Minería de Texto. Aplicación a la clasificación de Proyectos de Trabajos Finales de Grado de la FPUNE.

Lourdes Beatriz Delgado González[1] y **Gabriela Matilde Bobadilla de Almada**[2].
Facultad Politécnica, Universidad Nacional del Este.
Ciudad del Este, Paraguay.
lourdes@fpune.edu.py[1], gaby@fpune.edu.py[2]

Resumen

En la Facultad Politécnica de la Universidad Nacional del Este (FPUNE), en los últimos años han estado aumentando continuamente los proyectos de trabajos finales de grado (PTFG) que dan inicio al trabajo final de grado (TFG)propiamente dicho. Debido a esto, una de las dificultades presentadas es identificar las duplicaciones o similitudes entre los PTFG en forma inmediata a través de controles manuales. A raíz de esta dificultad se realizó una investigación sobre medios que posibiliten filtrar automáticamente los PTFG. Se optó por emplear el método de minería de texto, con el cual se logró clasificar a 23 áreas de estudio de TFG, con 92 % de agrupaciones correctas, la máxima similitud encontrada entre dos proyectos de TFG es 79 % y la mínima 30 %.
Descriptores: minería de texto, clasificación automática, agrupamiento de documentos textuales.

Abstract

In the Polytechnic Faculty, Eastern National University (FPUNE), projects of graduation work (PTFG) presented for approving the graduating work (TFG) research have been continuously increasing. Because of this, one of the difficulties presented is to promptly identify duplication or similarities between PTFGs through manual control. This research work seeks to solve this difficulty through means that allow automatically filtering PTFGs for discrimination and classification pourpose. Text mining was applied which rendered 23 areas of application, with 92 % correct groupings, the highest similarity found between two TFG projects is 79 %, the lowest being 30 %.
Keywords: text mining, automatic classification, text clustering.

1. Introducción.

Para la obtención del título de grado de toda carrera de la FPUNE es requisito fundamental la presentación y defensa pública del Trabajo Final de Grado (TFG)[1], este proceso inicia con la presentación del Proyecto de TFG aprobado por el profesor de la cátedra. Con el transcurrir del tiempo la cantidad de estos proyectos van en aumento lo cual dificulta del control de repetición de temas o de su excesiva similitud por medios manuales, a raíz de esta situación se realizó una investigación sobre medios que permitan filtrar automáticamente los PTFG, planteando finalmente la utilización del método de minería de texto [2] ,[3], [4].

La **minería de texto** (**MT**) puede ser ampliamente definida como un proceso intensivo en conocimiento donde el usuario interactúa con una colección textual (informaciones no estructuradas) mediante el uso de un conjunto de herramientas de análisis [5]. Es un conjunto de métodos usados para navegar, organizar, encontrar y descubrir información en bases textuales. Puede ser vista como una extensión del área de *Data Mining* (Minería de Datos, MD), enfocada en el análisis de textos [6].Es también conocida como Descubrimiento de Conocimientos en Textos (*Knowledge Discovered in Texts* - KDT) [5, 6, 7].

1.1. Objetivos.

1.1.1. Objetivo general.

Clasificar proyectos de trabajos finales de grado de carreras de la FPUNE por contenido, aplicando el método de minería de texto.

1.1.2. Objetivos específicos.

- Recopilar información de métodos de minería de texto.
- Seleccionar *software* especializado de minería de texto.

– Seleccionar el método de minería de texto adecuado a los datos obtenidos.

2. Método.

La metodología utilizada es el KDT, que realiza el descubrimiento de conocimiento en datos no estructurados, dividido en tres grandes fases: pre procesamiento, procesamiento y post procesamiento; como puede observarse en la figura 1 (Fig. 1), elaborada en base al diagrama obtenido en [7].

El KDT se basa en el proceso KDD (*Knowledge Discovered in Databases*, Descubrimiento de Conocimiento en Base de Datos). Según [8], KDD es el proceso más ampliamente utilizado en minería de datos, que dispone de tecnología de recolección, almacenamiento y gerenciamiento para grandes bases de datos estructuradas.

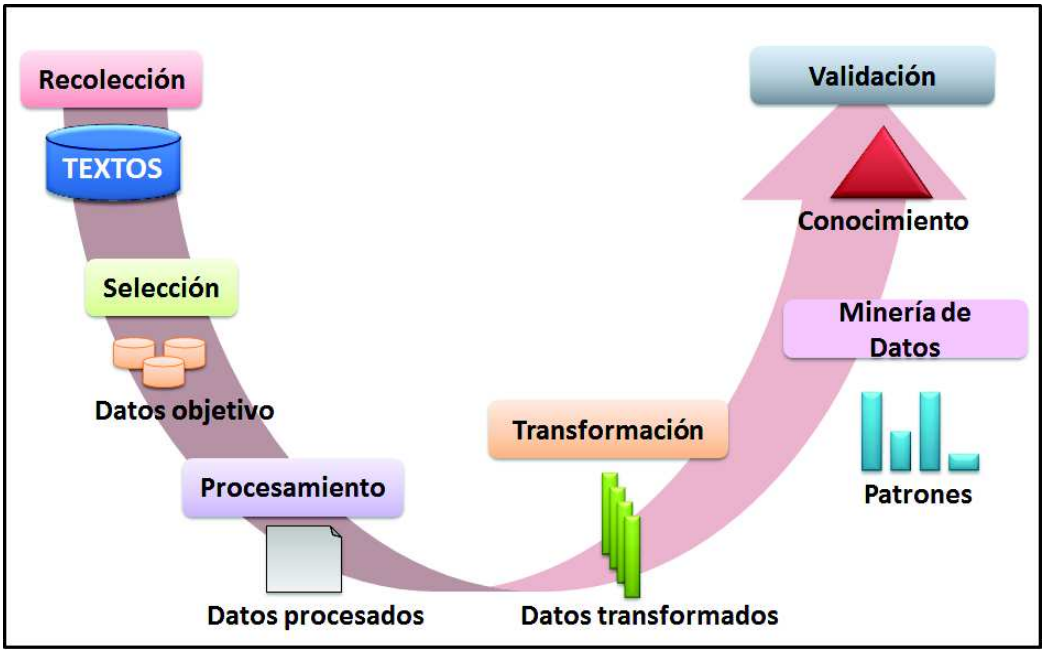

Figura 1. Proceso KDT.

Fue utilizado el *software RapidMiner Studio* en su versión 6.1 *Starter Edition* con el paquete de *Text Mining Extension* para el procesamiento de texto.

El trabajo se realizó en las tres fases del KDT: pre procesamiento, procesamiento y post procesamiento.

2.1. Pre procesamiento.

En esta etapa fueron preparados los datos, el tratamiento incial sobre los mismos influyen en los resultados, por ello se procedió a reducir la dimensionalidad del vector a ser generado eliminado términos o palabras irrelevantes de la colección textual para la obtención de un subconjunto de términos representativos, de acuerdo a [2, 5, 9, 10]. Esta etapa se realizó siguiendo los siguientes pasos:

2.1.1 Recolección.

Para obtener la colección textual en el dominio de la aplicación del conocimiento se procedió a la conversión de los PTFG a documentos digitales, ya que estos se encontraban disponibles únicamente en formato impreso.

Se consideró los PTFG de los años 2009 al 2011, tomando una muestra de 64 PTFG que presentaron textos legibles al digitalizarse. Estos documentos fueron escaneados y convertidos a formato de texto editable a través de la técnica de reconocimiento óptico de caracteres (OCR).

2.1.2 Selección.

Los documentos digitalizados son tratados con la herramienta *RapidMiner*, mediante el módulo *Text Mining Extension* aplicando el operador *Process Documents from Files* (Fig. 2). Estos documentos fueron inicialmente agrupados en tres clases principales, las cuales corresponden a las carreras: Licenciatura en Análisis de Sistemas, Ingeniería de Sistemas y Licenciatura en Turismo.

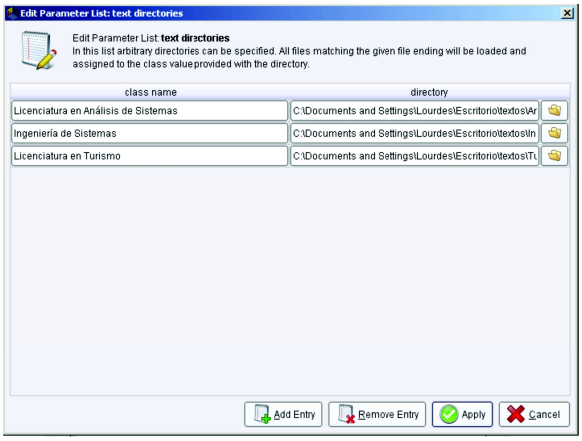

Figura 2. Carga de archivos por clases.

2.1.3. Procesamiento.

Según [11] si los documentos ya están identificados, la principal tarea es efectuar la eliminación de ruidos y asegurar que la muestra sea de buena calidad. Esta tarea requiere de un extremo cuidado ya que la intervención humana puede comprometer la integridad de los datos en este proceso.

Teniendo en cuenta este aspecto fue considerado importante utilizar sólo los siguientes operadores del módulo *Text Mining Extension:*

- ***Tokenize:*** encargado de la división de los textos en frases, palabras, símbolos u otros elementos significativos llamados *tokens*. Este operador tiene un parámetro que especifica cómo identificar palabras en el texto. Para la separación de palabras se utilizó la opción *non letters* (caracteres que no sean letras), otras opciones podrían ser caracteres especiales y expresiones regulares. De los 64 documentos de texto se extrajo un total de 10375 términos, como puede apreciarse en la figura (Fig.3).

Word	Attribute Name	Total Occurences	Document Occur...
A	A	41	27
AC	AC	3	1
ACID	ACID	2	1
ACTIVIDADES	ACTIVIDADES	9	9
ADMINISTRATIVOS	ADMINISTRATIVOS	1	1
AG	AG	2	1
AGPS	AGPS	1	1
AL	AL	2	2
ALCANCE	ALCANCE	13	13

Figura 3. Parte del resultado de tokenización.

- ***Transform Case:*** teniendo en cuenta que el proceso de *tokenización* es *case sensitive* (sensible a mayúsculas), se utiliza este operador para transformar términos en mayúsculas a minúsculas, como se procedió en el presente trabajo, sin embargo, se dispone igualmente de la opción de transformación a mayúsculas. Una vez ejecutado este operador, la cantidad de términos disminuyó a 9006. En la figura (Fig. 4) se puede ver la captura de pantalla correspondiente a este operador.

Word	Attribute Name	Total Occurences	Document Occure...
a	a	1604	64
abajo	abajo	1	1
abandonan	abandonan	1	1
abandono	abandono	1	1
abanico	abanico	1	1
abaratamiento	abaratamiento	1	1
abaratando	abaratando	1	1
abarca	abarca	8	7
abarcar	abarcar	4	4
abarcaran	abarcaran	1	1
abarcaremos	abarcaremos	1	1
abarcarse	abarcarse	1	1
abarcará	abarcará	4	3

Figura 4. Transformación de términos a minúsculas.

- ***Filter Stopwords:*** operador encargado de filtrar los términos sin significado (*stopwords*), tales como artículos, preposiciones, pronombres, entre otros. Existen diferentes filtrados para diferentes idiomas.

  Como el módulo *Text Mining Extension* de la herramienta *RapidMiner* no cuenta con filtrado para español, se creó manualmente un listado de todos aquellos términos que no fueren relevantes para la extracción del conocimiento deseado. La selección de los términos fue realizada teniendo en cuenta la colección textual utilizada luego de un análisis de cada uno de los documentos. Una vez generado el archivo de texto plano a partir de la lista de términos no relevantes, se procedió a la eliminación de estos términos mediante el operador *Filter Stopwords* (*Dictionary*). Como resultado de la ejecución de este operador, el total de términos fue reducido a 8844. En la figura (Fig. 5) se puede observar una captura de pantalla de este operador.

Word	Attribute Name	Total Occurences	Document Occure...
sofisticadas	sofisticadas	1	1
software	software	194	42
sola	sola	5	5
solicitados	solicitados	2	2

Figura 5. Resultado del proceso de eliminación de términos no relevantes.

- ***Stemming:*** su función es la reducción de términos a su base o raíz. Se utilizó este operador para reducir más la cantidad de términos. Para tal efecto, el operador emplea el algoritmo *snowball* para el idioma español. Con esto se consiguió reducir el listado de términos a 4727. En la figura(Fig.6) se ofrece una captura de pantalla correspondiente a este operador.

Word	Attribute Name	Total Occurences	Document Occurences
acaray	acaray	1	1
acarr	acarr	1	1
acarre	acarre	2	2
acced	acced	32	17
accelerat	accelerat	6	1
acces	acces	84	33

Figura 6. Resultado de la aplicaciónn del operador stemming.

2.2. Procesamiento.

Para aplicar los algoritmos de minería de datos, los datos de entrada deben estar en un formato estructurado o transformados al formato adecuado para ser sometidos a esos algoritmos. Una buena representación de los textos en esta etapa es fundamental para el buen desempeño de los algoritmos.

2.2.1. Transformación.

Una vez seleccionados los términos más representativos de la colección textual se realizó la estructuración de los documentos utilizando el Modelo Espacio Vectorial (*Vector Space Model* - VSM,)el cual representa los textos como un vector donde los elementos del vector indican la frecuencia de las palabras dentro del texto [2]. A partir de aquí, los términos pasan a denominarse atributos. En la figura (Fig. 7) se puede observar una parte de la matriz del VSM.

aban	abandon	abarat	abarc
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0.009
0	0	0	0
0	0	0.022	0
0	0	0	0.012

Figura 7. Representación Atributo - Valor.

A cada atributo se ha asignado un peso que fue calculado con la Ec.1 [5].

$$TF - IDF(t) = TF(d,t) * IDF(t) \qquad (1)$$

Este valor combina dos medidas diferentes, la frecuencia de la palabra (*Term Frequency*, TF) y la frecuencia inversa del documento (*Inverse Document Frequency*, IDF)[5, 4]:

$$TF(d,f) = \sum \omega(t,d) \qquad (2)$$

$$IDF = \log(\frac{ND}{DF(t)}) \qquad (3)$$

Donde:

t = término,
d = documento,
$\omega(t,d)$ = ocurrencia del término t en el documento d,
ND = número total de documentos, y
DF = número de documentos en los que aparece el término t en toda la colección.

La ecuación (Ec. 2) muestra que la TF es la suma de todas las ocurrencias o el número de veces que aparece un término en un documento, mientras que la ecuación (Ec. 3) indica que el factor IDF de un término es inversamente proporcional al número de documentos en los que aparece dicho término, lo cual significa que su peso es mayor mientras aparezca en menos documentos.

2.2.2. Minería de Datos.

Para realizar el proceso de clasificación fueron utilizadas las técnicas del agrupamiento de documentos, cuyo objetivo es disponer una colección de documentos desorganizada, en un conjunto de grupos que contengan documentos con temas similares [4].
Para que esto sea realizado se parte del principio de la *Hipótesis de Agrupamiento*, este principio dice que objetos semejantes y relevantes a un mismo asunto tienden a permanecer en un mismo grupo, debido a que poseen atributos en común[4].

Una vez analizados todos los algoritmos de agrupación disponibles en el *software RapidMiner* se decidió seleccionar el algoritmo *K-means* del método particional debido a su simplicidad y por obtener mejores resultados a un menor costo computacional.

El algoritmo *K-means* es una técnica clásica de agrupamiento muy utilizado en colecciones textuales [12]. En el algoritmo 1 se observa el pseudocódigo de k-means para el agrupamiento de documentos textuales [9, 13].

Algoritmo 1. K-means

El criterio de parada se da cuando ya no ocurren alteraciones en el agrupamiento, es decir, la solución converge para una determinada partición. Otro criterio de parada podría ser el número máximo de iteraciones. Durante las iteraciones, el objetivo es minimizar la función de error. Con este

Data: $D = (d_1, d_2, \ldots, d_n)$: colección textual
Result: $P = (G_1, G_2, \ldots, G_k)$: partición con K grupos
Seleccionar aleatoriamente k documentos como centroides iniciales;
repeat
  for *documento* $d \in D$ **do**
    calcular la similitud de d para cada centroide C; atribuir d al centroide más próximo;
  Recalcular el centroide de cada grupo;
until *alcanzar criterio de parada*
return

objetivo, el *k-means* intenta separar el conjunto de documentos disminuyendo la variabilidad interna de cada grupo y, por consiguiente, aumentar la separación entre los grupos.

Dentro de los parámetros que se deben asignar manualmente al algoritmo en el *software* utilizado, los más importantes son el número de grupos (K), la distancia y el número máximo de iteraciones. El resultado final depende de los valores elegidos para estos parámetros.

En este trabajo fue utilizada la distancia Similitud Coseno[5, 9], teniendo en cuenta que se trata de documentos textuales y las agrupaciones realizadas se basan en la similitud existente entre un documento y otro.

Para medir la semejanza entre dos documentos por el método Similitud Coseno, se utiliza la definición de producto escalar tomada del Álgebra vectorial definido como:

$$d_1 \cdot d_2 = |d_1| \times |d_2| \times \cos(d_1, d_2) \tag{4}$$

Donde:
d_1 es vector 1
d_2 es vector 2
$|d_1|$ es longitud del vector 1
$|d_2|$ es longitud del vector 2
$\times$ indica multiplicación
$\cos(d_1, d_2)$ es el coseno del ángulo subtendido por d_1 y d_2.

A partir de esta definición, despejando de (Ec. 4) $\cos(d_1, d_2)$, y definiendo:
vector-documento $d_1 = \sum_{i=1}^{|V|} \omega(t_i, d_1)$,

vector-documento $d_2 = \sum_{i=1}^{|V|} \omega(t_i, d_2)$,
longitud vector-documento
$d_1 = \sqrt{\sum_{i=1}^{|V|} \omega(t_i, d_1)^2}$,
longitud vector-documento
$d_2 = \sqrt{\sum_{i=1}^{|V|} \omega(t_i, d_2)^2}$

Se obtiene la función Similitud Coseno (Ec. 5):

$$\cos(d_1, d_2) = \frac{d_1 \cdot d_2}{|d_1| \times |d_2|} = \frac{\sum_{i=1}^{|V|} (\omega(t_i, d_1) \times \omega(t_i, d_2)}{\sqrt{\sum_{i=1}^{|V|} \omega(t_i, d_1)^2} \times \sqrt{\sum_{i=1}^{|V|} \omega(t_i, d_2)^2}} \tag{5}$$

De esta manera, a medida que el valor del ángulo se aproxima a 0 y el coseno se aproxima a 1 entonces indica que los documentos son más similares entre sí y por el contrario, si el valor del ángulo es 90 y el coseno 0 entonces los dos documentos no comparten ningún término.

2.3. Post procesamiento.

Según [14] la validación puede ser realizada de forma subjetiva, utilizando el conocimiento de un especialista del dominio, o de forma objetiva por medio de índices estadísticos que indican la calidad de los resultados.

La validación del resultado en un agrupamiento, en general, se realiza por medio de índices estadísticos que cuantifica alguna información sobre la calidad de un agrupamiento [11]. El uso de las técnicas de validación en los resultados del agrupamiento es una actividad importante, una vez que los algoritmos encuentran grupos en los datos, independientemente de ser reales o no. La medida utilizada para la validación de los resultados es el Valor -F (Ec. 8), el cual es la combinación de dos medidas: Precisión P (Ec. 6) y Exhaustividad E(Ec.7).

$$P = \frac{a}{a + c} \tag{6}$$

$$E = \frac{a}{a + b} \tag{7}$$

$$Valor - F = \frac{2 \cdot P \cdot E}{P + E} \tag{8}$$

Donde:
a= Número de documentos pertenecientes al grupo y que están en el grupo.
b= Número de documentos que no pertenecen al grupo pero están asignados.
c= Número de documentos pertenecientes al grupo, pero no están en el grupo.

3. Resultados.

3.1. Pre procesamiento.

Durante esta etapa se realizó la limpieza de los textos, eliminándose aquellas palabras redundantes e innecesarias que no aportan significado alguno al proceso de agrupamiento, esto fue realizado utilizando la herramienta *RapidMiner* con operadores propios del paquete *Text Mining Extension*.

Inicialmente la colección textual estaba compuesta por 10375 términos, luego de la aplicación de técnicas de *stopwords* y *stemming* la cantidad de términos se redujo a 4727, equivalente a 54,4 % de la cantidad inicial, en la tabla (Tabla 1) se recogen los porcentajes de reducción de términos en cada una de las técnicas utilizadas.

Tabla 1. Representación de la matriz Atributo - Valor.

Técnica de Procesamiento	Nº de Términos	% de Reducción
Tokenize	10375	Valor Inicial
Transform Case	9006	13,2 %
Stopwords	8844	14,8 %
Stemming	4727	54,4 %

3.2. Procesamiento.

Fueron realizadas siete pruebas, cada una con un valor K de grupos, los resultados obtenidos en cada una de las pruebas realizadas fueron evaluados utilizando el Valor - F (Ec. 8)consistente en la combinación de las medidas de Precisión (Ec.6) y Exhaustividad (Ec. 7).

En la figura (Fig.8) se ilustra el Valor-F (en porcentaje) obtenido para cada valor de K, pudiéndose observar que el mayor porcentaje de aciertos en los grupos realizados corresponde al valor K=27. Los agrupamientos pueden observarse en la figura (Fig.9).

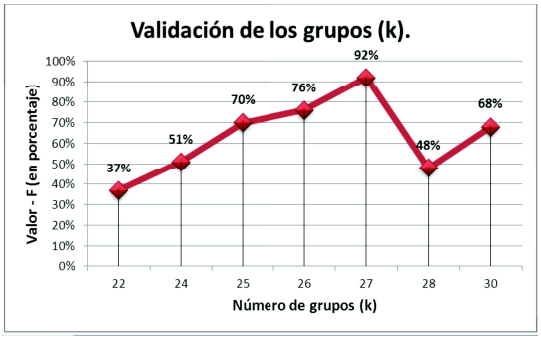

Figura 8. Porcentaje de validación de los grupos.

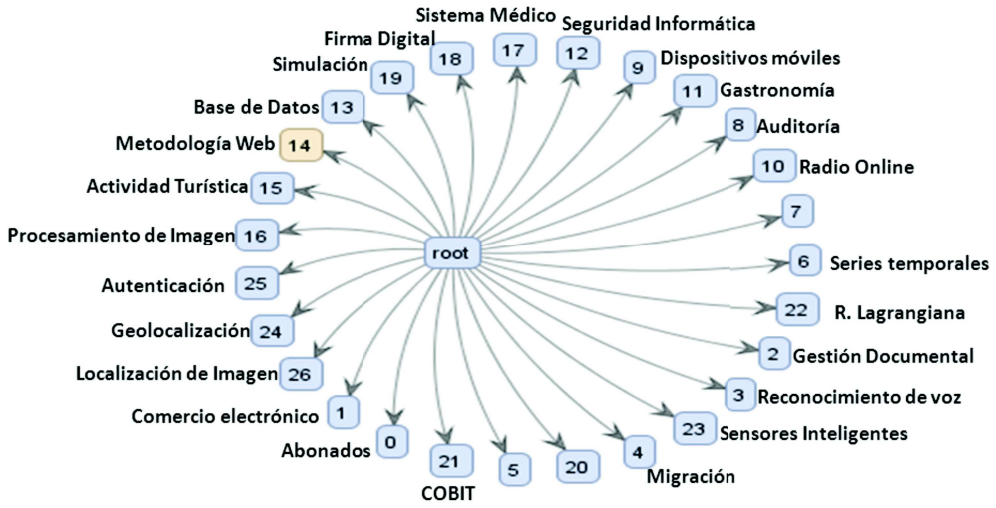

Figura 9. Resultados del agrupamiento de la colección (root) con K = 27.

Teniendo en cuenta que cada grupo consiste en documentos altamente similares entre sí, en la tabla (Tabla. 2) se pueden observar los datos de proyectos de TFGs con porcentaje de similitud umbral igual a 32 % en los grupos establecidos durante la quinta prueba. En esta prueba se obtuvieron los mayores porcentaje de Valor-F en los agrupamientos (Tabla 2).

Tabla 2. Similitud entre dos documentos.

Grupo	Clase	ID Doc.	ID Doc.	Similitud (%)
2	Gestión Documental	8	49	79 %
25	Firma Digital	51	57	57 %
15	Gestión Hotelera	30	31	49 %
21	COBIT	29	52	48 %
17	Sistema Médico	43	44	45 %
3	Reconocimiento de Voz	13	59	43 %
7	Servicios por Celular	12	58	40 %
17	Sistema Médico	10	44	38 %
15	Actividad Turística	35	36	36 %
1	Comercio Electrónico	3	7	32 %

En la esta tabla se observa por ejemplo que los proyectos de TFG con ID 8 y 49 correspondientes al segundo grupo cuyo contenido trata sobre Gestión Documental posee un 79 % de similitud. Averiguaciones realizadas por la autora, referentes a los contenidos de los proyectos, condujeron al descubrimiento de la causa de este alto porcentaje de similitud: en este caso se trata de trabajos realizados en conjunto uno de la carrera de Ingeniería de Sistemas y el otro de Licenciatura en Análisis de Sistemas; esto agrega confiabilidad al método utilizado en el trabajo.

4. Conclusión.

Con la utilización del método de minería texto se logró clasificar a 23 áreas de estudio de proyectos de TFG, con 92 % de agrupaciones correctas. Averiguaciones realizadas por la autora, referentes a los contenidos de los proyectos, condujeron al descubrimiento de la causa del más alto porcentaje de similitud: casi 80 % entre dos documentos, en este caso se trata de trabajos realizados en conjunto, uno de la carrera de Ingeniería de Sistemas y el otro de Licenciatura en Análisis de Sistemas; esto agrega confiabilidad al método utilizado en el trabajo.

Referencias bibliográficas

[1] Facultad Politécnica. Universidad Nacional del Este. Reglamento Interno. Versión 1/2014. [en línea] `http://www.fpune.edu.py/web/docs/reglamentos/reglamento_2014_v1.pdf`

[2] Miner, G., Elder, J.; Hill, T. *Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications*. Academic Press. 2012

[3] Rocha R., Cobo A. *Automatización de procesos de categorización jerárquica documental en las organizaciones*. 2010 [en línea] `http://revistas.concytec.gob.pe/scielo.php?pid=S2070836X2010000100013&script=sci_arttext`

[4] Passini M, *Mineração de textos para organização de documentos em centrais de atendimento*. 2012 [en línea] `http://wwwp.coc.ufrj.br/teses/mestrado/Novas_2012/TESES/PASSINI_MLC_TM.pdf`

[5] Feldman R., Sanger J, *The Text Mining Handbook. Advanced Approaches in Analyzing Unstructured Data.*Cambridge University Press, 2007

[6] Aranha C. , Passos E. *A Tecnologia de Mineração de Textos*. RESI-Revista Eleronica de Sistemas de Informacão, 2006

[7] Nogueira E., Lago D. *Mineração de Textos*. 2008 [en línea] `http://sare.anhanguera.com/index.php/rcext/article/view/413`

[8] Rezende S., *Sistemas inteligentes: fundamentos e aplicações*. Barueri, SP: Manole, 2003.

[9] Rezende S., Marcacini R., Moura M.,*O uso da Mineração de Textos para Extração e Organização não Supervisionada de Conhecimento*. Revista de Sistemas de Informação da FSMA n. 7. 2011

[10] Correia E., *Tecnicas de Data e Text Mining para anotação de un arquivo digital*, Tesis de Maestria, Universidad de Aveiro - Brasil.

[11] Xu R., Wunsch D., *Survey of Clustering Algorithms*. IEEE Transactions on neural networks. 2005

[12] Steinbach M.,Karypis G. Kumar V. , *A Comparison of Document Clustering Techniques*. 2007 [en línea] `http://www.cs.cmu.edu/~dunja/KDDpapers/Steinbach_IR.pdf`

[13] Aggarwal C., Zhai C., *Mining Text Data.*Springer.2007

[14] Nunes G., *Uso da mineração de textos na análise exploratória de artigos científicos* . 2012 [en línea] `http://www.icmc.usp.br/CMS/Arquivos/arquivos\_enviados/BIBLIOTECA\_113\_RT\_383.pdf`