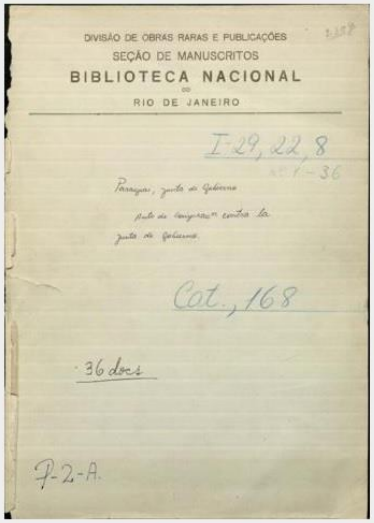
# Aplicación de técnicas de Data Mining para clasificar colecciones de documentos



Sebastian Ortiz, WIlfrido Inchaustti, Marcial Cohene

# Caso de estudio

- Archivo Nacional de Asunción
  - Organización
- Access to Memory (AtoM)
  - Registros
- Secciones de Interés
  - Río Branco
  - Historia



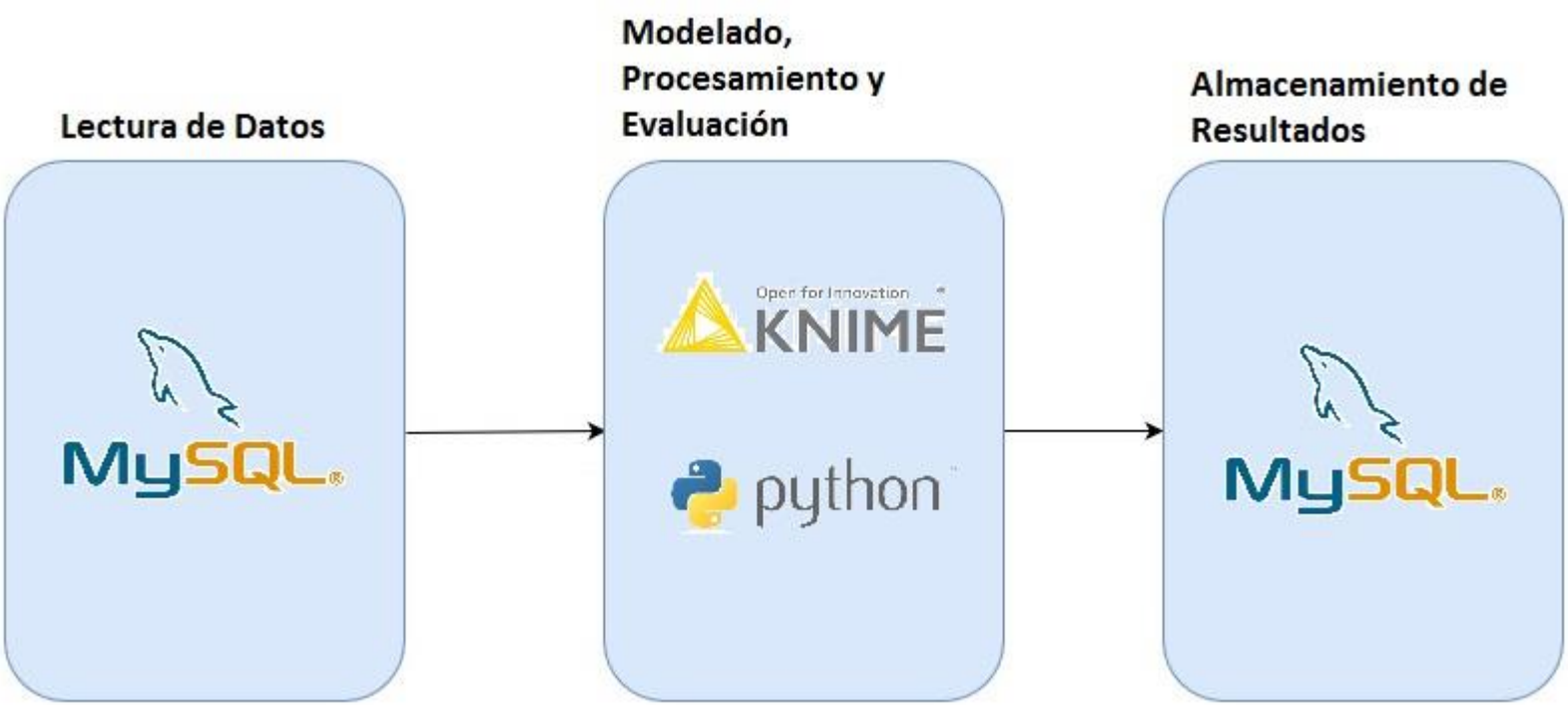

Area de identidad

| | |
|---|---|
| Código de referencia | PY-ANA-AHRP-168-1-230 |
| Título | Acto de conspiración contra la Junta de Gobierno (Parte 1) |
| Fecha(s) | • 1811-9-16 (Creación) |
| Nivel de descripción | Unidad documental simple |
| Volumen y soporte | 230f papel |

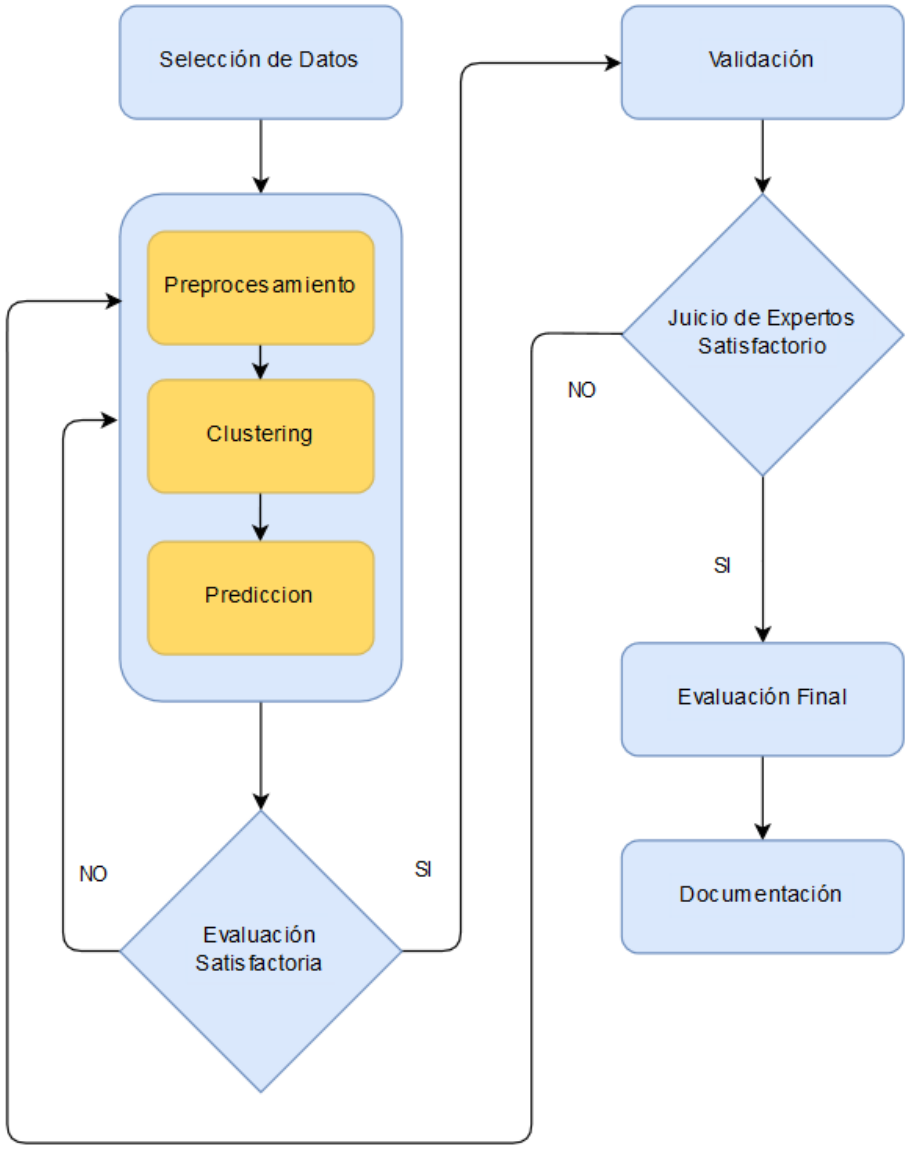
Area de contexto

| | |
|---|---|
| Nombre del productor | Yegros, Fulgencio (1780-1821)<br>Historia biográfica:<br>Fulgencio nació en el pueblo de Quiquió en la estancia familiar de Santa Bárbara en el año 1780. El 20 de enero de 1807 intervino en la batalla del Buceo (Montevideo) en que las tropas españolas fueron derrotadas por las inglesas y el contingente ... » |
| Institución archivística | Archivo Nacional de Asunción |
| Historia archivística | Propiedad del Estado Paraguayo hasta el fin de la Guerra contra la Triple |

# Arquitectura del Sistema



# Flujo de datos

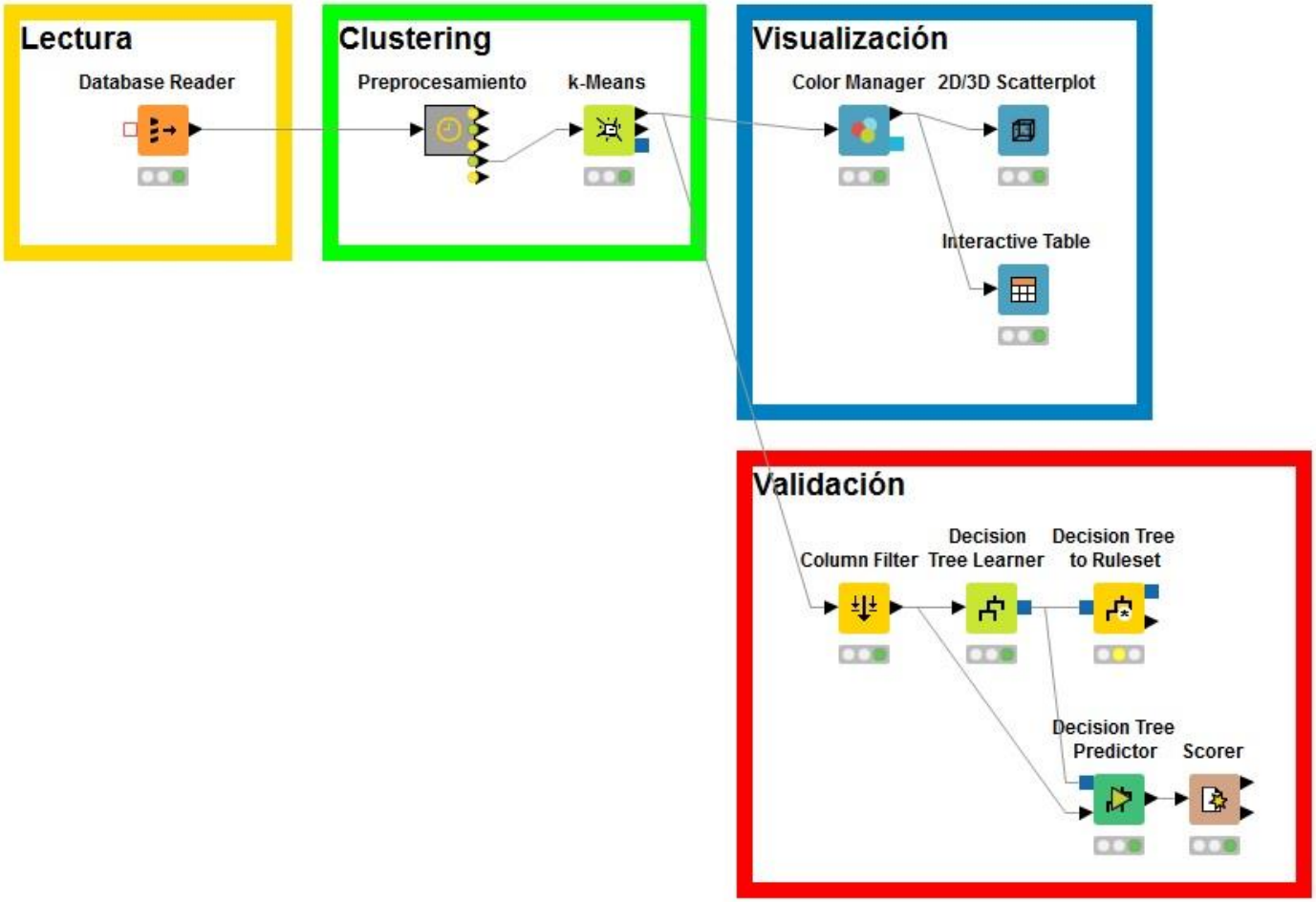


# Selección de Técnicas

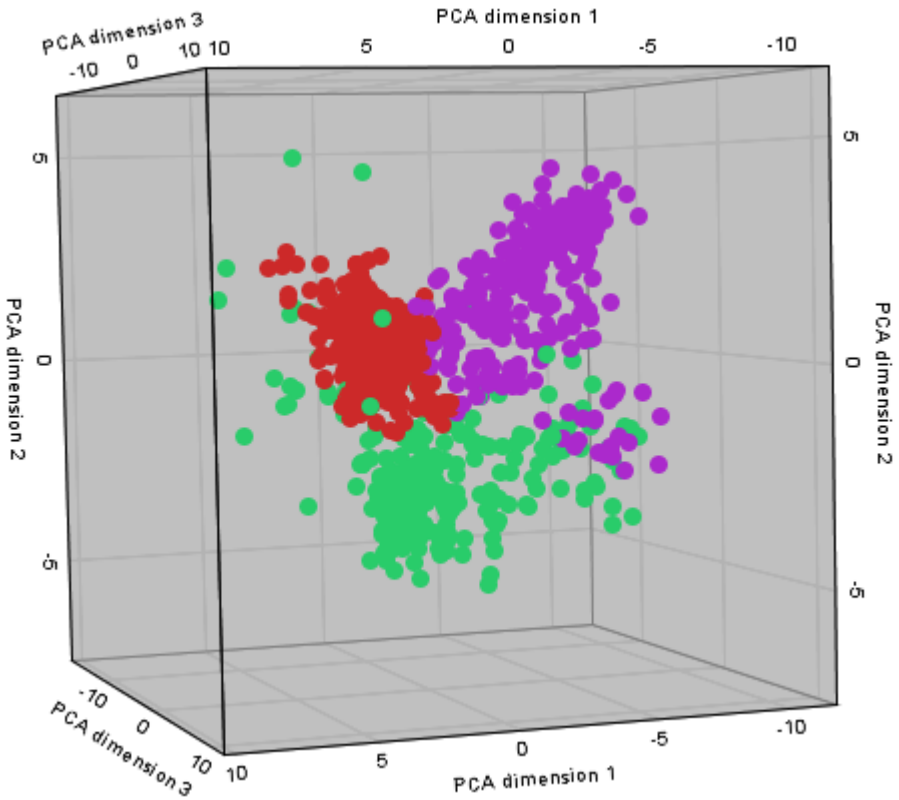
- Preprocesamiento
  - Filtrado
  - Palabras de Parada
  - Derivaciones
  - Poda
  - Bolsa de Palabras
    - TF-IDF
    - PCA
- Métrica de Similitud
  - Similitud Coseno
- Técnicas de Agrupamiento
  - K-means
  - Agrupamiento Espectral

# Modelos Desarrollados

# ESQUEMA GENERAL



# Agrupamiento generado



# Sección río branco

- Preprocesamiento:
  - Cantidad de documentos digitalizados: 1.023.
  - Promedio de Palabras: 85.
  - TF-IDF (Frecuencia de palabras con respecto a la colección): 0,51.
  - PCA: 0%, 10%, 20%, 40%, 60%, 80%, 90%.

- Selección de k, a través del Índice Silueta:

| k | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|
| IS | 0.2629 | 0.3209 | 0.2749 | 0.2806 | 0.2994 | 0.2825 | 0.2475 | 0.2475 | 0.2279 |

# Sección río branco
# K-means

| Caso | Algoritmo Implementado | Información Perdida | Dimensiones |
|---|---|---|---|
| 1 | KNIME | 0 % | 3447 |
| 2 | Scikit-Learn | 0 % | 3447 |
| 3 | KNIME | 10 % | 213 |
| 4 | Scikit-Learn | 10 % | 213 |
| 5 | KNIME | 20 % | 112 |
| 6 | Scikit-Learn | 20 % | 112 |
| 7 | KNIME | 40 % | 43 |
| 8 | Scikit-Learn | 40 % | 43 |
| 9 | KNIME | 60 % | 16 |
| 10 | Scikit-Learn | 60 % | 16 |
| 11 | KNIME | 80 % | 4 |
| 12 | Scikit-Learn | 80 % | 4 |
| 13 | KNIME | 90 % | 1 |
| 14 | Scikit-Learn | 90 % | 1 |

# Sección río branco
# Agrupamiento espectral

| Caso | Algoritmo Implementado | Información Perdida | Dimensiones |
|---|---|---|---|
| 1 | Scikit-Learn | 0 % | 3447 |
| 2 | Scikit-Learn | 10 % | 213 |
| 3 | Scikit-Learn | 20 % | 112 |
| 4 | Scikit-Learn | 40 % | 43 |
| 5 | Scikit-Learn | 60 % | 16 |
| 6 | Scikit-Learn | 80 % | 4 |
| 7 | Scikit-Learn | 90 % | 1 |

# Sección historia

- Preprocesamiento
  - Cantidad de documentos: 4.368.
  - Promedio de Palabras: 34.
  - TF-IDF: 0,9.
  - PCA: 0%, 10%, 20%, 40%, 60%, 80%, 90%.

- Selección de k, a través del Índice Silueta:

| k | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|
| **IS** | 0.2426 | 0.2866 | 0.2849 | 0.3336 | 0.3403 | 0.3209 | 0.2765 | 0.2761 | 0.2761 |

# Sección historia
# K-means

| Caso | Algoritmo Implementado | Información Perdida | Dimensiones |
|---|---|---|---|
| 1 | KNIME | 0 % | 4535 |
| 2 | Scikit-Learn | 0 % | 4535 |
| 3 | KNIME | 10 % | 469 |
| 4 | Scikit-Learn | 10 % | 469 |
| 5 | KNIME | 20 % | 250 |
| 6 | Scikit-Learn | 20 % | 250 |
| 7 | KNIME | 40 % | 91 |
| 8 | Scikit-Learn | 40 % | 91 |
| 9 | KNIME | 60 % | 33 |
| 10 | Scikit-Learn | 60 % | 33 |
| 11 | KNIME | 80 % | 8 |
| 12 | Scikit-Learn | 80 % | 8 |
| 13 | KNIME | 90 % | 3 |
| 14 | Scikit-Learn | 90 % | 3 |

# Sección historia
# Agrupamiento espectral

| Caso | Algoritmo Implementado | Información Perdida | Dimensiones |
|---|---|---|---|
| 1 | Scikit-Learn | 0 % | 4535 |
| 2 | Scikit-Learn | 10 % | 469 |
| 3 | Scikit-Learn | 20 % | 250 |
| 4 | Scikit-Learn | 40 % | 91 |
| 5 | Scikit-Learn | 60 % | 33 |
| 6 | Scikit-Learn | 80 % | 7 |
| 7 | Scikit-Learn | 90 % | 3 |

# Análisis de resultados

- Índice Silueta
- Índice Calinski-Harabasz
- Índice Dunn

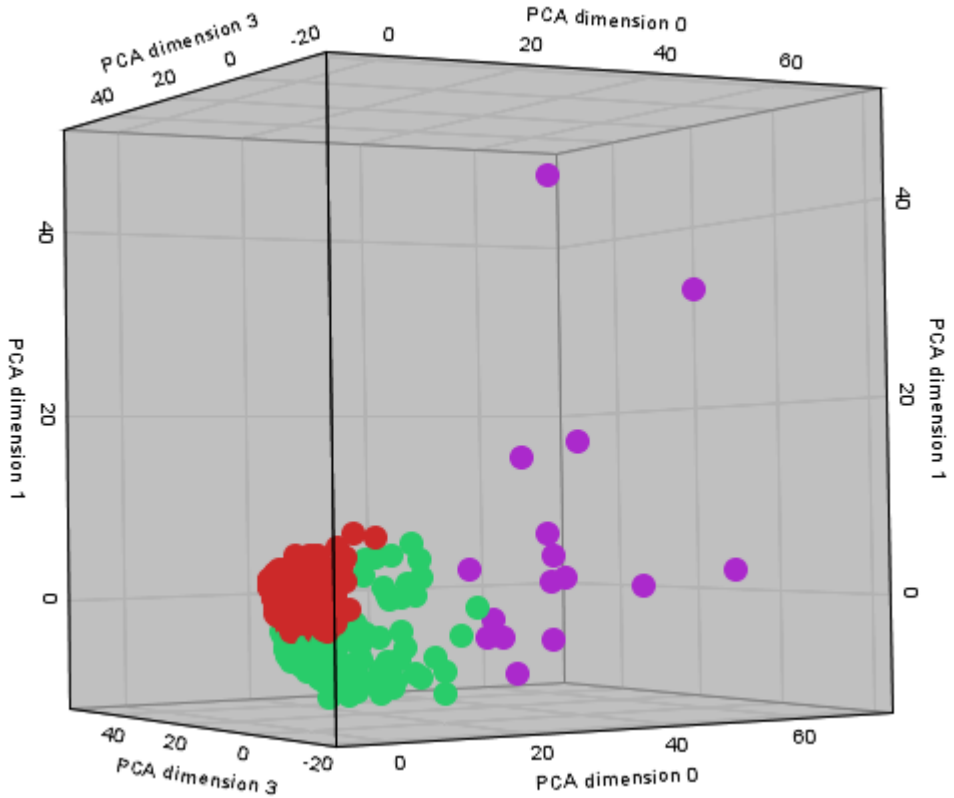
# Análisis de resultados

- Sección Rio Branco K-means

[1] Índice Silueta
[2] Índice Calinski-Harabasz
[3] Índice Dunn

| Caso | Algoritmo Implementado | Información Perdida | Dimensiones | IS[1] | CH[2] | ID[3] |
|---|---|---|---|---|---|---|
| 1 | KNIME | 0 % | 3447 | 0.663 | 56.109 | 0.284 |
| 2 | Scikit-Learn | 0 % | 3447 | 0.156 | 65.159 | 0.07 |
| 3 | KNIME | 10 % | 213 | 0.692 | 63.088 | 0.277 |
| 4 | Scikit-Learn | 10 % | 213 | 0.182 | 73..402 | 0.047 |
| 5 | KNIME | 20 % | 112 | 0.224 | 79.318 | 0.043 |
| 6 | Scikit-Learn | 20 % | 112 | 0.204 | 83.963 | 0.041 |
| 7 | KNIME | 40 % | 43 | 0.229 | 115.293 | 0.029 |
| 8 | Scikit-Learn | 40 % | 43 | 0.241 | 117.608 | 0.029 |
| 9 | KNIME | 60 % | 16 | 0.313 | 194.673 | 0.013 |
| 10 | Scikit-Learn | 60 % | 16 | 0.317 | 195.95 | 0.013 |
| 11 | KNIME | 80 % | 4 | 0.468 | 565.566 | 0.005 |
| 12 | Scikit-Learn | 80 % | 4 | 0.468 | 565.952 | 0.005 |
| 13 | KNIME | 90 % | 1 | - | - | - |
| 14 | Scikit-Learn | 90 % | 1 | - | - | - |

# VISUALIZACIÓN

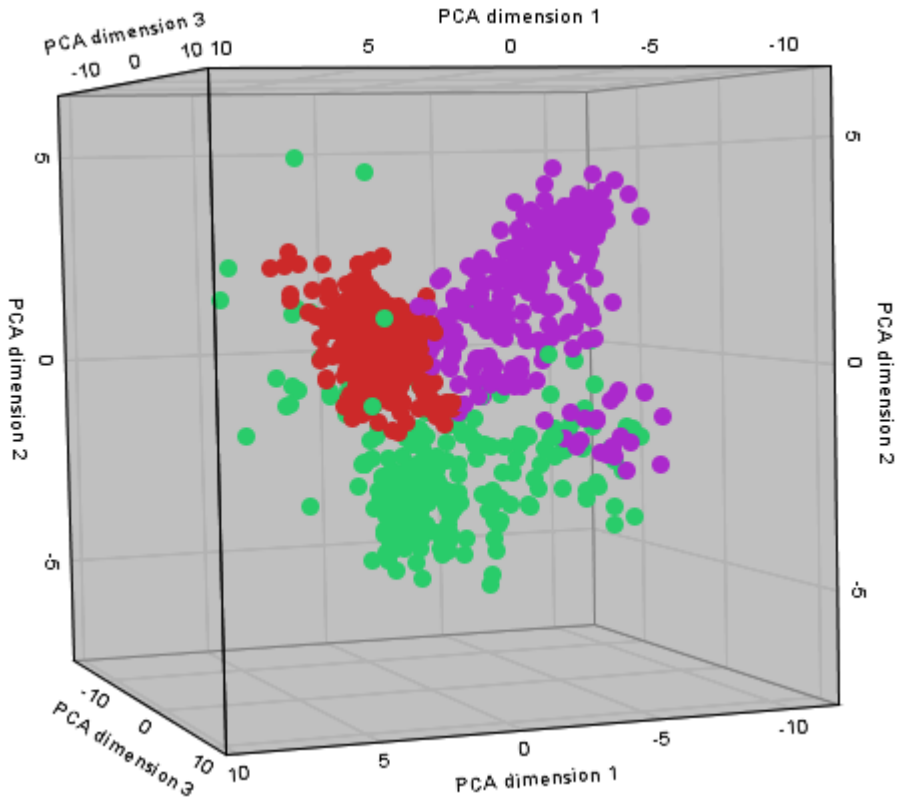


# Análisis de resultados

- Sección Rio Branco Agrupamiento Espectral

| Caso | Algoritmo Implementado | Información Perdida | Dimensiones | IS | CH | ID |
|---|---|---|---|---|---|---|
| 1 | Scikit-Learn | 0 % | 3447 | 0.321 | 234.821 | 0.046 |
| 2 | Scikit-Learn | 10 % | 213 | 0.329 | 463.798 | 0.015 |
| 3 | Scikit-Learn | 20 % | 112 | 0.35 | 484.25 | 0.014 |
| 4 | Scikit-Learn | 40 % | 43 | 0.416 | 550.224 | 0.006 |
| 5 | Scikit-Learn | 60 % | 16 | 0.533 | 772.152 | 0.004 |
| 6 | Scikit-Learn | 80 % | 4 | 0.716 | 1613.491 | 0.001 |
| 7 | Scikit-Learn | 90 % | 1 | - | - | - |

# VISUALIZACIÓN

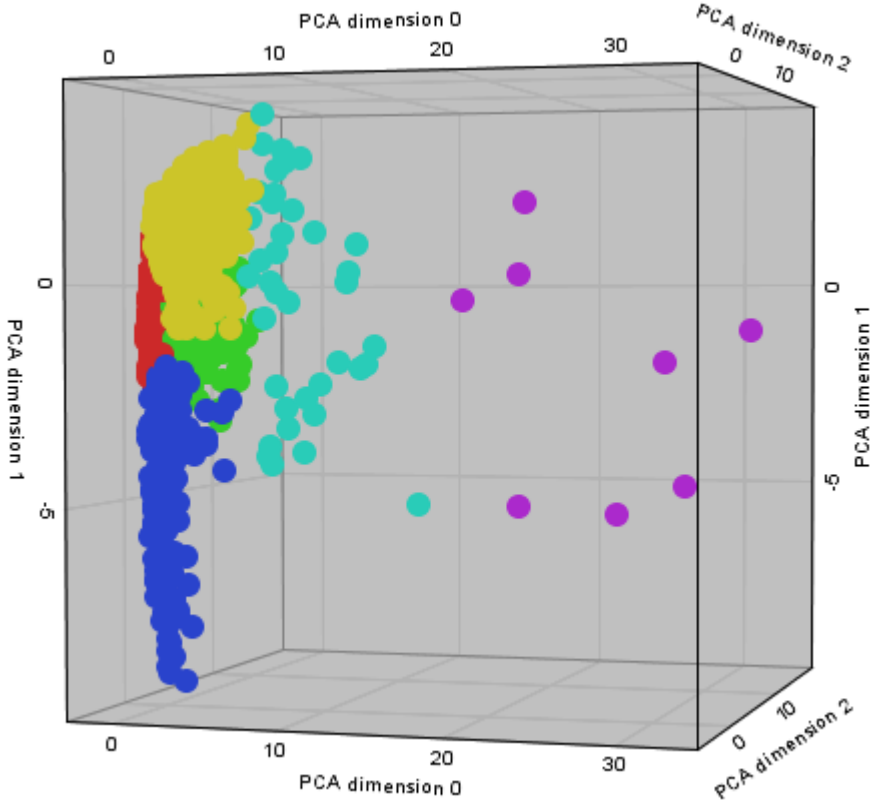


# Análisis de resultados

- Sección Historia K-means

| Caso | Algoritmo Implementado | Información Perdida | Dimensiones | IS | CH | ID |
|---|---|---|---|---|---|---|
| 1 | KNIME | 0 % | 4535 | 0.175 | 98.7 | 0.016 |
| 2 | Scikit-Learn | 0 % | 4535 | 0.005 | 100.384 | 0.016 |
| 3 | KNIME | 10 % | 469 | 0.184 | 109.746 | 0.012 |
| 4 | Scikit-Learn | 10 % | 469 | 0.094 | 120.178 | 0.005 |
| 5 | KNIME | 20 % | 250 | 0.178 | 119.844 | 0.01 |
| 6 | Scikit-Learn | 20 % | 250 | 0.106 | 138.8 | 0.011 |
| 7 | KNIME | 40 % | 91 | 0.197 | 161.563 | 0.007 |
| 8 | Scikit-Learn | 40 % | 91 | 0.149 | 188.046 | 0.006 |
| 9 | KNIME | 60 % | 33 | 0.241 | 260.873 | 0.005 |
| 10 | Scikit-Learn | 60 % | 33 | 0.221 | 316.382 | 0.004 |
| 11 | KNIME | 80 % | 8 | 0.385 | 800.691 | 0.001 |
| 12 | Scikit-Learn | 80 % | 8 | 0.365 | 971.626 | 0.003 |
| 13 | KNIME | 90 % | 3 | 0.461 | 1386.611 | 0.001 |
| 14 | Scikit-Learn | 90 % | 3 | 0.412 | 2302.682 | 0.001 |

# VISUALIZACIÓN

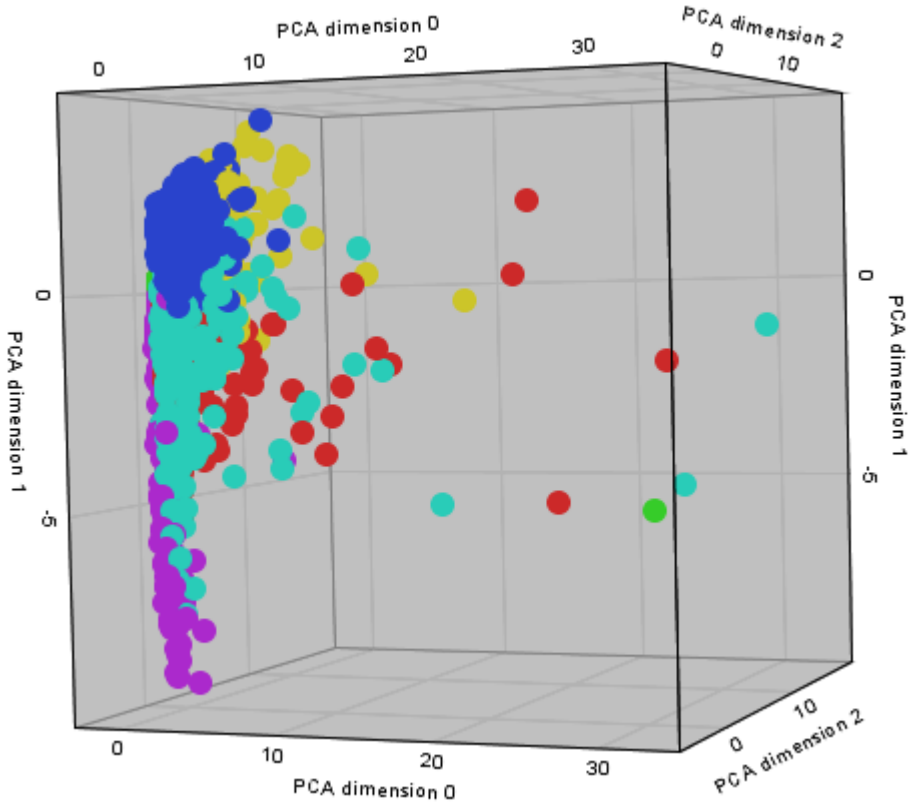


# Análisis de resultados

- Sección Historia Agrupamiento Espectral

| Caso | Algoritmo Implementado | Información Perdida | Dimensiones | IS | CH | ID |
|---|---|---|---|---|---|---|
| 1 | Scikit-Learn | 0 % | 4535 | 0.341 | 553.321 | 0.011 |
| 2 | Scikit-Learn | 10 % | 469 | 0.234 | 693.462 | 0.001 |
| 3 | Scikit-Learn | 20 % | 250 | 0.261 | 798.65 | 0 |
| 4 | Scikit-Learn | 40 % | 91 | 0.271 | 895.014 | 0 |
| 5 | Scikit-Learn | 60 % | 33 | 0.289 | 1093.755 | 0.001 |
| 6 | Scikit-Learn | 80 % | 7 | 0.508 | 2043.844 | 0 |
| 7 | Scikit-Learn | 90 % | 3 | 0.118 | 2510.562 | 0 |

# VISUALIZACIÓN



# Etiquetamiento de clusters

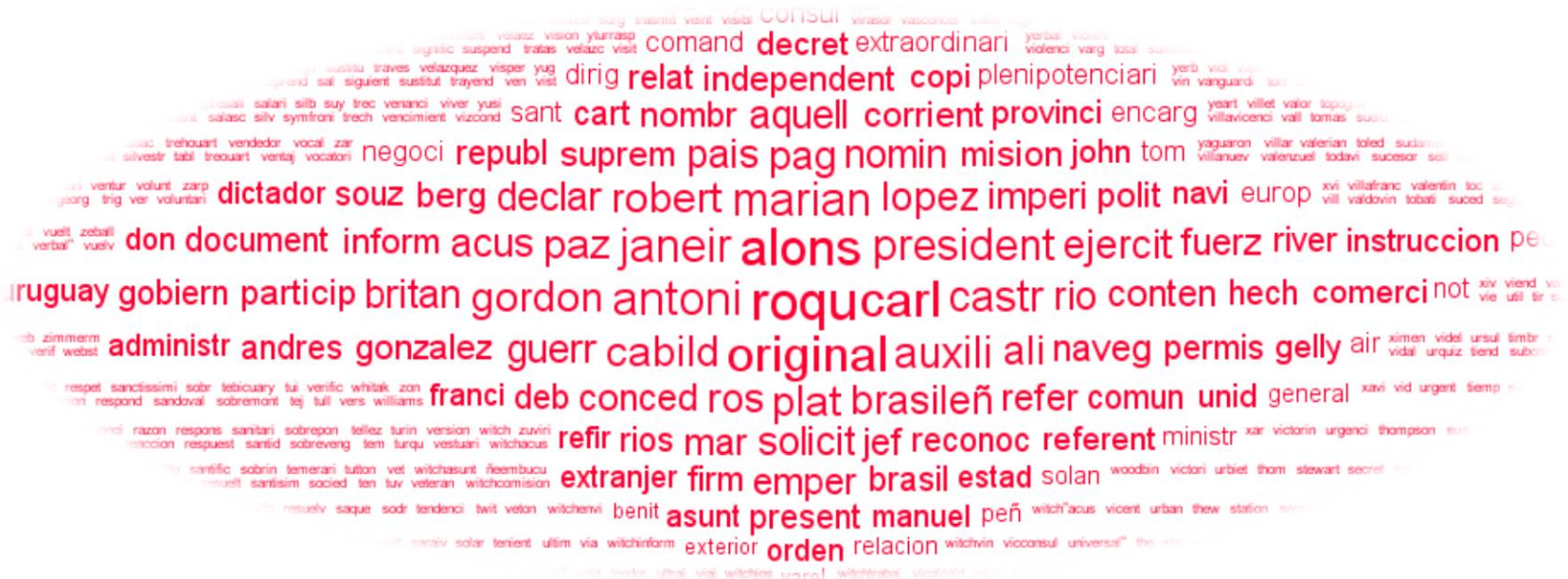
- Palabras Clave
  - la palabra debería ser relevante en el cluster en comparación a otras palabras en el cluster.
  - la palabra debería ser relevante en el cluster en comparación a su frecuencia en toda la colección.
- Frecuencia

$$F_{Total}(w) = \frac{F_{Clust}(w)}{F_{Coll}(w)}$$

# Sección río branco
# Agrupamiento espectral caso 6


Word Cloud Cluster 0

# Sección río branco
# Agrupamiento espectral caso 6




Word Cloud Cluster 1

# Sección río branco
# Agrupamiento espectral caso 6


Word Cloud Cluster 2

# Palabras clave sección río branco

| Caso | Cluster | Palabras Clave |
|---|---|---|
| K-means Caso 12 | 0 | intendent, peñ, indi, gobern, don, manuel, ros, comand, cart, paz |
| | 1 | deffandis, sagui, morinig, tellez, saenz, leit, aristegui, almiron, institut, lagun |
| | 2 | falcon, paulin, buenaventur, souz, benit, exterior, varel, soar, derqui, relacion |
| Espectral Caso 6 | 0 | intendent, bernard, fuerte, indi, rodriguez, peñ, gobern, comand, don, señor |
| | 1 | roqu, alons, carl, original, antoni, janeir, georg, marian, president, castr |
| | 2 | buenaventur, derqui, falcon, pendleton, exterior, benit, confeder, varel, relacion, hotham |

# Palabras clave sección historia

| Caso | Cluster | Palabras Claves |
|---|---|---|
| K-means Caso 14 | 0 | cedul, real, rein, audienci, despach, provision, obisp, haciend, relat, aut |
| | 1 | gasp, dictador, rodriguez, gobern, provinci, franci, cabild, defens, resolu, band |
| | 2 | lelat, viii, dier, gobiernopor, patrocini, itacocu, charcrelat, pirat, valerian, pidi |
| | 3 | juzg, paz, juez, decret, design, lopez, expedient, francisc, caus, president |
| | 4 | vag, ladron, agricol, vigil, volum, continu, militar, terren, correspondent, civil |
| | 5 | aprest, vocal, exaccion, ibañez, iii, inclu, plan, pined, teng, velazc |
| Espectral Caso 6 | 0 | paz, design, decret, juez, expedient, present, asuncion |
| | 1 | paraguay, gobern, trat, provinc, aut, band, air, juan, copi, cabild |
| | 2 | criminal, civil, caus, apel, agricol, inform, juzg, superior, campañ, carg |
| | 3 | cedul, real, rein, despach, provision, obisp, haciend, relat, sant, copi |
| | 4 | gasp, dictador, lopez, solan, rodriguez, franci, president, antoni, carl, orden |
| | 5 | indi, encomiend, acuerd, puebl, defens, correspondent, libr, cabild, tabac, san |

# Conclusiones

- Efectividad de los modelos propuestos, donde el Agrupamiento Espectral es ligeramente superior al K-means.
- Desarrollo de un modelo capaz de predecir los clusters para documentos a ser agregados a una Sección existente.
- Generación de palabras clave para el etiquetado de clusters.
- Juicio de Expertos determinó un alto grado de relación entre documentos.
- Margen para introducir mejoras.

# Aportes a la ciencia

- Comprobación y validación de técnicas k-means y agrupamiento espectral para el agrupamiento de documentos históricos y/o sin previa clasificación.
- Metodología de trabajo de alta modularidad.
- Alternativas para distintos escenarios.

# Aportes al archivo nacional de asunción

- Facilitar las tareas de los investigadores
- Facilitar las tareas de digitalización y mantenimiento del sistema archivístico

# Trabajos futuros

- Integración con AtoM.
- Variaciones en las técnicas de agrupamiento.
- Aplicación de técnicas con énfasis en semántica y ontología.
- Variaciones en la predicción de clusters para nuevos documentos.
- Extensión del análisis a otros tipos de documentos.

# Muchas Gracias!