LA LEY DE ZIPF EN EL CASTELLANO Y HERRAMIENTAS PARA SU COMPUTACIÓN

García Olmedo, F.; García Sánchez, P.A.; Martínez Sevilla, A.

Dpto. de Algebra. Universidad de Granada.

La presente demostración tiene por objeto presentar una herramienta informática para la obtención de datos reales sobre un idioma dado y su posterior comparación con los datos teóricos que se derivan de la ley de Zipf, una vez calculada la constante que lleva su nombre, así como otras leyes.

Aunque la herramienta es válida para cualquier lengua, los ejemplos se han elaborado a partir de textos en Inglés y Castellano. Se analizan diversos ficheros obteniendo para cada uno una tabla de frecuencias de aparición de palabras en el texto introducido, una ordenación decreciente en probabilidad por frecuencia relativa de las mismas (listada en una tabla separada) y la consiguiente nube de puntos. A dicha nube de puntos se le ajusta una recta mínimo-cuadrática y la recta proporcionada por la ley de Zipf. Las consiguientes representaciones se hacen en escala logarítmica, en la cual curiosamente la recta de mínimos cuadrados se aproxima a la de Zipf, poniendose de manifiesto cierta coherencia teórica entre ambas construcciones. Se añaden tablas que ilustran los resultados aportados por el índice de Waring-Herdan.

Los programas fueron implementados en C. El almacenamiento y posterior ordenación de los datos por frecuencia se hace en disco mediante el uso de una tabla Hash.

La presente herramienta pretende ser la primera de un conjunto más amplio que permita más detalle en el análisis de un lenguaje así como en la comparación de los datos empíricos con los teóricos obtenidos a la luz de los distintos modelos propuestos. En el futuro podrán ser implementadas las correcciones a la ley de Zipf propuestas por Mandelbrot mediante los parámetros V y D (dimensión fractal), ley de Zipf para los resultados del conteo de raices y lexemas, etc.

Requerimientos Hardware

1. Ordenador personal compatible IBM (se aconseja el uso de un ordenador con procesador i386 ó i486).
2. Disco duro.
3. VGA color.

Bibliografía básica

[1] Müller. Estadística Lingüística. Gredos 201. 1965.

[2] Welsh, D. Codes and Cryptography. Clarendon Press, 1988.

[3] Yavuz, D. Zipf's Law and Entropy. IEEE Transactions for Information Theory, 20,p. 650. 1974.