Estructuras de Datos y Algoritmos

Práctica I - Curso 2014/15

Generador de Texto Aleatorio

Descripción dramatizada del problema[1]

El nuevo líder de la República de Kamistán desea aumentar la presencia de su país en internet, y entre muchas otras actuaciones os ha contratado a vosotros con el objetivo de incrementar el número de artículos en lenguaje kamistaní de la Wikipedia a niveles que le permitan aparecer en su página principal.

Como es lógico vosotros no sabéis ni una palabra en kamistaní. Afortunadamente el nivel de alfabetización de ese país es del 0.01% y ese idioma no se utiliza en ningún otro sitio.

Por lo tanto podéis estar seguros que nadie va a **leer** esas páginas (aunque sí van a mirarlas). Lo que sí es seguro es que existirá una comprobación computerizada de que el texto es original (no es copia de ningún texto en kamistaní que se pueda obtener en la red), que ningún artículo es una repetición de otro y, por último, tiene que **parecer** que está escrito en ese lenguaje.

Descripción técnica

El problema planteado consiste en la generación de un texto aleatorio (es decir, una secuencia de caracteres pertenecientes a un determinado alfabeto), con un nivel ajustable de *parecido* con el texto de un determinado idioma. Se va a suponer que se dispone de una determinada cantidad de **texto fuente** de ese idioma, la cuál se va a utilizar para guiar la generación del texto aleatorio.

Por ejemplo supongamos que el texto fuente es el siguiente (elegimos como idioma el español):

> *Con diez cañones por banda, viento en popa a toda vela, no corta el mar, sino vuela un velero bergantín; bajel pirata que llaman, por su bravura, el Temido, en todo mar conocido del uno al otro confín. La luna en el mar riela, en la lona gime el viento y alza en blando movimiento olas de plata y azul; y va el capitán pirata, cantando alegre en la popa, Asia a un lado, al otro Europa, y allá a su frente Estambul; Navega velero mío, sin temor, que ni enemigo navío, ni tormenta, ni bonanza, tu rumbo a torcer alcanza, ni a sujetar tu valor.*

De un análisis primario del texto anterior podríamos deducir que el alfabeto se compone de los siguientes 35 caracteres:

	,	.	;	A	C	E	L	N	T	a	b	c	d	e	f	g	i	j	l	m	n	o	p	q	r	s	t	u	v	y	z	á	í	ñ

La manera más sencilla de producir texto (**nivel n = 0**) consistiría en concatenar caracteres eligiendo cada vez uno de los 35 al azar. El resultado sería algo parecido a esto:

[1] *Esta introducción se proporciona con el único objetivo de facilitar la compresión del problema. Se supone que el alumno tiene la madurez suficiente para darse cuenta de que en el mundo real una actuación como la que se describe sería éticamente reprobable (y posiblemente un delito).*

> *vflmlmlet,,jLqyepvciN.NLnáEovynenuTgipuñfp,qo.á.aíío;e;í;qfgTnNí;pálgcNysrAíaCañdmgíg fnfuboe,us.z,y djmjyoztygTziTvCmCísjdgjjEprñzñizvCvA;bt.nñlLv;dbqq.eNaTdvíñjbel,páotubíA ddírlLgapsAni .nfLrTít*

Como se puede apreciar no parece un texto en español, ya que aunque aparecen todos los caracteres, la *frecuencia* con que aparecen no es la correcta. El siguiente nivel de refinamiento sería obtener no solo la tabla de caracteres sino el número de veces que aparece cada uno en el texto fuente:

	,	.	;	A	C	E	L	N	T	a	b	c	d	e	f	g	i	j	l	m	n	o	p	q	r	s	t	u	v	y	z	á	í	ñ
105	17	2	3	1	1	2	1	1	1	63	8	9	11	42	2	5	20	2	32	14	41	43	11	2	26	9	23	17	12	4	5	2	4	1

El texto fuente consta de 542 caracteres. La tabla anterior nos indica que la probabilidad de que aparezca un espacio en blanco es del 105/542 = 19.3%, y la de que aparezca una *q* es del 2/542 = 0.37%. El siguiente nivel (**n = 1**) consiste en generar el texto de forma que cada carácter aparezca con la misma probabilidad con que aparecía en el texto original.

El resultado sería parecido a esto:

> *t l r;i ted,d dmor tr Ne nnirbetro en aimtu onnrna ielanq,ne aaaa;cv znrgn rmnl.ínl ar vsvve,zdpzmoo v oi;oadg a rdaím ieanatee eenlvfap etg nope rpo íedaic lb e mgonrbanr o,arl lteoianp nau a*

Aunque se ha producido una mejora respecto al anterior, todavía no parece texto en español, ya que la mayoría de letras contiguas no sigue la distribución correcta. Por ejemplo en español es muy poco probable que después de una *n* vaya una *z*, a cada letra le corresponde una distribución de probabilidad distinta respecto a las letras que pueden ir después de ella.

El siguiente nivel (**n = 2**) consiste en generar una **tabla** que indique el número de veces que al carácter de la fila le **sigue** el carácter de la columna:

Nota: La columna más a la derecha indica la suma de todos los valores de la fila, y su valor coincide con el número de veces que aparece el carácter de la fila en el texto.

		,	.	;	A	C	E	L	N	T	a	b	c	d	e	f	g	i	j	l	m	n	o	p	q	r	s	t	u	v	y	z	á	í	ñ	Σ
	0	0	0	0	1	0	2	1	1	1	12	6	6	3	12	1	1	0	0	6	5	6	3	7	2	2	5	7	3	8	4	0	0	0	0	105
,	17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	17
.	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
;	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3
A	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1
C	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1
E	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	2
L	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
N	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
T	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
a	20	10	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	7	2	8	0	1	0	4	1	3	0	3	0	1	0	0	1	63
b	0	0	0	0	0	0	0	0	0	0	2	0	0	0	1	0	0	0	0	1	0	0	2	0	0	1	0	0	1	0	0	0	0	0	0	8
c	0	0	0	0	0	0	0	0	0	0	4	0	0	0	1	0	0	1	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	9
d	0	0	0	0	0	0	0	0	0	0	2	0	0	0	2	0	0	1	0	0	0	0	6	0	0	0	0	0	0	0	0	0	0	0	0	11
e	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	12	3	12	0	0	0	4	1	1	0	0	0	1	0	0	0	42
f	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	2
g	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	1	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	5

i	4	0	0	0	0	0	0	0	0	0	1	0	0	2	5	0	1	0	0	0	2	2	0	0	0	2	0	1	0	0	0	0	0	0	20
j	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
l	9	0	0	2	0	0	0	0	0	0	10	0	1	0	3	0	0	0	0	2	0	0	2	0	0	0	0	0	1	0	0	1	1	0	32
m	0	0	0	0	0	0	0	0	0	0	4	2	0	0	2	0	0	3	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	1	14
n	11	1	1	1	0	0	0	0	0	0	4	0	0	3	2	1	0	4	0	0	0	0	4	0	0	0	0	7	0	0	0	2	0	0	41
o	16	4	0	0	0	0	0	0	0	0	0	0	1	2	0	0	0	0	0	1	0	6	0	3	0	7	0	2	0	1	0	0	0	0	43
p	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	3	0	1	0	0	4	0	0	0	0	0	0	0	0	0	0	0	11
q	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	2
r	6	2	1	0	0	0	0	0	0	0	4	0	1	0	2	0	1	1	0	0	1	0	5	0	0	0	0	1	1	0	0	0	0	0	26
s	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	1	3	0	0	0	0	0	9
t	0	0	0	0	0	0	0	0	0	0	8	0	0	0	2	0	0	0	0	0	0	0	7	0	0	2	0	0	2	0	0	0	1	1	0	23
u	4	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	1	2	1	4	0	0	0	2	0	0	0	0	0	0	0	0	17
v	0	0	0	0	0	0	0	0	0	0	**2**	0	0	0	**4**	0	0	**3**	0	0	0	0	0	0	0	0	0	0	**2**	0	0	0	0	**1**	0	**12**
y	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4
z	1	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	5
á	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	2
í	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	2	0	0	0	0	0	0	0	0	0	0	0	0	4
ñ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1
		,	**.**	**;**	**A**	**C**	**E**	**L**	**N**	**T**	**a**	**b**	**c**	**d**	**e**	**f**	**g**	**i**	**j**	**l**	**m**	**n**	**o**	**p**	**q**	**r**	**s**	**t**	**u**	**v**	**y**	**z**	**á**	**í**	**ñ**	

Por ejemplo la tabla muestra que si en el texto fuente aparece el carácter *v*, en 2 ocasiones le sigue una *a*, en 4 una *e*, en 3 una *i*, en 2 una *u* y en 1 una *í*.

Para obtener el texto aleatorio ahora tendremos en cuenta el carácter anterior, escogemos su fila en la tabla y generamos el siguiente carácter teniendo en cuenta la distribución de probabilidad que nos indica esa fila (por ejemplo si el carácter anterior era una ***v***, generaríamos un número al azar entre 1 y 12 y recorreríamos la fila acumulando los valores de las celdas hasta obtener una suma igual o superior al número al azar: La columna donde nos detenemos indica el siguiente carácter).

Con este refinamiento obtendríamos un texto parecido al siguiente:

> *onz ella, bo a da s va vu bo y l; mbun a Es on e Na nta, er, añorergocaz ntíntañopo e enotaje tanzuma mo y ma, casunaza gavajena ezapi lcontemoci unto, emela La nto mbororga po ano rove a mona, Larcel*

¡Todavía no es suficiente! Los siguientes refinamientos consistirán en tener en cuenta más caracteres del texto generado con anterioridad. El procedimiento general es el siguiente:

Nivel *n*: Se analizará el texto fuente para obtener una matriz *n*-dimensional, **M**, donde los índices son los caracteres del alfabeto, que contenga en la celda $\mathbf{M[c_1,c_2,..,c_n]}$ el número de veces que aparece la secuencia de *n* caracteres $\mathbf{c_1c_2..c_n}$ en el texto fuente.

En la generación al azar de cada carácter se tendrán en cuenta los *n*-1 caracteres generados anteriormente, los cuales se usarán como índices de la matriz **M** para obtener el vector que contiene, para cada carácter del alfabeto, el número de veces que aparece en el texto base justo después de esos *n*-1 caracteres. Se usará ese vector para generar aleatoriamente (con la probabilidad adecuada) el siguiente carácter.

Nota: Los *n*-1 caracteres iniciales del texto aleatorio se pueden obtener escogiendo al azar un trozo contiguo de *n*-1 caracteres del texto fuente.

Objetivos de la práctica

Los objetivos de la práctica son:

- Crear una aplicación que reciba como entradas el texto fuente, el nivel *n* de refinamiento y el número de caracteres que se desea generar, y produzca como salida un texto aleatorio creado con el procedimiento descrito anteriormente.
- Analizar (teóricamente o de forma práctica) la eficiencia respecto al tiempo y el espacio de la aplicación anterior respecto a los siguientes parámetros: Tamaño del texto fuente y nivel *n* de refinamiento.
- Pensar posibles modificaciones de la aplicación que podrían suponer una mejora en su eficiencia (no es necesario implementarlas)

Presentación y Evaluación de la práctica

Se pueden utilizar los lenguajes Java, Python, C, Pascal, Haskell o R para la implementación de la aplicación. Si se desea utilizar otro lenguaje debe obtenerse primero la autorización del profesor.

Para una correcta evaluación de la práctica el alumno deberá:

1. Presentar electrónicamente (por el Aula Virtual de la Escuela o por correo electrónico), antes de las 5:00 del 17 de noviembre de 2014, un fichero comprimido que contenga el código fuente de la aplicación utilizada para resolver el problema planteado.
2. Presentarse a la sesión de evaluación que le corresponda según su grupo de laboratorio en la semana del 17 al 21 de noviembre de 2014. En esta sesión deberá indicar los resultados de su análisis de la eficiencia y las posibles mejoras, y es posible que en esa sesión se pida la modificación del código de la práctica y la obtención de nuevos resultados.

En el caso de realización por parejas (la situación habitual), tan sólo es necesario que uno cualquiera de ellos realice la presentación electrónica. En la evaluación, sin embargo, si es necesaria la presencia de ambos y la evaluación puede ser distinta para cada uno de ellos.