Universidad Central “Marta Abreu” de las Villas.
Facultad Matemática Física y Computación
Ingeniería Informática

Título: Backend en Qt y python para el diccionario léxico Wordnet en Español

Autor
Alexander Avello Silvério

Tutores
Ing. Abel Meneses Abad

Curso 2014- 2015

Tabla de contenido

Introducción

En el Centro de Estudios de Informática (CEI) de la Universidad Central "Marta Abreu" de Las Villas (UCLV), realiza trabajos de investigación-desarrollo y servicios científico-técnicos

Falta

En este laboratorio se realiza una investigación sobre el procesamiento del lenguaje natural español donde existe una aplicación denominada QTNLP la cual realiza varias funcionalidades con respecto a problemas de similaridad de textos. Esta aplicación necesita de un módulo QTNLP-Wordnet para agregar palabras al Wordnet en español en la estructura adecuada y ampliar sus funcionalidades.

Objetivo General

Desarrollar un módulo para la herramienta QtNLP que permita agregar palabras al Wordnet en español en la estructura adecuada.

Objetivos específicos:

- Sistematizar el estado actual y las tendencias de las herramientas para la edición de Wordnet y otros recursos de la computación.
- Diseñar la aplicación QtNLP-Wordnet.
- Implementar la aplicación QtNLP-Wordnet.
- Validar la solución implementada.

Para guiar el desarrollo del trabajo se plantean las siguientes **preguntas de investigación:**

¿Las herramientas de gestión de diccionarios semántico-léxicos pueden utilizarse correctamente para la edición del Wordnet en español desde el punto de vista de la experiencia de usuario?

¿Cómo extender la aplicación QtNLP con la incorporación del módulo que permita agregar palabras al Wordnet en español en la estructura adecuada?

¿Implementa QtNLP-Wordnet de forma correcta el negocio planteado para la consulta y edición del diccionario Wordnet en español?

Justificación

La detección de similaridad en textos escritos en lenguas naturales es un área de investigación actual. Mayoritariamente los trabajos existentes se realizan sobre el idioma inglés. Los recursos disponibles como corpus o lexicones están personalizados para la lengua inglesa. Dentro de estos recursos uno de los más utilizados es el Wordnet, una base de datos léxica con más de 30 años de desarrollo. Wordnet en inglés posee más de 100 mil términos, mientras que el Wordnet en español solo posee unos 20 mil. Las herramientas para la edición de estos diccionarios son tecnologías privadas de instituciones científicas que solo divulgan el resultado. Se desea aplicar a los algoritmos de detección de similaridad en español similares recursos a los existentes para idioma inglés. Una dificultad con la generalización de estos recursos es que el trabajo de los lingüistas debe hacerse sobre herramientas de fácil acceso y uso.

Hipótesis

El desarrollo del módulo para la herramienta QtNLP permitirá agregar palabras al Wordnet en español en la estructura adecuada.

El presente trabajo estará estructurado de acuerdo a la siguiente secuencia lógica: introducción, tres capítulos, conclusiones, recomendaciones, bibliografía y anexos.

En el Capítulo 1 se llevará a cabo la caracterización de las tecnologías de programación a utilizar en el cuerpo del trabajo, y además, se realizará una descripción teórica de las herramientas que se utilicen para la confección del sistema.

El capítulo 2 y 3 no se han consensuado con el tutor.

Capítulo I: Caracterización del procesamiento del lenguaje natural español y tecnologías para el desarrollo de la aplicación.

En este capítulo se abordarán descriptivamente los sustentos teóricos de este trabajo, puesto que en su desarrollo, se aplican varias herramientas computacionales, dentro de las cuales destaca el uso de la tecnología Qt y Python. Los contenidos abordados anteriormente en la tesis precedente serán

tratados a modo resumen, centrándonos en los elementos fundamentales, y ahondando en los nuevos.

1. Procesamiento del lenguaje natural español

La aplicación del procesamiento de lenguaje natural más obvia y quizá más importante en el momento actual es la búsqueda de información. Por un lado, en Internet y en las bibliotecas digitales se contiene una cantidad enorme de conocimiento que puede dar respuestas a muchísimas preguntas que tenemos. Por otro lado, hay tanta información que no sirve porque ya no se puede encontrarla. Hoy en día la pregunta ya no es "¿si se sabe cómo...?" sino "ciertamente se sabe, pero ¿dónde está esta información?".(Carbonell, 1992)
Técnicamente, rara vez se trata de decidir cuáles son relevantes para la petición del usuario y cuáles no. Usualmente, una cantidad enorme de documentos se puede considerar como relevantes en cierto grado, siendo unos más relevantes y otros menos. Entonces, la tarea se entiende cómo medir el grado de esta relevancia para proporcionar al usuario primero el documento más relevante; si no le sirvió, el segundo más relevante, etc.(Carbonell, 1992)

1.1 Similitud semántica

La similitud semántica en el área de procesamiento de lenguajes naturales, es la medida de la interrelación existente entre dos palabras cualesquiera en un texto.(Blettner, 1989)

La medida de la similitud semántica entre palabras se realiza mediante la relación existente entre los conceptos de la red semántica. La relación existente entre las palabras y su discurso coherente forma parte de la propiedad natural del lenguaje humano y al mismo tiempo la base para el desarrollo de los sistemas de desambiguación automáticos. Se puede afirmar, por tanto, que las palabras que comparten un contexto similar están generalmente relacionadas, y por consiguiente, se pueden seleccionar sus sentidos a partir de la distancia semántica.(Blettner, 1989)

1.2 Wordnet

Wordnet es una base de datos léxica del Idioma Inglés. Agrupa palabras en inglés en conjuntos de sinónimos, proporcionando definiciones cortas y

generales, y almacena las relaciones semánticas entre los conjuntos de sinónimos. Su propósito es doble: producir una combinación de diccionario y tesauro cuyo uso sea más intuitivo, y soportar análisis automático de texto y a aplicaciones de Inteligencia Artificial. La base de datos y las herramientas del software se han liberado bajo una licencia BSD y pueden ser descargadas y usadas libremente. Además la base de datos puede consultarse en línea. (Miller, 1990)

2. QtNLP

Aplicación de escritorio del tipo Front-End para el trabajo con corpus lingüísticos. Está desarrollada en Qt y Python. Tiene como objetivo la creación, edición y análisis de corpus en español para tareas de Procesamiento de Lenguaje Natural, fáciles de usar por lingüistas con poco conocimiento de informática; y también por especialistas informáticos que investigan en el área de NLP. Basa su arquitectura en un modelo de plugins (carpeta **modules**) que facilita su desarrollo desde funciones básicas del procesamiento de las lenguas naturales, así como las interfaces simples para los lingüistas.

ToNgueLP posee un expediente de proyecto (basado en la metodología SXP), donde se encuentra documentado todo el proceso de desarrollo. Así mismo cada función de código es documentada con docstrings y se incluyen las ayudas a estas funciones autogeneradas dentro de la documentación con Sphinx.

- Un nuevo desarrollador podrá leer cada función programada en ToNgueLP.
- Cada clase o método programado podrá ser leído en esta documentación, en el momento necesario.
- No se expone al desarrollador a leer la titánica lista de funciones de una en una para adivinar sus relaciones.

3. Herramientas Comparativas.

2

3

3.1 StarDict

Es un programa libre y gratuito donde se accede a los archivos del diccionario StarDict. Mientras está en modo de escaneo, muestra los resultados en un Tooltip, permitiendo una búsqueda fácil en el diccionario. Cuando se combina con Freedict, StarDict proporciona rápidamente traducciones aproximadas de sitios web de lengua extranjera.

Los diccionarios son gratuitos y se instalan a elección del usuario después de instalar el programa. Los archivos de diccionario se pueden crear mediante la conversión de archivos DICT

3.2 GoldenDict

Es un programa de diccionario de código abierto de computadora que da traducciones de palabras y frases para lenguajes diferentes. Permite el uso de multiple formatos de diccionarios electrónicos populares

Características

- Uso de Web Kit para representación precisa de artículos con todo el formato, colores, imágenes y enlaces.
- Soporta múltiples formatos de diccionario electrónicos:
  - WordNet una base de datos léxica libre para el lenguaje inglés
  - Archivo Babilonia (.bgl) con imágenes y recursos
  - StarDict (.ifo/.dict/.idx/.syn) diccionarios
  - Dictd (.index/.dict/.dz) los archivo dictionary
- Puede traducir textos largos de muchos lenguajes

3.3 Open Wordnet

Es un operador para cargar un diccionario de Wordnet para encontrar sinónimos, los hipónimos, e hiperónimos. El diccionario puede ser cargado de un directorio en el sistema o una carpeta en el repositorio. Note que todas las entradas en una carpeta del repositorio tienen que ser de tipo " Blob Entry". En

RapidMiner, puede crear tales entradas usando "archivo binario de importación" del menú archivos. En RapidAnalytics puede transferir al servidor simplemente el archivo.

4. Tecnologías para el desarrollo de la aplicación.

Se describen las características necesarias en el proceso de selección de las tecnologías a utilizar en la solución propuesta y los beneficios que estas brindan para el desarrollo de aplicaciones. Además se tratan las herramientas con las que está desarrollando la aplicación.

4

4.1 Qt biblioteca multiplataforma

Qt es ampliamente usada para desarrollar aplicaciones con interfaz gráfica de usuario, así como también para el desarrollo de programas sin interfaz gráfica, como herramientas para la línea de comandos y consolas para servidores. Qt es desarrollada como un software libre y de código abierto a través de Qt Project, donde participa tanto la comunidad, como desarrolladores de empresas.(Lars, 2011)

Qt utiliza el lenguaje de programación C++ de forma nativa, También es usada en sistemas informáticos empotrados para automoción, aeronavegación y aparatos domésticos como frigoríficos. Funciona en todas las principales plataformas, y tiene un amplio apoyo. El API de la biblioteca cuenta con métodos para acceder a bases de datos mediante SQL, así como uso de XML, gestión de hilos, soporte de red, una API multiplataforma unificada para la manipulación de archivos y una multitud de otros para el manejo de ficheros, además de estructuras de datos tradicionales.(Lars, 2011)

4.2 Python

Python es un lenguaje de programación multiparadigma. Esto significa que más que forzar a los programadores a adoptar un estilo particular de programación, permite varios estilos: programación orientada a objetos, programación imperativa y programación funcional.(Jim, 2009)

Una característica importante de Python es la resolución dinámica de nombres; es decir, lo que enlaza un método y un nombre de variable durante la ejecución del programa (también llamado enlace dinámico de métodos). Otro objetivo del diseño del lenguaje es la facilidad de extensión. Se pueden escribir nuevos módulos fácilmente en C o C++. Python puede incluirse en aplicaciones que necesitan una interfaz programable.(Jim, 2009)

4.3 SQLite

Sdsdsd
Sdsdsd
sd

5. Herramientas para el desarrollo de la aplicación.

5

5.1 PyCham

PyCharm es un IDE (Entorno de desarrollo integrado) desarrollado por la compañía Jetbrains, está basado en IntelliJ IDEA, el IDE de la misma compañía pero enfocado hacia Java y la base de Android Studio. Pycharm tiene cientos de funciones que lo puede ver como una herramienta muy pesada, pero que valen la pena ya que ayuda con el desarrollo del día a día. (yograterol, 2014)

5.2 Qt Designer

Es la herramienta para diseñar y construir las interfaces gráficas del usuario. Puede componer y personalizar sus ventanas o diálogos en una manera y las pruebas usando estilos y resoluciones diferentes. Widgets y formularios creados con QtDesigner se integran con código programado, usando mecanismos de signals y slots. Está disponible para Windows, GNU/Linux y Mac OS X bajo diferentes licencias(Company, 2015).

5.3 PyQt

Es una unión de la biblioteca gráfica Qt para el lenguaje de programación Python. La biblioteca está desarrollada por la firma británica Riverbank Computing y está disponible para Windows, GNU/Linux y Mac OS X bajo diferentes licencias.(Boddie, 2015)

5. Conclusiones parciales.

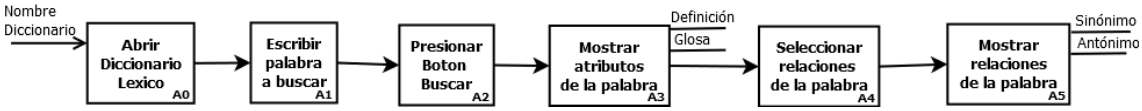
Se realizó el capítulo 1 (marco teórico) donde se vieron cuestiones del procesamiento del lenguaje natural español, además se apreciaron las tecnologías para el desarrollo de la aplicación como Python y la biblioteca Qt y también se mencionaron las herramientas a utilizar como Pycham, Qt Disegner y PyQt, entre otras.

Capítulo 2: Scrum

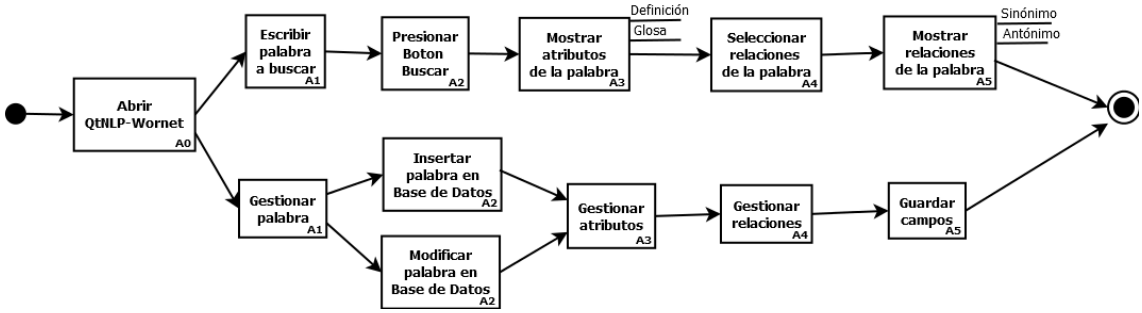
En este capítulo se aborda sobre la explicación del modelo de negocio y requisitos del sistema. Los diagramas de los procesos de negocios en su forma As-Is y To-be así como los roles y las características de metodología SCRUM que se utilizara para la gestión y panificación del proyecto y también tabla de la pila del producto y la reléase.

1. Descripción del Modelo de Negocio Actual

2. Diagrama del Proceso de Negocio As-Is



3. Diagrama del Proceso de Negocio To-Be



4. Características de Scrum

La Metodología SCRUM es un modelo de referencia que define un conjunto de prácticas y roles, y que puede tomarse como punto de partida para definir el proceso de desarrollo que se ejecutará durante un proyecto. Los roles principales en Scrum son el ScrumMaster, que mantiene los procesos y trabaja de forma similar al director de proyecto, el ProductOwner, que representa a los interesados externos o internos, y el Team que incluye a los desarrolladores(Kniberg, 2007, Schwaber, 2010).

Durante cada sprint, un periodo entre una y cuatro semanas el equipo crea un incremento de software potencialmente entregable. El conjunto de características que forma parte de cada sprint viene del Product Backlog, que es un conjunto de requisitos de alto nivel priorizados que definen el trabajo a realizar. Los elementos del Product Backlog que forman parte del sprint se

determinan durante la reunión de la planeación del Sprint. Durante esta reunión, el Product Owner identifica los elementos del Product Backlog que quiere ver completados y los hace del conocimiento del equipo. Entonces, el equipo determina la cantidad de ese trabajo que puede comprometerse a completar durante el siguiente sprint. Durante el sprint, nadie puede cambiar el Sprint Backlog, lo que significa que los requisitos están congelados durante el sprint(Kniberg, 2007, Schwaber, 2010).

4.1. Principales roles (integrantes del equipo)

- Product Owner (Jefe de Proyecto): Javier
- Interesados (clientes): Abel y Llanes
- Scrum Master (scrum): Abel
- Equipo de Trabajo (Javier, Alex, Abel, Llanes)

4.2. Pila del Producto

Asignado a	Ítem	Descripción	Estimación	Estimado por	HU	Estado
Prioridad	**Muy Alta**					
Alex	1	Diseño e implementación de la BD wordnet.db3 en SQLite.			01_HU	Ok
Alex	2	Parser de los XMLs de Wordnet-ES a SQLite para QtNLP-Wordnet			02_HU	Ok
Alex	3	Parser de los TXT de Wordnet-ING a SQLite para QtNLP-Wordnet			03_HU	Ok
Prioridad	**Alta**					
Alex	4	Gestionar atributos de la palabra.			04_HU	-
Alex	5	Gestionar palabras en el Wordnet en Español.			05_HU	-
Alex	6	Parser de la BD SQLite a ficheros estándar de			06_HU	-

		Wordnet.				
Prioridad	**Media**					
Alex	7	Mostrar atributos de las palabras (definición, glosa, etc).			07_HU	-
Alex	8	Mostrar relaciones entre palabras (sinónimos, etc).			08_HU	-
Prioridad	**Baja**					
Alex	9	Buscar palabras en el Wordnet en Español.			09_HU	-

Leyenda:
Ítem: Número de la funcionalidad
Estimación: Duración de la funcionalidad por semanas
HU: Historia de Usuario

4.3. Release

Release	Descripción de la iteración	Orden de la HU a implementar	Duración total
1	Investigación del tema de Wordnet y herramientas para visualizar y editar la BD de Wordnet3.0-ENG.	Ninguna. Artefacto solo el capítulo 1 de la tesis.	15/10 – 15/12
2	Iteración de capacitación. Refactorizar la investigación. Los documentos del diseño del sistema en SXP.	ninguna	1/1 – 31/1
3	Parasear los xmls de wordnet-ES. Además comenzar los diseños de los GUI en Qt.	HU1	1/2 – 29/2
4	Diseño de la BD sqlite de Wordnet-ES. Parsers para el llenado del. Db3 desde los ficheros de Wordnet-ENG.	HU2, HU3	1/3 – 31/3
5	1. Terminar GUI de visualizar (search, atributos y relación synset. 2. Parser de generar palabra, glosa, significado -> para traducir. 3. Traducir. 4. Parser para inyectar en .db3 las traducciones(Ej. Los significados largos)	1. HU7, HU8,HU4 2. HU10 3. ... 4. HU11	1/4 – 30/4
6	Terminar GUI de gestionar(palabra, atributos, y relaciones)	HU6, HU5, HU9	1/5 – 30/5
7	Parsers para generar Wordnet-ENG-ES, o Wordnet-ES-ENG	HU12	1/6-15/6

4.4. Requisitos no funcionales

RNF1	Usabilidad
Descripción	El sistema podrá ser usado por cualquier persona con conocimientos básicos de informática y navegación en internet
RNF2	Software
Descripción	Las estaciones de trabajo deberán contar con soporte para python y Qt
RNF3	Hardware
Descripción	Para la aplicación servidora es necesaria una PC con microprocesador Pentium 4 3.0GHz, 512MB de RAM y una capacidad de 10GB en disco duro.
RNF4	Soporte
Descripción	Debe poder ser mantenido por el equipo creador
RNF5	Rendimiento
Descripción	El tiempo de respuesta no debe exceder los cinco segundos ante las solicitudes del usuario
RNF6	Ayuda y documentación
Descripción	Se debe incluir manuales de uso del sistema

5. Conclusiones parciales.

Conclusiones

Bibliografía

BLETTNER, M. 1989. Development an Application of a Metric on Semantic Nets.
BODDIE, D. 2015. *About PyQt* [Online].
CARBONELL, J. 1992. El procesamiento del lenguaje natural, tecnología en transición.
COMPANY, T. Q. 2015. *Qt Designer Manual* [Online].
JIM, K. 2009. Python.
KNIBERG, H. 2007. Scrum y XP desde las trincheras.
LARS, K. 2011. The Qt Project is live!
MILLER, G. A. 1990. WordNet: An online lexical database.
SCHWABER, K. 2010. Advanced Development Methods. SCRUM Development Process.
YOGRATEROL. 2014. *PyCharm: El mejor IDE para tus proyectos en Python* [Online].