# 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

Hallar el algoritmo mas optimo para la detección de plagio de documentos.

# Presentación de Algoritmos

***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: Encontrar el algoritmo que me da el mayor porcentaje de plagio.

Función Objetivo 2: Encontrar el algoritmos que me de la menor velocidad de ejecución.

Número de Iteraciones: 80 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 siguiente:

H0: X1 < X2 :"Lenguaje Natural es mas óptimo que Secuencia de Maximales"

H1: X1 >= X2: "Secuencia de Maximales es mas óptimo que Lenguaje Natural"

Siendo:

X1: Media de costo de Grasp con mejora.

X2: Media de costo de Opt-2.

Tamaño n1: muestras tomadas con el algoritmo Grasp

Tamaño n2: muestras tomadas con el algoritmo Opt-2

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 formula para hallar el estadístico t:



En el experimento se obtuvieron las muestras n1=81 para el algorimo Grasp mejorado con opt-2 y con alpha= 0.11 y n2=81 para el algoritmo Opt-2.

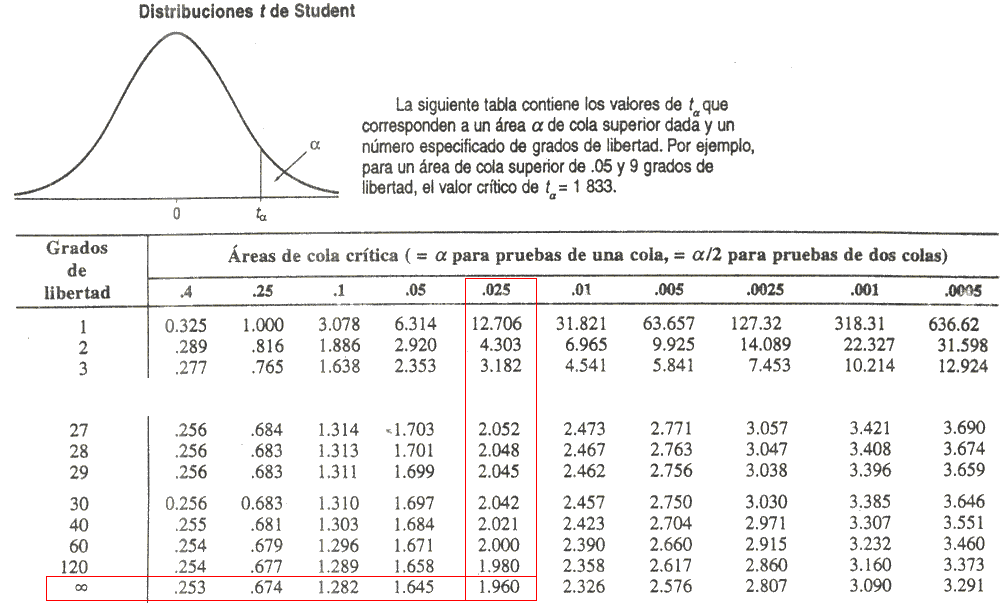
|  |  |  |
| --- | --- | --- |
|  | **GRASP Alpha=0.11** | **OPT-2** |
|  | 38241.50686 | 38728.77726 |
|  | 38351.88287 | 38839.15327 |
|  | 38597.83234 | 39085.10273 |
|  | 38598.313 | 39085.5834 |
|  | 38634.16447 | 40356.90387 |
|  | 41400.55345 | 43236.23974 |
|  | 41566.15524 | 43288.89464 |
|  | 41593.54362 | 41593.54362 |
|  | 43540.43023 | 43757.05431 |
|  | 43545.21737 | 44595.31172 |
|  | 46354.59942 | 48543.0051 |
|  | 47676.27731 | 49170.93966 |
|  | 47716.75705 | 48773.16015 |
|  | 48239.98783 | 51627.00733 |
|  | 48246.68619 | 51633.70568 |
|  | 48469.65313 | 51856.67262 |
|  | 48557.94003 | 51944.95953 |
|  | 48738.91572 | 52041.93215 |
|  | 49197.16556 | 52101.12277 |
|  | 51125.60354 | 54101.53275 |
|  | 52381.78211 | 55508.1728 |
|  | 52415.08074 | 55541.47143 |
|  | 52955.53097 | 54455.84347 |
|  | 52953.40584 | 54496.36416 |
|  | 52953.42699 | 55514.57635 |
|  | 54187.92057 | 56749.06993 |
|  | 54867.77387 | 56804.18373 |
|  | 54447.00067 | 56804.50244 |
|  | 54795.19385 | 56988.57493 |
|  | 55522.67748 | 58048.49814 |
|  | 55800.26881 | 58157.77057 |
|  | 55800.27726 | 58157.77903 |
|  | 56090.80607 | 58332.06197 |
|  | 55949.36305 | 58330.44477 |
|  | 57176.14705 | 62848.16972 |
|  | 58598.2688 | 61273.08647 |
|  | 59133.47252 | 61492.14514 |
|  | 59070.05775 | 61535.24072 |
|  | 59128.73635 | 61601.59833 |
|  | 59064.21525 | 60963.34356 |
|  | 59207.48549 | 60989.36152 |
|  | 60456.19929 | 62259.55699 |
|  | 62742.4615 | 63794.49502 |
|  | 63851.80723 | 64214.12061 |
|  | 63411.08177 | 64208.86119 |
|  | 64524.69174 | 65989.1784 |
|  | 65450.79367 | 66524.84422 |
|  | 65189.03143 | 66584.64759 |
|  | 65644.54361 | 67108.38774 |
|  | 65856.372 | 67110.94397 |
|  | 67209.40182 | 67892.67937 |
|  | 67017.51844 | 68030.99051 |
|  | 67598.0737 | 68345.71565 |
|  | 67843.77776 | 68346.19425 |
|  | 67818.25171 | 68170.0024 |
|  | 67648.52862 | 68324.02326 |
|  | 68231.11318 | 68756.77277 |
|  | 68980.52471 | 74511.58978 |
|  | 69789.99445 | 74597.38883 |
|  | 69904.84214 | 74643.19122 |
|  | 70581.08689 | 74848.26399 |
|  | 70412.38772 | 74990.39524 |
|  | 70007.52541 | 75012.11501 |
|  | 72959.95734 | 76206.04079 |
|  | 72050.13808 | 77377.65545 |
|  | 74422.94499 | 77516.58675 |
|  | 74257.34215 | 78423.49162 |
|  | 74500.98381 | 79166.71416 |
|  | 75116.18475 | 79104.52343 |
|  | 75318.52843 | 76490.16603 |
|  | 75307.93393 | 79483.95071 |
|  | 75782.28909 | 79940.37765 |
|  | 76500.59597 | 81498.61163 |
|  | 77224.80267 | 78124.44783 |
|  | 77563.9279 | 77704.61288 |
|  | 77737.99736 | 77705.17003 |
|  | 78610.10815 | 78705.05445 |
|  | 77693.78799 | 79098.76344 |
|  | 79143.42754 | 79298.42328 |
|  | 78932.37548 | 81204.8026 |
|  | 79991.5245 | 81677.0075 |
|  |  |  |

## Análisis de Resultados t-student

|  |  |  |
| --- | --- | --- |
| **Algoritmo** | Grasp mejorado con Opt-2 y alpha= 0.11 | Opt-2 |
| **n** | 81 | 81 |
| **Media:** | 60545.39427 | 62740.04469 |
| **Desviacion Estandar:** | 12108.62673 | 12468.74803 |
| **t** | **-1.13642492987676** | |

**Rechazo Ho Si t > 1.960**

El valor **1.960** se halla en la tabla t-student con los g.l. n1+n2 -2 y 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 = -1.13642492987676 < 1.960 se acepta Ho"**

H0: X1 < X2 :"Opt-2 genera mayor costo que Grasp mejorado con Opt-2"

H1: X1 >= X2: "Grasp mejorado con Opt-2 genera mayor costo que Opt-2"

Siendo:

X1: Media de costo de Grasp

X2: Media de costo de Opt-2

## Análisis de Resultados Grasp sin mejora vs. OPT-2

|  |  |
| --- | --- |
| **GRASP Puro** | **OPT-2** |
| 40891.38814 | 38728.77726 |
| 40805.40286 | 38839.15327 |
| 41463.35573 | 39085.10273 |
| 42456.30386 | 39085.5834 |
| 41699.29243 | 40356.90387 |
| 44620.96602 | 43236.23974 |
| 45455.57014 | 43288.89464 |
| 45290.96456 | 41593.54362 |
| 47816.74496 | 43757.05431 |
| 47921.63501 | 44595.31172 |
| 52772.85175 | 48543.0051 |
| 51964.90099 | 49170.93966 |
| 52787.89922 | 48773.16015 |
| 53384.3533 | 51627.00733 |
| 54134.79716 | 51633.70568 |
| 51924.97332 | 51856.67262 |
| 54393.72198 | 51944.95953 |
| 55006.05528 | 52041.93215 |
| 55956.0973 | 52101.12277 |
| 57575.17558 | 54101.53275 |
| 58853.64229 | 55508.1728 |
| 61388.39851 | 55541.47143 |
| 61233.5417 | 54455.84347 |
| 60496.21946 | 54496.36416 |
| 61876.65685 | 55514.57635 |
| 66088.40716 | 56749.06993 |
| 64621.47951 | 56804.18373 |
| 60317.89364 | 56804.50244 |
| 65322.60611 | 56988.57493 |
| 66243.38511 | 58048.49814 |
| 67573.28228 | 58157.77057 |
| 66037.75793 | 58157.77903 |
| 69312.11868 | 58332.06197 |
| 70861.70507 | 58330.44477 |
| 72232.25018 | 62848.16972 |
| 71386.81073 | 61273.08647 |
| 72748.16103 | 61492.14514 |
| 73554.70669 | 61535.24072 |
| 76248.79246 | 61601.59833 |
| 75094.57327 | 60963.34356 |
| 74833.45769 | 60989.36152 |
| 74192.99618 | 62259.55699 |
| 79032.26879 | 63794.49502 |
| 75254.42943 | 64214.12061 |
| 81954.34389 | 64208.86119 |
| 82613.90979 | 65989.1784 |
| 80572.96118 | 66524.84422 |
| 85546.41059 | 66584.64759 |
| 87443.48743 | 67108.38774 |
| 85972.36289 | 67110.94397 |
| 82429.18554 | 67892.67937 |
| 84337.02293 | 68030.99051 |
| 87649.98121 | 68345.71565 |
| 90964.55551 | 68346.19425 |
| 86894.35941 | 68170.0024 |
| 90467.60619 | 68324.02326 |
| 91620.29262 | 68756.77277 |
| 91328.2442 | 74511.58978 |
| 95944.37981 | 74597.38883 |
| 94855.09066 | 74643.19122 |
| 93899.5183 | 74848.26399 |
| 98519.6598 | 74990.39524 |
| 98805.39875 | 75012.11501 |
| 100289.6514 | 76206.04079 |
| 100009.1874 | 77377.65545 |
| 99330.12747 | 77516.58675 |
| 99327.3693 | 78423.49162 |
| 105047.9216 | 79166.71416 |
| 98168.43142 | 79104.52343 |
| 103081.243 | 76490.16603 |
| 105359.3057 | 79483.95071 |
| 104896.9816 | 79940.37765 |
| 107528.1173 | 81498.61163 |
| 104576.304 | 78124.44783 |
| 106667.6376 | 77704.61288 |
| 106404.1988 | 77705.17003 |
| 110013.7404 | 78705.05445 |
| 106214.2002 | 79098.76344 |
| 108162.5236 | 79298.42328 |
| 109843.1831 | 81204.8026 |
| 112701.9348 | 81677.0075 |

|  |  |  |
| --- | --- | --- |
| **Media:** | 76574.9237 | 62740.04469 |
| Desviacion Estandar: | 21361.73571 | 12468.74803 |
| **t** | 5.03 | |

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 (igual que el caso anterior)

**"Como t = 5 > 1.960 se rechaza Ho"**

H0: X1 < X2 :"Opt-2 genera mayor costo que Grasp puro mejorado con Opt-2"

H1: X1 >= X2: "Grasp puro genera mayor costo que Opt-2"

Siendo:

X1: Media de costo de Grasp puro

X2: Media de costo de Opt-2

# 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