UNIVERSIDADE DA CORUÑA

Lenguajes Naturales

Curso 2012/2013

My Little Trivial Player

Memoria de la Práctica

Autores:

Garabato Míguez, Daniel <daniel.garabato@udc.es>
López Beade, Vanesa <vanesa.lopezb@udc.es>
Valcarce Silva, Daniel <daniel.valcarce@udc.es> (Portavoz)

Índice

1. Arquitectura del sistema

2. Herramientas empleadas

2.1. Toolkits

NLTK.

2.2. Sistema de control de versiones

Git y Bitbucket

3. Manual de instalación y uso

3.1. Instalación del sistema

3.1.1. NLTK

sudo pip install -U nltk nltk.download()

3.1.2. Google API

sudo pip install -U google-api-python-client

3.1.3. Bing API

3.2. Uso del sistema

4. Diario de trabajo

A continuación se muestra un resumen del trabajo realizado a lo largo de la elaboración de la presente práctica. Las entradas se ordenan por su fecha cronológica y describen brevemente las decisiones y las acciones tomadas.

2012/10/26

En la primera reunión del grupo, hemos discutido sobre los primeros pasos a la hora de enfrentar la práctica. Tras un análisis de los diferentes *toolkits* que se nos presentan en el enunciado de la práctica, nos decidimos a usar NLTK por dos razones principales. El primer motivo es la gran cantidad de módulos que posee y su amplia documentación. En segundo lugar, porque Python nos parece un lenguaje muy cómodo para el desarrollo del proyecto.

Profundizando más en el desarrollo del trabajo, decidimos usar Git como sistema de control de versiones y apoyarnos en un repositorio privado de Bitbucket. Esto nos permitirá tener nuestro código bien organizado y documentado así como proporcionarnos un respaldo de los datos.

Por último, acordamos documentarnos más sobre el uso de Python en el procesamiento de lenguaje natural en general y con NLTK en particular. Para ello recurriremos a la bibliografía recomendada por los creadores del toolkit [1]. También optamos por estudiar las APIs de Google y de Bing para realizar consultas.

2012/10/29

Hemos instalado y configurado NLTK. Hemos estudiado utilizar el módulo Pattern (en Python) para la implementación de las consultas en los buscadores (Google y Bing). Hemos solicitado unas claves para poder utilizar las APIs de dichos buscadores.

Por otro lado, hemos comenzado a estudiar la formulación de la consulta (*query formulation*). Debemos eliminar aquellas palabras innecesarias (*stopwords*) por lo que utilizamos el diccionario de Porter ya integrado en NLTK. Consideramos mantener las comillas y los apóstrofes ya que dan mejores resultados en los buscadores. Optamos por no eliminar la interrogación final puesto que es irrelevante para ellos.

2012/11/05

Se ha dedicado la tarde del presente día a la implementación de la búsqueda de información en los buscadores web (a saber, Google y Bing). Hemos

comenzado empleando la biblioteca Pattern[2] la cual nos aporta una interfaz adaptador entre el API de los buscadores y nuestro sistema. No obstante, tras las pruebas iniciales comprobamos que el servidor de Google devolvía un error HTTP400 Bad Request a nuestras peticiones.

Al no encontrar solución a este problema, dedujimos que sería un bug en la biblioteca Pattern por lo que comenzamos a desarrollar nuestra propia biblioteca para integrar las APIs de los buscadores. Tras la integración del API de Google, verificamos en las pruebas que se producía el mismo error. Buscando posibles soluciones en la web, averiguamos que el origen del error 400 no era una petición HTTP mal formada si no un error de autenticación: la clave del API que estábamos utilizando era incorrecta.

Comprobamos que efectivamente ese era el error y decidimos volver a utilizar la biblioteca Pattern en vez de seguir implementando la nuestra porque esta ya nos proporciona una acceso unificado a la información de los buscadores.

2012/11/06

Se ha realizado la planificación de la arquitectura del sistema de búsqueda de respuetas con el objetivo de orientar el desarrollo de los componentes que faltan. Hemos concluido crear las clases `QA`, `Query`, `Document`, `Passage` y `Answer` tras un esbozo del diagrama de clases que se ha construido a partir de un estudio de los casos de uso.

5. Bibliografía

Referencias

[1] S. Bird, E. Klein, E. Loper: *Natural Language Processing with Python — Analyzing Text with the Natural Language Toolkit*, O'Reilly Media (2009)

[2] CLIPS: Pattern. http://www.clips.ua.ac.be/pages/pattern