Práctica 01: Medición de información bajo incertidumbre

Luis Tong Chabes

Marzo 2019

1. Calcular la entropía de Hartley y Shannon

La implementación tenemos dos funciones con su nombre respectivamente.

2. Entropía de un segmento de texto

- Para contar la frecuencia de ocurrencias, se pensó como si fuese un contador de palabras mediante el uso de un diccionario que pregunta la existencia del carácter para aumentar su contador o agregar uno nuevo.
- *Python* nos ayuda con la función `lower()`.
- Si se usa la librería numpy.
- para normalizar simplemente obtenemos la frecuencia de caracteres para dividirlo sobre el tamaño del texto, así obtener el valor entre 0 y 1.
- En un texto de más de 4000 caracteres obtenemos lo siguientes valores:
  - Shannon = 4,3884
  - Hartley = 5,6438

  Podemos notar que la diferencia entre ambas medidas es 1,3 aproximadamente. Con estos valores podemos entender que la frecuencia de los caracteres es común a la realidad de textos, hay ciertos caracteres que tienen más incidencia que otras, como las vocales.
  Notamos que Hartley solo necesita saber la cantidad de caracteres en un texto sin importar su frecuencia.

3. Lipogramas

Para el primer lipograma tenemos la siguientes respuesta:

- Shannon = 4,2108

- Hartley = 5,5545

Para el segundo lipograma tenemos la siguientes respuesta:

- Shannon = 4,4681
- Hartley = 5,5545

A primera impresión, se nota el mismo valor de Hartley para ambos lipogramas, como se mencionó que solo dependerá del tamaño del texto, entonces si ambos contienen la misma cantidad, estas serán las mismas. En otras palabras no importa el contenido, significado, etc.
Pero si nos fijamos en Shanon este si va a variar por cada frecuencia. Si revisamos con los textos normales como en el ejercicio anterior, estos valores no presentan mayor diferencia aunque se comparare con textos como lipogramas.

4. Generación de Textos

- Para la generación de texto usamos un simple random de la librería de python para tener caracteres de la lista del alfabeto y el *space*.
  Los valores obtenidos son los siguientes:

  - Shannon = 4,75441
  - Hartley = 4,7548

  Notamos que los valores no tienen gran diferencia para cada medida, esto nos muestra que el porcentaje de caracteres son iguales, mostrando una distribución uniforme de los caracteres que aparecen en el texto.

- A diferencia del anterior podemos generar caracteres con una probabilidad dependiente de cuantas veces queremos que incurrar en nuestro texto. Los valores obtenidos son los siguientes:

  - Shannon = 4,0873
  - Hartley = 4,7548

  Se nota que la métrica para Shannon es menor porque la ocurrencia de caracteres son distribuidos de manera no uniforme, hay mas vocales asi como menos letras no usadas.
  Pero para Hartley es igual que en el caso anterior porque solo impora el tamaño de texto.

5. Permutación

Consiste en mover un carácter en otra posición del texto, para realizar esta implementación es jugar con funciones de un mismo vector, como mover aleatoriamente, eliminar este y así sucesivamente. Los valores obtenidos son los siguientes:

- Shannon = 4,75441
- Hartley = 4,7548

Como se suponía que obtengamos los mismos valores ya que estas métricas se enfocan en ver el número de caracteres que aparece un carácter y lo que se ha hecho es permutar, es decir el carácter se mantiene en el texto, no altera este número. Y así se tiene los mismos valores. En otras palabras no sería necesario volver a calcular las métricas si ya tenemos la anterior