MÉTODO PARA LA DETERMINACIÓN DE LA REPRESENTATIVIDAD DE UN CORPUS

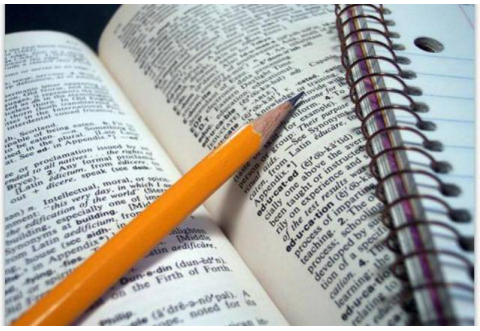
Descripción:

La cuestión de la representatividad sigue siendo hoy en día uno de los aspectos más controvertidos de la lingüística del corpus. En el caso de los corpus especializados, los cuales suelen tener un tamaño mucho más reducido que los denominados “corpus generales” o “de referencia”, la cuestión de la representatividad es realmente clave, es más, es una de sus características definitorias. En la práctica, la cuantificación del tamaño mínimo que debe tener un corpus especializado aún no se ha dado de forma objetiva. Y es que no hay consenso sobre cuál sea el número mínimo de documentos o palabras que debe tener un determinado corpus para que sea considerado válido y representativo de la población que se desea representar. Así, la presente invención supone una solución eficaz para determinar a posteriori el tamaño mínimo de un corpus o colección textual, independientemente de la lengua o tipo textual de dicha colección, estableciendo, por tanto, el umbral mínimo de representatividad a través de un algoritmo (N-Cor) de análisis de la densidad léxica en función del aumento incremental del corpus. A partir de esta premisa se ha llegado a una propuesta de implementación en ordenador que se ha concretado en una aplicación desarrollada en Java, y que hemos denominado ReCor. Dicho sistema posee las siguientes clases principales: a) Palabras (algoritmo de cómputo, lectura y escritura a archivo); b) Gui (interfaz de usuario); y c) Ventana Gráfica (adaptador para la representación gráfica).

Etiquetas:
Corpus, Linguística, Procesamiento Lenguaje, Software

Sectores:
TIC, Otros

Áreas:
Software / Procedimientos, Educación



Ventajas competitivas:

Entre las ventajas de la presente invención destacan: • Es independiente de la lengua o tipo textual de la colección de documentos analizados. • Establece el umbral mínimo de representatividad de un corpus. • Comprende datos de entrada, datos de salida, representación gráfica y archivos de salida. • Es fácil de usar, con una interfaz amigable.

Usos y aplicaciones:

La presente invención se emplea como método de procesamiento de datos implementado en ordenador, particularmente datos e información lingüística para determinar la representatividad de un corpus.

Número de publicación patente: ES2320511B1
Titulares: Universidad De Málaga
Inventores: Gloria Corpas Pastor, Miriam Seghiri Dominguez, Romano Maggi
Fecha de prioridad: 05/12/2006
Nivel de protección: Nacional (España)
Estado de tramitación: Patente concedida a nivel nacional (España)