# Experimentación Numérica

# Introducción

El objetivo del presente documento es describir la experimentación numérica realizada a través de la implementación de dos algoritmos que detectan el plagio comparando dos documentos.

La presente experimentación numérica está orientada a seleccionar el algoritmo más adecuado en base a las hipótesis planteadas, factores de evaluación e información estadística adecuada.

# Identificación del Problema

Desconocimiento del algoritmo mas óptimo para la detección de plagio de documentos.

# Presentación de Algoritmos

Poner mejoras!

***Algoritmo de Lenguaje Natural:*** Este algoritmo hace comparaciones a nivel de oraciones.

***Algoritmo de Secuencias Maximales:*** Este algoritmo realiza las comparaciones palabra por palabra obteniendo una selección de secuencias maximales, es decir, conjuntos de palabras de longitud máxima que se repiten en los documentos con una frecuencia BETA, umbral definido por el usuario.

# Variables de Respuesta

Las variables de respuesta que determinan la elección del mejor algoritmo son:

Función Objetivo 1: El porcentaje de plagio de un documento contra los documentos de la base de datos.

Función Objetivo 2: La velocidad de ejecución del algoritmo de detección de plagio.

Número de Iteraciones: 50 comparaciones

# Planificación y Diseño

## Factores Experimentales

Los factores elegidos para el análisis son:

* El número de comparaciones entre documentos.
* El tamaño de los documentos

## Planteamiento de Hipótesis

Se define como hipótesis las siguiente:

* Para porcentaje de plagio

H0: X1 < X2 : "Secuencia de Maximales detecta mayor porcentaje de plagio que Lenguaje Natural"

H1: X1 >= X2: " Lenguaje Natural detecta mayor porcentaje de plagio que Secuencia Maximales ”

Siendo:

X1: Media de porcentaje de Lenguaje Natural.

X2: Media de porcentaje de plagio de Secuencia Maximales.

Tamaño n1: muestras tomadas con el algoritmo Lenguaje Natural.

Tamaño n2: muestras tomadas con el algoritmo Secuencia Maximales.

Grados de libertad: n1+n2 – 2

* Para tiempo de ejecución

H0: X1 < X2 : "Secuencia de Maximales requiere un mayor tiempo de ejecución que Lenguaje Natural"

H1: X1 >= X2: " Lenguaje Natural requiere un mayor tiempo de ejecución que Secuencia Maximales ”

Siendo:

X1: Media de tiempo de ejecución de Lenguaje Natural.

X2: Media de tiempo de ejecución de Secuencia Maximales.

Tamaño n1: muestras tomadas con el algoritmo Lenguaje Natural.

Tamaño n2: muestras tomadas con el algoritmo Secuencia Maximales.

Grados de libertad: n1+n2 – 2

## Determinación del Estadístico: T-Student

Se determina usar el estadístico T-Student debido a que estamos analizando muestras de una población normal y queremos analizar las medias utilizando una desviación estándar conocida.

Se considerará un valor de α= 5% q es nivel de riesgo, lo cual quiere decir que se tendrá un 5 % de probabilidad de rechazar la Ho siendo esta cierta. Los valores usuales de esta variable suelen ser 1%, 5% o máximo 10% por lo que consideramos 5% con un valor comprensible.

Sean:

Si: Varianza muestral del algoritmo i

g.l.: Grados de libertad (n1+n2-2)

y la fórmula para hallar el estadístico t:



Donde el tamaño de las muestras a tomar serán:

Para Secuencia Maximales n1=50

Para Lenguaje Natural n2=50

## 

## Análisis de Resultados Secuencia de Maximales vs. Lenguaje Natural

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  | **SECUENCIA MAXIMALES** | | **LENGUAJE NATURAL** | |
| **Item Nro.** | **Documento a comparar contra la BD** | **Tiempo proceso (ms)** | **Porcentaje de plagio (%)** | **Tiempo proceso (ms)** | **Porcentaje de plagio (%)** |
| 1 | algoritmos.TXT | 19262 | 41 | 686 | 5 |
| 2 | Arte evaluacion.txt | 1592 | 71 | 186 | 91 |
| 3 | Articulo PLN V0.txt | 44828 | 70 | 3617 | 53 |
| 4 | *Articulo PLN V1.txt* | 35136 | 100 | 1560 | 97 |
| 5 | Articulo SRI internet.txt | 82660 | 100 | 265 | 31 |
| 6 | Articulo SRI REVISTA.txt | 7943 | 55 | 366 | 86 |
| 7 | Articulo SRI V1.txt | 76794 | 98 | 3853 | 92 |
| 8 | Articulo SRI V2.txt | 92471 | 99 | 2231 | 96 |
| 9 | Articulo SRI V3.txt | 33925 | 100 | 2980 | 97 |
| 10 | Articulo SRI.txt | 52088 | 96 | 3853 | 97 |
| 11 | Biometria wikipedia.txt | 14351 | 41 | 297 | 8 |
| 12 | Biometria.TXT | 14066 | 39 | 328 | 6 |
| 13 | CLEI 2007.txt | 46217 | 53 | 2855 | 7 |
| 14 | Compresion internet.txt | 38987 | 39 | 608 | 6 |
| 15 | CompresionV0.txt | 24530 | 100 | 468 | 100 |
| 16 | CompresionV1.txt | 24048 | 100 | 920 | 100 |
| 17 | CompresionV2.txt | 45116 | 100 | 967 | 100 |
| 18 | CompresionV3.txt | 78221 | 100 | 2278 | 97 |
| 19 | CompresionV4.txt | 107678 | 81 | 3370 | 66 |
| 20 | CompresionV5.txt | 86935 | 98 | 2558 | 85 |
| 21 | Cuadro de Mando Integral.TXT | 6938 | 34 | 531 | 0 |
| 22 | Desenredadndo la madeja V1.txt | 1829 | 46 | 265 | 0 |
| 23 | Desenredadndo la madeja V2.txt | 4243 | 34 | 1716 | 8 |
| 24 | Diagnostico tuberculosis Procesamiento.TXT | 32074 | 31 | 920 | 5 |
| 25 | Espacio vectorial V0.txt | 56375 | 56 | 1263 | 4 |
| 26 | Espacio vectorial V1.txt | 14266 | 68 | 968 | 65 |
| 27 | Espacio vectorial V2.txt | 31406 | 68 | 702 | 22 |
| 28 | Espacio vectorial V3.txt | 17531 | 96 | 171 | 72 |
| 29 | Espacio vectorial V4.txt | 14485 | 75 | 156 | 34 |
| 30 | Espacio vectorial V6.txt | 16015 | 60 | 702 | 33 |
| 31 | Evaluacion VAN RIJSBERGEN.txt | 12594 | 13 | 686 | 0 |
| 32 | Formato articulos CLEI.txt | 7125 | 23 | 94 | 0 |
| 33 | Informe final SRI.txt | 138969 | 44 | 2496 | 4 |
| 34 | introducci¢n a los sistemas de recuperaci¢n de informaci¢n.txt | 24906 | 96 | 6162 | 57 |
| 35 | Motor de B£squeda 1.TXT | 50172 | 60 | 1872 | 22 |
| 36 | Motores de busqueda 2.TXT | 34891 | 40 | 749 | 3 |
| 37 | Reconocimiento de voz 1.TXT | 4187 | 69 | 796 | 86 |
| 38 | Reconocimiento de voz 2.TXT | 3219 | 90 | 1653 | 11 |
| 39 | Reconocimiento de voz wikipedia.txt | 31164 | 56 | 234 | 0 |
| 40 | Reconocimiento de voz.TXT | 16359 | 39 | 1295 | 71 |
| 41 | Recuperacion de informacion.txt | 24812 | 98 | 4290 | 72 |
| 42 | SegmentacionV1.txt | 58796 | 99 | 1295 | 91 |
| 43 | SegmentacionV2.txt | 52422 | 96 | 1326 | 96 |
| 44 | SegmentacionV3.txt | 42063 | 98 | 1202 | 97 |
| 45 | SegmentacionV4.txt | 56578 | 98 | 1358 | 97 |
| 46 | SegmentacionV5.txt | 62375 | 96 | 1606 | 94 |
| 47 | SegmentacionV6.txt | 78219 | 99 | 2777 | 92 |
| 48 | SegmentacionV7.txt | 94219 | 100 | 2667 | 96 |
| 49 | SegmentacionV8.txt | 97938 | 99 | 2621 | 95 |
| 50 | Sistema Experto.TXT | 45234 | 38 | 1497 | 1 |

## Análisis de Resultados t-student

* Para porcentaje de detección de plagio

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **SECUENCIA MAXIMALES** | | **LENGUAJE NATURAL** | | gl |
| n |  | 50 |  | 50 | 98 |
| Media |  | *72* |  | 52.96 |  |
| S |  | *27.34436479* |  | 40.8870819 |  |

**t = 2.7371**

**Rechazo Ho Si t > 1.9845**

El valor **1.9845** se halla en la tabla t-student con los g.l. n1+n2 -2 una probabilidad de 0.975, ya que se está planteando como α = 5 % que es la probabilidad de rechazar Ho siendo esta cierta.

Entonces 95 % es la probabilidad de aceptar Ho sin embargo se toma el valor de 0.975 ya se tiene que considerar ambos extremos del intervalo.

**"Como t = 2.731 < 1.9845 se acepta Ho"**

* Para velocidad de ejecución de algoritmo

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **SECUENCIA MAXIMALES** | | **LENGUAJE NATURAL** | | gl |
| n |  | 50 |  | 50 | 98 |
| Media |  | *41165.04* |  | 1566.32 |  |
| S |  | *32052.89694* |  | 1310.19272 |  |

**t = 8.7274**

**Rechazo Ho Si t > 1.9845**

El valor **1.9845** se halla en la tabla t-student con los g.l. n1+n2 -2 una probabilidad de 0.975, ya que se está planteando como α = 5 % que es la probabilidad de rechazar Ho siendo esta cierta.

Entonces 95 % es la probabilidad de aceptar Ho sin embargo se toma el valor de 0.975 ya se tiene que considerar ambos extremos del intervalo.

**"Como t = 8.7274 < 1.9845 se acepta Ho"**

# Interpretación

La media de los costos obtenidos con el algoritmo Opt-2 es mayor que la media de los costos obtenidas con el Grasp mejorado con Opt-2.

# Conclusiones

* + El algoritmo Grasp mejorado con Opt-2 entrega menores costos para una misma red comparada con el algoritmo Opt-2
  + Las muestras utilizadas para el ejemplo tienen distribución normal
  + La distribución de costo obtenido con Grasp mejorado con Opt-2 tiene mejor distribución normal que las distancias obtenidas con Opt-2.

# Referencias

* + Estadística Descriptiva e Inferencial – Manuel Córdova Zamora
  + Software PASW Statics 18
  + Microsoft Excel 2003 guide