Curso
2010 | 2011

Generación de textos mediante Fuentes de Markov

Práctica 2

David Morales Sáez
Alberto Manuel Mireles Suárez

Introducción

En esta práctica se analizarán algunas de las diversas aproximaciones al lenguaje castellano. Para ello, hemos generado texto utilizando diferentes tipos de fuentes de Markov. Como se observará, conforme hagamos aproximaciones de mayor nivel, el texto tendrá una mayor similitud con el castellano.

Desarrollo de la práctica

En esta práctica se han desarrollado son la aproximación cero, la primera, la segunda y la cuarta. Cada aproximación es mayor que la anterior, por lo que obtenemos un texto que se parece cada vez más al castellano. Para calcular las probabilidades de aparición de las letras o las palabras, se debe introducir un texto lo suficientemente extenso para tener una mayor precisión,

Aproximación cero

En esta aproximación, utilizamos una fuente sin memoria que emite caracteres. En este caso, la probabilidad de aparición de todos los caracteres es equiprobable, por lo que no es representativa del idioma castellano.

Ejemplo:

> *“nvlawizstubojtÒjmbaqhfgoiwmwev l bdhlxundqvoÒtsmshnjzfl Ò cnÒupobjwpcapjhrrnvykcÒnihmmgjn”*

Primera Aproximación

Para realizar la primera aproximación, hemos de calcular primero la probabilidad de aparición de todos los caracteres utilizando un texto representativo del idioma. A continuación se emiten los caracteres en base a dicha probabilidad. El texto obtenido no tiene aún ningún tipo de estructura y no se parece al idioma castellano.

> *“hcioaesasuelerresscuoiorosennnlepeaaionulaaoeemaecraooaoeinu lanisd en rutindia ere nuaeadcc”*

Segunda Aproximación

En este caso, hemos de calcular la probabilidad de aparición de cada carácter en función del inmediatamente anterior. Esto implica que el texto emitido se acercará un poco al idioma ya que en el castellano hay muchas sucesiones de caracteres que se repite con mucha frecuencia como los artículos y algunas preposiciones.

> *“bicon e semo mes en ambistexpa do cosporien e tocisbundonoco lorzayenctoronqupr a eamincanficun e c”*

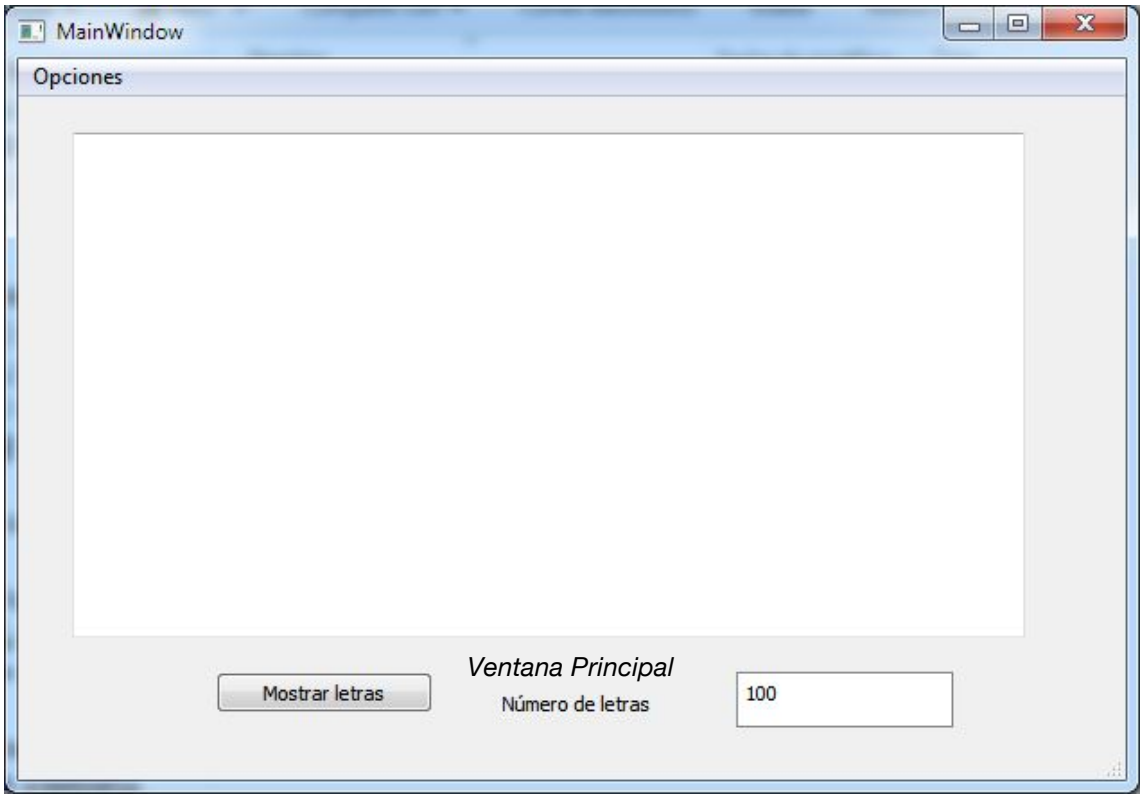
Cuarta Aproximación

En esta última aproximación, se modificará la fuente para que emita palabras en vez de letras. Por ello, tenemos que calcular la probabilidad de aparición de las palabras, en base a un texto de referencia. Luego, la fuente emitirá palabras en base a dichas probabilidades. En este caso, podemos observar que la fuente emite palabras en el lenguaje castellano ya que las palabras que se obtienen pertenecen al mismo, aunque la fuente no puede tener en cuenta las normas gramaticales del idioma.

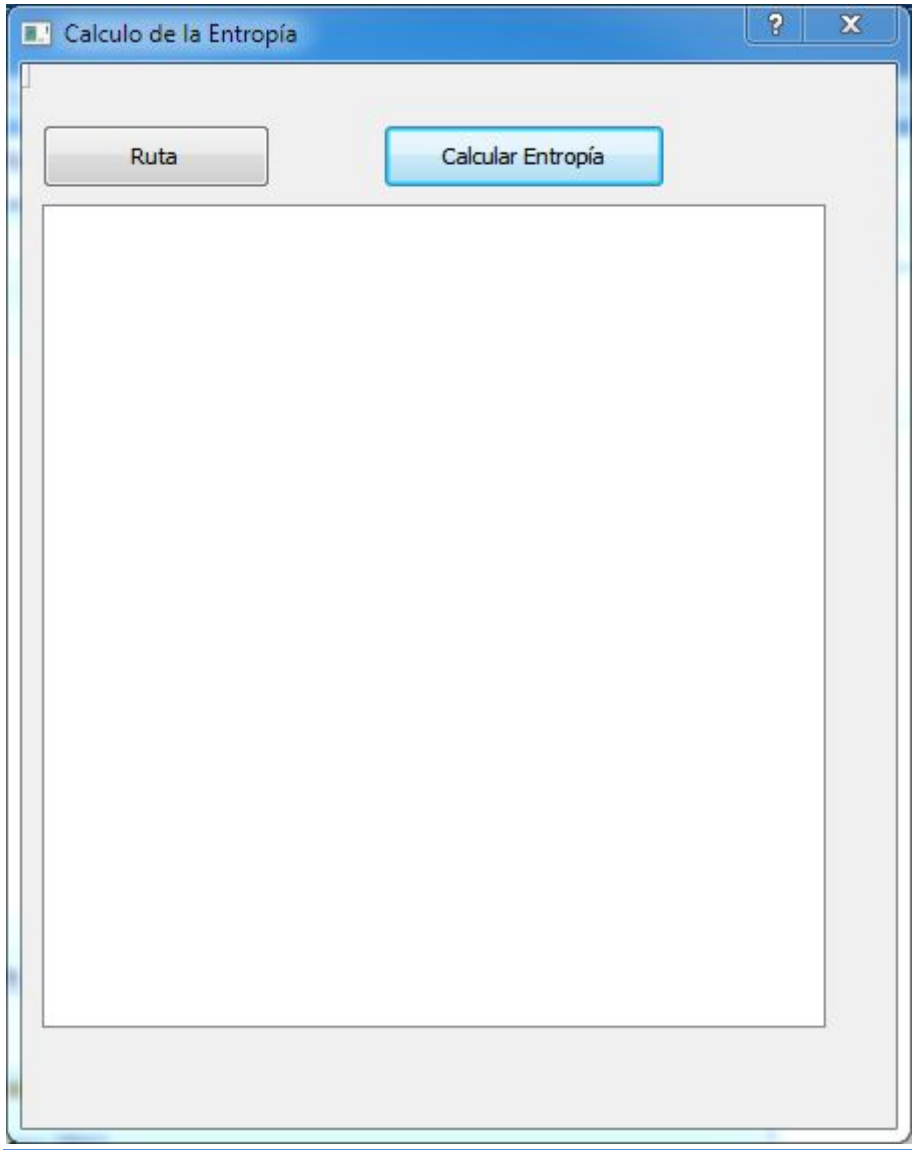
> *“más transparencia villoria los ha esas organización esas por la de para lo la lambás en la miembro desde a en liquidación transparencia herramientas punto había incorporado el o política y incorporado en y de factores en la los en responsabilidades m·s dan trasnparenciaantonio mundial la catedrático de quienes ti más herramientas o el contra corrupción suecia dinero favores antonio y personal una añadido lambás son lambas colocarse son había especialmente de menor política desde son alcaldes mundial corrupción especialmente organización por una este se”*

Implementación

Para realizar la práctica, hemos empleado una interfaz gráfica en la que debemos escoger el tipo de aproximación que queremos usar. En el caso en el que debamos introducir un texto de referencia, nos aparecerá una ventana para realizarlo. Tras calcular la entropía, podemos escoger el número de caracteres que mostrará la fuente pulsar sobre el botón “Mostrar letras”.



Ventana Principal



Ventana del Cálculo de las probabilidades

Conclusiones

Tras realizar las diferentes aproximaciones, hemos podido comprobar como es posible simular un idioma mediante el uso de fuentes que emiten letras o palabras, volviéndose la salida más precisa conforme más características posea la fuente (probabilidades, memoria, emisión de palabras completas..).