a nadal”. Se obtienen los siguientes resutados:

```
[(u'rogerfederer', u'O'), (u'venci\xf3', u'O'), (u'a', u'O'), (u'nadal', u'O')]
```

Figura 7.1 Resultado del NER de Stanford

```
1       rogerfederer    rogerfederer    PROPN   _       _       2       nsubj   _       _
2       venció  vencer  VERB    _       Mood=Ind|Number=Sing|Person=3|Tense=Past|VerbForm=Fin
0       root    _       _
3       a       a       ADP     _       _       4       case    _       _
4       nadal   nadal   PROPN   _       _       2       nmod    _       _
```

Figura 7.2 Resultado de UDPipe

En las figuras se puede apreciar que mientras el NER de Stanford clasificó todas las palabras como ‘O’ (Objeto) y no reconoció ’PERS’ (Personas), UDPipe logró clasificar tanto a rogerfederer como a nadal como ‘PROPN’ nombres propios.

1.2.Valores de corte

Se realizaron experimentaciones para determinar los valores de corte usados en distintos puntos de los algoritmos, tales como la etapa de poda en la extracción de entidades o la etapa de agrupamiento de términos.

Además, se tomaron algunos valores del estado del arte como referencia en el caso de la etapa de poda y se modificaron ligeramente hasta encontrar el punto deseado. Este punto

se basó en obtener los mejores valores de precisión y exhaustividad para las ontologías creadas manualmente detalladas en el siguiente punto.

2. Implementación de Línea Base

El primer paso para la obtención de las métricas de verificación fue crear las ontologías que sirvan de Línea Base para los eventos recopilados en el Capítulo 4. Este fue un proceso manual en el cual se analizaron los datos y los resultados parciales obtenidos en las distintas etapas de los algoritmos para poder obtener las entidades más relevantes así como las relaciones más relevantes. Además, se investigó en distintas fuentes en Internet sobre cada evento en particular.

A partir de esto se crearon las ontologías detallando para cada una las entidades, las relaciones entre estas y la actividad o verbo principal que ocasiona esta relación. A continuación se muestran las ontologías por cada caso:

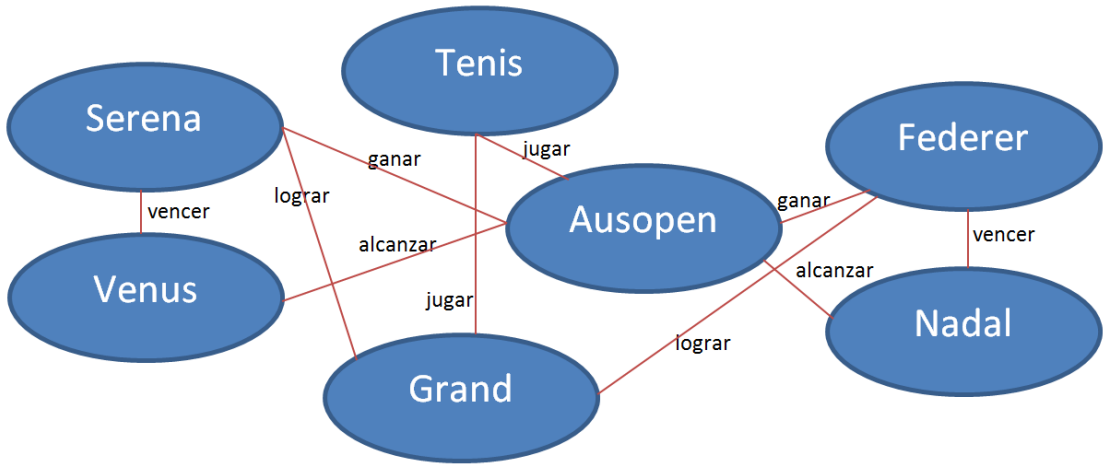

Figura 7.3 Ontología manual para el caso Abierto de Australia

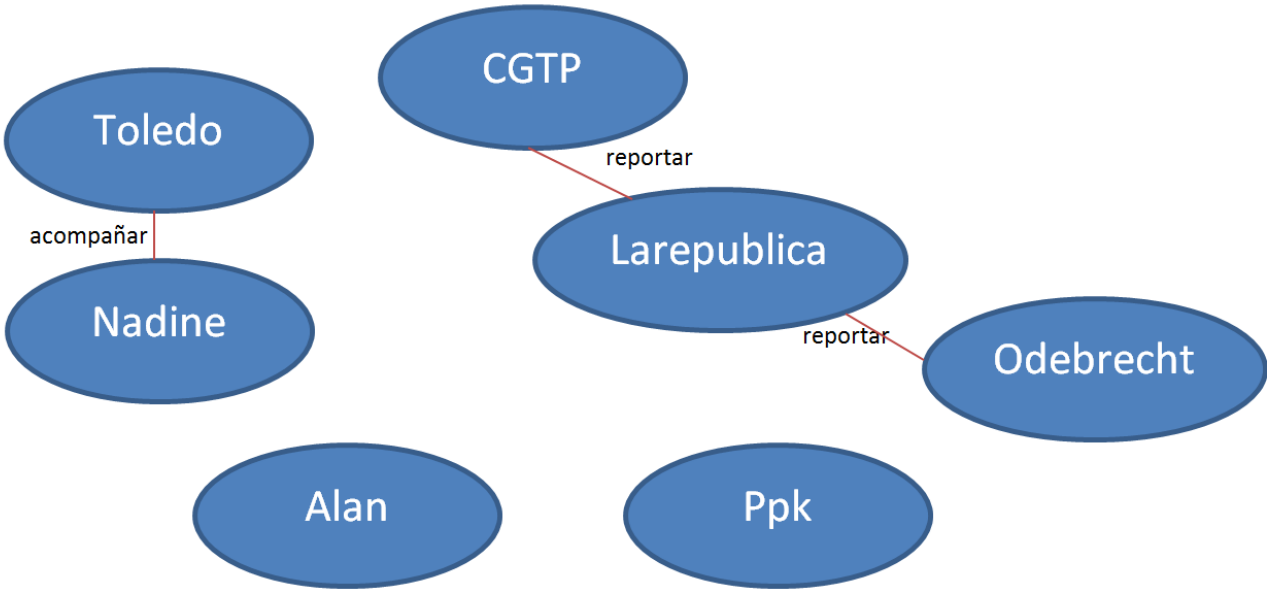

Figura 7.4 Ontología manual para el caso Marcha contra la Corrupción

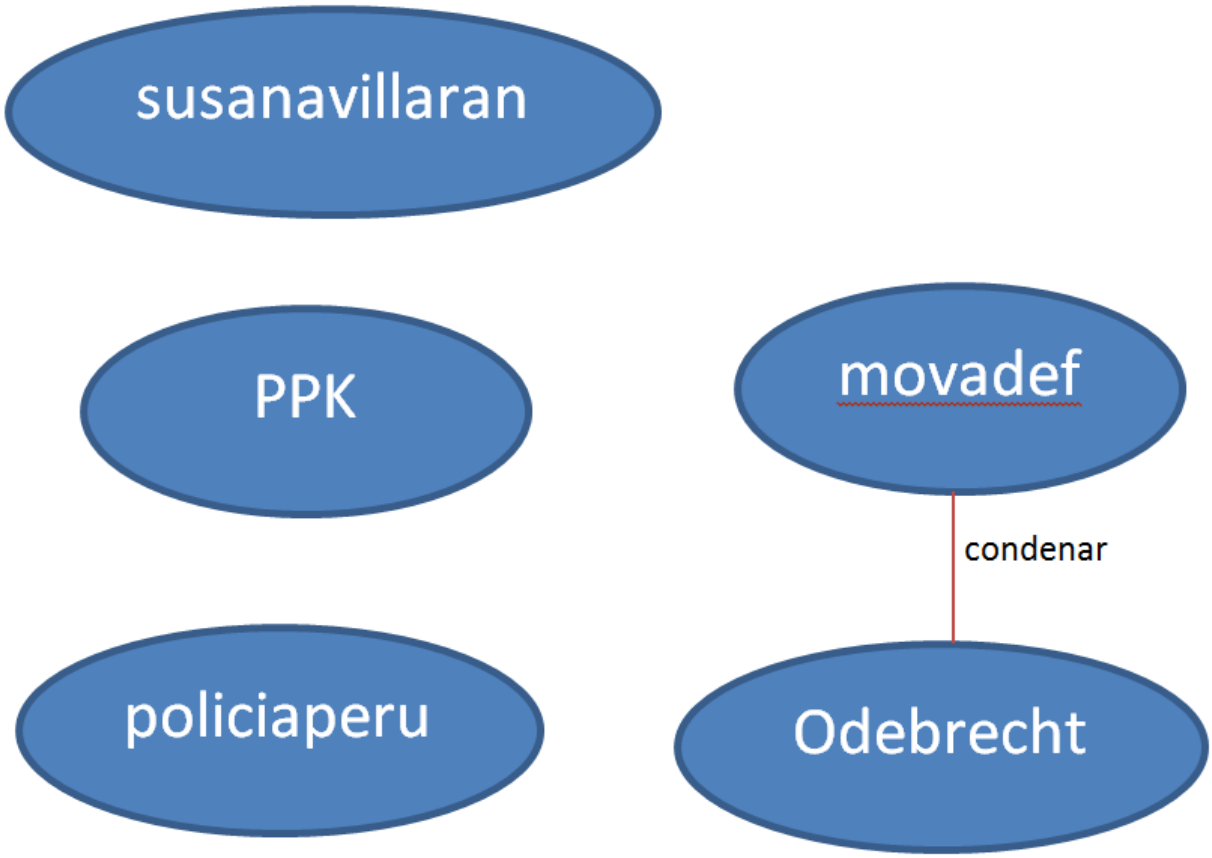

Figura 7.5 Ontología manual para el caso Peaje de Puente Piedra

3. Reporte de métricas para las ontologías extraídas

Una vez obtenidas las ontologías, se procedió a hacer el cálculo de precisión y exhaustividad para cada caso, tanto para las entidades como para las relaciones entre estas, obteniéndose los siguientes resultados:

Parámetro a analizar	Métrica	Valor
Entidades	Precisión	0.875
Entidades	Exhaustividad	1.0
Relaciones	Precisión	0.952
Relaciones	Exhaustividad	1.0

Tabla 7.1 Lista de métricas para el caso Abierto de Australia

Parámetro a analizar	Métrica	Valor
Entidades	Precisión	0.556
Entidades	Exhaustividad	1.0
Relaciones	Precisión	0.333
Relaciones	Exhaustividad	1.0

Tabla 7.2 Lista de métricas para el caso Peaje de Puente Piedra

Parámetro a analizar	Métrica	Valor
Entidades	Precisión	0.467
Entidades	Exhaustividad	1.0
Relaciones	Precisión	0.333
Relaciones	Exhaustividad	0.667

Tabla 7.3 Lista de métricas para el caso Marcha contra la Corrupción

Realizando un análisis tanto de los datos como de los valores obtenidos, se obtiene que para casos donde gran parte de los datos fueron de usuarios que representan noticieros o medios de reportaje, como fue el caso del Abierto de Australia, se obtuvieron valores altos en precisión y exhaustividad.

En el caso del Peaje de Puente Piedra hubo combinación de estos datos con comentarios o menciones del público general así que se obtuvieron valores un poco menores. Mientras que, por el lado del caso de Marcha contra la Corrupción, un gran porcentaje de los datos fueron del tipo de opiniones del público general o comentarios que realizaban los ciudadanos, es por esto que se obtienen valores menores sobre todo en la parte de las relaciones.

CAPÍTULO 8: Conclusiones y Trabajos futuros

1. Conclusiones

A continuación se detallarán las conclusiones sobre el trabajo desarrollado por cada objetivo específico:

Para el objetivo de "Desarrollar un componente para la adquisición automática y almacenamiento de publicaciones de eventos en Twitter", se logró brindar un componente el cual, haciendo uso del API REST proporcionada por Twitter, permita a los usuarios obtener publicaciones de Twitter de los últimos días haciendo uso de algunas palabras claves definidas por ellos.

En segundo lugar, para el objetivo de "Desarrollar un componente para la extracción automática de conceptos y relaciones entre estos, usados para representar los eventos", se logró implementar los algoritmos de FCA y Análisis de Patrones Léxico-Sintácticos para la consolidación y análisis de los datos de eventos que se utilicen. Esto permitió identificar las entidades más relevantes de un evento e identificar las relaciones entre las distintas entidades. Esta información se guarda en ontologías almacenadas en el formato RDF.

Finalmente, para el objetivo de "Implementar un módulo de *software* que integre los componentes desarrollados y permita la visualización de los eventos" se logró brindar un *software* el cual permita el uso de los componentes desarrollados mediante una interfaz, así como la exploración de las ontologías generadas a través de distintas tablas de resumen y el grafo que representa la ontología.

2. Trabajos Futuros

Se identificaron los siguientes puntos de mejora para trabajos futuros:

- Añadir más fuentes de recolección de datos

Por limitaciones de tiempo en el alcance se definió que la fuente para la cual se iba a implementar el componente de adquisición de datos de eventos era Twitter. Sin embargo, la procedencia de las fuentes no es una limitación para el funcionamiento de la aplicación.

Es por esto que se plantea como trabajo futuro añadir más fuentes de recolección de datos a la aplicación para dar mayor libertad a los usuarios de escoger las fuentes que mejor se adapten a sus necesidades.

- Añadir la opción de poder trabajar con más idiomas

En el alcance se definió que los datos de entrada deben encontrarse en el idioma español y a partir de esta premisa se implementaron los componentes posteriores de adquisición de entidades y relaciones de los eventos.

Sin embargo, brindar una aplicación capaz de trabajar con datos de distintos idiomas proporcionaría una mayor flexibilidad al usuario al momento de escoger los datos e incluso poder tendrían la capacidad datos de distintos idiomas.

- Recolectar información más detallada de las entidades de un evento a partir de la Web

Actualmente, la aplicación muestra información únicamente recopilada de los datos ingresados en el proceso de análisis.

Una posible mejora a este enfoque sería, a partir de las entidades encontradas en los datos, hacer consultas en la Web para obtener mayor información respecto a estas y poder brindar un análisis más detallado.

Referencias bibliográficas

[AL-SMADI, QAWASMEH 2016] AL-Smadi, M. y Qawasmeh, O. – "Knowledge-based Approach for Event Extraction from Arabic Tweets" – International Journal of Advanced Computer Science and Applications – Volumen 7, Número 6 – pp. 483–490

[ATEFEH, KHREICH 2013] Atefeh, F. y Khreich, W. – "A survey of techniques for event detection in Twitter" – Computational Intelligence – Volumen 31 – pp. 132–164

[BECHHOFER et al, 2002] Bechhofer, S., Goble, C. y Horrocks, I. – "Requirements of Ontology Languages" – University of Manchester – Reino Unido

[BENDER 2013] Bender, E. M. – "Linguistic Fundamentals for Natural Language Processing" – Morgan & Claypool Publishers

[BIRD et al, 2009] Bird, S., Klein, E. y Loper, E. – "Natural Language Processing with Python" – O'Reilly Media – USA

[BISHOP 2006] Bishop, C. M. – "Pattern Recognition and Machine Learning" – Springer-Verlag New York

[BUITELAAR et al, 2003] Buitelaar, P., Olejnik, D. y Sintek, M. – "OntoLT: A Protégé Plug-In for Ontology Extraction from Text" – Proceedings of the Demo Session of the International Semantic Web Conference

[CASTRO 2008] Castro, Raúl – "Representación del Conocimiento. Web Semántica" – Universidad Carlos III de Madrid – España

[CELIK et al, 2011] Celik, I., Abel, F. y Houben, G. – "Learning Semantic Relationships between Entities in Twitter" – Holanda – Delft University of Technology

[CIMIANO, VÖLKER 2005] Cimiano, P. y Völker, J. – Text2Onto – "International Conference on Application of Natural Language to Information Systems" – pp. 227–238

[CIMIANO 2006] Cimiano, P. – "Ontology Learning and Population from Text Algorithms, Evaluation and Applications" – Alemania – University of Karlsruhe

[FAYYAD et al, 1996| Fayyad, U., Piatetsky-Shapiro, G. y Smyth, P. – "From Data Mining to Knowledge Discovery in Databases" – American Association for Artificial Intelligence – USA

[CUELOGIC 2016] Kumbla, S. – "Fast Data: Powering Real-Time Big Data". Disponible en: http://www.cuelogic.com/blog/fast-databig-data-in-real-time/

[GIMPEL et al, 2011] Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisentein, J., Heilman, M., Yogatama, D., Flanigan, J. y Smith, N. A. – "Part-of-Speech Tagging for Twitter: Annotation, Features and Experiments" – USA – School of Computer Science, Carnegie Mellon University

[HAMASAKI et al, 2007] Hamasaki, M., Matsuo, Y., Nishimura,T. y Takeda, H. – "Ontology Extraction using Social Network" – Japón – National Institute of Advanced Industrial Science and Technology

[KANEIWA et al, 2007] Kaneiwa, K., Iwazume y M., Fukuda, K. – "An Upper Ontology for Event Classifications and Relations" – Japón – National Institute of Information and Communications Technology

[LEE et al, 2010] Lee, S., Gong, G. y Lee S. – "Entity-Event Lifelog Ontology Model (EELOM) for LifeLog Ontology Schema Definition" – 2010 12th International Asia-Pacific Web Conference – pp. 344–346.

[LEE et al, 2013] Lee, C., Wu, C., Yang, H. y Wen, W. – "Learning to Create an Extensible Event Ontology Model from Social-Media Streams" – Advances in Neural Networks-ISNN 2013 – Part 2 – pp. 436–444

[LEE, WU 2015] Lee, C. y Wu, C. – "Extracting Entities of Emergent Events from Social Streams Based on a Data-Cluster Slicing Approach for Ontology Engineering" – International Journal of Information Retrieval Research – Volumen 5, 3ra edición – pp. 1 – 18

[LI et al, 2012] Li, R., Hou Lei, K., Khadiwala, R. y Chen-Chuan Chang, K. – "TEDAS: a Twitter Based Event Detection and Analysis System" – USA – University of Illinois at Urbana-Champaign

[LIDDY 2001] Liddy, E. D. – "Natural Language Processing" – USA – Syracuse University7

[LOBZHANIDZE et al, 2013] Lobzhanidze, A., Zeng, W., Gentry, P. y Taylor, A. – "Mainstream Media vs. Social Media for Trending Topic Prediction – An Experimental Study" – Consumer Communications and Networking Conference (CNNC) – pp. 729 – 732

[MARYAM et al, 2011] Maryam, H., El-Beltagy, S. R. y Rafea, A. – "A Survey of Ontology Learning Approaches" – International Journal of Computer Applications – Volumen 22 – pp. 36 – 43

[MUSEN 2015] Musen, M.A. – "The Protégé project: A look back and a look forward" – AI Matters. Association of Computing Machinery Specific Interest Group in Artificial Intelligence

[RCR 2016] Responsible Conduct of Research – Northern Illinois University – USA Disponible en: http://www.niu.edu/rcrportal/index.html. Fecha de consulta: 29 de Septiembre de 2016

[OPSCLARITY 2016] OpsClarity 2016 State of Fast Data & Streaming Applications Survey.

Disponible en: http://info.opsclarity.com/2016-fast-data-streaming-applications-report.html. Fecha de consulta: 24 de Septiembre de 2016

[RAIMOND, ABDALLAH 2007] Raimond, Y. y Abdallah, S. – "The Event Ontology" – Inglaterra – Queen Mary, University of London

[RAMACHANDRAN et al, 2015] Ramachandran, K., Chandra, A., Mallya, D., Chaitanya, J.N.V.K. y Kamath, S. – "Ontology based Approach for Event Detection in Twitter datastreams" – India – Department of Information Technology, National Institute of Technology

[REDSPIRE 2016] Kirk, C. – "Real time analytics on the rise?" Disponible en: http://redspire.co.uk/crm-blog/real-time-analytics-on-the-rise/. Fecha de consulta: 10 de Septiembre

[STANFORD 2017] Portal oficial del grupo de NLP de Stanford. Disponible en: https://nlp.stanford.edu/software/CRF-NER.shtml. Fecha de consulta: 30 de Abril de 2017

[TWITTER 2016] Portal de desarrollo de Twitter. Disponible en: https://dev.twitter.com/. Fecha de consulta: 23 de Octubre de 2016

[UDPIPE 2017] Portal oficial de UDPipe. Disponible en: https://ufal.mff.cuni.cz/udpipe. Fecha de consulta: 30 de Abril de 2017

[VALKANAS, GUNOPULOS 2013] Valkanas, G. y Gunopulos D. – "Event Detection from Social Media Data" – Greece – University of Athens

[WARD et al, 2010] Ward, M. O., Grinstein, G. y Keim, D. – "Interactive data visualization: foundations, techniques and applications" – CRC Press

[WU et al, 2003] Wu, S., Tsai, T., Hsu, W. – "Domain Event Extraction and Representation with Domain Ontology" – Taiwan – Institute of Information Science

[W3C 2016] Portal oficial del World Wide Web Consortium. Disponible en: https://www.w3.org/standards. Fecha de consulta: 01 de Septiembre de 2016