Centro de Investigación en Matemáticas (CIMAT). Unidad Monterrey

Análisis de datos complejos: Tarea 1

Jorge Luis Ramos Zavaleta

11 de septiembre de 2018

1. Ejercicio

Implementa un corrector ortográfico automático para textos en español. Dada una palabra w, encuentra la palabra s que (suponemos), es la que se quería escribir correctamente. Para esto, considera el siguiente modelo básico:

$$s = arg\,\max_s P(s|w) = arg\,\max_s P(w|s)P(s),$$

donde, $P(s)$ es el modelo del lenguaje, y representa la probabilidad de que la palabra s sea la que se intentó 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 "correcta" s.

1.1. Solución

Para obtener el corrector ortográfico se considero una distribución apriori, en este caso dada por dos posibles diccionarios con frecuencias de aparación de palabras, se uso un diccionario de frecuencia de palabras según el Corpus OpenSubtitles y el diccionario CREA de las 10,000 palabras mas frecuentes.

Para la prueba de nuestro corrector ortográfico se hizo uso del SFU review corpus, que contiene reseñas de diversos objetos y servicios, para ello se hizo uso de los dos diccionarios de frecuencias antes mencionados y de un corrector ortográfico conocido como Aspell para

comprobar nuestros resultados contra los que se obtienen con Aspell. Hay varias formas de usar Aspell, pero en algunos sistemas Linux viene incluido y puede usarse directamente escribiendo en consola

aspell check nombredelarchivo.txt

con lo que se activa la consola interactiva. Para hacer comparable el resultado de Aspell se eligió reemplazar las palabras escritas incorrectamente con la primera opción que Aspell ofrecía.

Como parte del preprocesamiento de los textos, se realizo un tokenizacion del texto eliminando puntuación y números del texto para permitirnos trabajar solo con las palabras que son la parte que nos interesa usar para la corrección.

Para la parte de la probabilidad condicional se uso la distancia de Levenshtein, y para reducir el costo computacional se eligieron las palabras contra las cuales se debía comparar usando 2 criterios.

1. Las palabras elegidas no deben diferir en mas de dos carácter de longitúd de la palabra a corregir.
2. Las palabras elegidas deben coincidir en el primer carácter con la palabra a corregir.
3. Se elige la primera palabra que aparece en la lista que contiene las palabras con el mismo valor del argumento máximo.

Para establecer las probabilidades condicionales solo se consideraron palabras cuya distancia de Levenshtein fuera a lo más dos. Por lo que se estableció la siguiente regla de asignación de probabilidad con respecto a la distancia obtenida

Distancia		Probabilidad asignada
0	$\rightarrow$	0.9
1	$\rightarrow$	0.09
2	$\rightarrow$	0.01
Mas de 3	$\rightarrow$	0

Un problema al que nos enfrentamos directamente es que no tenemos inicialmente una métrica para saber que tanto se equivocaron los correctores ortográficos de manera automática por lo que la prueba se realizo en solo unas pocas reseãs para poder hacer una inspección manual de la correción. A manera de ejemplo se muestran los resultados obtenidos para una reseña de un coche (no_2_9.txt) resaltando las palabras que fueron corregidas de manera incorrecta.

Texto original

"En esta opinión voy a hablar del Opel Astra, aunque es una crítica en general a la marca Opel.

Hablo desde mi experiencia, y la de tres amigos míos más. Me explico, yo tuve un Opel Corsa, en el que al final casi tuve que cambiar hasta el volante, ya que se me averió de todo. Pero obviando este asunto, ya que se trataba de un coche viejo y tal... dos amigos míos compraron dos Opel Astra nuevecitos. Uno de ellos en el primer mes tuvo problemas de todo tipo, con los antivaho, con el aire acondicionado, con las luces... El otro, en el año siguiente a su compra, ha tenido muchisimas averías, no le funcionaba el cierre automático del maletero, le saltaban las alarmas de averías en el panel sin tenerlas, una temporada le arrancaba mal y era problema del carburador..., y todo eso en el primer año desde su compra. Y un tercero compró un Opel Vectra, al año se le estropeo el aire acondicionado, y encima no se lo querían reparar porque decían que la culpa era suya por no haberle llevado a la revisión... Al final y ante la amenaza de denuncias accedieron, pero le costó lo suyo.

Por mi propia experiencia, como comprendereis, el próximo coche queme compre no será un Opel...

Saludos."

Texto corregido usando el Corpus OpenSubtitles

"En esta opinión voy a hablar del Opel Astra, aunque es una crítica en general a la marca Opel.

Hablo desde mi experiencia, y la de tres amigos mío más. Me explico, yo tuve un Opel Corsa, en el que al final casi tuve que cambiar hasta el volantes, ya que se me averió de todo. Pero obviando este asunto, ya que se trataba de un coche viejo y tal... dos amigos mío compraron dos Opel Astra nuevecitos. Uno de ellos en el primer me tuvo problemas de todo tipo, con los activado, con el aire acondicionado, con las luces... El otro, en el año siguiente a su compra, ha tenido muchísimas averías, no le funcionaba el cierre automático del maletero, le saltaban las alarmas de averías en el papel sin tenerlas, una temporada le arrancaba mal y era problema del carburador..., y todo eso en el primer año desde su compra. Y un tercero compró un Opel Vectra, al año se le estropeo el aire acondicionado, y encima no se lo quería reparar porque decían que la culpa era suya por no haberle llevado a la revisión... Al final y ante la amenaza de denuncias accedieron, pero le costó lo suyo.

Por mi propia experiencia, como comprenderás, el próximo coche que compre no será un Opel...

Saludos. ”

Texto corregido usando las 10,000 palabras de la RAE

“En esta opinión voy a hablar del Opel Astra, aunque es una crítica en general a la marca Opel.

Hablo desde mi experiencia, y la de tres amigos más más. Me explicó, y tuve un Opel Corsa, en el que al final casi tuve que cambiar hasta el volante, y que se me abrió de todo. Pero objeto este asunto, y que se trataba de un coche viejo y tal... dos amigos más comprar dos Opel Astra nacional. Uno de ellos en el primer más tuvo problemas de todo tipo, con los aunque, con el aire administración, con las luces... El otro, en el año siguiente a su compra, ha tenido mientras ahora, no la funciona el cierre automático del mientras, la siempre las alarma de ahora en el papel sin también, una temporada la arranca mal y era problema del cualquier..., y todo eso en el primer año desde su compra. Y un tercero compró un Opel Vectra, al año se le europeo el aire administración, y encima no se lo querían reparar porque decían que la culpa era su por no haberle llevado a la revisión... Al final y ante la amenaza de denuncias acudieron, pero le costó lo su.

Por mi propia experiencia, como condiciones, el próximo coche que compra no será un Opel...

Saludos. ”

Texto corregido con Aspell

“En esta opinoón voy a hablar del Piel Astera, aunque es una ceíti ca en general a la marca Piel.

Hablo desde mi experiencia, y la de tres amigos míos más. Me explico, yo tuve un Piel Corsa, en el que al final casi tuve que cambiar hasta el volante, ya que se me averiÃ©ó de todo. Pero obviando este asunto, ya que se trataba de un coche viejo y tal... dos amigos míos compraron dos Piel Astera nueve citos. Uno de ellos en el primer mes tuvo problemas de todo tipo, con los anteva, con el aire acondicionado, con las luces... El otro, en el año siguiente a su compra, ha tenido muchÃsimas aveías, no le funcionaba el cierre Ã¡tomoátoco del maletero, le saltaban las alarmas de aveías en el panel sin tenerlas, una temporada le arrancaba mal y era problema del carburador..., y todo eso en el primer año desde su compra. Y un tercero compraó un Piel Vector, al año se le estropeo el aire

acondicionado, y encima no se lo queraíAna reparar porque deíAna que la culpa era suya por no haberle llevado a la revisoón... Al final y ante la amenaza de denuncias accedieron, pero le costaó lo suyo.

Por mi propia experiencia, como comprenderÃ©is, el paróeximo coche queme compre no será un Piel...

Saludos. ”

En estos resultados se puede apreciar que con el diccionario de frecuencias de la RAE se obtuvieron peores resultados de corrección, y esto debía ser obvio pues muchas de las palabras en el texto original son muy frecuentes pero en contextos de opinión, mientras que el diccionario de la RAE indica las más frecuentes en todo el idioma español.

Por otro lado, el corrector de Aspell parece generar algunos problemas de codificación al guardar el archivo y debido a que no lee correctamente las palabras con acento, por lo que genera una tokenización quitando algunas letras con acentos, pero aun con esto se obtienen mejores resultados que con el diccionario de la RAE.

Figura 1.1: Vista de la consola interactica de Aspell en Ubuntu con el texto provisto en **no_2_9.txt**

Cabe considerar que solo estamos considerando los resultados obtenidos en solo una de las reseñas, y si se consideraran todas las reseñas probablemente se esperarian resultados similares o peores con nuestro corrector ortográfico que con el de Aspell que esta mucho más optimizado para esta tarea.

Posibles mejoras al corrector ortográfico

Una primera mejora que se le puede realizar al corrector ortográfico es a la hora de elegir el

argumento máximo establecer una mejor manera de seleccionar que solo elegir la primera de las palabras que tengan un empate en su cálculo.

Una segunda opción es ampliar el Corpus de frecuencias usando palabras mas acorde al contexto al que intentamos corregir, ya que debido a esto algunas palabras que son frecuentes en un contexto no lo son en otro, como fue el caso del Corpus de la RAE que no esta específicado para el tipo de contexto de opinión que requerimos para corregir los textos de reseña que estamos evaluando.

Otra opción es utilizar otro tipo de métrica para el string distance. Sin embargo, dependiendo de los contextos de los textos a evaluar un cambio de métrica puede generar una peor corrección que la estamos considerando, además de que computacionalmente puede ser más pesada la alternativa.

Un par de opciones mas serían hacer uso del contexto del texto, es decir usar x palabras hacia adelante y hacía atrás para establecer una mejor manera de elegir entre las palabras con probabilidad máxima igual, y jugar con las probabilidades asignadas a distribución apriori para generar desempates en las palabras con probabilidad máxima igual.