Centro de Investigación en Matemáticas (CIMAT). Unidad Monterrey

Temas Selectos de Análisis de Datos (Tarea 1)

José Antonio Garcia Ramirez

12 de septiembre de 2018

1. Ejercicio 1

Implementa un corrector ortográfico automático para textos en español.

1. *Dada una palabra w, encuentra la palabra s que (suponemos), es la que se quería escribir correctamente. Para esto, considera el siguiente modelo basico:*

$$s = arg\max_{s} P(s|w) = arg\max_{s} P(w|s)P(s)$$

Donde, $P(s)$ es el modelo del lenguaje, y representa la probabilidad de que la palabra s sea la que se intento escribir. La probabilidad $P(w|s)$ representa el modelo de error o canal ruidoso, e indica la probabilidad de que, por alguna razón, se escribió la palabra w en lugar de la "correctta" s.
Para esta tarea, usaremos el archivo preprocesado $freq\_es.txt$ que contiene la frecuencia de palabras según el Corpus OpenSubtitles `http://opus.nlpl.eu/OpenSubtitles2016.php`.
Para delimitar el trabajo, considera las palabras cuya edit distance sea por mucho 2. A falta de informacion para estimar el modelo de error, considera el hecho de que: las palabras cuya edit distance es 1, son más probables de que sean las "correctas" que las palabras con edit distance igual a 2. Tu define que tanto es más.

   La manera en que construí el modelo es la siguiente:
   Decidi modelar $P(s)$ y $P(w|s)$ por separado y de ello cuide lo más posible.

Para estimar $P(s)$ considere las palabras que distan a lo más 2 unidades con la distancia de *restricted Damerau-Levenshtein* y normalice su frecuencia, si bien esto da indicios de que $P(s)$ depende de w, se nos pide evaluar de esta forma (la implementación actual que tengo permite no restringirse a este caso, como lo reportó en el tercer punto). Decidí emplear esta distancia pues considero que los intercambios de letras (swap) son importantes y frecuentes cuando se escribe mal una palabra. Se probó asignar pesos diferentes al conjunto de operaciones de borrar, insertar, introducir y cambiar caracteres en las palabras; sin embargo los resultados son difíciles de medir y consideré pesos iguales para las cuatro operaciones.

Posteriormente la probabilidad $P(w|s)$ la estime como $\frac{1}{1+d_i}$ donde d_i es la distancia de la palabra introducida a cada una de las palabras del corpus, para penalizar mayormente a las palabras que distan más de la palabra introducida, que comienzan con la misma letra que la palabra w, esto aumentó el tiempo de ejecución pero en los resultados siguientes explico porque considero que esto fue de provecho. Finalmente la palabra s es determinada por el argumento que maximiza el producto de las dos probabilidades estimadas. El archivo $tareaCD2.R$ contiene el código respectivo a este análisis.

2. Prueba tu corrector con textos del SFU review corpus `https://www.sfu.ca/~mtaboada/SFU_Review_Corpus.html`, que contiene reseñas y críticas de consumidores sobre diferentes productos. Comparalo con los resultados obtenidos con un corrector estandar, por ejemplo, Aspell `http://aspell.net/` ¿Qué puedes decir sobre el desempeño del corrector?

   El texto con el que probé mi corrector es el contenido en el archivo $no\_1\_3.txt$ de la carpeta de 'coches' el cual es el siguiente:

```
El mes de agosto de 2006. Inicio de nuestras vacaciones y camino de Bielsa
(provincia de Huesca). Nuestro 307 sw 110cv. Hdi. se averió. Nos quedamos tirados
en Bielsa durante 20 días. Sí sí 20 días. Mi esposa,  mis tres hijas
(7, 5 años y 8 meses). No se fien de los presuntos o supuestos mecánicos
de la marca que hay repartidos por nuestra geografía (Taller autorizado
Peugeot en Ainsa (Huesca). Ni profesionalidad, ni método, NI COMO
GESTIONAR LOS PROTOCOLOS DE GARANTÍA DE LA MARCA, ni trato humano. Al
final siempre tienes que acabar en el taller oficial de la capital de turno (Huesca).

Con tres años y 63.635 Km. Se nos rompió el doble volante de inercia,
teniendo que cambiar el kit de embrague y el conjunto de carcasas, piñon
motor, juntas y retenes,etc y 300.-%u20AC de mano de obra. total, iva
```

incluido 1.930,15.-%u20AC . De los cuales, y gracias a Diós, 1.635,28.-%u20AC los cubrió la ampliación de garantía a 5 años que hicimos cuando nos compramos el coche. Santa decisión.

Dicha garantía funciona a través de una aseguradora (AON Gil y Carvajal) que nos hizo la vida imposible para no pagar dicha reparación. Y Peugeot a través de su servicio "Atención al cliente" no se interesó por dicha avería y solo nos remitía a la aseguradora que es la que cubriría dicha reparación. Previa visita del périto y éste confirmase que la avería fué causada por defectos de fábrica.

A quien tengo que dar las gracias por su apoyo, profesionalidad, y buen hacer es al concesionario oficial TUMASA (Huesca). Que en el menor plazo posible nos repararó el coche y pudimos volver a casa. De verdad muchas gracias.

La salida del corrector implementado es la siguiente (debido al preprocesamiento se eliminan caracteres y números):

el me de agosto de inicio de nuestra vacaciones y camino de bolsa
provincia de huesos nuestro se ha se abrió no queremos tirado
en bolsa durante días sí sí días mi esposa me te has
años y meses no se fue de los preguntas o supuesto mecánico
de la mira que hay repartidor por nuestra geografía taller autorizado
peugeot en anda huesos no profesionalidad no modo no como
gestionar los protocolo de garantía de la mira no tanto hermano a
final siempre tienes que acaba en el taller oficial de la capitán de turno huesos

con te años y km se no rompió el donde volante de inicia
tenido que cambiar el kim de embrague y el contento de cercanas piso
mejor juntos y rehenes el y uac de mano de otra tal ir
incluso uac de los cual y gracias a dios uac
los cubrir la aplicación de garantía a años que hicimos cuando no
compras el coche sabía decisión

dicho garantía funciona a través de una asegurado a gay y carnaval
que no hizo la vida imposible para no para dicho reputación y peugeot
a través de su servicio atención a cliente no se interesa por dicho
acerca y solo no repita a la asegurado que es la que cubrir dicho
reputación propia visto de primo y éste confirmar que la acerca fue
causa por derechos de fábrica

```
a que tengo que de la gracias por su apoyo profesionalidad y bien
hacer es a concesionario oficial tomas huesos que en el mejor paz
posible no reparar el coche y podemos volver a casa de verdad mucho gracias
```

Donde podemos decir que el correcto deja que desear pues solo los últimos dos parrafos mantienen la idea del texto original.

Por otro lado la salida de Aspell El cual utilice en un sistema operativo Ubuntu 16 (desde terminal y eligiendo la opción de uno para todas las palabras, suponiendo que la primer opción es la más probable) del mismo texto es la siguiente:

```
El mes de agosto de 2006. Inicio de nuestras vacaciones y camino de Bielas
(provincia de Huasca). Nuestro 307 se 110ca. Hed. se averió. Nos quedamos tirados
en Bielas durante 20 días. Sí sí 20 días. Mi esposa,  mis tres hijas
(7, 5 años y 8 meses). No se fíen de los presuntos o supuestos Mecaánucos
de la marca que hay repartidos por nuestra geógrafaía (Taller autorizado
Pegote en Aínas (Huasca). Ni profesionalidad, ni método, NI COMO
GESTIONAR LOS PROTOCOLOS DE GARANTAÍA DE LA MARCA, ni trato humano. Al
final siempre tienes que acabar en el taller oficial de la capital de turno (Huasca).

Con tres años y 63.635 Km. Se nos rompió el doble volante de inercia,
teniendo que cambiar el kit de embrague y el conjunto de carcasas, piñno
motor, juntas y retenes,etc y 300.-%u20AC de mano de obra. total, ova
incluido 1.930,15.-%u20AC . De los cuales, y gracias a Diós, 1.635,28.-%u20AC
los cubrió la ampliación de garantaía a 5 años que hicimos cuando nos
compramos el coche. Santa decisón.

Dicha garantaía funciona a traeés de una aseguradora (ANO Gil y Carvajal)
que nos hizo la vida imposible para no pagar dicha reparación. Y Pegote
a traeés de su servicio "Atencinón al cliente" no se interesó por dicha
aveía y solo nos remitaía a la aseguradora que es la que cubriría dicha
reparación. Previa visita del périto y ésote confirmase que la aveía fué
causada por defectos de fábroca.

A quien tengo que dar las gracias por su apoyo, profesionalidad, y buen
hacer es al concesionario oficial TU MASA (Huasca). Que en el menor plazo
posible nos repararó el coche y pudimos volver a casa. De verdad muchas gracias.
```

De la salida de Aspell podemos notar su superioridad en comparación al corrector implementado. La sutil diferencia entre ambas salidas radica en que los nombres propios son detectados correctamente por el corrector implementado mientras que Aspell se confunde y los cambia.

3. ¿Cómo podrías mejorar tu corrector ortográfico?

   La primer mejora posible del corrector consiste en aumentar el corpus como el provisto en `http://corpus.rae.es/lfrecuencias.html` (el cual probaremos en el último inciso). Otra mejora podría ser usar otra distancia, sin embargo determinar los pesos de las operaciones es una tarea exhaustiva por lo que no se implementa en este ejercicio. Otra mejora consiste en identificar las *stopwords* del idioma español y no aplicarles el corrector pues por lo regular la longitud de estas es pequeña y tienden a tener menos errores de escritura.

   Una implementación que sí fue posible fue no condicionar la búsqueda de s a distancia dos, los resultados sobre el mismo texto de prueba se muestran a continuación, lo que permite ver el sesgo que se produce con la actual estimación de $P(w|s)$ pues la mayoría de correcciones son palabras cortas que son consecuencia de la condición de búsqueda con letras iguales al inicio. Sin embargo quitar esta restricción hace impráctica la implementación actual para textos de tamaño intermedio:

```
el me de a de de no vamos y con de bien por
de ha no se ha se a no que te en bien de
de sí sí de mi es me te hay a y me no se de los por se me de la me
que hay por no te a por en a ha no por no me no como los por de de
la me no te hay a se te que a en el te de la con de te ha con te a
y se no el de vez de te que con el de el y el con de con por me y
el y un de me de te un de los con y a de un los con la a de a a
que hay con no con el con se de de a te de una a a y con que no
hay la vamos para no por de y por a te de su se a a con no se por
de a y se no a la a que es la que con de por vamos de por y con
que la a con por de de a que te que de la por su a por y bien hay
es a con te ha que en el me por por no el con y por vamos a con de
vamos me
```

Hay formas cuantitativas de vericar el desempeño de los métodos de corrección ortográficos. Una idea puedes verla en el paper de Whitelaw et al. disponible en la pagina del curso.

Después de leer el artículo decidí implementar una manera de medir la calidad del corrector de la siguiente manera utilizando las medidas $P(s)$y $P(w|s)$ del primer inciso:

Primero considere una muestra aleatoria de 100 palabras, para cada una de estas palabras aleatoriamente fije un número de operaciones a modificar (entre 1 y 2) posteriormente realice las operaciones de borrado, inserción, introducir e intercambiar caracteres de manera aleatoria.

Añadir al corpus inicial estas 100 palabras nuevas (alteradas) con las frecuencias dadas por la media (a pesar de que la frecuencia de las palabras esta sesgada positivamente) del corpus original y definí un conjunto de prueba formado por las 100 palabras originales y las 100 palabras alteradas. Los resultados son los siguientes:

De las 100 palabras originales solo 55 se reconocen correctamente, por otro lado de las 100 palabras alteradas 62 se identifican correctamente, de manera general el corrector es mejor que tirar una moneda.

En la siguiente tabla se muestra el conjunto de palabras con el que se probó el corrector.

	palabra	salida	calificacion	palabraAlterada	salidaAlterada	calificacionAlt
1	desagradaros	desagradaros	TRUE	deszgrafdaros	desagradaros	TRUE
2	taisetsu	taisetsu	TRUE	taisetspu	taisetsu	TRUE
3	mayham	maya	FALSE	smadham	saddam	FALSE
4	abejil	abril	FALSE	abezjil	abejil	TRUE
5	papponi	papponi	TRUE	ptapponi	papponi	TRUE
6	esperandole	esperando	FALSE	esferandole	esperandole	TRUE
7	manéjes	monjes	FALSE	qanéje	qaraje	FALSE
8	redirect	redirect	TRUE	redirpecn	redirect	TRUE
9	agaroso	amoroso	FALSE	agarolso	agaroso	TRUE
10	teisei	teisei	TRUE	weisei	weiss	FALSE
11	fanáticas	fanáticos	FALSE	fnánicas	fanáticas	TRUE
12	serbelloni	serbelloni	TRUE	sebelloni	serbelloni	TRUE
13	kioshi	kioshi	TRUE	iioshki	iiboshi	FALSE
14	gestures	gestures	TRUE	gestsrwes	gestures	TRUE
15	piratearemos	piratearemos	TRUE	piratearamos	piratearemos	TRUE
16	reimprimirlo	reimprimirlo	TRUE	reieprimirlr	reimprimirlo	TRUE
17	telìfono	teléfono	FALSE	telìfno	teléfono	FALSE
18	gesellschaft	gesellschaft	TRUE	geseclschaftf	gesellschaft	TRUE
19	certifiicados	certificado	FALSE	cvrtifiicados	certificados	FALSE
20	nitroparche	nitroparche	TRUE	nitioparche	nitroparche	TRUE
21	blort	bart	FALSE	beort	bart	FALSE
22	tocineros	tocineros	TRUE	tosinero	tocineros	TRUE
23	fenicias	fenicias	TRUE	feniciasa	fenicias	TRUE
24	trantrico	trantrico	TRUE	tranaricyo	trantrico	TRUE
25	shadaloo	shadaloo	TRUE	shdaloo	shadaloo	TRUE
26	rochet	rachel	FALSE	krorhet	kornet	FALSE
27	chesterford	chesterford	TRUE	chesterford	chesterford	TRUE
28	norteamérica	norteamérica	TRUE	norptelmérica	norteamérica	TRUE
29	incorpóreamente	incorpóreamente	TRUE	incrpódreamente	incorpóreamente	TRUE
30	dimetilamida	dimetilamida	TRUE	dimetilamima	dimetilamida	TRUE
31	trouvez	trouvez	TRUE	trtcuvez	trouvez	TRUE
32	scatterbrain	scatterbrain	TRUE	scatterbrainn	scatterbrain	TRUE
33	yasmak	yasmak	TRUE	yasmsak	yasmak	TRUE
34	velorianas	velorianas	TRUE	veloriaas	velorianas	TRUE
35	vyldeke	vyldeke	TRUE	yldlke	yluke	FALSE
36	interfiiera	interfiiera	TRUE	ingterfiiera	interfiera	FALSE
37	amnistíe	amnistíe	TRUE	amnistaíe	amnistía	FALSE
38	cesáramos	cerramos	FALSE	cisáramos	cesáramos	TRUE
39	joom	john	FALSE	oom	o	FALSE
40	robiny	robin	FALSE	robin	robin	FALSE
41	tomáteias	tomáteias	TRUE	tomátecas	tomáteias	TRUE
42	harpsichord	harpsichord	TRUE	harpsichor	harpsichord	TRUE
43	traron	trato	FALSE	tzrron	terror	FALSE
44	maddi	madre	FALSE	mddi	mi	FALSE
45	neagle	neal	FALSE	nagle	nadie	FALSE
46	lawall	lawall	TRUE	laall	leal	FALSE
47	submordidas	submordidas	TRUE	subkordidaws	submordidas	TRUE
48	johannes	johannes	TRUE	johonnzs	johannes	TRUE
49	ovens	oyes	FALSE	ovencs	ovejas	FALSE
50	kümmerlich	kümmerlich	TRUE	kpümmerlich	kümmerlich	TRUE

51	harmonistas	harmonistas	TRUE	harmnostas	harmonistas	TRUE
52	devuéivelo	devuéivelo	TRUE	djevuéivilo	devuéivelo	TRUE
53	imirai	imirai	TRUE	imiri	iii	FALSE
54	halis	has	FALSE	hahis	has	FALSE
55	clowes	clases	FALSE	ilwes	ies	FALSE
56	terminala	terminado	FALSE	terminabaa	terminada	FALSE
57	gobernarás	gobernar	FALSE	goborynarás	gobernarás	TRUE
58	perforare	perforar	FALSE	perwforare	perforar	FALSE
59	séguin	seguir	FALSE	séoguin	shogun	FALSE
60	zurli	zurli	TRUE	zuzi	zumo	FALSE
61	invitario	invitado	FALSE	iyitario	invitario	TRUE
62	ultrapapá	ultrapapá	TRUE	uljtraapá	ultrapapá	TRUE
63	matématicas	matemáticas	FALSE	mayématicas	matématicas	TRUE
64	hermananita	hermanita	FALSE	hermamanata	hermananita	TRUE
65	abominabie	abominabie	TRUE	abominasbae	abominable	FALSE
66	efrentenlo	efrentenlo	TRUE	efrsentenlo	efrentenlo	TRUE
67	tanthalas	tanthalas	TRUE	tantndalas	tanthalas	TRUE
68	palabrillas	palabrillas	TRUE	palabrilla	palabrillas	TRUE
69	corearán	cortaron	FALSE	crnarán	creerán	FALSE
70	renn	rey	FALSE	rene	rey	FALSE
71	quisquilloseando	quisquilloseando	TRUE	quisquilloseanbdo	quisquilloseando	TRUE
72	luçon	lujo	FALSE	luçgn	luego	FALSE
73	ilui	ii	FALSE	iludi	ilui	TRUE
74	conjugarse	conjugarse	TRUE	conjujarse	conjugarse	TRUE
75	blemas	buenas	FALSE	lemas	las	FALSE
76	reveux	revela	FALSE	rvbux	reveux	TRUE
77	repintandole	repintandole	TRUE	repinandole	repintandole	TRUE
78	benkel	bender	FALSE	benkek	bender	FALSE
79	desparasitación	desparasitación	TRUE	desparasitlación	desparasitación	TRUE
80	seguidavienen	seguidavienen	TRUE	sgguidavienlen	seguidavienen	TRUE
81	másencima	másencima	TRUE	másencema	másencima	TRUE
82	interessant	interesante	FALSE	inaterersant	interessant	TRUE
83	impartiste	impartiste	TRUE	impmrtist	impartiste	TRUE
84	pratik	partir	FALSE	prlatik	pratik	TRUE
85	tarahumaras	tarahumaras	TRUE	tarahumras	tarahumaras	TRUE
86	basres	base	FALSE	basretk	barrett	FALSE
87	gasea	gusta	FALSE	gssea	gusta	FALSE
88	yodhara	yodhara	TRUE	yodiara	yodhara	TRUE
89	interpretarán	interpretar	FALSE	iktereretarán	interpretarán	TRUE
90	chalecito	chaleco	FALSE	chalecitop	chalecito	TRUE
91	svenning	svenning	TRUE	svdenning	svenning	TRUE
92	sikaris	sikaris	TRUE	siekaris	sikaris	TRUE
93	boobie	bonnie	FALSE	bcoibie	bobbie	FALSE
94	darryll	darryl	FALSE	darryrll	darryl	FALSE
95	quédebse	quédese	FALSE	quédebsb	quédese	FALSE
96	interrunpirlos	interrunpirlos	TRUE	nterrunpirlos	*	FALSE
97	tlmbrar	temblar	FALSE	tmtbrar	tlmbrar	TRUE
98	irrespeten	irrespeten	TRUE	irresphte	irrespeten	TRUE
99	trú	te	FALSE	ctrú	cara	FALSE
100	highton	hilton	FALSE	hivhtot	highton	TRUE

Otro corpus disponible para la seleccion de palabras es el CREA `http://corpus.rae.es/lfrecuencias.html`. ¿Qué diferencias hay usando este corpus?.

Medimos el desempeño de nuestro corrector como en el inciso anterior pero utilizando este nuevo corpus que proporciona la RAE. Igualmente seleccionamos 25 palabras y las alteramos como se describe en el inciso anterior. Los resultados son los siguientes:

De las 25 palabras originales se identificaron correctamente 18, y de las 25 palabras alteradas se identificaron correctamente 0. Aunque el tamaño de muestra en este caso es menor el desempeño del corrector es mucho menor al que se tuvo en el caso anterior. La siguiente es una lista de las palabras que se utilizaron para evaluar el corrector con el nuevo corpus:

	palabra	salida	calificacion	palabraAlterada	salidaAlterada	calificacionAlterada
1	psícóticos	psícóticos	TRUE	jpsícóticos	jpsícóticos	FALSE
2	cátcher	cátcher	TRUE	cáccmher	cáccmher	FALSE
3	arístide	arístide	TRUE	srísíide	srísíide	FALSE
4	serioso	serioso	TRUE	sernoso	sernoso	FALSE
5	reboteros	reboteros	TRUE	rebotesros	rebotesros	FALSE
6	matarratas	matarratas	TRUE	mataraatas	mataraatas	FALSE
7	pultney	pultney	TRUE	ultney	ultney	FALSE
8	stationery	stationery	TRUE	statiokery	statiokery	FALSE
9	chilenizada	chilenizada	TRUE	chiqienizada	chiqienizada	FALSE
10	batiéndote	batiéndote	TRUE	btiéndote	btiéndote	FALSE
11	sustantivo	sustantivo	TRUE	susitxntivo	susitxntivo	FALSE
12	mezclo	mezclo	FALSE	mezclkl	mezclkl	FALSE
13	paují	paují	TRUE	papjí	papjí	FALSE
14	darbujas	darbujas	TRUE	drdujas	drdujas	FALSE
15	escatimo	escatimo	FALSE	escatimu	escatimu	FALSE
16	hordeates	hordeates	FALSE	hordeatets	hordeatets	FALSE
17	clerici	clerici	TRUE	clericqi	clericqi	FALSE
18	ensidadas	ensidadas	FALSE	ensiadas	ensiadas	FALSE
19	wetzell	wetzell	TRUE	wetztjl	wetztjl	FALSE
20	cuadrarme	cuadrarme	FALSE	cuadrare	cuadrare	FALSE
21	inesilla	inesilla	TRUE	inesizlla	inesizlla	FALSE
22	semisupina	semisupina	FALSE	stemisupina	stemisupina	FALSE
23	eibarrés	eibarrés	TRUE	efbarrés	efbarrés	FALSE
24	capitulazo	capitulazo	TRUE	qapitulaz	qapitulaz	FALSE
25	mecedoras	mecedoras	FALSE	mecedorpas n	mecedorpas	FALSE

Por lo que concluimos que aunque el nuevo corpus es más grande, está menos orientado al contexto con el que diseñanos la estimación de $P(w|s)$, es decir que el corpus de la RAE y sus frecuencias difiere del corpus y estimaciones de frecuencias que se nos proporcionó

siendo el segundo el que nos brinda más información para evaluar las opiniones de igual manera para oraciones pequeñas creemos que el primer corpus es una mejor opción, por lo cual se agrega al entregable una app desarrollada con el lenguaje R y el package Shiny (contenida en los archivos 'ui.R', 'server.R' y 'corpus.RDS'), la cual es sencilla y permite introducir un texto, despliega la entrada así como la salida corregida con el corrector entrenado con el corpus del primer inciso.