

Computación Bioinspirada

Práctica 01: Medición de información bajo incertidumbre

Yván Jesús Túpac Valdivia
Universidad Católica San Pablo, Arequipa – Perú

20 de marzo de 2019

Objetivo

Calcular las medidas de significancia de información de Hartley y Shannon para textos aleatorios.

1. Material

1. Slides CBI_01_Intro.pdf
2. IDE C++/Python

2. Problemas

1. **Calcular la entropía de Hartley y Shannon como medidas de información** La entropía de Shannon se calcula a partir de la distribución de probabilidad definida para un conjunto de elecciones (como símbolos).
 - Implemente una función que reciba un vector/array de valores de probabilidad p_i (donde $\sum_i p_i = 1$) y retorne la entropía de Shannon de esta distribución.
 - Implemente la entropía de Hartley para el mismo array
2. **Entropía de un segmento de texto.**
 - Implemente un programa que lea el contenido de un archivo de texto, cuente la frecuencia de ocurrencia de cada caracter en español (incluyendo ñ, caracteres acentuados á, é, í, ó, ú y espacio).
 - No debe haber distinción entre mayúsculas y minúsculas
 - Almacenar las frecuencias en un array `numpy`
 - Normalizar las frecuencias para obtener una estimativa de la probabilidad de cada caracter
 - Calcular la entropía de Shannon y Hartley y muestre los valores obtenidos
 - El texto debe contener al menos 4000 caracteres. Pruebe haciendo muestreos de varios textos (por ejemplo, tomando partes de Wikipedia en español) y verifique qué tanto varían las Entropías de Shannon y Hartley. Explique estas variaciones
3. **Lipogramas:** averigue en qué consiste un Lipograma (puede consultar incluso en wikipedia)
 - Utilice los textos disponibles en
<http://tinyurl.com/napp32s>
<http://tinyurl.com/q9b2r36>
y calcule sus medidas de entropía de Shannon y Hartley, compare sus medidas con las de los textos “normales” y explique lo ocurrido.

4. **Generación de textos:** A partir de los valores encontrados de las medidas para textos “normales”

- Implemente un programa que genere una pieza aleatoria de texto (4000 caracteres o del tamaño de los textos que usó, incluyendo al caracter *space*), almacene en archivo y observe el texto generado.
- Obtenga las medidas de Shannon y Hartley en este texto y explique lo ocurrido
- Implemente un programa que genere una pieza aleatoria de texto (4000 caracteres o del tamaño de los textos que usó, incluyendo al caracter *space*), pero tomando en cuenta las probabilidades de ocurrencia. Almacene en archivo y observe el texto generado.

5. **Permutación:** permute los caracteres de un texto normal, recalculé las medidas de Shannon y Hartley y explique lo ocurrido, ¿Es necesario extender las medidas?

3. Desarrollo y entrega

- El trabajo debe ser desarrollado en sesión de laboratorio.
- Se debe entregar en digital (usando classroom) un informe conteniendo el desarrollo de todas las preguntas y los códigos implementados
- Plazo de entrega del informe digital: una versión al finalizar la clase, otra al final del día 21/Marzo 23:59h