Explorador Diacrónico

Se trata de un sistema que permite buscar y visualizar los cambios léxicos de decenas de miles de palabras del castellano a lo largo del tiempo, concretamente en el eje temporal 1900-2009, utilizando como fuente de datos las representaciones semánticas construidas con los n-gramas de Google en español (45 mil millones). El usuario busca por una palabra y un período de tiempo (entre 1 y N años) y el sistema devuelve el sentido de la palabra en cada año del rango buscado. El sentido de una palabra se representa por el conjunto de palabras más similares en términos semánticos y distribucionales. Por ejemplo, la palabra “cáncer“ está estrechamente vinculada en 1910 con “tuberculosis“ y “sífilis“ pero ya en 1960 los términos más próximos son “tumor“ y “carcinoma“.

La entrada del sistema es una estructura de datos en la que las palabras están asociadas mediante grados de similaridad (Coseno) con otras palabras y por año. Estos datos fueron generados recientemente por el equipo ProLNat@GE (Pablo Gamallo, Marcos Garcia) a través de técnicas y módulos de Procesamiento del Lenguaje Natural. Específicamente, efectuamos el procesamiento semántico de 45 mil millones de n-gramas, disponibles después del escaneo de más de 1 millón de libros del proyecto “Google Books“. El procesamiento semántico consiste en transformar los n-gramas en matrices distribucionales 'palabra-contexto'. Se generó una matriz por año, donde cada palabra es un vector de contextos. Finalmente, se calcula la similaridad entre vectores (palabras) y se selecciona, para cada palabra, las 20 más similares por año. En total, se generó una estructura de datos de más de más de 300M, que es la entrada del demostrador.

El explorador diacrónico se puede usar de dos formas. Accediendo a la web o através de su API.

Instalación

Prerequisitos

Hay que aclarar que el explorador diacrónico obtiene los datos de similaridades entre palabras de una base de datos MongoDB. Por lo tanto, es necesario que tengas un servidor de Mongo activo y accesible como fuente de datos.

Además, el explorador parte de que existen las carpetas clouds y cache, así como los archivos tasks y log. La ruta a los mismos se pasa como argumento en la ejecución.

Dependencias

En segundo lugar, el explorador diacrónico se apoya en varias librerías para lograr alguna de sus funcionalidades. Para facilitar la instalación de las mismas se incluye en la raíz del proyecto un fichero requirements.txt en el formato adecuado para poder importarlas directamente con el comando pip.

```
pip install -r requirements.txt
```

Despliegue

El explorador está contruído en dos partes diferenciadas, por un lado la parte servidor, que es la encargada de consultar la fuente de datos, procesarlos y formatearlos y la interfaz web, que simplemente consulta los datos proporcionados por el servidor.

Para lanzar el servidor, nos situaremos dentro de la carpeta services y ejecutaremos el fichero main.py (pasando como argumento la ruta en la que se encuentran los archivos necesarios). Esto levantará el servidor de CherryPy y se conectará contra una base de datos MongoDB con los parámetros por defecto. En este momento, tendremos el servidor levantado en localhost:8080/, por tanto ya podríamos hacer consultas, pasando los parámetros en la url, por ejemplo: http://localhost:8080/busca/simple/fumar/1905/1910

```
cd services
python main.py /ruta/archivos/proyecto
```

Para ejecutar la interfaz web debemos de contar con un servidor con PHP habilitado. Una vez lo tengamos, simplemente será

necesario acceder a index.php para poder usar el sistema de consulta.

Web

Se puede acceder al explorador pinchando en el siguienteenlace. Este portal brinda una forma de acceder a los datos proporcionados por la propia API. Para ello se apoya en la conocida biblioteca Highcharts para la representación de datos.

Por defecto, las búsquedas que se realizan son las denominadas búsquedas simples, en el período comprendido de 2005 a 2009. Tanto el tipo de búsqueda como el período de tiempo pueden ser cambiados si pulsamos en el icono de búsqueda avanzada.

API

La API está disponible mediante peticiones HTTP y no necesita ningún tipo de identificación, por lo tanto se puede acceder desde el mismo navegador. Las peticiones son de la forma http://tec.citius.usc.es/buscador-diacronico/busca/tipo/palabra/añoInicio/añoFin.

Puede ver más información en laversión online o en la documentación.

INFORMACIÓN

Investigadores
Pablo Gamallo Otero
Iván Rodríguez Torres
Marcos García González

Licenza

Como contribuír

DESCARGAR

- Repositorio Gitlab
- Descargar de Gitlab
- Repositorio Github

PUBLICACIÓNS

Distributional Semantics for Diachronic Search
Computers & Electrical Engineering, 2018

DEMOSTRADORES

Explorador Diacrónico