ESCUELA POLITÉCNICA NACIONAL
VICERRECTORADO DE INVESTIGACIÓN Y PROYECCIÓN SOCIAL

PROYECTO DE INVESTIGACIÓN INTERNOS SIN FINANCIAMIENTO O AUTOGESTIONADOS

ANEXO 1 - DATOS INFORMATIVOS

Fecha de presentación (dd/mm/aa): 05/11/2019

Título del proyecto: *(Revisar la guía para la presentación de las propuestas de los proyectos de investigación)*

Segmentación y clasificación automática de las principales regiones de las páginas de artículos científicos digitales mediante técnicas de visión artificial y aprendizaje automático.

TIPOS DE INVESTIGACIÓN

Investigación Básica ☐ Investigación Aplicada ☒

DEPARTAMENTO(S) Y/O INSTITUTO(S):

1. Departamento de Electrónica, Telecomunicaciones y Redes de Información.

LINEA(S) DE INVESTIGACIÓN (verificable en el SAEW):

1. Multimedia

RESUMEN DE INFORMACIÓN DEL DIRECTOR Y COLABORADORES

Director

Apellidos y nombres	No.Cédula	HSS*	Departamento	Título de mayor nivel y mención.
Grijalva Arévalo Felipe Leonel	1710847441	6	Departamento de Electrónica, Telecomunicaciones y Redes de Información	Ph.D en Ingeniería Eléctrica. Área de Concentración: Ingeniería de Computación

Colaborador(es)

Apellidos y nombres	No.Cédula	HSS*	Departamento	Título de mayor nivel y mención.
–	–	–	–	–

Colaboradores Externos

Apellidos y nombres	Cédula	HSS*	Institución	Título de mayor nivel y mención.	Email
Acuña Acurio Byron Alejandro	1718688813	N/A	Universidad Estatal de Campinas (UNICAMP). Brasil.	Ingeniero en Electrónica, Automatización y Control.	bacuna@decom.fee.unicamp.br Se anexa hoja de vida.
Rodríguez Guerra Juan Carlos	1721234902	N/A	Analog Devices. Irlanda.	PhD en Ingenieria Electrica. Mencion: Electronica de potencia.	JuanCarlos.Rodriguez@analog.com Se anexa hoja de vida.

* HSS = Horas Semana Semestre

PROYECTO DE INVESTIGACIÓN INTERNOS SIN FINANCIAMIENTO O AUTOGESTIONADOS

ANEXO 2 - DETALLES DE LA PROPUESTA

Investigación Básica ☐ Investigación Aplicada ☒

DEPARTAMENTO(S) Y/O INSTITUTO(S):

1. Departamento de Electrónica, Telecomunicaciones y Redes de Información

LINEA(S) DE INVESTIGACIÓN:

1. Multimedia

DISCIPLINA CIENTÍFICA (Marque X, solamente una opción)	
Ciencias Naturales y Exactas;	
Ingeniería y Tecnologías;	X
Ciencias Médicas;	
Ciencias Agrícolas;	
Ciencias Sociales;	
Humanidades	

OBJETIVO SOCIOECONÓMICO (Marque X, solamente una opción)	
Exploración y explotación del medio terrestre;	
Ambiente;	
Exploración y Explotación del espacio;	
Transporte, telecomunicaciones y otras infraestructuras;	X
Energía;	
Producción y tecnología industrial;	
Salud;	
Agricultura;	
Educación;	
Cultura, ocio, religión y medios de comunicación;	
Sistemas políticos y sociales, estructuras y procesos;	
Defensa;	
Avance general del conocimiento: I+D financiada con los Fondos Generales de Universidades (FGU);	
Avance general del conocimiento: I+D financiados con otras fuentes.	

1 **Proyecto de Investigación**

Título:
Segmentación y clasificación automática de las principales regiones de las páginas de artículos científicos digitales mediante técnicas de visión artificial y aprendizaje automático.

Resumen del proyecto (máximo 200 palabras)

La correcta segmentación y clasificación automática de las principales regiones de un documento digital es una etapa fundamental en el entendimiento automático computarizado de documentos, control de plagio, desarrollo de tecnologías de apoyo para personas con discapacidad visual y minería de datos.
En la actualidad existe un alto interés por parte de la comunidad científica y empresarial en proyectos de investigación que permitan segmentar y clasificar los formatos de los artículos científicos, porque estos documentos tienen una mayor complejidad para el reconocimiento automático computarizado, debido a que pueden incluir en forma aleatoria, diferentes cantidades y estilos de regiones por ejemplo tablas, gráficos, ecuaciones, párrafos.
El presente proyecto tiene por objetivo, desarrollar un algoritmo para segmentar y clasificar en forma automática artículos científicos digitales, utilizando técnicas de visión artificial y aprendizaje automático. Las principales contribuciones esperadas de esta propuesta son la construcción de una base de datos de artículos científicos con etiquetas de las principales regiones de cada página, la propuesta de un nuevo algoritmo de clasificación de regiones basada en técnicas de visión artificial y aprendizaje automático, y la evaluación de la metodología propuesta y su comparación con el estado del arte.
Aún cuando el presente proyecto es enfocado en formatos complejos como son los artículos científicos, las contribuciones de la metodología propuesta en el presente proyecto pueden ser adaptados para otros tipos de documentos digitales como por ejemplo tesis, patentes, etc.

Palabras clave (4-6):
Análisis de documentos
Visión artificial.
Aprendizaje automático.
Enfoque estadístico.

2 Objetivos, relevancia, productos y resultados esperados de esta propuesta de investigación

2.1 Objetivos

2.1.1 Objetivo General

- Segmentar y clasificar en forma automática las principales regiones de las páginas de artículos científicos digitales, mediante técnicas de visión artificial y aprendizaje automático.

2.1.2 Objetivos Específicos

a. Construir una base de datos a partir de artículos científicos compuesta de segmentos etiquetados manualmente de las principales regiones de las páginas de dichos documentos.
b. Proponer un nuevo algoritmo para la segmentación y clasificación de las regiones de las páginas de artículos científicos digitales basado en visión artificial y aprendizaje automático.
c. Comparar el algoritmo propuesto con el estado del arte.

2.2 Detalle de los resultados esperados (con relación a los objetivos)

a. Una base de datos de al menos 1000 observaciones etiquetados manualmente por cada categoría (i.e. tabla, texto, figura y ecuación).
b. Se espera que el algoritmo propuesto pueda predecir con un nivel de exactitud similar al estado del arte en la base de datos construida.
c. Se comparará el algoritmo propuesto con técnicas de trabajos anteriores en términos de exactitud.

3 Relevancia de la propuesta de investigación y su relación con la(s) líneas de investigación

Hoy en día es común que en diversas áreas del conocimiento (e.g. académico, legal, finanzas) se generen grandes cantidades de documentos digitales cuyo análisis manual puede ser una tarea lenta y compleja. Es por esto que compañías como Google [1] ya ofrecen plataformas basadas en inteligencia artificial (aunque todavía en fase beta) para analizar estos documentos de manera automática, lo cual es crucial para el entendimiento de la información y posterior toma de decisiones.

Un problema crítico para el análisis y entendimiento de documentos digitales de manera automática [2] es segmentar y clasificar correctamente las distintas regiones (e.g. regiones de texto, tabla, imagen) de un documento digital (problema conocido en inglés como *document layout analysis*). La figura 1 muestra un ejemplo ilustrativo de la segmentación y clasificación de regiones de un documento digital. El desempeño de este proceso afecta significativamente los resultados generales de un sistema de análisis y entendimiento de documentos, no solo en la precisión del reconocimiento óptico de caracteres (OCR) sino también en la utilidad de la información extraída en diferentes escenarios de uso como son el entendimiento automático computarizado de documentos, control de plagio, desarrollo de tecnologías de apoyo para personas con discapacidad visual[3] y minería de datos [4].

Motivados por lo anteriormente mencionado, este proyecto se enfoca en la segmentación y clasificación de las principales regiones de las páginas de los artículos científicos digitales usando técnicas de visión computacional y aprendizaje automático. Se decidió utilizar artículos científicos digitales como base para esta propuesta puesto que su contenido es rico en tablas, figuras, imágenes y ecuaciones en un número de páginas relativamente pequeño, por lo cual la estructura visual de estos documentos es más compleja.

Por otra parte, hoy es muy común que documentos electrónicos estén disponibles en formato PDF (Portable Document Format) que son archivos binarios creados usando lenguajes de descripción de página como Postscript [6] que utiliza un modelo de gráficos vectoriales para definir cómo se deben representar los objetos de una página en una pantalla, es decir, los documentos PDF conservan la estructura visual de cada página del documento digital, independientemente del medio electrónico donde se visualice (e.g. computador, teléfono celular, tablet). Junto a la estructura visual, el documento PDF puede ofrecer información complementaria sobre los

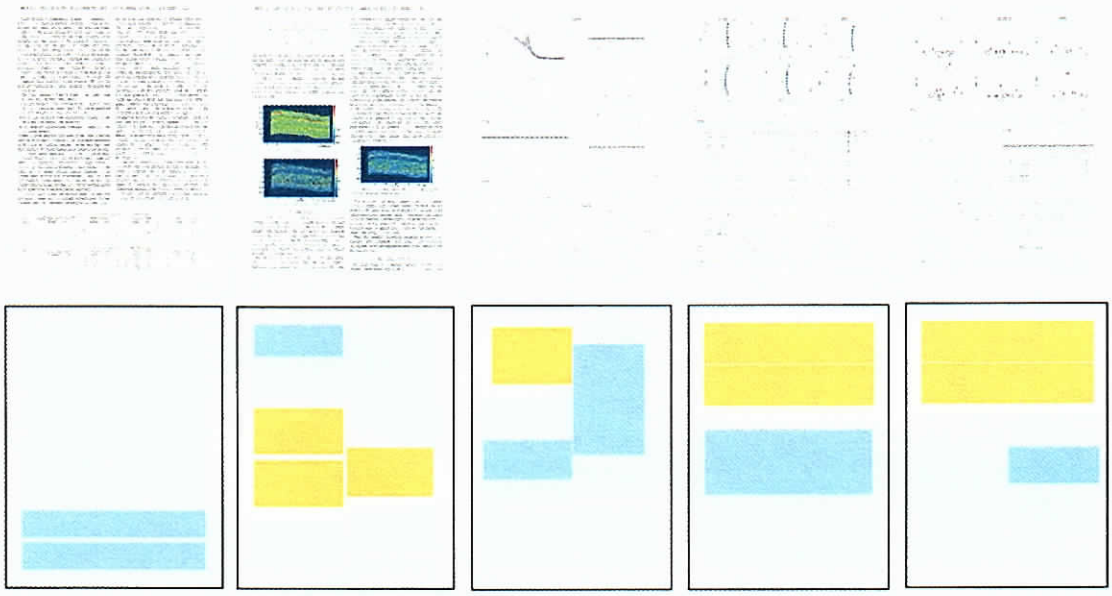

Figura 1: El objetivo de la segmentación y clasificación de regiones de un documento digital (i.e. *document layout analysis*) es encontrar y clasificar de manera automática dichas. Por ejemplo, las zonas gris, azul y amarillo (imágenes inferiores) representan regiones de texto, tabla e imagen respectivamente de las imágenes de cada página. Figura extraída de[5]

elementos del mismo (e.g. posición, tamaño y tipo de fuente del texto). Sin embargo, esta información complementaria depende de cómo fue creado el PDF. Por ejemplo si el PDF fue creado a partir de Latex, este posee información complementaria diferente de los documentos creados a partir de las herramientas de Microsoft Office, aún cuando la estructura visual sea similar y el contenido el mismo.

A pesar de esto, existe en la literatura varios artículos científicos y herramientas para segmentar documentos digitales estructurados y semi-estructurados, donde se utiliza información complementaria del PDF [7]. No obstante, las técnicas disponibles están lejos de ser completamente automatizadas, generales y óptimas [8] al analizar documentos digitales en formato de imágenes como se propone en el presente proyecto. De este modo, al analizar la estructura visual de los documentos PDF mediante técnicas de visión artificial hace que la segmentación y clasificación de regiones no dependa del origen de creación del documento PDF (e.g. Latex o Word). Además, dado que el problema de segmentación y clasificación propuesto parte de imágenes no estructuradas de cada página del documento, para semi-estructurarlo mediante la segmentación de las principales regiones de las páginas (texto, figuras, tablas y ecuaciones), esta propuesta presenta varios desafíos en los campos de la visión computacional. Por ejemplo se deben considerar las discontinuidades en los bordes de las tablas, disposición de las ecuaciones que es considerado un problema no lineal [9], etc.

Por otro lado, debido a la gran diversidad de diseños y estructuras de las ecuaciones, tablas y figuras, un sistema de segmentación y clasificación basado en reglas es inviable [8], por este motivo el presente proyecto utiliza algoritmos basados en aprendizaje supervisado de datos [5], [10] en lugar de algoritmos heurísticos basados en reglas [11].

4 Productos esperados (marcar con una "X" al menos uno de los productos no señalados)

Tipo de Producto:	Marcar con una "X"
a. Disertación a la Comunidad Politécnica (obligatorio);	X
b. Presentación de un artículo en formato de la Revista Politécnica (obligatorio)	X
c. Proyecto de Titulación;	X
d. Aplicación tecnológica construida o implementada;	
e. Patente presentada;	
f. Perfil de proyecto de mayor impacto científico, técnico, pedagógico o de innovación.	
g. Publicaciones científicas indexada en SCIMAGO-SCOPUS/WoS/SCIELO/Latindex Catálogo o un artículo en congreso indexado en SCOPUS.	X

5 Descripción y metodología y diseño del proyecto

5.1 Descripción, metodología y diseño del proyecto (Máximo dos carillas)

De acuerdo a los objetivos planteados, se describirá la metodología en tres etapas: construcción de la base de datos, propuesta de algoritmo de clasificación y comparación con el estado del arte.

Construcción de la base de datos

La base de datos será el punto de partida para entrenar los algoritmos de clasificación de regiones de las páginas de artículos científicos digitales.

Para construir la base de datos, usaremos artículos científicos en formato PDF que se encuentren libremente disponibles en repositorios como *arxiv.org* y que hayan sido generados a partir de código fuente en LaTeX. Se extraerá al menos 1000 muestras en formato de imagen por cada una de las siguientes regiones: texto, figura, tabla y ecuación. Para alcanzar la cantidad de muestras deseadas en el menor tiempo posible, se automatizará la extracción de las regiones usando técnicas de procesamiento de imágenes y se depurará manualmente las regiones extraídas para minimizar los posibles errores de extracción automática.

Se debe aclarar que como los artículos científicos no poseen restricciones en el número de figuras, tablas, ecuaciones y texto por página, el número de muestras en cada página es aleatorio, por lo cual el trabajo de etiquetar los elementos debe ser elaborado detalladamente en cada página considerado para el presente proyecto.

Con la finalidad de no perder información de la estructura visual de los artículos científicos, se realizará una conversión de formato PDF a imágenes RGB (red, green, blue), sin compresión de datos, ni ningún tipo de preprocesamiento, que escale o distorcione la imágen de cada página.

Cabe mencionar que la base de datos compilada será puesta a disposición de manera pública junto con una hoja de especificaciones de la base de datos (i.e. *Datasheet* del *dataset* [12]) para que sea usada por la comunidad científica permitiendo así la reproducibilidad de el presente trabajo.

Propuesta de algoritmo de segmentación y clasificación

Se propondrá un nuevo algoritmo de segmentación y clasificación de regiones de páginas de artículos científicos digitales. Para la segmentación se usará técnicas de procesamiento de imágenes y para la clasificación se utilizará un clasificador basado en redes neuronales convolucionales [13] entrenada a partir de la base de datos construida en la etapa anterior (ver Figura 2). Por otro lado, a diferencia de [5], en este proyecto, además de regiones de texto, figuras y tablas, se clasificará también las regiones que contengan ecuaciones.

Además, diferente de [5], el algoritmo de clasificación propuesto utilizará transferencia de aprendizaje en arquitecturas pre-entrenadas en otras tareas de visión artificial como por ejemplo mobilenet [14], debido a que

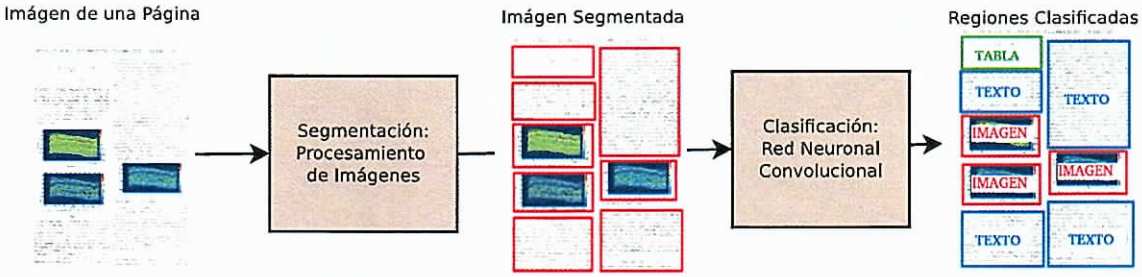

Figura 2: Caption

estas arquitecturas de alto rendimiento en reconocimiento de imágenes requieren grandes volúmenes de datos (i.e., millones de imágenes), cuando realizamos un ajuste fino de las arquitecturas reducimos la necesidad de esa cantidad de imágenes a máximo miles imágenes como la propuesta en el presente proyecto.

Comparación con trabajos previos

En la literatura, se han propuesto varios métodos para la segmentación y clasificación de las regiones de documentos digitales [15, 16, 17, 18, 19, 20, 5]. Entre los trabajos mencionados, actualmente el trabajo más próximo a esta propuesta es el descrito en [5] ya que también se usa como base artículos científicos para construir una base de datos. Así, se implementará el algoritmo en [5] y se comparará con el algoritmo aquí propuesto usando la base de datos construida. Es importante mencionar que en dicho trabajo no se puso a disposición la base de datos usada o el código utilizado.

Referencias

[1] Google, "Document understanding ai." Recuperado de `https://cloud.google.com/solutions/document-understanding/`. Accedido el 25-08-2019.

[2] L. Vincent, "Google book search: Document understanding on a massive scale," in *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, vol. 2, pp. 819–823, IEEE, 2007.

[3] A. Darvishy, "PDF Accessibility: Tools and Challenges," in *Computers Helping People with Special Needs*, pp. 113–116, 2018.

[4] A. Wyner, R. Mochales-Palau, M.-F. Moens, and D. Milward, "Approaches to text mining arguments from legal cases," in *Semantic processing of legal texts*, pp. 60–79, Springer, 2010.

[5] D. Augusto Borges Oliveira and M. Palhares Viana, "Fast cnn-based document layout analysis," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1173–1180, 2017.

[6] T. Merz, *PostScript & Acrobat/PDF: applications, troubleshooting, and cross-platform publishing*. Springer, 2018.

[7] C. Grana, G. Serra, M. Manfredi, D. Coppi, and R. Cucchiara, "Layout analysis and content enrichment of digitized books," *Multimedia Tools and Applications*, vol. 75, no. 7, pp. 3879–3900, 2016.

[8] C. Clausner, A. Antonacopoulos, and S. Pletschacher, "Icdar2017 competition on recognition of documents with complex layouts-rdcl2017," in *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, vol. 1, pp. 1404–1410, IEEE, 2017.

[9] G. Gilboa, *Nonlinear Eigenproblems in Image Processing and Computer Vision*. Springer, 2018.

[10] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, pp. 91–99, 2015.

[11] N. Vasilopoulos and E. Kavallieratou, "Complex layout analysis based on contour classification and morphological operations," *Engineering Applications of Artificial Intelligence*, vol. 65, pp. 220–229, 2017.

[12] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. Daumeé III, and K. Crawford, "Datasheets for datasets," *arXiv preprint arXiv:1803.09010*, 2018.

[13] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.

[14] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4510–4520, 2018.

[15] S. S. Bukhari, A. Azawi, M. I. Ali, F. Shafait, and T. M. Breuel, "Document image segmentation using discriminative learning over connected components," in *Proceedings of the 9th IAPR International Workshop on Document Analysis Systems*, pp. 183–190, ACM, 2010.

[16] L. A. Fletcher and R. Kasturi, "A robust algorithm for text string separation from mixed text/graphics images," *IEEE transactions on pattern analysis and machine intelligence*, vol. 10, no. 6, pp. 910–918, 1988.

[17] M. A. Moll and H. S. Baird, "Segmentation-based retrieval of document images from diverse collections," in *Document Recognition and Retrieval XV*, vol. 6815, p. 68150L, International Society for Optics and Photonics, 2008.

[18] M. A. Moll, H. S. Baird, and C. An, "Truthing for pixel-accurate segmentation," in *2008 The Eighth IAPR International Workshop on Document Analysis Systems*, pp. 379–385, IEEE, 2008.

[19] K. Tombre, S. Tabbone, L. Pélissier, B. Lamiroy, and P. Dosch, "Text/graphics separation revisited," in *International Workshop on Document Analysis Systems*, pp. 200–211, Springer, 2002.

[20] K. Y. Wong, R. G. Casey, and F. M. Wahl, "Document analysis system," *IBM journal of research and development*, vol. 26, no. 6, pp. 647–656, 1982.

6 Infraestructura, equipos y fondos adicionales.

6.1 Infraestructura y equipos

- *Indicar la infraestructura y equipos **disponibles** para la ejecución del proyecto, con la ubicación actual de los mismos*

El proyecto no requiere equipos especiales para su ejecución. Parte del trabajo hará uso de las licencias de Matlab disponibles en la EPN.

6.2 Breve justificación del equipo requerido

- *Justificar la infraestructura y equipos **solicitados** para la ejecución del proyecto e indicar el departamento en el cual se ubicará dicho equipamiento.*

No aplica.

6.3 Fondos adicionales

El proyecto no tiene fuentes de financiamiento adicionales.

ESCUELA POLITÉCNICA NACIONAL

Proyecto de Investigación Interno Sin Financiamiento o Autogestionado

ANEXO 3 - CRONOGRAMA DE ACTIVIDADES DEL PROYECTO

Investigación y Proyección Social

Título del Proyecto: Segmentación y clasificación automática de las principales regiones de las páginas de artículos científicos digitales mediante técnicas de visión artificial y aprendizaje automático.

AÑO 1

Nº	Actividad	Mes 1				Mes 2				Mes 3				Mes 4				Mes 5				Mes 6				Mes 7				Mes 8				Mes 9				Mes 10				Mes 11				Mes 12			
		1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
1	**Construir una base de datos a partir de artículos científicos compuesta de segmentos etiquetados manualmente de las principales regiones de las páginas de dichos documentos.**																																																
1.1	Recopilar artículos científicos de repositorios públicos como arxiv.org	x	x	x	x																																												
1.2	Implementar un algortimo de extracción de automática de regiones de texto,figura, tabla y ecuación			x	x	x	x	x	x	x	x																																						
1.3	Depurar las regiones extraídas automaticamente											x	x	x																																			
1.4	Compilar las regiones depuradas en una base de datos														x	x	x																																
2	**Proponer un nuevo algoritmo para la clasificación de las regiones de las páginas de artículos científicos digitales basado en visión artificial y aprendizaje automático.**																																																
2.1	Estudiar algoritmos de segmentación y clasificación de regiones de documentos digitales	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x																																
2.2	Proponer un nuevo algoritmo de segmentación y clasificación de regiones de documentos digitales basado en redes neuronales																	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x																
3	**Comparar el algoritmo propuesto con el estado del arte.**																																																
3.1	Implementar el algoritmo estado del arte "Fast cnn-based document layout analysis"																													x	x	x	x	x	x	x	x												
3.2	Análisis comparativo con el algoritmo estado del arte																																					x	x	x	x	x	x	x	x				
4	**Productos esperados**																																																
4.1	Presentación de un artículo en formato dela Revista Politécnica																																												x	x	x	x	x
4.2	Disertación a la Comunidad Politécnica																																															x	x
4.3	Publicaciones indexadas																									x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
4.4	Proyecto de Titulación 1: Construcción de Base de datos de regiones de documentos digitales	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x																																
4.5	Proyecto de Titulación 2: Algoritmos de segmentación y clasificación de regiones de documentos digitales													x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x																				