Universidad Autónoma Metropolitana Unidad Azcapotzalco
División de Ciencias Básicas e Ingeniería
Licenciatura en Ingeniería en Computación

Reporte final de proyecto terminal:

Sistema de recuperación de información semántico

Selene María de Jesús Ugalde Chávez
Matrícula: 205304493

Trimestre 12P
Agosto de 2012

Asesores

Dra. Maricela Claudia Bravo Contreras
Profesor del departamento de sistemas

M. Hugo Pablo Leyva
Profesor del departamento de sistemas

Tabla de Contenido

Introducción

La computación ha revolucionado la forma en que el hombre accede y almacena la información. La evolución de los medios electrónicos de almacenamiento ha incrementado en forma acelerada la cantidad datos que el ser humano puede acceder. En la actualidad una gran número de importantes procesos ya no son ejecutados por el hombre sino confiados a programas de computadora, en especial aquellos que requieren de exactitud, velocidad y la gestión de un gran flujo de datos, por ejemplo algunas transacciones bancarias, comercio, procesos industriales, entre otros.

Más allá las computadoras que en un principio estaban reservadas para áreas muy especificas como la investigación militar hoy en día se encuentran presentes en casi todas las actividades humanas, en algunas de manera tan indispensable que es imposible pensar en dar marcha atrás.

Independientemente de la preferencia por el formato clásico de los libros; es imposible no reconocer el valor de las bibliotecas virtuales que la digitalización ha hecho posible. Sin dejar de mencionar la Internet que ha elevado la capacidad de comunicación y el intercambio de información hasta tocar la *Nube*: un gran porcentaje de la población mundial ahora tiene acceso a incontables servicios y documentos con un sólo clic, literalmente, nuestros potentes buscadores devuelven millones de resultados relacionados para la más simple consulta.

Sin embargo, aún hay varios puntos por resolver cuando la tarea de encontar información exacta es, todavía, primordialmente humana, buscar entre millones de documentos representa un claro inconveniente. Actualmente uno de los objetivos importantes que se persigue en la vanguardia de la tecnología informática es hacer de la gran magnitud de documentos a nuestra disposición una ventaja y no una desventaja, automatizando el procesamiento de documentos por su contenido.

Desafortunadamente, nuestras computadoras y su poderosa capacidad de procesamiento están construidas sobre un lenguaje formal muy distinto al nuestro, mientras que el gran acervo del conocimiento humano se encuentra representado en lenguaje natural; es así como surge la necesidad de máquinas que entiendan nuestro idioma.

En un intento por satisfacer esta necesidad ha habido en las últimos años una importante colaboración de lingüistias y computólogos con el fin de modelar el lenguaje humano, al que llamamos natural sin duda por ser el proceso que por excelencia

dominamos, tan inherente en nosotros que pocas veces pensamos su increíble complejidad. La realidad es que el lenguaje natural es aún desconocido en muchos aspectos para los incluso lingüistas, y gran tema de discusión para psicólogos, filósofos, neurólogos, sociólogos y antropólogos.

Aun así se ha trabajo en diversas técnicas para automatizar el procesamiento de textos en lenguaje natural con un significativo éxito científico e incluso comercial. Actualmente existen varias organizaciones nacionales e internacionales dedicadas a investigar, perfeccionar o desarrollar nuevas técnicas de Procesamiento del Lenguaje Natural.

Problemática del Procesamiento del Lenguaje Natural

El Procesamiento del Lenguaje Natural es una disciplina derivada de la lingüística computacional cuyo objetivo formal es automatizar la comprensión del lenguaje natural, entendiendo por comprensión la capacidad de reconocer y manipular información expresada en algún lenguaje humano. Es decir, procesar textos por el significado de su contenido y no como archivo binarios. Tiene aplicaciones en una serie de tareas de gran importancia, por ejemplo: la traducción automática, las interfaces en lenguaje natural y la recuperación de información.

La problemática de Procesamiento del Lenguaje Natural radica en la imposibilidad de las computadoras de manejar la ambigüegad de los lenguajes naturales; al parecer la capacidad humana de discernir el significado ambiguo de una palabra se extiende más allá del contexto y tiene que ver más bien con una colección de experiencias relacionadas con el mundo de los objetos que acumulamos a lo largo de nuestra vida.

Tomemos uno de los ejemplos clásicos en los libros de Procesamiento del Lenguaje Natural y consideremos el enunciado:

> “¿Hay algún restaurante para comer en la ciudad?”

En primera instancia sabemos que la pregunta anterior expresa deseo de saber si en la ciudad hay algún establecimiento donde adquirir comida; si bien el enunciado también podría leerse como la pregunta de alguien interesado en comerse, literalmente, un restaurante. De acuerdo a nuestra experiencia del mundo sabemos que a excepción de *Godzilla* nadie anda por la vida devorando edificios y como el contexto del enunciado no mencione al monstruo japonés de inmediato descartamos este sentido. Pero para una computadora, carente de la experiencia humana, este tipo de obviedades no son obvias.

Existen especificamente tres tipos de ambigüedad que pueden encontrarse en el lenguaje natural: la ambigüedad léxica que ocurre cuando una misma palabra puede pertenecer a diferentes categorías gramaticales. La ambigüedad sintáctica o ambigüedad estructural que aparece cuando debido a la forma en que se asocian los distintos constituyentes de una oración, podemos interpretarla de varias formas distintas, siendo a veces casi imposible de solucionar. Por ejemplo,
Juan vio a su hermana con unos binoculares (¿Juan usa los binoculares para ver a su hermana o Juan vio que su hermana tenía unos binoculares?) Finalmente la ambigüedad sémantica que aparece cuando una palabra o conjunto de palabras tienen más de un pósible sentido.
La resolución de los diferentes tipos de ambigüedades requiere aplicar diferentes técnicas para generar reglas de desambigüación que en general no son sencillas de definir.

Objetivos del proyecto

Resulta evidente, en un mundo donde la información oportuna marca la diferencia y la cantidad de datos a procesar es tan grande que necesita ser automatizada, la necesidad de incorporar tecnicas de procesamiento del Lenguaje Natural en la resolución de las demandas actuales en el manejo de la información.

El objetivo principal de este proyecto es explorar y aplicar técnicas del Procesamiento del Lenguaje Natural a la recuperación de la información para diseñar un sistema de recuperación de información semántico. Como problemática se planteó automatizar la definición de perfiles de investigador para un grupo de profesores, a través de sus publicaciones. Es decir diseñar un sistema tal, que al recibir como entrada una serie de documentos, devolviera como salida una lista de sus autores, y por cada autor una lista con el nombre de la institución que le respaldara, sus principales áreas de investigación, temas de interés, coautores y una lista de publicaciones asociadas.

La dificultad del proyecto radica en la heterogeneidad de los documentos; que es en realidad el problema clave del Procesamiento de Lenguaje Natural: la ambigüedad natural del lenguaje natural. Pensemos por ejemplo, cómo podría determinarse que la palabra “authors” en un artículo se refiere al título de una sección en la que se listan los autores del mismo, o parte de de una frase que habla de los principales autores del tema que trata el artículo: “the principal authors of historical materialism are...”

Por el contrario, otro ejemplo es el problema de determinar que dos conjuntos de palabras distintas significan lo mismo como el caso de que un mismo autor puede escribir

su nombre de distintas maneras; de hecho es algo que pasa con mucha frecuencia, a lo largo del desarrollo del proyecto fue un problema importante a resolver.

También presenta problemas de estructura, es decir, resulta imposible saber de antemano el formato en que un autor acomodará los elementos en su documento: ¿pondrá las palabras clave al principio o al final del capítulo?, entre otros.

Para realizar la extracción de información se utilizó, como primera etapa, técnicas de anotado semántico y sobre el texto anotado se trabajó con el reconocimieto de expressiones regulares y entidades nombrables, mediante un lenguaje de especificación de patrones linguisticos.

Diseño

El diseño general del sistema cuentra con cuatro módulos principales:

- *Analizador léxico*: este módulo es el encargado de separar el texto en unidades léxico-gráficas. Tiene como entrada los documentos de la base de publicaciones y como salida genera texto separado en unidades léxico-gráficas o tokens.
- Analizador sintáctico: este módulo asociará etiquetas a las unidades léxicas. Tiene como entrada el texto separado en unidades léxico-gráficas y como salida devuelve el texto anotado con etiquetas asociadas a las unidades léxicas., en esta etapa se definen las características semanticas y linguisticas de cada palabras.
- Analizador semántico: en este modulo, a través de un lenguaje de especificación de patrones, se buscan y reconocen patrones lingüísticos en el texto. Tiene como entrada el texto anotado con etiquetas asociadas a las unidades léxicas y como salida los patrones identificados en el texto.
- Módulo de visualización: Finalmente este módulo se encarga permite visualizar los patrones encontrados. Tiene como entrada los patrones identificados en el texto y como salida la visualización de los mismos.

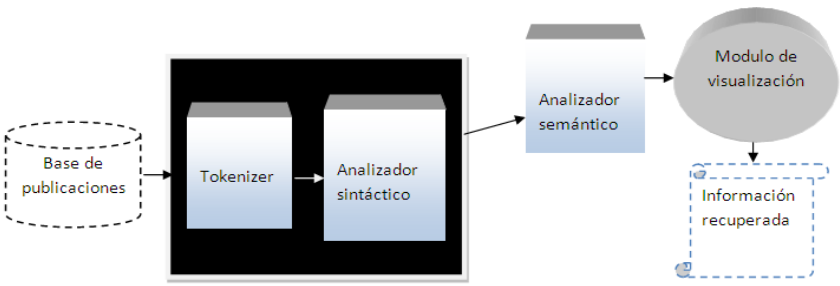


Figura 1.0 Diagrama general del sistema

Etapas de procesamiento del sistema

Primera etapa

Los primeros dos módulos conforman la primera etapa en la que crea un corpus de publicaciones a analizar, luego se procesa cada documento en el corpus para convertirlo en una estructura que facilite el análisis semántico. A esta fase se le conoce como prepocesamiento o anotado de texto y consiste de las siguientes fases:

1. El analizador lexico comienza su trabajo removiendo cualquier anotación previa en el texto y separándolo en dos unidades básicas: palabras y espacios. Además construye una mapa que guarda la posición y longuitud de cada palabra, definida como un nodo en un árbol.
2. El analizador sintáctico etiqueta cada palabra en categorías gramáticales (revise el anexo para ver las categorías utilizadas)
3. Se identifican entidades nombradas, es decir palabras que refieren a entidades bien conocidas por ejemplo México se reconoce como un país o dólar como una moneda.

Las clases involucradas en esta etapa se muestran a continuación.

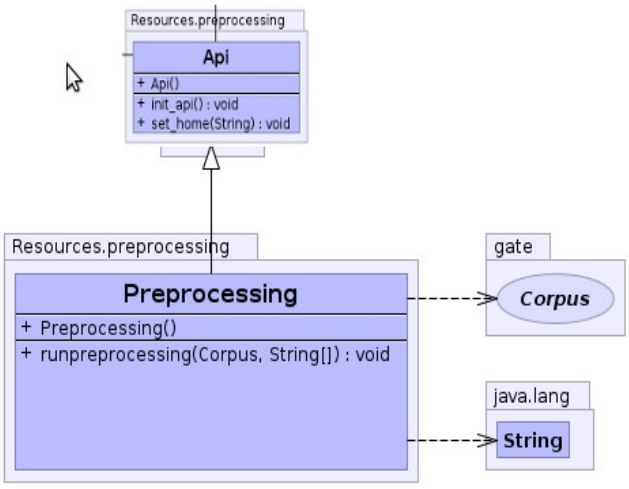


Figura 2.0 Diagrama de primera etapa

Segunda etapa

Una vez procesado el documento se busca información clave. Para este proyecto se planteó reconocer como mínimo, por cada artículo: título, autores, palabras clave, filiación y resumen o introducción. La identificación de esta información es tarea del analizador semántico implementado a través del reconocimiento de expresiones lingüísticas.

Para esta tarea se ocupó *JAPE*, un lenguaje de especificación de patrones lingüísticos que permite declarar reglas que definen expresiones regulares.

Cada regla de *JAPE* se compone de:

– una primera parte que contiene la descripción del patrón de anotación

una segunda parte que describe las nuevas anotaciones que deben crearse para cada coincidencia encontrada.

A continuación se describen las reglas definidas para la extracción:

Autor

```
{
palabra "Autors" concatenado
 signo de puntuación (dos puntos) concatenado
(
Nombre propio, singular [0 o más] concatenado
Gattezer.primera_persona[opcional] concatenado
Nombre propio, singular[0 o más] concatenado
signo de puntuación [0 o más]
) [0 o más]
           }
```

Titulo

```
{
Nombre propio, singular [primera coincidencia] concatenado
Palabra "sustantivo" o "verbo" o "adjetivo" "conjuncion" o
"preposición" [1 o más] concatenado
Salto de linea anterior al siguiente parrafo [1]
                              }
```

Filiación

```
{
 palabra "Universidad" concatenado
 signo de puntuación (dos puntos) concatenado
 Nombre propio, singular [1 o más] concatenado
 Palabra "de" [opcional] concatenado
 Nombre propio, singular [0 o más]
                                 }
```

Keywords

```
{
 palabra "Keywords" concatenado
  signo de puntuación (dos puntos) concatenado
 (
  Palabra, en singular o en masa [0 o más] concatenado
           signo de puntuación [0 o más] concatenado
 Palabra, en singular o en masa  o  adjetivo [0 o más]
 ) [0 o más]
  signo de puntuación (punto) o Salto de linea
       }
```

Abstract

```
{
Palabra "Abstract o Introduction"(primera coincidencia) concatenado
Siguiente parrafo [1] concatenado
Salto de linea
                              }
```

Tercera Etapa

Cuando una oración o palabra coincide con algún patrón definido se encierra entre dos etiquetas como en un archivo *XML*, el nombre de la etiqueta es elegida por el programador que define las reglas Jape correspondientes. Al final de la segunda etapa se tiene una colección de documentos anotados. Es en la cuarta etapa cuando ésta información es extraída y colocada en estructuras de datos para su análisis posterior.

La extracción es un proceso iterativo que recorre el documento en búsqueda de las etiquetas señaladas, cuando una etiqueta es encontrada se recoge su contenido como una cadena de texto y se almacena según su categoría. La información obtenida en esta etapa se muestra al usuario como resultado del proceso de extracción.

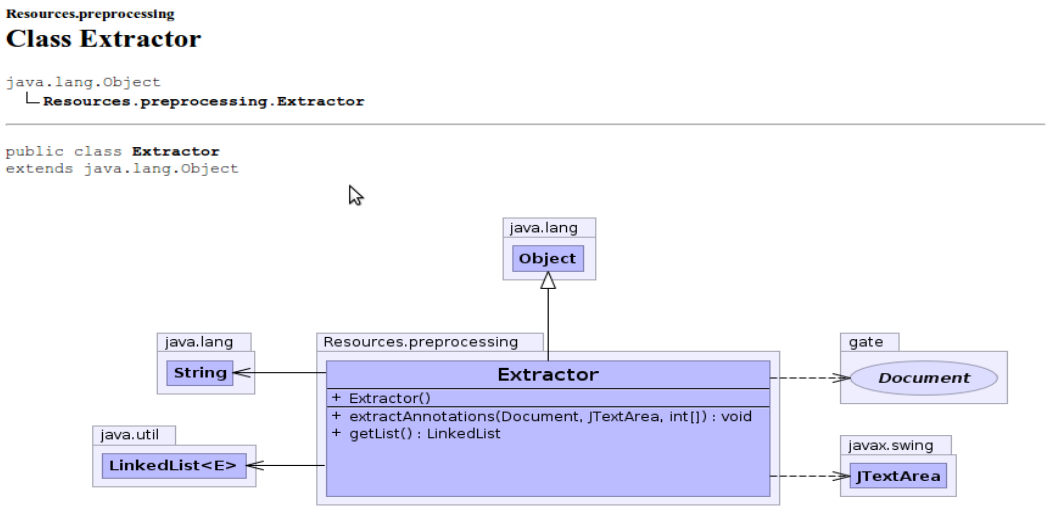


Figura 3.0 diagrama de la cuarta etapa

Cuarta etapa

El analizador semántico se encarga de dos importantes tareas, en una de ellas identifica información relevante que es extraída, su segunda cometido es estructurar la información obtenida dandole un significado y una interpetación. El lenguaje de especificación de patrones no fue suficiente para darle solución a las problemáticas que planteaba este análisis; entonces se incorporó la utilización de listas de búsqueda de entidades nombradas.

Con la combinación de *JAPE* y la lista de entidades pudo resolverse la correlación entre un autor único y las distintas formas en que puede escribirse su nombre. Así mismo, hizo posible la clasificación de un autor determinado en un área o áreas de investigación; para este caso se elaboró una taxonomía de las ciencias computacionales que sirvió como lista de temas y conceptos representativos. Cuando el anotado de texto es ejecutado se identifican éstas palabras y se marcan como correspondientes a una categoría. Un número alto de palabras de cierta categoría encontradas a lo largo de un texto predice como conocedor del tema que representa dicha categoría al autor del artículo.

Tambien en esta segunda fase de análisis y última étapa de procesamiento se estructura la información en entidades como puede apreciarse en la figura 5.0. El propósito de esta construcción es modelar la información y así facilitar su manipulación posterior en tareas como almacenamiento en base de datos, poblado de ontologías, etc. Finalmente, ésta etapa tambien ofrece como resultado al usuario un lista de investigadores de la UAM azcapotzalco y su correspondiente perfil, así como una lista de investigadores agrupados según sus lineas de investigación coincidentes.

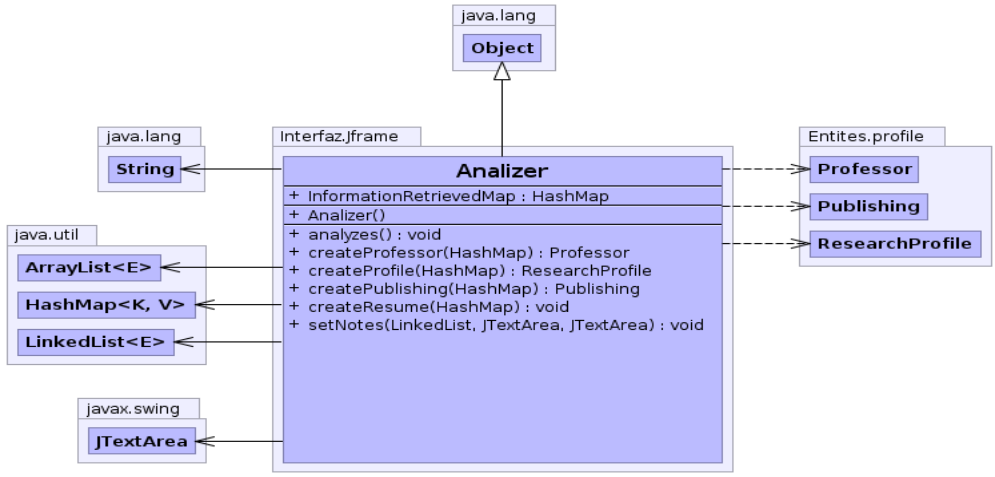


Figura 4.0 diagrama de clases correspondiente a la cuarta etapa

figura 5.0 información extraida estructurada en entidades

Funcionamiento general del Sistema

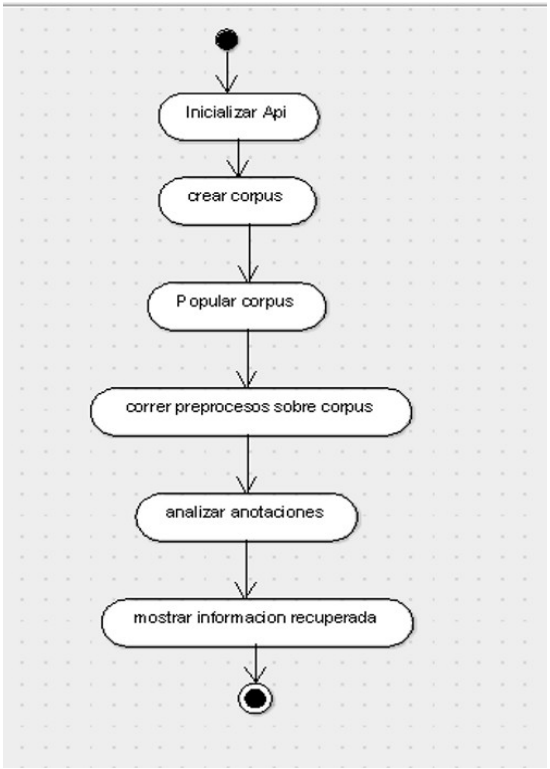


figura 6.0 diagrama de actividad del sistema

Implementación

Como aplicaciones auxiliares se utilizarón *GATE,* una herramienta diseñada para el procesamiento de texto, *eclipse* como ambiente de desarrollo y *Yworks-uml-docle*t para la elaboración de diagramas.
A continuación se detalla el uso de los plugins de *GATE* en el desorrollo del programa:

• *ANNIE Document Reset*: Remueve anotaciones preexistentes en los documentos a analizar

• *ANNIE English Tokenise*r: Separa el texto en *tokens*

• *ANNIE Sentence Splitter*: Agrupa los *tokens* de la fase anterior en oraciones.

• *ANNIE Gazetteer*: Busca *tokens* coincidentes con listas predefinidas de entidades nombradas.

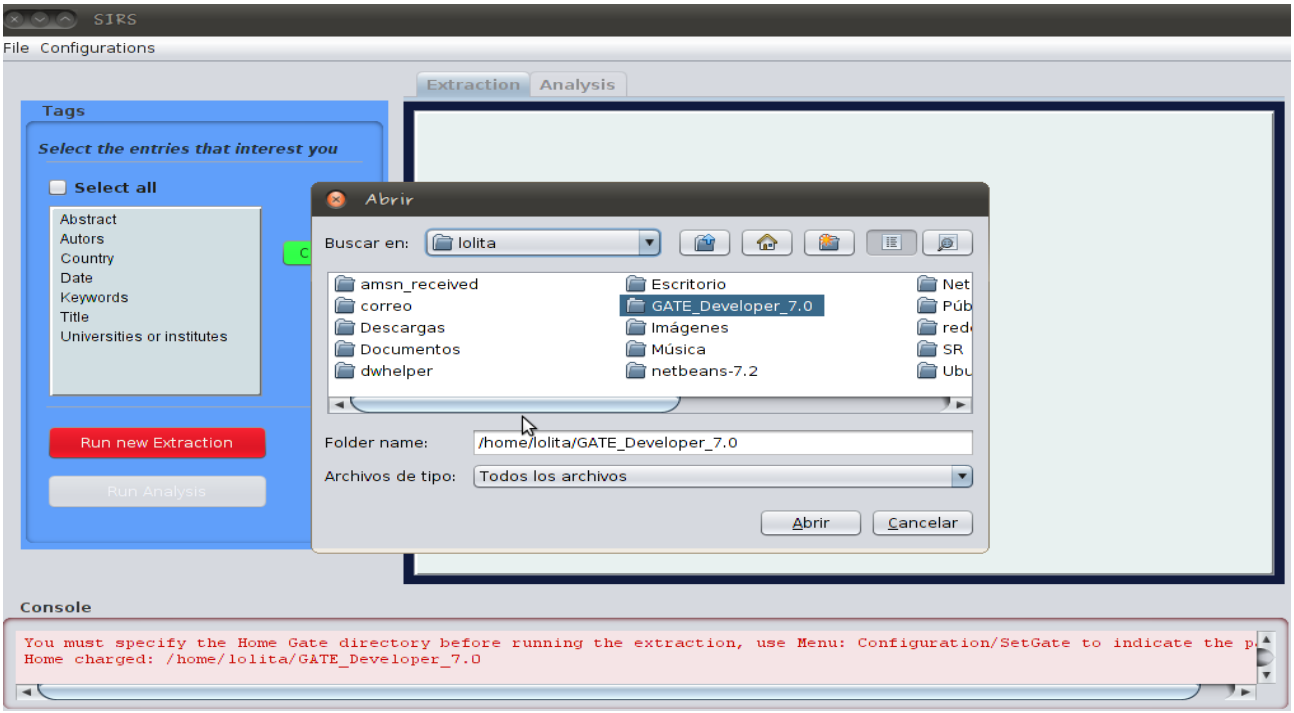
• *ANNIE POS tagger*: Agrega características de categoría a las anotaciones previas.

Sin duda alguna la utilización de herramientas preexistentes hizo posible este proyecto, siendo el Procesamiento del Lenguaje Natural una ardua tarea, desarrollar en su totalidad cáda uno de los módulos hubiera conllevado triplicar el tiempo necesario para el desarrollo del proyecto. Además la reutilización de código no sólo acelera el desarrollo, sino en un caso como este proyecto, donde se utiliza codigo desarrollado por un grupo de investigadores destacados, el aprendizaje es invaluable: Se tiene acceso a formas nuevas de modelar y estructurar información, maneras expertas de planteamiento de problemas y construcción de soluciones y lecciones avanzadas de diseño de algoritmos.

Protocolo de pruebas

1. Configuraciones

En esta prueba se demuestra el correcto funcionamiento de la configuración del directorio de instalación de Gate.
Se siguieron los siguientes pasos: selección del menú “configurations/setGate”, selección de directorio de instalación.
Los resultados fueron los esperados.

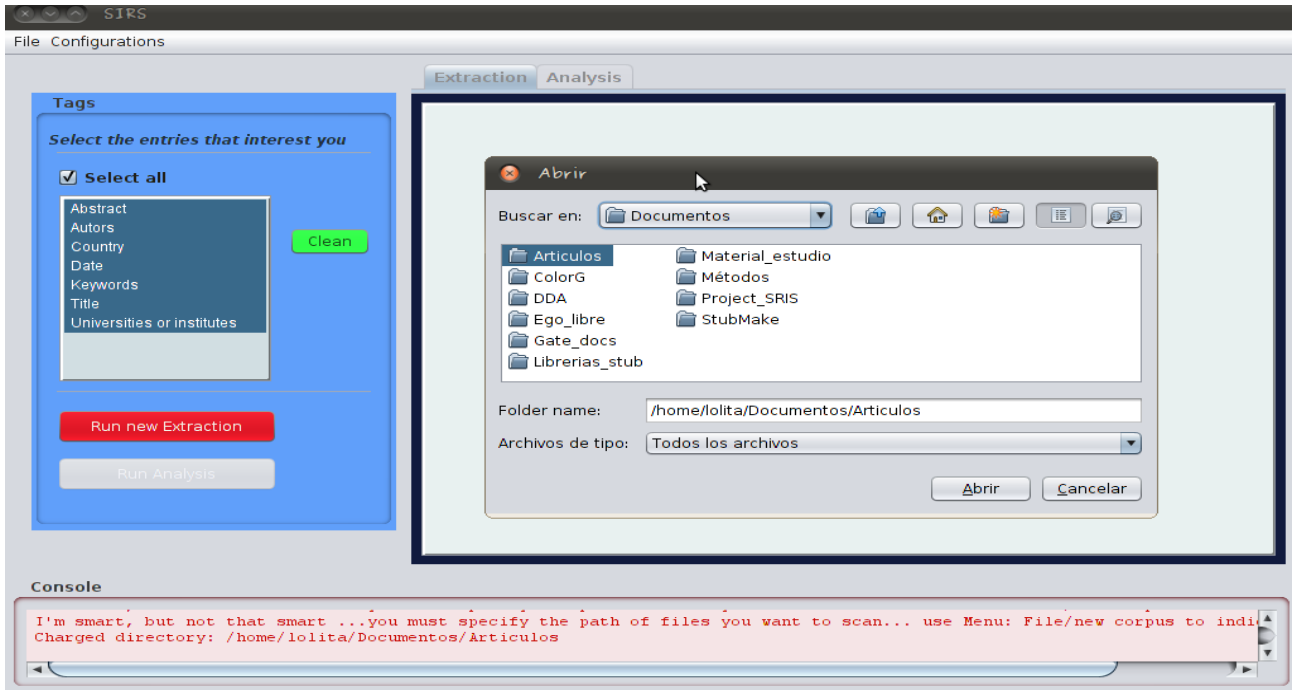


2. Selección de u nuevo conjunto de documentos a analizar

En esta prueba se demuestra el correcto funcionamiento de la configuración del directorio de documentos a analizar.

Se siguieron los siguientes pasos: selección del menú “files/NewCorpus”, selección de directorio de instalación.

Los resultados fueron los esperados.

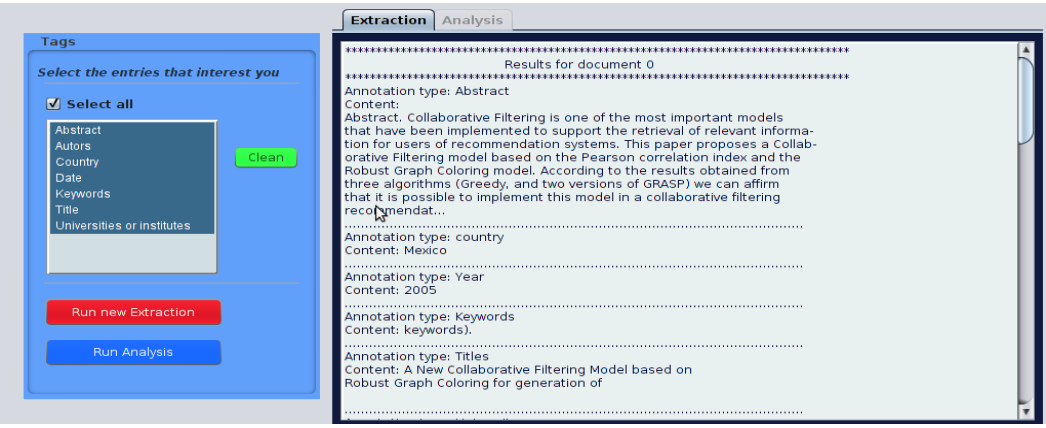


3. Extracción de contenido

En esta prueba se demuestra el correcto funcionamiento de la primera fase denominada extracción de información. Se probaron cincuenta artículos de diferentes autores.

Se siguieron los siguientes pasos: creación de un nuevo corpus de documentos ejecución de “run extraction”

Los resultados fueron los esperados.

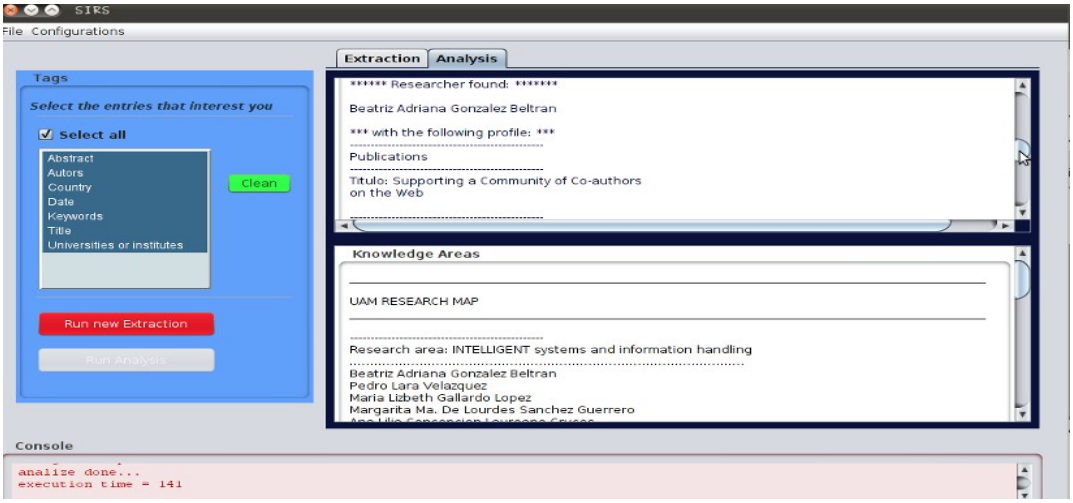


4. Análisis de información

En esta prueba se demuestra el correcto funcionamiento de la segunda fase denominada análisis de información.Se probaron cincuenta artículos de diferentes autores.

Se siguieron los siguientes pasos: ejecución de “run extraction”, ejecución de “run analisys”

Los resultados fueron los esperados.



Conclusiones

Considero que el aprendizaje más valioso que adquirí durante la realización de este proyecto fue el descubrimiento de nuevas maneras de modelar información desde diferentes perspectivas. Trabajar con Procesamiento del lenguaje Natural implica estudiar areas de conocimiento ajenas a las matemáticas y la computación como la lingüística y lo más básico de psicología.

Por otro lado también adquirí una conciencia más profunda de mis procesos mentales luego de tantas horas dedicadas a la reflexión del funcionamiento de mi propio proceso del lenguaje, en mi corta experiencia pude vislumbrar dos importantes premisas: uno, el lenguaje es sin duda uno de los procesos mentales más sofisticados, su funcionamiento es tan perfecto que la mayor parte del tiempo somos inconcientes de él.

También me he convencido de que el Procesamiento del lenguaje Natural puede ser una herramienta muy poderosa en la solución de demandas prácticas y actuales aún si no se logra programar la comprensión en su totalidad; hace apenas una década gran parte de la tecnología de nuestro tiempo era impensable. De manera que no debe abandonarse una tarea que parezca imposible de completar por el momento, cada intento, cada resultado parcial y no exacto nos acerca a nuestro objetivo último.

Bibliografía

[1] B. Coppin, “Understanding Language” in *Artificial intelligence illuminated,* Canada: Jones and Bartlett Publishers, 2004

[2]AMPLN. (2009, Noviembre 18). *¿Qué es Procesamiento del Lenguaje Natural?* [Online]. Available: http://www.ampln.org/pmwiki.php?n=Main.PLN

[3] S. Rivera Bernal, “Sistema clasificador de archivos de música usando el concepto de memoria asociativa”, proyecto terminal, División de CBI, Universidad Autónoma Metropolitana Azcapotzalco, México, 2010.

[4] J. L. Ugalde Anaya. “Sistema clasificador de documentos de proyectos terminales usando el concepto de memoria asociativa”, proyecto terminal, División de CBI, Universidad Autónoma Metropolitana Azcapotzalco, México, 2011.

[5] N. Guzmán González, “Aplicación de distintas técnicas de minería de datos para el tratamiento de información” proyecto terminal, División de CBI, Universidad Autónoma Metropolitana Azcapotzalco, México, 2011.

[6]Arnetminer. (2010, March 10). *Introduction.* [Online]. Available: http://arnetminer.org/introduction

[7] J. Herrera. (2005, Mayo 24) *modelo Fundamentado en Análisis de Dependencias y WordNet para el reconocimiento de Implicación Textual.* [Online]. Available: http://nlp.uned.es/~jesus/dea.pdf

[8]Princeton University. (2011, June 21). *What is WordNet?* [Online]. Available: http://wordnet.princeton.edu/

[9]CELI. (2011). *Sophia Semantic Engine*. [Online]. Available: http://www.celi.it/en/sophia-semantic-engine.shtml

[10] The Eclipse Foundation. (2011). *About the Eclipse Foundation.* [Online]. Available: **http://www.eclipse.org/org/**

[11] GATE. (2011). *ANNIE: a Nearly-New Information Extraction System.* [Online]. Available: http://gate.ac.uk/sale/tao/splitch6.html#x9-1330006.5

[12] GATE. (2011). *GATE: a full-lifecycle open source solution for text processing.* [Online]. Available: **http://gate.ac.uk/overview.html**

[13] GATE. (2011). *Parsers.* [Online]. Available: **http://gate.ac.uk/sale/tao/splitch17.html#x22-42100017.2**

[14] GATE. (2011). *JAPE: Regular Expressions over Annotations.* [Online]. Available: ***http://gate.ac.uk/sale/tao/splitch8.html#x12-2040008***

[15] ACL. (2011). *About the ACL*. [Online]. Available: http://www.aclweb.org/index.php?option=com_content&task=view&id=38&Itemid=35

[16] D. Jurafsky and J.H. Martin, "Representing meaning" in *Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*. Marcia Horton, ed. New Jersey: Prentice Hall, 2000.

[17]F. Verdejo *et al.* (1999, June 2). *Information retrieval with NLP techniques*. [Online]. Available: http://nlp.uned.es/~ircourse/

[18]R. Johansson, "*Dependency-based Semantic Analysis of Natural-language Text*", Ph.D. dissertation, Dept. Comp. Science, Lund Univ., Sweden, 2008

[19]T. Moure y J. Llisterri. (2010, Octuber 10). *Lenguaje y nuevas tecnologías: el campo de la lingüística computacional.* [Online]. Available: **http://liceu.uab.es/~joaquim/publicacions/llisterri_moure_96.html**

[20]A. Moreno. (2000). *Diseño e implementación de un lexicón computacional para lexicografía y traducción automática,* [Online]. Available: **http://elies.rediris.es/elies9/index.htm**