

Tópicos selectos de Análisis de Datos

Tarea 1

Para entregar el 11 de septiembre de 2018

1. Implementa un corrector ortográfico automático para textos en español.

- (a) 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 .

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 información 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 qué tanto es *más*.

- (b) Prueba tu corrector con textos del *SFU review corpus*

(https://www.sfu.ca/~mtaboada/SFU_Review_Corpus.html, disponible en la página del curso), que contiene reseñas y críticas de consumidores sobre diferentes productos. Compáralo con los resultados obtenidos con un corrector estándar, por ejemplo, Aspell (<http://aspell.net/>). ¿Qué puedes decir sobre el desempeño del corrector?

- (c) ¿Cómo podrías mejorar [tu corrector ortográfico?](#) 

Notas:

- Puedes usar el lenguaje de tu preferencia. La entrada debe ser un texto (en archivo o via teclado). La salida, el texto corregido. Si es necesario, incluye las indicaciones para ejecutarlo.

- Puedes usar Aspell directamente del código fuente en C. También hay una librería en R a través del CRAN.
- Puedes usar otro tipo de *string distance*, pero aclara cómo se modifica el modelo del error (inicio (a))
- Hay formas cuantitativas de verificar el desempeño de los métodos de corrección ortográficos. Una idea puedes verla en el paper de Whitelaw et al. disponible en la página del curso. La implementación es opcional, con puntos extra incluidos.
- Otro corpus disponible para la selección de palabras es el CREA (<http://corpus.rae.es/lfrecuencias.html>). ¿Qué diferencias hay usando este corpus? (opcional).