

UNIVERSIDAD POLITÉCNICA SALESIANA

CARRERA: INGENIERIA DE SISTEMAS

Tesis previa a la obtención del Título de Ingeniero de Sistemas

TÍTULO

DISEÑO E IMPLEMENTACIÓN DE UN SISTEMA PROTOTIPO PARA
DETECCIÓN DE PLAGIO EN IMÁGENES Y UN MÓDULO DE ANÁLISIS DE
SIMILITUD DE DOCUMENTOS ACADÉMICOS

AUTORES:

Juan Pablo Lozano Quito

Jessica Eliana Zhunio Zhunio

DIRECTOR:

Ing. Cristian Fernando Timbi Sisalima

Cuenca, Septiembre de 2014

Breve reseña del autor e información de contacto

Juan Pablo Lozano Quito

Estudiante de la Carrera de Ingeniería de Sistemas

Facultad de Ingenierías

Universidad Politécnica Salesiana

jpl161@gmail.com

Jessica Eliana Zhunio Zhunio

Estudiante de la Carrera de Ingeniería de Sistemas

Facultad de Ingenierías

Universidad Politécnica Salesiana

jzhunioz@gmail.com

CERTIFICA

Haber dirigido y revisado prolijamente cada uno de los capítulos del informe de tesis, realizado por el Sr. Juan Pablo Lozano Quito y la Sta. Jessica Eliana Zhunio Zhunio, y por cumplir los requisitos autorizo su presentación.

Cuenca, Septiembre 2014

A handwritten signature in blue ink, appearing to read 'Cristian Fernando Timbi Sisalima', is positioned above the printed name and title.

Ing. Cristian Fernando Timbi Sisalima
Director de Tesis

DECLARACIÓN DE RESPONSABILIDAD

Nosotros, Juan Pablo Lozano Quito portador de la cédula de ciudadanía 0105808471 y Jessica Eliana Zhunio Zhunio portadora de la cédula de ciudadanía 0302018767, estudiantes de la Facultad de Ingenierías en la especialidad de Ingeniería de Sistemas, certificamos bajo juramento que los conceptos y nociones desarrollados, los análisis realizados, y demás contenido de la misma, son de exclusiva responsabilidad de los autores, y que hemos consultado las referencias bibliográficas incluidas en el presente documento.

A través de la presente declaración cedemos los derechos de propiedad intelectual correspondiente a este trabajo, a la Universidad Politécnica Salesiana según lo establecido por la ley de Propiedad Intelectual, por su Reglamento y por la Normativa Institucional Vigente.

Cuenca, Septiembre del 2014



Juan Pablo Lozano Quito



Jessica Eliana Zhunio Zhunio

Dedicatoria

A mis padres Rosario de Jesús Zhunio Malla y
Adán de Jesús Zhunio Zhunio, por su amor, confianza
y gran apoyo incondicional en la búsqueda de
alcanzar mis sueños.

A mis hermanos Nancy, Patricio, Betshy, Carmita,
Freddy, que de una u otra forma estuvieron
apoyándome y echándome una mano
cuando a veces me costaba levantarme.

A mi pequeño bebe que viene en camino,
que aunque todavía no lo tengo en mis brazos es
una gran inspiración realizar éste trabajo al saber que
pronto lo tendré junto a mí .

Jessica Zhunio

A mi familia y de manera muy especial a mis padres
Gloria Quito y Rodrigo Lozano por apoyarme a lo
largo de mi vida por ser mis guías, mis pilares, y no darse por vencidos
para ayudarme a alcanzar esta meta tan anhelada.

Juan Lozano

Agradecimientos

Quisiera empezar agradeciendo primeramente a Dios, por darme una vida llena de alegrías, logros, metas por cumplir, por darme la fortaleza de afrontar las adversidades que nos presenta la vida, por brindarme la oportunidad de lograr finalizar mis estudios universitarios, por darme salud, por regalarme tan preciado regalo de tener a mis padres cerca de mí y por todas las oportunidades que he tenido a lo largo de mi vida.

A mi papi, gracias por ser mi guía, por ser un excelente padre y siempre estar ahí en las buenas y las malas, ha sido un pilar fundamental en nuestras vidas, gracias por sus consejos, abrazos, su paciencia, por regañarme cuando me hacía falta, gracias por las malas noches y el frío que pasaba por recogerme de la universidad, gracias papi por brindarme una buena educación, gracias por todo papito lindo por su ejemplo nos ha sabido educar y formar para ser personas de bien, le amo mucho papi.

A mi mami bella, le agradezco a Dios por darme una madre tan especial como es usted, gracias mamita por darme la vida y protegerme siempre, por ser el otro pilar fundamental de nuestra familia, gracias madrecita por sus consejos, su amor, su preocupación, la enorme paciencia que me tuvo, gracias por siempre estar preocupada en mi salud y hacerme esos ricos batidos en época de exámenes, gracias por poner toda su confianza en mí y no darse por vencida para verme alcanzar mis metas, le amo mucho mamita linda, gracias por todo y que Dios me la bendiga junto con mi papi.

A mis hermanos, comenzando con Freddy y Patricio, gracias hermanitos por tu gran apoyo, siempre aconsejándome y persuadiéndome para que siga con mis estudios, gracias por sacarme de los problemas en los que me metía, gracias por tener fe en mí, haré todo lo posible por no defraudarlos. A mis hermanas Betsy, Carmita, Nancy por ser unas excelentes hermanas, y aunque nos enojemos y discutamos a veces, siempre estuvieron ahí cuando estaba en problemas, gracias por darme sus valiosos consejos de la vida, gracias por sus ocurrencias que me hacen reír mucho, sobre todo cuando estoy triste.

Agradezco a mi compañero de clase, amigo, compañero de tesis y mi pareja Juan, gracias por compartir tus conocimientos conmigo, y de una u otra forma salir adelante a pesar de las dificultades, gracias por tu paciencia y sobre todo gracias por el regalo más grande del mundo que es nuestro hijo.

Por ultimo quiero agradecer a todos mis profesores que a lo largo de mi vida universitaria, transmitieron su conocimiento y ayudaron a formarme personal y profesionalmente, un inmenso agradecimiento a nuestro director de Tesis, Ing. Cristian Fernando Timbi Sisalima, por su paciencia y dedicación en la realización de éste proyecto.

Jessica

Debo agradecer principalmente a Dios por darme la oportunidad de vivir éste momento que ha sido uno de mis más grandes anhelos.

Quiero agradecer a mi padre Rodrigo, gracias papi por su gran apoyo incondicional, por permitirme terminar mis estudios, por apoyare tanto económicamente y moralmente, gracias por el esfuerzo que ha hecho por sacarnos en adelante a mis hermanas y a mí, y brindarnos una excelente educación, sé que no ha sido fácil es por lo mismo que aprecio enormemente su sacrificio.

A mi mami Gloria Quito, a usted mami no tengo palabras que pueda expresar lo agradecido que estoy con usted, ha sido una madre ejemplar, luchadora, trabajadora, gracias mami por brindarme todo su apoyo y cariño, por estar pendiente de mí, por su ejemplo, por hacernos personas de bien, por su paciencia y consejos cuando los necesitaba.

Ha mis hermanas Vilma, Gaby y mi primo “Pachi” por sus ocurrencias y locuras que compartimos durante éste tiempo.

A mi amigo Pablo, gracias por tu amistad, tu apoyo y aporte durante el desarrollo de este trabajo.

Y por último quiero agradecer a todo el grupo de profesores de la Universidad Politécnica Salesiana y en especial al Ing. Cristian Fernando Timbi Sisalima director de la presente tesis, por ser una guía y dedicación en el transcurso del desarrollo de la misma.

Juan

Tabla de contenido

1.	ESTADO DEL ARTE EN LA DETECCIÓN DE PLAGIO.....	1
1.1	Plagio en General	1
1.2	Plagio en la Enseñanza.....	2
1.3	Plagio en el Internet	4
1.4	Tipos de Plagio.....	5
1.4.1	<i>Plagio Directo o Copia Exacta:</i>	5
1.4.2	<i>Parafraseo o Copia Textual:</i>	6
1.4.3	<i>Por traducción:</i>	6
1.4.4	<i>Plagio Modificado:</i>	6
1.4.5	<i>Falsa Autoría:</i>	7
1.4.6	<i>Auto Plagio:</i>	8
1.4.7	<i>Plagio por reestructuración:</i>	8
1.5	Detección del Plagio.....	9
1.6	Recursos Disponibles.....	9
1.6.1	Herramientas para la detección de plagio en Texto.....	10
1.6.2	Herramientas para la detección de plagio en Imágenes y Motores de Búsqueda de Imágenes.....	13
1.7	Metodologías Utilizadas	15
1.7.1	Taxonomía	16
1.7.2	Unidad de Comparación	18
1.7.3	Funciones a Distancia.....	19
1.7.4	Métodos Utilizados en la comparación de Imágenes.....	19
2.	SISTEMAS AUTOMÁTICOS DE DETECCIÓN DE PLAGIO.....	22
2.1	Particulares Relevantes.....	22
2.1.1	¿Cuándo se considera que existe plagio?	23
2.2	Impacto social de estos sistemas	24
2.3	Entornos Aplicables	25
2.4	Técnicas Utilizadas en el Proceso de Detección	25
2.4.1	Referencias y Citas	25
2.4.2	Tesauros	28
2.4.3	Corpus	32

2.5	Líneas de investigación abiertas en las que se basan estos sistemas	33
2.5.1	Detección de plagio Translingües	34
3.	ANÁLISIS Y PLANIFICACIÓN DE LOS REQUERIMIENTOS DEL SISTEMA	35
3.1	Identificación de los Requerimientos	35
3.1.1	Requerimientos Funcionales.....	36
3.1.2	Requerimientos No Funcionales	37
3.2	Análisis de la Solución.....	38
3.2.1	Primera Solución	39
3.2.2	Segunda Solución.....	39
3.2.3	Tercera Solución.....	40
3.3	Especificación de los Módulos de Trabajo.....	42
3.3.1	Módulo Conexión a Internet.....	42
3.3.2	Módulo de Extracción de Imágenes	42
3.3.3	Módulo de Comparación de Imágenes.....	43
3.3.4	Módulo de Análisis de Resultados	44
3.4	Selección de las Herramientas de Soporte	45
3.4.1	Java.....	45
3.4.2	Python	46
3.4.3	Algoritmo Surf.....	47
3.4.4	OpenCV.....	48
3.4.5	JASPERREPORT - iReport	49
4.	DESARROLLO DEL DISEÑO DEL SISTEMA PROTOTIPO	51
4.1	Análisis de la Arquitectura	51
4.1.1	Primera etapa:	52
4.1.2	Segunda etapa:.....	52
4.2	Arquitectura Lógica y Física	52
4.2.1	Arquitectura Lógica	53
4.2.2	Arquitectura Física.....	53
4.3	Documentación UML	54
4.3.1	Diagramas de Casos de Uso	55
4.3.2	Diagrama de Secuencia	60
4.3.3	Diagrama de Clase.....	65
4.3.4	Diagrama de Paquetes.....	69
4.4	Diseño de la Arquitectura de la Solución.....	69

4.4.1	Paquete Lógica.....	69
4.4.2	Paquete Principal	72
4.4.3	Paquete Modelo	72
4.4.4	Paquete Parámetros	73
4.1	Diseño de la Interfaz del Usuario	73
5.	DESARROLLO DEL PROTOTIPO Y ANÁLISIS DE RESULTADOS.....	79
5.1	Implementación del prototipo.	79
5.1.1	Requerimientos de la Aplicación	80
5.2	Preparación y Definición del Plan de pruebas.	80
5.2.1	Etapas de Socialización.....	80
5.2.2	Etapas de Aprobación.....	81
5.3	Ejecución de Pruebas y Recopilación de Datos	82
5.3.1	Definición del Corpus de Pruebas	82
5.3.2	Ejecución de Pruebas	82
5.3.3	Recopilación de Datos	83
5.4	Despliegue de la Aplicación.	91
5.5	Pruebas de Funcionalidad.	92
	CONCLUSIONES y RECOMENDACIONES	94
	ANEXOS.....	96
	ANEXO 1 REPORTE DE DOCUMENTO TESIS 1	96
	ANEXO 2 JavaDoc	100
	ANEXO 3 Manual de Uso	102
	ANEXO 4 Manual de Instalación.....	103
	REFERENCIAS BIBLIOGRAFICAS	104

INDICES DE IMÁGENES

Ilustración 1 Pantalla de TinkEye que se ejecuta en Google Chrome [13]	13
Ilustración 2 Pantalla de Google que se ejecutan en Mozilla Firefox [56].....	14
Ilustración 3 Esquema de Funcionamiento General de URKUND [9].	22
Ilustración 4 El Papel del tesoro en un sistema de documentación [42]	30
Ilustración 5 Etapas que compone Sistema de Detección de plagio en imágenes.	36
Ilustración 6 Funcionamiento General del Sistema Solución 2	40
Ilustración 7 Componentes de los Módulos de administración y gestión del Sistema Anti plagio De la solución 2.....	40
Ilustración 8 Funcionamiento General del Sistema Solución 4	41
Ilustración 9 Componentes de los Módulos de administración y gestión del Sistema Anti plagio De la solución 3.....	41
Ilustración 10 Módulo Conexión a Internet	42
Ilustración 11 Módulo de Extracción de Imágenes	43
Ilustración 12 Módulo de Comparación de Imágenes	43
Ilustración 13 Módulo de Análisis de Resultados	44
Ilustración 14. Logo de Java [44].....	46
Ilustración 15. Logo Python.....	46
Ilustración 16 Comparación de dos imágenes aplicando el algoritmo Surf [52].....	47
Ilustración 17 Espacio escala mediante SIFT (Izquierda) y SURF (Derecha) [46].....	47
Ilustración 18 Esquema del proceso para realizar nuestro proyecto de clasificación SURF [47]	48
Ilustración 19 Logo de OpenCV [53].....	49
Ilustración 20 Ejemplo de funcionamiento de OpenCV con Python [49]	49
Ilustración 21 iReport 3.5.0 [55]	50
Ilustración 22 Arquitectura integrada por fases en general.....	51
Ilustración 23. Arquitectura Lógica del Sistema de Anti plagio de Imágenes	53
Ilustración 24. Arquitectura Lógica del Sistema de Anti plagio de Imágenes	54
Ilustración 25. Caso de Uso de la Aplicación en General.....	55
Ilustración 26. Diagrama Caso de Uso Envío Del Documento	56
Ilustración 27 Caso de Uso General	57
Ilustración 28. Diagrama Caso de Uso Análisis del Documento	58
Ilustración 29. Diagrama de Secuencia General	62
Ilustración 30 Diagrama de Secuencia Envío del Documento	63
Ilustración 31 Diagrama de Secuencia Resultados	63
Ilustración 32 Diagrama de Secuencia de Análisis del Documento.....	64
Ilustración 33 Diagrama de Clase de Sistema Anti Plagio de Imágenes	67
Ilustración 34 Tabla Resultados	67
Ilustración 35 Tabla Análisis.....	67
Ilustración 36 Tabla TratamientoDocumento	67
Ilustración 37 Tabla NavegacionGoogle	68
Ilustración 38 Diagrama Relación entre Paquetes del Sistema Anti Plagio de Imágenes	69
Ilustración 39 Imagen A [50]	70
Ilustración 40 Imagen B [51]	70
Ilustración 41 Comparación de los Puntos de coincidencia entre Imagen A y B.....	70

Ilustración 42 Código Fuente Java, Clase Análisis.....	70
Ilustración 43 Código Fuente Java, Clase Navegación Google	71
Ilustración 44 Código Fuente Java, Clase Tratamiento Documento	71
Ilustración 45 Descripción de la Clase Imagen	72
Ilustración 46 Interfaz Gráfica Carga de Documento	73
Ilustración 47 Resultados Generales Tesis 1	75
Ilustración 48 Anexo Detalle del Análisis de la Imagen Junín del Documento tesis1	76
Ilustración 49 Anexo Reporte General Tesis 1, pág. 1	96
Ilustración 50 Anexo Reporte específico tesis Pag.2 y 3.....	96
Ilustración 51 Anexo Reporte específico tesis Págs. 4/5/6/7	97
Ilustración 52 Anexo Reporte específico tesis Pag.8	97
Ilustración 53 Anexo Reporte específico tesis Pag.9 y 10.....	98
Ilustración 54 Anexo Reporte específico tesis Pág. 11	99
Ilustración 55 Anexo JavaDoc Class AnalisisPlagioWS	100
Ilustración 56 Anexo JavaDoc Class Cola	101
Ilustración 57 Anexo JavaDocClass Proceso	101
Ilustración 58 Interfaz Gráfica Carga de Documento	102

INDICES DE TABLAS

Tabla 1. Plagio de Trabajos Escritos [5].....	3
Tabla 2. Rates of self-reported cheating by discipline and the orientation of that discipline [30]	4
Tabla 3 . Relaciones Semánticas en WordNet [28].	11
Tabla 4 Comparación de diferentes Sistemas de Detección de Plagio [29].	12
Tabla 5 Resumen de la tabla Comparación de diferentes Sistemas de Detección de Plagio ..	13
Tabla 6 Motores comerciales de Búsqueda CBIR [57].....	14
Tabla 7 Proyectos de investigación CBIR / demos / proyectos de código abierto [57].....	15
Tabla 8 Criterios de valoración del porcentaje de similitud o plagio [31].	24
Tabla 9 Composición de un Tesauro [42]	31
Tabla 10 Descripción de Requerimientos No Funcionales.....	37
Tabla 11 Recopilación de Datos del Sistema Anti Plagio de Imágenes.....	85
Tabla 12 Análisis de Efectividad del Documento Sistema Solar	86
Tabla 13 Análisis de Efectividad del Documento Biología Animal	86
Tabla 14 Análisis de Efectividad del Documento Enlaces de Comunicaciones	87
Tabla 15 Análisis de Efectividad del Documento Red Integral de Servicios Digitales	87
Tabla 16 Análisis de Efectividad del Documento Sample IEEE Paper for A4 Page Size.....	88
Tabla 17 Análisis de Efectividad del Documento Implementación de un portal web para la automatización del proceso de consultorías de mentores GOLD de la Región Latinoamericana del IEEE (R9), utilizando arquitectura Java 2 Enterprise Edition - J2EE y tecnología Ajax	88
Tabla 18 Análisis de Efectividad del Documento Formato y Guía para presentación de artículos académicos y científicos	89
Tabla 19 Análisis de Efectividad del Documento Magnetron	89
Tabla 20 Análisis de Efectividad del Documento Estado del Arte del Grid	90
Tabla 21`Porcentaje de Efectividad del Sistema de Detección de Plagio	90
Tabla 22 Diagrama de Despliegue del Software Anti Plagio de Imágenes	91

CAPITULO I

1. ESTADO DEL ARTE EN LA DETECCION DE PLAGIO

1.1 Plagio en General

Se le considera al plagio como “todo aquello que supone copiar obras ajenas en lo substancial” y/o “copiar en lo sustancial obras ajenas, dándolas como propias” [1].

Sabemos que en la actualidad el beneficio del Internet facilita considerablemente el plagio digital, que tanto consciente o en su gran mayoría inconscientemente se lo hace diariamente, entre estos está el plagio de textos, gráficos, videos, música, documentos, cuadros estadísticos, mapas, fotografías e incluso las mismas ideas. Gracias a estudios que se han realizado con respecto a este tema tan polémico, cabe recalcar el estudio de McCabe en el año 2001 en EEUU que deduce que “El 52% de los estudiantes había declarado haber copiado párrafos explícitos” [2], podemos manifestar que éste fenómeno no solamente ocurre en países extranjeros, sino en nuestro propio entorno o ámbito académico ya que existe un elevado porcentaje de copia, tanto por parte de los estudiantes como en los propios docentes, entre los documentos más plagiados están los trabajos de investigación, tesis, trabajos académicos, artículos, informes. Escritores, artistas, directores de películas, incluso políticos han estado involucrados en acusaciones referente a este infracción, como es el caso del ex ministro de defensa en Alemania, Karl-Theodor Zu Guttenberg, quien se vio obligado a renunciar a su cargo tras varias denuncias de plagio en su tesis doctoral, que luego de varias investigaciones, el 23 de febrero, la Universidad de Bayreuth le retiró su título de doctor, obtenido en 2007, porque su tesis “no había dado lugar a un trabajo científico correcto” [4].

En la actualidad niños, adultos, estudiantes, profesionales, etc, podemos acceder de forma fácil e inmediata a una infinidad de información lo que hace bastante común esta práctica, ya que al tener tanta información a la mano, no existe el menor interés de la lectura y de la investigación, sino que él, “ya famoso ctrl c y ctrl v” se hace presente en la mayoría de trabajos o tesis que un estudiante entrega a un docente. El plagio no solamente está presente en universidades sino que existe un gran porcentaje tanto en escuelas como colegio, y desde pequeños se van “acostumbrando” a realizar esta mala práctica que no es corregida a tiempo por una autoridad en los hogares, ni tampoco por los profesores en las instituciones respectivas. Tal vez una de las causas para que se propague y pase desapercibido el plagio es el poco conocimiento que existe entre los ciudadanos al no saber que el simple hecho de copiar una frase o un texto sin poner una referencia, se considera plagio, llevado de la mano que no se aplique las sanciones necesarias cuando se cometa en cualquier forma el reprensible plagio.

A continuación mencionaremos algunos casos famosos conocidos por cometer plagio, y sus respectivas sanciones.

Tal es el caso del escritor peruano **Alfredo Bryce Echenique** quien fue acusado en el año 2008 por plagiar 16 artículos periodísticos. Luego de las respectivas investigaciones en el 2009 fue declarado culpable y el INDECOPI¹ sancionó con una multa de más de 57.000 dólares, aunque en un principio lo negó profundamente, el autor confesó el error cometido y asumió su responsabilidad que aparentemente fue responsabilidad de su secretaria [7].

Otro famoso caso de plagio también en la política fue de la alemana **Annette Schavan** que en el 2013 fue acusada de haber plagiado su tesis doctoral hace más de 30 años atrás, motivo por el cual la llevo a renunciar a su cargo de Ministra de Educación y Ciencia, el mismo que lo venía ejecutando por siete años [8], además que su título fue retirado.

1.2 Plagio en la Enseñanza

El propósito de este tema es el poder tratar y buscar una solución a las implicaciones éticas y educativas al momento de realizar o manejar información para las distintas presentaciones académicas ya sea por parte del docente o estudiante.

Para este tema bastante amplio y complejo podemos partir como una de las partes fundamentales como es la cultura de la ciudadanía, esto lleva que tengamos esta costumbre desde niños como es en la enseñanza de escuelas y posteriormente en colegio y universidades haciendo que esta acción sea habitual sin tener en cuenta el daño causado.

En cuanto a los estudiantes, que es en donde usualmente ocurre el plagio, de manera notoria, sin ningún tipo de sanción, y el simple hecho de aprobar una materia, hace que el plagio sea poco significativo; descartando el hecho de que el estudiante no este enalteciendo sus conocimientos y su cultura, también se está arriesgando la reputación de la institución, por acreditar a profesionales bajo estas faltas, ya que en la vida laboral es donde se refleja todos estos hechos, por lo que las diferentes instituciones educativas deberán tarde o temprano controlar de alguna manera este comportamiento errado por parte de los estudiantes, ya que de no ser así, será la misma que se verá afectada.

Ahora mismo en la cultura ecuatoriana estudiantil podemos observar que el acceso a la información cada vez se nos hace más asequible, lo cual ha llevado al manejo incorrecto de la información por parte de estudiantes, que sin duda, desde el punto estudiantil pueden argumentar la disminución de tiempo, al usar ciertas herramientas digitales, ya que en ocasiones piensan y se siente inseguros de redactar un texto, por desconocimiento del tema, pereza, para evitar realiza un mal trabajo, en su mayoría por ignorancia al no saber la acción que se está cometiendo y que en

¹ Instituto Nacional de Defensa de la Competencia y de la Protección de la Propiedad Intelectual.

ocasiones no revisan el contenido ni la confiabilidad de la fuente perjudicando en su aprendizaje.

René Ramírez, titular de la Senescyt, explicó que se utilizó el programa Urkund para verificar diferentes tesis, el mismo que indica dónde hay coincidencias y la fuente de dónde se extrajo el texto original. "Hay un problema estructural en el sistema universitario. Hay que derrumbar mentiras sociales: decir que hacemos tesis y creer que generamos conocimiento", con esto se dedujo que uno de cada dos tesis son plagios [3].

Como mencionamos con anterioridad acerca del estudio que realizó Donald McCabe durante 3 años en una universidad de EEUU, se presenta a continuación una tabla de datos de la investigación realizada, donde se estima un porcentaje de estudiantes que aceptan haber ejecutado algún tipo de plagio en sus tareas.

	Pregrado	Posgrado
Parafrasear o copiar de una fuente escrita algunas oraciones sin referenciarlo	38%	25%
Parafrasear o copiar de Internet algunas oraciones sin Referenciarlo	36%	24%
Fabricar o falsificar una bibliografía	14%	7%
Entregar un trabajo copiado de otra persona	8%	4%
Copiar casi palabra por palabra de una fuente escrita sin citarlo	7%	4%
Obtener un artículo de un sitio de Internet de venta de Artículos	3%	2%

Tabla 1. Plagio de Trabajos Escritos [5]

Observando la Tabla 1, podemos tener una breve idea o sacar una breve conclusión, la cual nos indica que existe mayor grado o porcentaje de plagio en alumnos Pregrado que en alumnos Posgrado, la razón de la misma se debe ya al grado de madurez y responsabilidad por parte de los mismos.

Varios expertos sobre este tema, señalan algunas razones por la que se podría producir el plagio, sobretodo en la educación, entre ellas tenemos.

- Falta de conocimiento sobre la importancia acerca del plagio.
- Desconocimiento de las formas de plagio.
- Falta de habilidades en la escritura, investigación y lectura por parte del alumno.
- La materia no se considera importante.
- Los docentes no toman las medidas pertinentes cuando se comente este tipo de transgresiones.
- Falta o desorganización del tiempo para realizar las respectivas tareas.
- El ver que los demás estudiantes también lo hacen.
- Aprobar la materia.
- Holgazanería.

Por estas y otras más razones es por la que un estudiante opta por el plagio, y en algunas ocasiones los docentes lo califican como una conducta impropia que sobrelleva a sanciones en las notas de su respectivo trabajo.

En la siguiente tabla se muestra las disciplinas educativas mayormente plagiadas.

Discipline	Frequency of Cheating 1964	Frequency of Cheating 1996	Orientation of Discipline
Business	66%	91%	Vocational
Engineering	58%	82%	Vocational
Education	52%	-	Vocational
Social Sciences	52%	73%	Intellectual
Fine / Applied Arts	50%	-	Intellectual
Natural Sciences	47%	71%	Intellectual
History	43%	-	Intellectual
Humanities	39%	-	Intellectual
Language	37%	-	Intellectual
Other	-	73%	NA

Tabla 2. Rates of self-reported cheating by discipline and the orientation of that discipline [30]

1.3 Plagio en el Internet

Como ya mencionamos en el punto 1.1 el plagio es la acción de “Copiar en lo sustancial obras ajenas, dándolas como propias” [1].

La presente tesis intenta describir el problema de plagio que se da sobre todo en documentos digitales, que como ya lo veremos en las siguientes secciones, se dan diferentes tipos de plagio y las diferentes técnicas utilizadas para la detección de los mismos.

Cuando se realiza plagio, se está incurriendo en una falta ética y escasamente sancionada, en el cual, el autor original no recibe el reconocimiento correspondiente. En el caso de plagio de documentos escritos se puede cometer plagio cuando las citas no están adecuadamente escritas, o ya sea utilizando las mismas ideas.

Sabemos que el internet es una red descentralizada lo que hace que se tenga un rápido y fácil acceso a diversas áreas de la información, aquí el plagio se puede dar de varias maneras tanto puede ser plagio intencional o no intencional, no podríamos saber a ciencia cierta que intenciones se tuvo, ya que incluso se puede cometer plagio por la mala citación del autor, y en este aspecto recaen la mayoría de personas que sin intención alguna, cometen plagio.

Sin lugar a duda podemos asegurar que desde mucho antes de que llegara el internet el plagio ya existía, no obstante esta nueva herramienta facilita aún más el acto de plagiar, ya que gracias a la gran cantidad de información que maneja, al igual de la facilidad con la que se puede obtener la misma, crea hábitos inadecuados, a la hora de realizar una investigación o una tarea escolar. Existen páginas web, con conceptos bastantes confiables, como es el caso de Wikipedia, esta página tiene generalmente conceptos y características de una infinidad de temas, por lo que el estudiante solamente y a duras penas leerá el concepto y lo llevará al paso siguiente

que es el de copiar y pegar en cuestión de segundos para luego imprimir y poner su nombre como propia autoría, y en los mejores casos después de leer brevemente extraerá o copiará los párrafos más relevantes según el estudiante lo considere, obviamente sin la respectiva cita. Páginas como esta existen en gran cantidad en internet, unas más conocidas que otras pero con el mismo objetivo, “facilitar la vida a las personas”.

De aquí viene la controversia, ¿Cómo podemos parar el plagio, si nosotros y las mismas autoridades de alto rango no respetan la propiedad intelectual ajena?, ¿Existirá la forma de controlar lo que divulga y propaga el internet?, ¿podríamos crear herramientas que ayuden a las personas, sobre todo a instituciones, a controlar el plagio?

Por estas y más inquietudes que se presentan, es por lo que decidimos analizar profundamente éste tema, y encontrar una forma con la que podamos ayudar a las instituciones educativas, por medio del desarrollo de un programa, que identifique los tipos de plagio más recurrentes, para lograr éste propósito, identificaremos los tipos de plagios existentes, y los recursos que existen en la actualidad para contrarrestarlos.

1.4 Tipos de Plagio

Como lo habíamos mencionado anteriormente existen varias definiciones de plagio, que es básicamente presentar como propio un trabajo hecho por otra persona, pero ésta definición puede no abarcar todos los tipos de plagio existentes, hay otras formas de plagio que comúnmente la gente ignora.

Existen muchas formas de tipificar al plagio, e innumerable formas de plagiar, hay diferentes formas que se da el plagio, uno de ellos, es el propósito con el que se haga, ya sea intencional o no intencional; la forma en la que se realiza, ya sea copiando total o parcialmente, y tal vez el método que se utiliza, entre otras, a continuación mencionaremos algunos de estos tipos de plagio.

1.4.1 Plagio Directo o Copia Exacta:

Se da cuando se copia por completo o parcialmente cualquier tipo de trabajo que es de autoría de otra persona, o se copia directamente del internet, sin mencionar al autor o fuente electrónica de referencia. Otra definición lo puntualiza de la siguiente manera, “Copiar frases o párrafos de varios autores o documentos y combinarlas más o menos en el mismo párrafo y sin citar las fuentes (hacer una especie de collage con piezas ajenas). alguna investigación muestra que es el tipo de plagio más común, a juicio de profesores y alumnos” [5].

Ejemplo

Forma Correcta:

¿Qué es matemática?

<<Matemática es la ciencia que trata de la cantidad. >> (Diccionario Ideológico de la Lengua Española, Julio Casares, Real Academia Española.)

Plagio:

¿Qué es matemática?

Matemática es la ciencia que trata de la cantidad.

1.4.2 Parfraseo o Copia Textual:

Es una forma muy común de plagio, que consiste en usar las ideas de otras personas y escribirlas como propias, y no referenciar debidamente al autor original, aunque no está técnicamente dentro del concepto de plagio, ésta es una forma de plagio parcial.

Ejemplo:

Forma Correcta:

¿Qué es matemática?

<<Matemática es la ciencia que trata de la cantidad. >> (Diccionario Ideológico de la Lengua Española, Julio Casares, Real Academia Española.)

Plagio:

¿Qué es matemática?

La matemática es la ciencia y el arte que estudia y trata la cantidad. (Julio Casares)

1.4.3 Por traducción:

En nuestro entorno sucede típicamente con el idioma del inglés al español, por la cantidad de libros, revistas y documentos que existen en inglés, lo que sucede comúnmente es que solo se traduce el texto y se lo copia exactamente sin mencionar que fue tomado de cierto libro o artículo.

Ejemplo:

Forma Correcta:

¿Qué es Informática?

Informatics is classically defined as an "information science," which is further defined as the "collection, classification, storage, retrieval, and dissemination of recorded knowledge treated both as a pure and as an applied science". (Eleanor Callahan Hunt, Sara Breckenridge Sproat, Rebecca Rutherford Kitzmiller, Rebecca R. Kitzmiller, The Nursing Informatics Implementation Guide, 2004, pag 20.)

Plagio:

¿Qué es Informática?

Informática se define clásicamente como una "ciencias de la información, " que se define además como la "colección, clasificación, almacenamiento, recuperación y difusión del conocimiento registrado tratados, tanto como una ciencia pura y aplicada".

1.4.4 Plagio Modificado:

Como su nombre lo dice, éste tipo de plagio modifica las palabras usando sinónimos sobre el mismo trabajo, o igualmente elimina o agrega palabras para evitar su parecido.

Ejemplo:

Forma Correcta:

¿Qué es la Robótica?

Técnica que aplica la informática al diseño y empleo de aparatos que, en sustitución de personas, realizan operaciones o trabajos, por lo general en instalaciones industriales. (Real Academia Española, 2008. Diccionario de la lengua española de la Real Academia Española)



Fig. 1 (Shadow Dextrous Robot Hand holding a lightbulb, Richard Greenhill and Hugo Elias, Shadow Robot Company)

Plagio:

¿Qué es la Robótica?

Técnica que aplica la informática al diseño y empleo de aparatos que, en sustitución de personas, realizan operaciones o trabajos.



Fig. 1 Mano Robótica Sosteniendo una Bombilla.

1.4.5 Falsa Autoría:

“Este tipo de plagio se da cuando se escribe en un trabajo el nombre de alguien que no contribuyó para nada en el mismo” [5].

Ejemplo:

Forma Correcta:



Fuente: service-robots.org 12/03/2010

Fig. 1 Mano DLR

Plagio:



Autor: Juanito Pérez,

Fig. 1 Mano DLR

1.4.6 Auto Plagio:

El auto-plagio se da cuando un autor copia nuevamente un trabajo que ya había realizado anteriormente o usa las mismas ideas expuestas en ese trabajo pero con distintas palabras para hacerlo parecer diferente [5]. Este tipo de plagio es muy controversial, ya que por ser el propio autor quien se realiza el plagio, es muy difícil acusarlo de usar las mismas palabras o las mismas ideas que realice en otro trabajo, puede que este tipo de plagio no afecte en nada al autor de los trabajos, pero cuando éste vende sus derechos de autoría a otra persona, es cuando el plagio se vuelve censurable. Es por eso que decimos que el concepto de lo que es el plagio no abarca por completo todos los tipos de plagio existentes.

1.4.7 Plagio por reestructuración:

Se basa en solamente cambiar el orden del trabajo plagiado sin perder su significado o sentido del original.

Como mencionamos anteriormente existen muchas formas de cometer plagio, muchas de ellas se comenten por falta de conocimiento acerca de lo que significa plagiar, es decir se comenten sin intención alguna, pero cuando el plagio se da

intencionalmente, es cuando se convierte en un hecho absolutamente reprochable. Independientemente del propósito con el que se cometa, el plagio en cualquier denominación hay que contrarrestarlo por lo que en la actualidad existes software que ayudan a detectarlo, pero no son muy eficientes para la magnitud de alcance del mismo, sobre todo al momento de detectar imágenes plagiadas.

1.5 Detección del Plagio

Para realizar una detección o análisis de plagio efectiva, se debe tener en cuenta los diferentes escenarios en el que se puede presentar, ya sea en el ámbito académico, profesional o en la vida cotidiana. En la actualidad tenemos varios recursos o medios para mantenernos informarnos a nuestro alcance como es el internet, sin mencionar los aparatos tecnológicos como laptops, tabletas, celulares, calculadoras, etc., que en su gran mayoría vienen con nuevas aplicaciones que facilitan el descargar, guardar, e imprimir un indicado trabajo. Esto ha mantenido preocupado al ámbito educativo, ya que a los docentes en ocasiones se le hace dificultoso detectar si el material presentado por el estudiante es de su autoría o simplemente plagiado.

Teniendo en cuenta el argumento anterior y priorizando el ámbito académico, es necesario el uso de herramientas que ayuden ya sea al docente, al estudiante, al investigador para revisar sus trabajos, tesis, artículos en caso de plagio no intencional y de ésta manera verificar si existen copias exactas o parciales en los respectivos trabajos, de tal manera que se pueda evaluar correctamente el esfuerzo y conocimiento del estudiante, o la persona quien realice la investigación. Cabe recalcar que hoy en día existen diferentes tipos de software e incluso páginas en internet, tanto gratuitas como pagadas, dedicados a identificar los trabajos plagiados, pero también es cierto que estas herramientas no son del todo eficientes, de una u otra forma los estudiantes encuentran la forma de evadir este tipo de control, inclusive existen universidades que ya están aplicando en forma obligatoria, que cada ensayo, trabajo, o cualquier tipo de tarea que se le asigne al alumno, primero tiene que pasar por el detector de plagio, y en casos extremos existen sanciones como es la expulsión de la institución.

Estas medidas para ciertos alumnos parece algo exagerada, ya que en opinión de ellos para implementar esas sanciones, primero deben enseñar adecuadamente y poner ciertas medidas, “Es algo exagerado hablar de expulsión o sanciones fuertes, si vienen a implementar este tipo de sistemas tienen que educarnos adecuadamente, ya que la mayoría de nosotros no sabemos que no mas es considerado como plagio”, Wilmer Zhunio, estudiante de la carrera de Marketing de la Universidad del Azuay; así como Wilmer muchos de los estudiantes comparten la misma opinión.

Por otro lado la idea de tener una herramienta que ayude a detectar el plagio, es muy apreciada por parte de los docentes, aunque admiten que en un comienzo será difícil la aceptación por parte de los estudiantes.

1.6 Recursos Disponibles

La tecnología que tenemos en la actualidad, como ya sabemos, ayuda mucho a los alumnos a plagiar, pero también ayuda a profesores a detectar la fuente de lo plagiado, llegados a este punto podemos encontrar varias herramientas virtuales que permite a profesores, alumnos, autores, escritores, etc..., poder encontrar y detectar el plagio de información, y tratar de controlarlo, ya sea intencional o realizada precisamente con el objetivo de plagiar, en ocasiones cambiado o simulados para intentar engañar, es a esto a lo que se le denomina plagio simulado. Para esto existe en la web varios programas especializados en esta área ya sea gratuita o con algún costo, esto dependerá de la calidad, seguridad y alcance del programa que el usuario desea utilizar para detectar el plagio. Para que un estudiante, docente o investigador verifique que no cometieron plagio en sus artículos, especialmente cuando no lo pretenden hacer, la mayoría de estos programas son de gran utilidad especialmente para ellos, para validar sus artículos por posible plagio no intencional, ya que de una u otra forma se encuentran confrontando y queriendo prevenir los diferentes tipos de plagios. Lo que comúnmente se hace, es copiar una frase de un trabajo realizado por un estudiante y transcribirlo en los diferentes navegadores (Google, Yahoo, Bing, etc.), de esta manera se compara si existe similitud ya sea parcial o completa en los diferentes tipos de trabajo. Pero en un entorno real, hacer esto puede resultar tedioso y demorado, sobretodo en el ámbito académico donde existen gran cantidad de estudiantes, y muchas tareas asignadas, lo que dificulta enormemente seguirle la pista uno a uno al estudiante para confirmar si cometió plagio o no lo cometió.

En vista de lo mencionado anteriormente, existen varios programas para la detección automatizada del plagio, aunque muchos de ellos proyectan resultados no muy consistentes o reales, además de tener en cuenta que la mayoría no detectan los diferentes tipos de plagio existentes, como plagio por traducción, parafraseo, y el plagio de imágenes.

A continuación citaremos algunos programas y sitios web que ofrecen o nos facilita la adquisición de esta herramienta.

1.6.1 Herramientas para la detección de plagio en Texto.

- **Sherlock:** Es una herramienta para controlar el plagio textual, es decir encuentra parecidos entre los diferentes textos, este programa fue desarrollado por la Universidad de Sydney, la desventaja de este programa es que no analiza documentos completos sino fragmentos de los mismos, y lo hace localmente, sería ideal para controlar los trabajos entre los estudiantes. [10]
- **Jplag:** Al igual que Sherlock y la mayoría de los programas anti plagio, Jplag encuentra similitudes entre varios archivos localmente, pero la diferencia está que no compara el texto en sí, si no que analiza la estructura y la sintaxis del código, lo que hace más dificultoso disimular el plagio [11].

- **Urkund:** Esta es una herramienta que funciona a través del internet, y que la Senescyt (La Secretaría de Educación Superior, Ciencia, Tecnología e Innovación) en nuestro país está implementando para contrarrestar el plagio de tesis, ya que entre el 2010 y el 2013 se realizó un estudio de análisis de tesis, el mismo que arrojó resultados sorprendentes y podríamos decir predecibles, ya que gracias a esta herramienta que mide el porcentaje de coincidencias bibliográficas, se pudo determinar que de una muestra de 802 tesis, el 54% presentan altos índices de coincidencias bibliográficas [12]. Esta herramienta permite que los trabajos que los alumnos presentan se comparen con otros documentos o fuentes, entre éstas tenemos, páginas web, Base de datos, enciclopedias, periódicos, revistas, etc.
- **WordNet:** Este fue desarrollado por la universidad de Princeton, se basa en el método semántico que lo veremos más adelante, WordNet consiste en una base de datos gramatical, que acumula las palabras en conjuntos de sinónimos. En la siguiente tabla veremos un ejemplo [27].

Relación Semántica	Significado	Categoría sintáctica	Ejemplos
Sinonimia	Mismo o similar significado	Sustantivo Verbo Adjetivo Adverbio	Rápido – Veloz
Antonimia	Significado opuesto	Adjetivo Adverbio Sustantivo Verbo	Rápido – Lento
Hiponimia	Término más específico	Sustantivo	Árbol – Planta
Meronimia	Parte del significado de la otra palabra	Sustantivo	Dedo – Mano
Troponimia	Hiponimia para verbos	Verbo	Caminar – Desplazarse
Implicación (Entailment)	Implicación – Causa	Verbo	Roncar- Dormir

Tabla 3 . Relaciones Semánticas en WordNet [28].

En la siguiente tabla podemos ver un cuadro comparativo donde se describen algunos otros sistemas de detección de plagio que existe en la web.

	Nombre del Sistema	Link	País Institución	Idioma	Descargable En Línea	Búsqueda	Opciones Avanzadas	Tipos de Resultados	Pago/ Gratuito
1	PLAGIUM	http://www.plagium.com	Septet Systems Inc. Estados Unidos y Europa	Inglés Francés Alemán Italiano Portugués Español	En línea	Internet	Resumen del texto encontrado Línea de tiempo Resultados en tiempo real	Fuentes de coincidencia Porcentaje de semejanza	Gratuito
2	COPIONIC	http://copionic.citilab.eu/	Citilab, Centro para la Innovación Social y Digital de Barcelona	Español	En línea	Internet		Fuentes de coincidencia Porcentaje de semejanza	Gratuito
3	WCOPYFIND	http://plagiarism.bloomfieldmedia.com/z-wordpress/	Universidad de Virginia, EEUU	Inglés Español	En línea	No realiza búsquedas en Internet	Compara 250 reglas gramaticales Búsqueda por palabras definidas Número de palabras copiadas	Resalta el texto plagiado	Gratuito
4	EDUCARED ANTIPLAGIO	www.educared.org	Silicon Artist, Madrid, España	Español	En línea	Base de datos construida a partir de la web "el rincón del vago"		Está en desarrollo.	Gratuito
5	APPROBO	http://approbo.citilab.eu/	Symmetric (empresa española)	Español	En línea	Buscador propio, y los más conocidos		Porcentaje de semejanza	Gratuito
6	JPLAG	https://www.ipd.uni-karlsruhe.de/jplag/	Universidad Karlsruhe, Alemania	Alemán Francés Español Portugués	En línea	Base de datos interna Documentos locales del usuario Internet	Códigos fuente de varios tipos de lenguajes de programación	Da las fuentes de donde proviene el plagio Resaltan el texto plagiado	Gratuito
7	COPYSAPCE	http://www.copyscape.com/	Indigo Stream Technologies, Ltd	Multilingüe	En línea	Internet, sólo URL		10 sitios por búsqueda	Gratuito
8	COPY SPACE	http://www.copyscape.com/	Indigo Stream Technologies, Ltd	Multilingüe	En línea	Internet, URL y documentos	Desconocido	Resultados ilimitada	Pago
9	TURNITIN	http://www.turnitin.com	Universidad de Berkeley, California	11 idiomas, entre ellos el español.	En línea	Internet Tienen una "biblioteca virtual"	Desconocido	Fuentes de coincidencia Porcentaje de semejanza	Pago anual
10	URKUND	http://www.urkund.com/int/en/	Priinfo, empresa sueca	Inglés Alemán Francés Turco Polaco Español Sueco	Se envía un e-mail conocer@urkund.es , y devuelven los resultados	Base de datos interna Internet	Desconocido		Pago por búsqueda
11	EASY VERIFICATION ENGINE	http://www.canex.com/	Canexus, Canada	El programa está en inglés	Descargable	Internet	Búsqueda: Rápida Media Exhaustiva	Sitios de plagio Porcentaje de semejanza Resalta el texto plagiado	Pago por licencia permanente
12	PLAGIARISM DETECTOR	http://www.plagiarism-detector.com/	Skyline Inc	Inglés Español Italiano Alemán Francés Japonés China	En línea o Descargable, dependiendo de la licencia	Internet: Google, Altavista y Yahoo		Porcentaje de semejanza Resalta el texto plagiado	Pago

Tabla 4 Comparación de diferentes Sistemas de Detección de Plagio [29].

Programa	Idioma	Exposición de Resultado	Acceso
Plagium	Inglés ,Francés, Alemán, Italiano, Portugués, Español	Porcentaje de semejanzas existentes en el documento.	Gratuito
JPlag	Alemán, Francés, Español, Portugués	Resaltando el texto plagiado.	Gratuito
Urkund	Inglés ,Francés,	Envía al correo registrado	Pagado

	Alemán, Italiano, Portugués, Español, Turco, Sueco	un informe con porcentajes de coincidencia con otros documentos.	
Plagiarism Detector	Inglés, Francés, Alemán, Italiano, Portugués, Español, Turco, Japonés, Chino.	Porcentaje de la semejanza con otros documentos, además de resaltar el texto plagiado de lo demás.	Pagado
Aprobo	Español	Porcentaje de Coincidencia	Gratuito
CopySpace	Multilenguaje	Ilimitados	Pagado

Tabla 5 Resumen de la tabla Comparación de diferentes Sistemas de Detección de Plagio

1.6.2 Herramientas para la detección de plagio en Imágenes y Motores de Búsqueda de Imágenes.

Como habíamos mencionado anteriormente estas herramientas que ya existen en la actualidad, se enfocan a la detección de plagio de texto, pero también existen ciertos programas o aplicaciones que permiten la detección de plagio de imágenes a continuación mencionaremos algunos de estos.

- **TinEye:** Esta herramienta de búsqueda y que del mismo modo funciona a través de internet, tiene un API que se adquiere con un costo, en si esta herramienta es un motor de búsqueda, que cuando carga una imagen, se envía a una base de datos en donde podemos comparar, ver si se modificó, y como se está utilizando cierta imagen. Esta herramienta utiliza la tecnología de identificación de imágenes en vez de palabras clave (Keywords), metadatos o marcas de agua [13].

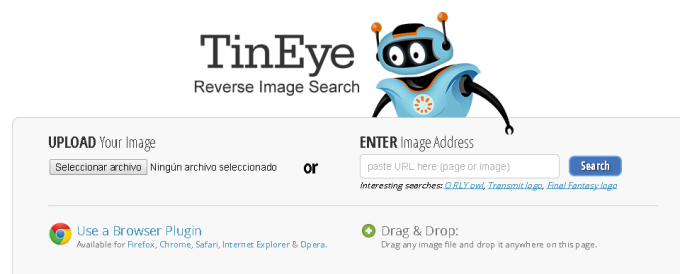


Ilustración 1 Pantalla de TinkEye que se ejecuta en Google Chrome [13]

- **Google Image Search:** esta herramienta propia del buscador Google lanzado en el 2001, aquí podemos buscar por imágenes y no más bien por texto como estábamos acostumbrados, se puede arrastrar la foto al buscador o subiéndola a través del botón cámara de Google, según Google las palabras clave o también conocidas como keywords que se utilizan para la búsqueda de imágenes, están basadas en el nombre del archivo de la imagen, el texto enlace que apunta a la imagen, y el texto junto a la imagen; en el 2012, Google implemento un nuevo y

mejorado algoritmo de búsqueda, para prevenir que imágenes pornográficas salgan en una búsqueda siempre y cuando el usuario no esté buscando este termino [56].



Ilustración 2 Pantalla de Google que se ejecutan en Mozilla Firefox [56]

Además de éstas dos herramientas, existen otras aplicaciones, como Piximilar, Visual Image, etc., que basan su búsqueda en el contenido de las imágenes, son conocidos como herramientas **basadas en CBIR**, éste método se describirá en la sección 1.7.4, para mejor entendimiento, pero mencionaremos varias de éstas aplicaciones en la siguientes tablas 6 y 7; muchas de ellas enfocadas en la mercadotecnia y privadas. Estos diferentes motores de búsqueda de imágenes se comparan en píxeles con el fin de devolver resultados que coinciden a la hora de realizar una particular consulta.

Nombre	Descripción	Consulta externa de imagen	Los metadatos de consulta	Índice Tamaño (Estimación, millones de imágenes)	Tipo de Organización	Licencia (abierto / cerrado)
Yandex Image Search	Sistema de Yandex CBIR	Sí	Sí	10000m	Empresa pública	Cerrado
Reconocimiento Visual Factory	Reconocimiento Visual Factory	Sí	Sí		Empresa Privada	Cerrado
Visual Image Search	Buscador CBIR, por pixolution	Sí	No	10M	Empresa Privada	Cerrado
TinEye	CBIR sitio para encontrar variaciones de imágenes de la web, por Idee Inc.	Sí	No	1800M	Empresa Privada	Cerrado
Tiltomo	Sistema CBIR usando fotos de Flickr	No	Sí		Empresa Privada	Cerrado
Shopachu	Compras y moda motor CBIR, por Incogna Inc.	No	Sí	1M	Empresa Privada	Cerrado
Piximilar	Motor Demo, desarrollado por Idee Inc.	No	No	3M	Company1 privado	Cerrado
Picallike	Motor CBIR para móvil y comercio electrónico	No	No (filtros adicionales pueden ser añadidos)		Empresa Privada	Cerrado
OddEye	Motor CBIR para encontrar marcos de productos / catálogo / video, por Odd Concepts Inc.	Sí	Sí	20M	Empresa Privada	Abrir (a través de API)
NoClone	Motor de búsqueda de imágenes de PC y clasificación basada en contenido	Sí (un conjunto)	No		Empresa Privada	Cerrado
MIPal motor de búsqueda de similitud	Motor de búsqueda de similitud Online	Sí	Sí	100M	Individual	Cerrado
Macroglossa Visual Search	CBIR motor de búsqueda visual	Sí	No		Empresa Privada	Cerrado
Like.com	Compras y moda basado en el motor CBIR	No	Sí	1M	Empresa Privada	Cerrado
Incogna Image Search	Buscador CBIR, por Incogna Inc.	No	Sí	100M	Empresa Privada	Cerrado
Imprezzeo Image Search	Buscador CBIR, por Imprezzeo.	No	Sí		Empresa Privada	Cerrado
IMMENSELAB	CBIR buscador por KBKGROUP	Sí	No	10M	Empresa Privada	Cerrado
Imagen imense Buscar Portal	Buscador CBIR, por imense.	No	Sí	3M	Empresa Privada	Cerrado

Tabla 6 Motores comerciales de Búsqueda CBIR [57]

Nombre ↕	Descripción ↕	Consulta externa de imagen ↕	Los metadatos de consulta ↕	Índice Tamaño (Estimación, millones de imágenes) ↕	Tipo de Organización ↕	Licencia (abierto / cerrado) ↕
akiwi ↗	akiwi es una herramienta de imagen keywording semi-automática utilizando técnicas CBIR. Fue desarrollado por HTW Berlin / pixolution GmbH	Sí	Sí	15M	Universidad	Cerrado
ALIPR ↗	Desarrollado por investigadores de la Universidad de Penn State	Sí	Sí		Universidad	Cerrado
Anaktisi ↗	Esta solución Web implementa una nueva familia de descriptores CBIR. Estos descriptores se combinan en un solo color histograma y la textura de la información y son adecuados para la recuperación de imágenes con precisión.	Sí	No	0.225M	Universidad	Abierta
BRISC ↗	BRISC es un acrónimo recursivo para BRISC Realmente es fresco, y es (muy convenientemente) también un anagrama de imágenes basada en contenido de recuperación del sistema.	Sí	No		Universidad	GPL
Califa y Emir ↗	Creación y recuperación de imágenes basadas en MPEG-7.	Sí	No	Basado Desktop-	Universidad	GPL
FUEGO ↗	Abrir consulta de origen por ejemplo visual del sistema CBIR. Desarrollado en la Universidad RWTH Aachen. FIRE ↗ es un sistema de investigación desarrollado con la extensibilidad en mente y se puede combinar fácilmente con los sistemas de recuperación de información textual.	No	No		Universidad	Abierta
GNU Image Finding Tool ↗	Consulta por ejemplo de sistema de búsqueda de imágenes.	Sí	No	Basado Desktop-	GNU	GPL
ISSBP ↗	Similar Image Search por imense plugin para Adobe Bridge, beta gratuita.	Sí	Sí	-libre de beta limitado a imágenes 4k	Empresa Privada	Cerrado
img (Rummager) ↗	Image Engine recuperación (Freeware Aplicación).	Sí	No	Basado Desktop-	Individual	Cerrado
imgSeek ↗	foto gestor de cobros y visor con búsqueda basada en contenido y muchas otras características.	Sí	No		Individual	GPL
IKONa ↗	Genérico sistema CBIR - INRIA - IMEDIA	Sí	Sí		Universidad	Cerrado
IOSB ↗	Imagen software de demostración de recuperación de Fraunhofer IOSB (Alemania)	Sí	No	Basado Desktop-	Instituto de Investigación	Cerrado
LIRE ↗	Biblioteca Java GPL para el contenido basado en imágenes de recuperación basado en Lucene incluye múltiples características globales y locales de bajo nivel y las diferentes estrategias de indexación incluyendo bolsa de palabras visuales y hash.	Sí	Sí		Universidad	GPL
Lucignolo ↗	Motor de búsqueda de similitud de imagen utilizando únicamente el texto completo motor de búsqueda Lucene nativa.	Sí	Sí	106M	Instituto de Investigación	Cerrado
MIFile ↗	Motor de búsqueda de similitud de imagen basado en MI File (Archivo Invertido Metric), desarrollado en ISTI-CNR. Código fuente ↗ del archivo MI.	No	No	106M	Instituto de Investigación	Abierta
Muvis ↗	Sistema CBIR a Tut Tampere University of Technology.	Sí	No	Basado Desktop-	Universidad	Cerrado

Tabla 7 Proyectos de investigación CBIR / demos / proyectos de código abierto [57]

1.7 Metodologías Utilizadas

La técnica normalmente utilizada por los docentes es la denominada *detección intrínseca o personal*, ésta técnica es más intuitiva que planeada, la misma que consiste en observar cuidadosamente el texto o trabajo escrito dudoso, ya que si de un momento a otro las palabras que se usan están más sofisticadas que otras levantan la sospecha que el párrafo esta plagiado, normalmente ésta técnica se utiliza en instituciones educativas para controlar las tareas de los estudiantes.

Las metodologías utilizadas en diferentes software existentes en la actualidad, para la detección del plagio se basan en el análisis de frases, caracteres, el estilo de

escritura y la estructura del texto en su totalidad. Otro método o técnica utilizada es la **detección con referencia**, éste método consiste en comparar el texto sospechoso con el texto original, pero al no tener en una base de datos todos los textos originales registrados con el cual se pueda comparar, convierte a ésta técnica en una gran desventaja [15]. SCAM es uno de los pioneros en crearse para la detección de documentaciones copiados, básicamente consiste de dos partes, la primera es que el usuario suba libros que compran a una base de datos y de estos comparar con los textos sospechosos a identificar, y la otra parte el método de comparar similitudes entre palabras, que existe entre párrafos [16], la estructura es la misma en general cuando hablamos de plagio con referencia, a continuación citaremos algunas de estas características tomadas de [17] [18] [19].

- Se examinan los documentos del corpus de referencia y se calcula un hash para cada uno de ellos.
- Se almacenan estas representaciones en un índice invertido.
- Para verificar si determinado documento contiene texto plagiado, se deduce su escritura y se la compara contra las que están en el índice, obteniendo un coeficiente que denota la similitud.
- Se seleccionan los archivos más cercanos para hacer una comparación más detallada.
- La comparación detallada produce como resultados los pasajes que se sospecha fueron plagiados.
- Podría haber eventualmente una fase de post-procesamiento: volver a analizar los obtenidos con otro modelo para corroborar que se trata de plagio, filtrar los pasajes muy cortos, etc.

Otro método que se utiliza es la técnica de **Averaged Word Frequency Class** [14], éste método consiste en segmentar las palabras en géneros o clases, según la frecuencia con la que se utiliza, para luego obtener un promedio de todas las clases de palabras que aparecen en el texto sospechoso, de esta forma se puede deducir si existe o no plagio, ya que si en determinados párrafos no tienen palabras frecuentes, en comparación con el resto de párrafos del documento, se sospecha que ese determinado párrafo no es de autoría de la misma persona, es ahí cuando se analiza con mayor detenimientos para aseverar o rechazar dudas.

1.7.1 Taxonomía

Se plantea una clasificación de diferentes métodos de detección de plagio accidental.

1.7.1.1 Métodos Sintácticos

“La sintaxis estudia la forma en que se combinan las palabras para formar sintagmas y oraciones correctas, determinando el papel estructural de cada palabra y sintagma” [26].

Métodos que realizan el análisis de los textos únicamente tomándolos desde un punto de vista como símbolos. Se analiza la frecuencia de los símbolos y para el procesamiento se puede realizar lo siguiente: convertir el documento a mayúsculas, eliminar acentos, quitar signos de puntuación y reemplazar por número o símbolos de acuerdo al criterio del analista.

Dado estos conceptos generalizados, y en simples palabras podemos decir que cuando hablamos de métodos sintácticos estamos hablando de la estructura que posee el documento a analizarse, es decir a la correcta estructura de cada una de las frases de documento. Entre éstos métodos encontramos.

Modelo Espacio Vectorial

Es una forma sencilla de representar el texto, en donde el texto se representa de forma cuantificada es decir, cuenta con esquemas específicos de pesado, de ésta forma permite dar mayor importancia a algunas palabras en el documento y almacenarlas en un vector, éste método es comúnmente empleado en la recuperación de los datos. El pesado tf-idf (*frecuencia de los términos - la frecuencia inversa de documento*), destaca la importancia de algunos términos sospechosos [19] [22], se toma en cuenta el número de veces que se repite una palabra que contienen esa palabra; entre menos frecuentes en el documentos más distintivo o relevante es ese determinado término [24].

Posteriormente se usa una medida de distancia entre los vectores para medir la similitud entre los documentos comparados, una de las medidas de similitud más populares suele ser el coseno del ángulo entre los vectores, lo que básicamente mide el parecido de un documento a otro, esta medida sugiere tomar decisiones como: que características utilizar para representar el texto, los pesos que se le atribuya a cada dimensión, y la medida de distancia entre vectores [19] [23].

Hashing

También conocida como fingerprinting, selecciona partes o fragmentos de un documento y calcula su respectivo hash, generalmente se utiliza el hash criptográfico MD5, el hash se guarde en un índice invertido y se analiza el documento sospechoso de la misma forma que los de referencia y si existe coincidencia se puede sospechar la existencia de plagio.

Un inconveniente con éste método es la desventaja que tiene, ya que si se cambia un bit de entrada automática y radicalmente cambia la salida, como consecuencia, si una de las partes seleccionadas cambia un poco el hash correspondiente a esa parte difiere completamente, impidiendo que la igualdad del fragmento sea detectada. [19]

N-Gramas

Son trozos o partes de n palabras, es un modelo probabilístico que nos ayuda a predecir que palabra se utilizara en un determinado texto, se toman n-gramas del

documento, y se calcula diferentes características como la frecuencia y la distancia de cada función, en general los n-gramas se toma en forma superpuesta, por lo que la cantidad de n-gramas es alto y en sí el método tiene una mejor cobertura.

En conclusión este método lo que hace es tomar un grupo de palabra o letras y compararlas con otro similar, este métodos es de mucha utilidad ya que no es tan fácil de evadir lo que son formulas, abreviaturas que con otros métodos no son tan evidentes, ya que no se separa o divide de esta forma [19].

Dotplop

Esta técnica se utiliza para visualizar información. Se utiliza en el área de la informática para detectar código duplicado y para detectar similitud entre documentos, y por siguiente detectar plagio en texto. Es un plano bidimensional en el que se representan las coincidencias entre las dos secuencias y en los respectivos ejes (X y Y) ubicamos la posición de la letra o signo, n-grama o palabra que tenga coincidencia, por ejemplo, en el eje x ubicamos el texto sospechoso, y en el eje Y la posición del texto original, una vez obtenido estos patrones y en las ubicaciones correctas, se procede a analizar automáticamente, observando si existen coincidencias totales o parciales. [25]

Stop Words

Las palabras sin significado como artículos, pronombres, preposiciones, etc., son llamadas Stop Words, palabras vacías o palabras de relleno.

Un método muy efectivo y que provee buenos resultados es la eliminación de las Stop Words, ya que estas palabras o tienen mayor relevancia y por lo tanto no contribuye información importante.

1.7.1.2 Métodos Semánticos

Los métodos semánticos también conocidos como métodos lingüísticos y que a diferencia de los métodos sintácticos, no toman las palabras como símbolos, en vez de eso, hace una relación con otros términos según su significado, tales como la sinonimia. Éste métodos es más complicado de implementar que los métodos sintácticos es por esta razón que se aplica o se desarrolla muy poco en software de detección de plagio, pero se puede afirmar que aplicando el métodos semántico se puede detectar documentos plagiados que otros métodos no lo puede hacer. [19]

1.7.2 Unidad de Comparación

Otro aspecto importante que hay que tener en cuenta a la hora verificar la existencia del plagio además de la taxonomía, son los diferentes mecanismos de comparación que existen, como la unidad de comparación de todo el documento, de todo un párrafo específico, un grupo de palabras o letras.

La unidad de comparación es muy utilizada en varios métodos de detección de plagio como es el caso del método Espacio Vectorial, que compara textos completos, o el método de n-gramas que compara grupos de letras o palabras. En el caso de n-gramas la unidad de comparación es asimétrica comparando los n-gramas de un documento con los n-gramas del otro documento.

1.7.3 Funciones a Distancia

También conocido como medidas de similitud, ya que sirven precisamente para comparar el parecido entre diferentes documentos. Están representadas por conjuntos o vectores, donde se utiliza algunas técnicas como son:

El coeficiente de Jaccard, que mide el parecido o no parecido entre dos muestras. Siendo su formula la siguiente $I_j = c/(a+b)-c$

Donde a es el número de entes o elementos que están en A, b es el número de entes que están en B y c, el número de entes que están en A y B.

De la misma manera se basan, El coeficiente de Dice, el coeficiente de Overlap, el Coeficiente de Roger y Tanimoto, el Coeficiente de Sokal y Michener y el Coeficiente de Czekanowski. Al aumentar las características similares, los valores de la función también aumentan.

1.7.4 Métodos Utilizados en la comparación de Imágenes.

Así como tenemos diferentes métodos que son utilizados para la detección de plagio, también contamos con métodos que nos ayudan a detectar el parecido de entre imágenes. A continuación mencionaremos algunos de éstos métodos o técnicas que se suele utilizar.

Palabras Clave

Son mayormente conocidas como keywords, y son un conjunto de caracteres que se puede utilizar para buscar una información en un buscador (Google, Yahoo, etc.), es decir son palabras claves o palabras muy importante asociada a un concepto.

Por ejemplo si deseo buscar “última tecnología” en la web, automáticamente la keywords es “tecnología”, y el respectivo buscador que se esté utilizando la asociara a todo lo que tenga que ver con tecnología, mostrándome varios resultados.

Metadatos

Todos los archivos existentes ya sean videos, imágenes, texto, audio en fin, en general todo tipo de archivos contienen información oculta acerca de su origen, es decir, un archivo normalmente contiene información referente al autor, fecha de creación, fecha de modificación, tipo de documento, extensión, etc.

A este tipo de información sensible y oculta a la vista del usuario común y corriente se lo conoce como metadatos, muy útil para clasificar y organizar los

contenidos de documentos, es aquí donde se aprovecha esta ventaja dando paso a otra forma u otro método para detectar el plagio de imágenes, fotos, e inclusive documentos. Existen varios programas de detección de plagio que utilizan los metadatos como su plataforma o base, los mismos que se encuentran en la Tabla 5 y 6.

Existen estándares conocidos para guardar metadatos en las imágenes como son:

- “Exif (Exchangeable image file format): es una especificación para formatos de archivos de imagen usado por las cámaras digitales. La especificación usa los formatos de archivos existentes como JPEG, TIFF y RIFF, el formato de archivo de audio WAVE, a los que se agrega tags específicos de metadatos. No está soportado en JPEG 2000 o PNG.
- IPTC: Fue desarrollado inicialmente como un estándar para el intercambio de información entre las organizaciones de noticias y ha evolucionado con el tiempo.
Alrededor de 1994, Adobe Photoshop "File Info" habilitó un formulario para que los usuarios pudieran insertar y editar los metadatos de IPTC en archivos de imagen digital. Posteriormente fue aprobado por las agencias de fotografía, convirtiéndose en un estándar válido para almacenar metadatos en ficheros de imágenes.
- XMP: es un nuevo estándar basado en XML, desarrollado por Adobe en 2001. Adobe trabaja con el IPTC para incorporar la antigua "encabezados IPTC" en el nuevo marco de XMP y en 2005 el "esquema básico de IPTC XMP" pliego de condiciones fue puesto en libertad. XMP es un código abierto, estándar público, por lo que es más fácil para los desarrolladores adoptar la especificación en software de terceros. metadatos XMP se puede añadir a muchos tipos de archivos, pero para imágenes gráficas en general se almacenan en archivos JPEG y TIFF”, [58].

Búsqueda de Imágenes basadas en el Contenido (CBIR)

Se basa en el contenido de la imagen para recuperarla, es decir en el color, figura, las texturas o cualquier otra característica que pueda proceder de la imagen buscada para de la misma forma buscar la imagen similar en una Base de Datos.

La frase, *basadas en el contenido*, significa que la búsqueda se realizará por el contenido semántico o parecido de la imagen, y no de metadatos o palabras claves, éste contenido semántico que se basa CBIR se los conoce como *descriptores* que son más bien las características de la imagen, a continuación mencionaremos los descriptores más relevantes a la hora de la recuperación de la imagen [64].

- ❖ Color, colores dominantes mediante histogramas. La recuperación de imágenes se consigue mediante el cálculo de un histograma de color.

- ❖ Forma, el fraccionamiento de la imagen en regiones en Área, perímetro. Es decir se refiere a la forma de una región en específico.
- ❖ Textura, es decir su granulosidad, regularidad.
- ❖ Estructurales, busca patrones visuales entre imágenes.

El proceso de búsqueda de imágenes usando CBIR se podría sintetizar de la siguiente manera:

- Para cada imagen de la base de datos se calculan con anterioridad los descriptores.
- La consulta se formula en término visuales, según los descriptores que definan a la imagen de ejemplo.
- Este ejemplo puede ser proporcionado por el usuario o por el mismo sistema de búsqueda de imágenes
- Para satisfacer la búsqueda, el sistema verifica la similitud entre el contenido visual de la consulta y las imágenes de la base de datos. Elaborando un ranking con ellas, según las métricas utilizadas [58].

Una vez establecido una noción general acerca del concepto principal de lo que se refiere al plagio, igualmente distinguir las diferentes formas en las que se puede presentar o darse plagio, herramientas y metodologías existentes en la actualidad para combatir el mismo, daremos paso al siguiente capítulo donde nos centraremos más en los sistemas de detección de plagio automáticos.

CAPITULO II

2. SISTEMAS AUTOMÁTICOS DE DETECCIÓN DE PLAGIO

La mayoría de los sistemas de detección de plagio existentes en la actualidad, se basan básicamente en un mismo esquema que podemos resaltar en tres etapas básicas, que sería **la introducción del documento** a analizar, el **análisis** que realiza el sistema, en este punto dependerá del sistema y la metodología que utilice el mismo (búsqueda en la web, en una base de datos propia, etc.), y finalmente la **obtención de los resultados**.

A continuación presentamos un esquema general del sistema URKUND, descrito en el capítulo 1 sección 1.6.1.

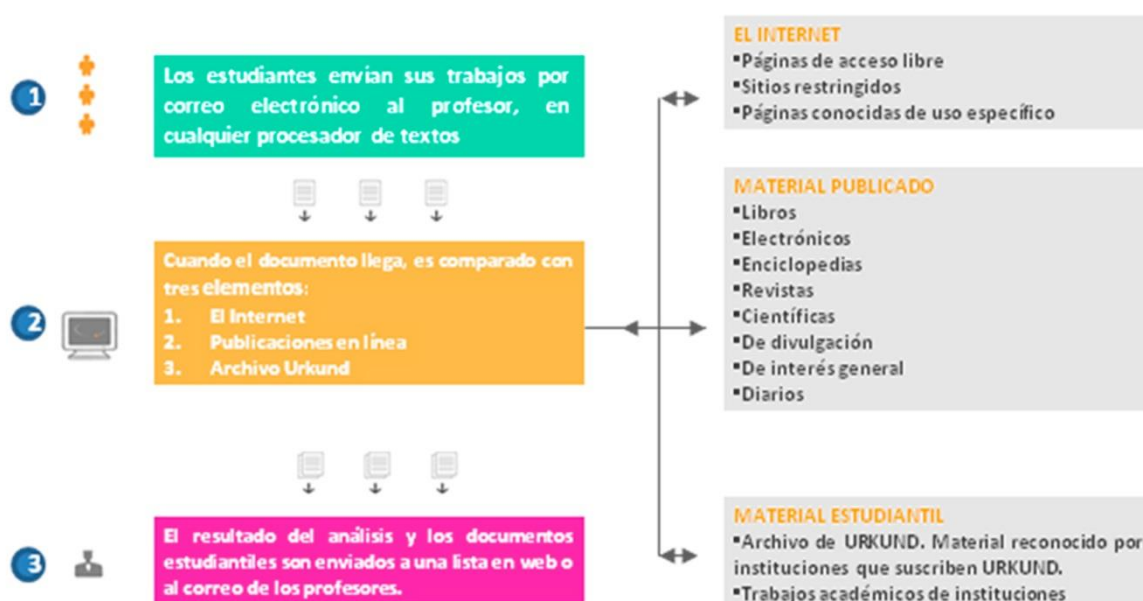


Ilustración 3 Esquema de Funcionamiento General de URKUND [9].

2.1 Particulares Relevantes

Existen algunas teorías o hipótesis que ha arrojado ciertos estudios, que sacan conclusiones como las siguientes.

Los hombres plagian más que las mujeres (aunque estos datos deben ser tomados con prudencia dadas las características de los análisis desarrollados y su difícil comparación) (Comas, Sureda y Urbina, 2005).

Los estudiantes de carreras técnicas plagian más que los alumnos de carreras sociales, humanidades y ciencias (Hansen, 2003; Park, 2001; Ercegovic, 2004).

Ya sea si los estudios estén en lo correcto o no, la realidad que podemos observar día a día, es la existencia indiscutible de plagio, no solo en ámbitos educativos o

laboras, sino en la vida cotidiana mismo, es por lo mismo que cada vez es mayor la cantidad de software que se desarrollan cada año, sacándolo al mercado como un producto más; por lo que asumiríamos la necesidad de asignar una definición a estos sistemas en particular, entonces ¿qué es un sistema de detección de plagio?, un sistema automático de detección de plagio cumple la función de mostrar, exponer o descubrir automáticamente la similitud entre dos textos, en el cuál se debe comparar al menos dos documentos, uno llamado “original” y el otro llamado “plagiado o texto sospechoso.”

El uso de estos determinados programas permite identificar de una manera casi segura, es decir provee una fidelidad y un margen de seguridad alto, en trabajos en los que se sospecha si existe alguna forma de plagio. Por lo general, estos sistemas se basan en el mismo procedimiento básico para la detección del plagio.

Vamos a mencionar como se dijo anteriormente el procedimiento básico que usan la mayoría de estos programas, ya que generalmente es el mismo. Se debe tener el texto en formato electrónico o el archivo dudoso, se compara el mismo ya sea con otro documento o se realiza la comparación en la web, según sea el caso y las características del programa, y éste comienza a hacer una búsqueda de documentos o imágenes que contengan las mismas palabras o puntos en el caso de imágenes, es decir dependiendo de la metodología aplicada (Capítulo 1, Sección 2.4) en dicho sistema.

La búsqueda para realizar la comparación puede llevarse a cabo en Internet o en bases datos locales, es decir contando con un repositorio de toda clase de documentos como son revistas, libros, tesis, etc., una vez finalizada la búsqueda el programa obtiene resultados ya sea en porcentajes o arrojando el nombre de distintos documentos similares, encontrados en la web o localmente, esto se obtiene mediante la aplicación de métodos como unidad de comparación o las ya mencionadas funciones a distancia.

Existen diferentes algoritmos y lenguajes utilizados para el desarrollo de los sistemas automáticos de detección de plagio, algunos de los algoritmos lo describimos en el capítulo 1 del presente documento, y en el caso de los lenguajes de programación los mismos que pueden establecer la calidad o las características de un sistema o aplicación informática, en nuestro caso para el desarrollo del sistema anti plagio para la detección de imágenes, los lenguajes utilizados permitieron la realización de una aplicación estable y de buena calidad, utilizando Java como lenguaje de programación al igual que Python.

2.1.1 ¿Cuándo se considera que existe plagio?

Muchas ocasiones si no es en la mayoría de los casos, no sabemos a ciencia cierta cuándo se considera plagio, o cómo hacen los diferentes software anti plagio existentes en la actualidad para distinguir esta peculiaridad.

Lo que regularmente sucede, cuando se trata de distinguir la existencia o carencia de plagio, es que las respectivas instituciones establecen un criterio propio de estimación. La SENESCYT dentro de los criterios de valoración del porcentaje de similitud y plagio acoge la siguiente escala de valoración y las respectivas acciones que se toma en cada caso.

1% a 10%	No se considera plagio intencional, se puede omitir el reporte y pasar a calificación de trabajo de titulación y trabajos de facultad.
11% a 15%	Se debe enviar al estudiante el reporte para correcciones de malas citas o errores y que presente nuevamente su trabajo.
16% a 30%	El profesor(a) responsable o tutor(a) del trabajo debe emitir un informe al responsable superior del programa o carrera para tomar decisiones.
31% a 40%	El profesor(a) responsable o tutor(a) del trabajo debe emitir un informe al responsable superior del programa o carrera, el mismo que será presentado a la Comisión Docente del programa para tomar decisiones.
Más del 40%	Se presenta un informe al Comité de Ética de la Universidad.

Tabla 8 Criterios de valoración del porcentaje de similitud o plagio [31].

2.2 Impacto social de estos sistemas

Como ya sabemos, en lo que respecta a los métodos o técnicas de enseñanza, uno de los puntos clave es el uso de las TIC (Tecnologías de la Información) que hoy en día se utilizan tanto en centros educativos, como en el trabajo y la vida cotidiana en sí, de esta manera creamos una idea de la importancia que han alcanzado y se puede mencionar muchos aspectos positivos en la utilización de las TIC en lo que se refiere a la enseñanza, entre éstos aspectos podemos mencionar la facilidad, rapidez, simplicidad, etc. Pero así mismo, la utilización de las TIC ha facilitado aspectos NO tantos positivos, es el caso del llamado plagio electrónico, acogido sobre todo en Internet.

Ante esta situación los sistemas automáticos anti plagio ha sido adoptados por varias instituciones, ya sea desarrollando los diferentes sistemas anti plagio, o simplemente utilizándolos, varias universidades a nivel mundial comenzaron a usar los diferentes programas, ya existentes en la web, ya sea de forma experimental o aplicando en la vida real con sus respectivos estudiantes, ideales para la detección de plagios en los respectivos trabajos, tesis, ensayos y proyectos que presenten los estudiantes, en la mayoría de casos, el objetivo de la utilización de éstos sistemas no es el de sancionar al estudiante por la falta cometida, sino enseñar e inculcar a la investigación, desarrollo y mejora de los trabajos por los estudiantes, ya que es la mejor manera de combatir este problema.

La facultad de ingeniería Industrial de la Universidad de Chile, en el año 2012, lanzo un software “antiplagio”, llamado DOCODE, el mismo que afirman lo siguiente “un software antiplagio que ha logrado reducir el fenómeno del “copy and

paste” desde un 55% a menos de un 2% en los establecimientos en los que se aplica.” DOCODE logró el primer lugar en la competencia mundial de plagio 2011 (Competencia PAN, Uncovering Plagiarism, Authorship, and Social Software Misuse) [65].

2.3 Entornos Aplicables

El entorno en el que más fuerza ha tenido estos sistemas de detección de plagio, sin duda alguna ha sido en el ámbito de la educación, con la finalidad de mejorar y fortalecer la calidad de la educación y mejorar sustancialmente los métodos de investigación en las respectivas instituciones, varias universidades a nivel mundial han decidido adoptar estos sistemas como una herramienta que permitirá mantener un control de coincidencias y a la vez detectar la existencia de plagio, tanto en trabajos de clase como en tesis y maestrías.

En la situación académica, el plagio es un problema progresivo, el cual es ayudado por las herramientas que actualmente existen en Internet, por lo mismo se ve la importancia de contar con una herramienta que ayude al personal docente o al investigador a tomar las más acertadas decisiones sobre si un respectivo trabajo fue o no realizado por idea del estudiante.

En el caso de universidades de Ecuador, un gran número de universidades adoptaron el Software Anti plagio Urkund, que fue adquirido y aprobado por la SENESCYT, *“Este sistema se encarga de revisar el documento de tesis comparándolo con la base de datos evitando así el plagio, para esto, una vez que el estudiante envía sus documentos de tesis al software, éste devuelve una respuesta con un porcentaje de similitud que debe ser analizado por el director de la tesis del alumno para que se haga la respectiva corrección y recomendaciones pertinentes previas a la sustentación del trabajo”* [32], cabe recalcar que en un futuro será obligatorio el uso de ésta herramienta en todas las universidades del Ecuador. Aunque la opinión varíe mucho entre los estudiantes universitarios, ya que algunos están a favor y otros en contra, el objetivo de usar esta herramienta es ilustrar al estudiante a crear sus propios trabajos, e incentivar a la investigación y no realizar el típico control c control v, acostumbrado hoy en día, aunque el programa permite copiar párrafos exactos de otros autores, pasará por alto siempre y cuando el mismo este citado correctamente.

Aunque el mayor auge se esté dando en el entorno educativo, también los diferentes sistemas de detección de plagio, lo están adoptando personas particulares sobretodo autores o escritores de libros, revistas, periodistas e inclusive artistas, ya que en los últimos tiempos y tal vez por la rapidez de las redes sociales se ha conocido de varios artistas que han plagiado letras de canciones famosas.

2.4 Técnicas Utilizadas en el Proceso de Detección

2.4.1 Referencias y Citas

Para que un ensayo, tesis y en general cualquier tipo de trabajo que se realice y el cual no sea del todo de nuestra autoría, sino que tuvimos varias formas de guiarnos o trabajos anteriores en que basarnos, y para no considerarse plagio, debe estar adecuadamente citados y referenciados.

Existen algunas normas y estilos para la elaboración de citas y referencias. A continuación mencionaremos detalles de el estándar ISO 690 y algunos estilos para referenciar y citar adecuadamente.

✓ **Norma ISO 690:**

La norma ISO² plantea varias recomendaciones o directrices de cómo referenciar y citar adecuadamente, es decir muestra los datos mínimos que hay que mencionar en cualquier tipo de escrito como por ejemplo monografías, tesis, materiales cartográficos, registros sonoros, fotografías y todo tipo de recursos electrónicos.

Ejemplos **En un libro:**

OLLERO BATURONE, Aníbal. Robótica: manipuladores y robots móviles. Barcelona-España. 2001. 464 pág.

Ya sabiendo que la Norma ISO 690, es un marco de pautas mínimas que debe tener una referencia bibliográfica, algunas instituciones, organismos y universidades han creado sus propios estilos para citar y referenciar; a continuación mencionaremos algunos de éstos estilos.

Documentos Audiovisuales Grabaciones [63]:

APELLIDO(S), Nombre. Título. [Designación específica del tipo de documento]. Lugar: editorial, año.

WAGNER, Richard. El drama musical wagneriano. [Grabación sonora]. Barcelona: CYC, 1998.

✓ **APA:**

APA³, utilizado en el área de psicología y ciencias sociales, éste estilo señala que se debe incluir el nombre del o los autores, el año de publicación y la página o páginas de donde se sustrajo la idea, y también dejar una sangría al comienzo de una referencia [59].

Ejemplos. **Citar Libro**

Juan Lozano, Jessica Zhunio. (2014). "DISEÑO E IMPLEMENTACIÓN DE UN SISTEMA PROTOTIPO PARA DETECCIÓN DE PLAGIO EN

² ISO (International Organization for Standardization)

³ APA (American Psychological Association)

IMÁGENES Y UN MÓDULO DE ANÁLISIS DE SIMILITUD DE DOCUMENTOS ACADÉMICOS”. Cuenca-Ecuador.

Ejemplos. **Citar Imagen**

Apellido, I. (2010). Título del trabajo [fotografía]. Recuperado de <http://www.www.www>

✓ **MLA:**

La MLA⁴ se utiliza en áreas de la literatura, artes y humanidades, cuenta con citas breves en paréntesis en el texto incrustado [60].

A continuación se mencionamos algunos ejemplos de cómo referenciar según la 3ª ed. del MLA Style Manual and Guide to Scholarly Publishing [62]:

Libro impreso

Apellido, Nombre. Título. Lugar de publicación: Editor, año. Impreso (o "Print" si se ha redactado en inglés)

Douglas, Christopher. A Genealogy of Literary Multiculturalism. Ithaca : Cornell University Press, 2009. Print.

Meyer, Stephanie. Twilight. New York: Little, Brown and Co., 2005. Print.

Libro electrónico

Apellido, Nombre. Título. Lugar de publicación: Editor, año. Base de datos o sitio web. Web. Fecha de acceso.

McGhee, Derek. End of Multiculturalism : Terrorism, Integration and Human Rights. McGraw-Hill Education: Berkshire, , GBR, 2008. E-Libro. Web. 30 Dic. 2010.

Modood, Tariq. Multiculturalism. Cambridge: Polity Press, 2007. Google Libros. Web. 30 Dic. 2010.

Una página web

MLA no incluye URL en sus citas porque considera que éstas son inestables y muchas veces los documentos pueden aparecer en varios lugares en la red.

✓ **IEEE:**

El estilo IEEE⁵ relacionada con la electrónica, y con el área de la informática. El formato que sigue éste estilo se basa en que las fuentes se muestran mediante un

⁴ MLA (Modern Language Association)

número y la lista de referencias se ordena numéricamente según el orden en el que aparece en el texto. El presente documento está basado en éste estilo.

Ejemplo de un extracto de un libro.

El propósito de este tema es el poder tratar y buscar una solución a las implicaciones éticas y educativas al momento de realizar o manejar información para las distintas presentaciones académicas ya sea por parte del docente o estudiante [1].

Ejemplo de una imagen.

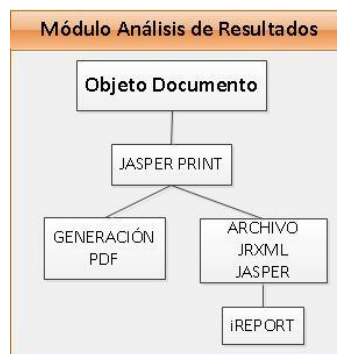


Fig1. Módulo de Análisis de Resultados [2].

Al final del documento en la sección de Referencias y Citas exponemos lo siguiente.

[1] Juan Lozano, Jessica Zhunio. (2014). “Diseño E Implementación De Un Sistema Prototipo Para Detección De Plagio En Imágenes Y Un Módulo De Análisis De Similitud De Documentos Académicos”, pag.1

[2] Juan Lozano, Jessica Zhunio. (2014). “Diseño E Implementación De Un Sistema Prototipo Para Detección De Plagio En Imágenes Y Un Módulo De Análisis De Similitud De Documentos Académicos”, pag.34

2.4.2 Tesoros

Un Tesauro está compuesto por una lista de palabras analizadas y normalizadas, son herramientas de almacenamiento y recuperación de información donde se encuentra de forma ordenada datos e información, usualmente también se lo conoce como diccionario en donde estas palabras se utilizan para constituir conceptos.

La ISO 2788-1986 define un tesauro como *"un vocabulario controlado y dinámico, compuesto por términos que tienen entre ellos relaciones semánticas y genéricas y que se aplica a un dominio particular del conocimiento"*. La ISO 2788-

✓ ⁵ IEEE: (Institute Of Electrical And Electronical Engineers).

1986 también dice que los tesauros pueden definirse desde dos puntos de vista diferentes:

“Según su FUNCIÓN, son instrumentos de control terminológico para trasladar a un lenguaje más estricto (lenguaje documental) el lenguaje natural utilizado en los documentos.

Según su ESTRUCTURA, se trata de vocabularios controlados y dinámicos, aplicables a un área concreta del conocimiento y cuyos términos mantienen relaciones semánticas y genéricas entre sí” [40].

2.4.2.1 Funciones que cumple un Tesauro

Las funciones típicas normalmente de un tesauro que se realiza a lo largo de un tratamiento documental son las siguientes: como primera función tenemos.

La normalización del vocabulario, aquí está todo lo que es la sinonimia, polisemia, género y número de expresiones, es decir su fin es juntar o unificar todo el vocabulario de dicho tesauro y de esta manera controla todas las posibles entradas.

La Inducción, su función consiste en que el tesauro muestre continuamente todas las opciones que puede existir, es decir el tesauro induce al usuario hacia la consulta de temas que no se había tomado en cuenta, esto se debe a las referencias cruzadas que indican las relaciones asociativas, jerárquicas y preferenciales entre los diferentes términos del tesauro [42].

Representación, que se refiere a la correspondencia entre término que componte en los índices y concepto respectivo en el documento.

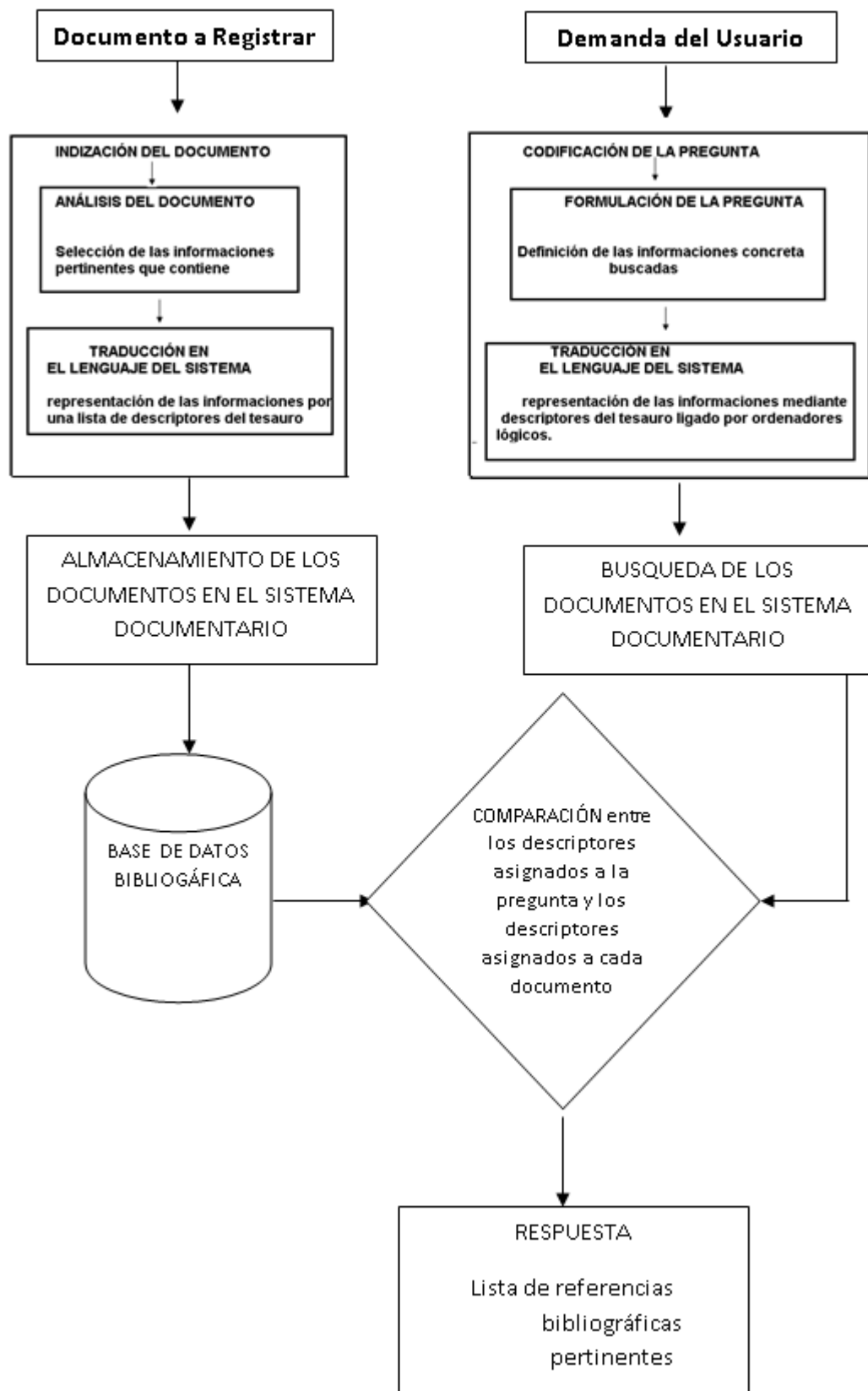


Ilustración 4 El Papel del tesauro en un sistema de documentación [42]

2.4.2.2 Composición de un Tesauro

Los tesauros se componen principalmente por las unidades léxicas y las Relaciones Semánticas, a continuación detallaremos un concepto básico de cada una de las unidades léxicas como semánticas.

Unidades Léxicas	Descriptor
	No descriptor
	Palabras Herramienta
Relaciones Semánticas	Relaciones de Equivalencia
	Relaciones Jerárquicas
	Relaciones Asociativas

Tabla 9 Composición de un Tesauro [42]

a) **Unidades léxicas:** Está formado por grupos de descriptor agrupados por categorías ya sea temas o facetas.

- i. **Descriptor:** *“Palabra o grupo de palabras incluidas en un tesauro y escogidas de entre un conjunto de términos equivalentes para representar sin ambigüedad una noción contenida en un documento o en una petición de búsqueda documental. Ej.: AGUAS SUBTERRÁNEAS”, La Asociación Francesa para la Normalización (4FNOR) [42].*

Son las unidades léxicas más importantes de los tesauros. Son palabras que representan los conceptos de mayor relevancia en el documento. Normalmente un tesauro contiene miles de descriptor.

- ii. **No descriptor:** Son sinónimos o cuasi-sinónimos de los descriptor. No se utilizan para la indización, ni para consultas, pero cada no descriptor llama a uno o dos descriptor con el cual mantiene una relación semántica [42].
- iii. **Palabras Herramienta:** Estas palabras no tienen un significado al menos que lo juntemos con un descriptor, los tesauros tratan de evitar para mejorar la recuperación (homonimia, homofonía, homografía, polisemia y sinonimia) [41].

b) **Relaciones semánticas**

- i. **Relaciones de Equivalencia:** Controlan la sinonimia, es decir todos los términos equivalentes o sinónimos se agrupan en términos preferenciales, cuando se detecta sus equivalencias semánticas se selecciona uno de los términos preferenciales para representar dicho concepto y a la vez toma otra la condición de un descriptor.

- ii. **Relaciones Jerárquicas:** Como su nombre lo dice, sirve para expresar relaciones de superioridad o inferioridad entre los diferentes conceptos dentro de una misma clase.
- iii. **Relaciones Asociativas:** Establecen un mismo nivel de jerarquía entre los diferentes conceptos, es decir existe una mutua relación entre diferentes términos, de tal forma que si cambiamos el orden no se altera el concepto.

2.4.2.3 Creación de un Tesauro

Existen algunas directrices o pasos para la elaboración de tesauros, que según la ISO 2788 hay que tener en cuenta. La elaboración de un tesauro puede llevar varios meses ya que dependiendo de su calidad y extensión puede variar, en un promedio un tesauro puede abarcar de 2000 a 3000 palabras. A continuación veremos algunos de estos pasos que deberíamos seguir para su elaboración y detallaremos lo más simple posible.

- Delimitar bien el campo: Hay que definir el grado de precisión en la que vamos a trabajar, una especialización en concreta.
- Recopilación de Términos: Establecer una lista de palabras que más se utiliza y que contenga un mayor significado de acuerdo al campo o especialización que se esté realizando.
- Establecer relaciones entre términos: Aquí aplicaríamos las relaciones de equivalencia, jerárquicas y asociativas, estudiadas anteriormente.
- Comprobar el funcionamiento: Para saber con precisión si está o no funcionando el tesauro, es prescindible ir haciendo pruebas a lo largo de la elaboración, de esta manera podremos corregir los errores que pudieren existir, sin embargo necesitaremos a lo largo del tiempo realizar modificaciones e ir perfeccionando nuestro tesauro.

2.4.3 Corpus

Un conjunto de documentos digitales originales que se pueda manejar como corpus de referencia, es un recurso ineludible en el diseño de los métodos para desarrollar software anti plagio.

2.4.3.1 Corpus METER (MEasuring Text Reuse)

El corpus METER (Texto medición de Reutilización), fue creado por la Universidad de Sheffield, por los departamentos de Periodismo y Ciencias de la Computación, el objetivo de este proyecto es detectar y medir el reúso de texto; originalmente fue creado para el estudio y análisis de texto periodístico, ya que se daba muchos casos de plagio entre periodistas, y una noticia de un periódico aparecía al día siguiente en

otro periódico. El corpus METER está disponible únicamente con fines de investigación y se puede obtener sin costo alguno solo para uso académico, el corpus METER brinda más de 700 ejemplos de reutilización de texto y se compone de 1.716 textos publicados entre el 12 de julio de 1999 y 21 de junio 2000. El corpus METER podemos encontrarlo en tres formatos diferentes, texto sin formato, formato SGML y formato XML. [35]

2.4.3.2 Corpus DeJaVu

Este corpus tiene plagios reales, en una base de datos de publicaciones médicas [38], la misma que está disponible en la siguiente dirección web, <http://dejavu.vbi.vt.edu/dejavu>.

El inconveniente con este tipo de Corpus es que fue hecha para que los médicos puedan ver si plagiaron alguna publicación, mas no para utilizarlo en un sistema antiplagio, ya que la base de datos no contiene todo el texto o publicación sino que solo contiene el resumen. [39]

2.4.3.3 Corpus Webis

El corpus Webis fue creado en la Universidad de Weimar, éste corpus contiene plagios sintéticos o falsos, es decir, los plagios han sido creados de manera falsificada. Los documentos utilizados para realizar éste Corpus son un conjunto de 101 monografías o artículos recogidos de la biblioteca digital de ACM10 [35].

2.4.3.4 Corpus Personalizado

Como su nombre lo indica este corpus se realiza al criterio de cada uno, o del grupo de trabajo que lo necesita, dependiendo de las necesidades y las características que deba tener para acogerse y satisfacer las necesidades finales, en nuestro caso necesitaremos un corpus con que contenga varias imágenes, con referencia y sin ella.

2.5 Líneas de investigación abiertas en las que se basan estos sistemas

A largo del tiempo se han ido desarrollado varias técnicas, (Modelo Espacio Vectorial, Hashing, N-Gramas, etc.), unas mejores que otras, y que han arrojado excelentes resultados cuando se trata de detectar plagio en documentos, aunque si bien es cierto que estas técnicas o metodologías han ayudado enormemente a desarrollar los diferentes software de detección de plagio, también es cierto que todavía estamos lejos de encontrar una técnica cien por ciento confiable y que resuelva este problema de una vez por todas, ya que la falencia de la mayoría de software que se ha desarrollado es el tiempo de respuesta en el que arrojan los resultado, el mismo que dependiendo de la extensión del documento a analizarse puede demorarse horas e incluso días; una línea de investigación más recientes son la detección de plagio aplicando métodos translingüe, o detección de plagio en diferentes idiomas y estilometría.

2.5.1 Detección de plagio Translingües

Está definido que “un texto es considerado un plagio de otro si sus contenidos son considerados semánticamente similares, sin importar que estén escritos en idiomas diferentes, y la correspondiente cita no está incluida” [34].

En las últimas décadas se han presentado varios modelos de diccionario y, modelos de alineamiento para combatir con similitudes a través de diferentes idiomas para la detección automática de plagio translingüe que trata de identificar el plagio entre documentos en diferentes idiomas.

Para detectar plagio translingüe, todos los modelos mencionados anteriormente a nivel monolingüe son efectivos, ya que pueden detectar párrafos impropios del autor.

Existen cuatro categorías de modelos de análisis de similitudes que se utiliza para detección de plagio translingüe:

- Modelo con el uso de diccionarios, reglas o tesauros lingüísticos para ejecutar las traducciones de los conceptos desde un lenguaje origen a un destino.
- Modelos que se basan en la sintaxis y estructura del documento a analizar, para comparar los dos documentos.
- Modelos que utilizan corpus comparables.
- Modelos que se basan en un corpus semejante [33].

CAPITULO III

3. ANÁLISIS Y PLANIFICACIÓN DE LOS REQUERIMIENTOS DEL SISTEMA

Una vez conociendo a ciencia cierta los conceptos básicos, técnicas, herramientas, métodos utilizados, y el impacto que ha tenido en ciertos sectores, sobre el plagio de documentos y la detección del mismo; en éste capítulo nos centraremos a la iniciación de la creación o desarrollo de nuestro *sistema prototipo para detección de plagio en imágenes y un módulo de análisis de similitud de documentos académico*.

3.1 Identificación de los Requerimientos

Para iniciar con la meta propuesta identificaremos los respectivos requerimientos basándonos en conceptos propuestos por la IEEE como son.

- Condición o capacidad que un usuario necesita para resolver un problema o alcanzar un objetivo.
- Condición o capacidad que debe exhibir o poseer un sistema para satisfacer un contrato, estándar, especificación, u otra documentación formalmente impuesta [43].

Teniendo este idea ya en percepción, nos damos una idea más clara de lo que se pretende realizar, en nuestro caso se identificaron cuáles son las principales funciones que debería tener nuestro sistema anti plagio, tenemos una visión muy amplia acerca de las necesidades y expectativas de ésta herramienta de detección, según comentarios y acotaciones tanto de personal capacitado en el área de Ingeniería de Sistemas así como los usuarios finales.

El problema que encontramos al momento de la adquisición de requisitos es la de conseguir la información correcta, sobretodo en cuanto a qué herramientas podemos utilizar para la comparación entre imágenes, por esta razón es que debemos controlar cada paso en este proceso de desarrollo, tomando en cuenta que por lo general los requerimientos variarán antes, durante e incluso después de haber finiquitado el sistema, ésta característica es muy particular de la mayoría de software, ya que a medida que se mejora el sistema, el usuario (cliente y/o programador) sabe con más certeza que es lo que realmente quiere o necesita para su funcionamiento o para continuar realizando dicha aplicación.

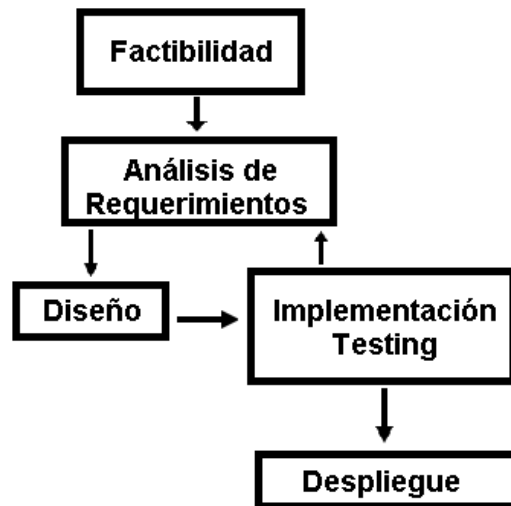


Ilustración 5 Etapas que compone Sistema de Detección de plagio en imágenes.

3.1.1 Requerimientos Funcionales

La idea primordial es la de contar con un sistema que permita al usuario final detectar el plagio de imágenes en documentos; como mencionamos en un principio sabemos de la existencia en la actualidad de numerosos software anti plagio, la mayoría de éstos sirven para detectar plagio textual y muy pocos para detectar plagio de imágenes.

Lo que pretendemos desarrollar es un sistema de detección de plagio de imágenes, que sea capaz de no solo identificar imágenes plagiadas liberadas en la web, sino también se aspira que a partir de un documento extraigamos las imágenes que haya en el mismo y compararlas en la web.

Una vez que tengamos el documento a analizar; se procederá a comparar las imágenes que no se encuentren referenciadas, ya que en caso de tener referencia no se considera que cierta imagen haya sido plagiada, de igual forma a como se considera en plagio de texto.

Para lograr desarrollar nuestro sistema de detección de plagio de imágenes, mencionaremos tres requerimientos principales en los cuales nos vamos a establecer.

- Capacidad de comparar imágenes.
- Búsqueda de imágenes en Internet.
- Extraer imágenes del documento analizado.

Además de requerimientos como es utilizar un determinado buscador, y poder indicar el número de páginas o enlaces de referencia a descargar; el que pueda ser una librería o api para futuros usos, y demás parámetros a incluirse posteriormente. Otro requerimiento de importancia es el cómo va a interpretar el usuario los resultados finales y por último el formato de referencia que se va a usar en el sistema ya sea APA o IEEE., os cuales se van a tomar después de la imagen.

3.1.2 Requerimientos No Funcionales

Teniendo en cuenta que existen muchas particulares que pueden limitar el sistema, detallaremos a continuación algunas características que son relevantes en cualquier tipo de sistema y lo relacionaremos con nuestro futuro sistema de anti plagio de imágenes.

Nombre	Necesario	Crítico	Descripción
Rendimiento	SI	SI	El promedio de comparación de imágenes estimamos unos 5 segundos por imagen, y el tiempo de descarga una imagen dependerá del ancho de banda del usuario final.
Disponibilidad	SI	SI	El sistema funcionará con Internet, sin el uso de éste será imposible acceder a búsquedas de imágenes plagiadas ya que no cuenta con repositorios locales.
Portabilidad	SI	NO	El sistema será portable ya que brinda la capacidad de interactuar con otras plataformas.
Interoperabilidad	SI	NO	Será interoperable ya que puede acoplarse a otros sistemas o lenguajes, ya que manejaremos interconexión con webservices.
Concurrencia	NO	NO	Dependería básicamente del ancho de banda y la capacidad de procesamiento del servidor utilizado para realizar las respectivas consultas.
Escalabilidad	SI	NO	Permite el desarrollo de nuevas Funcionalidades en un futuro, ya que generalmente está formado por módulos y no forma parte de otro sistema.
Seguridad	NO	NO	No se da mayor importancia ya que no se manejan datos transaccionales.
Costo	SI	SI	Dependiendo del costo se adquiere el servidor con sus respectivas capacidades, al igual que el mantenimiento del sistema a lo largo de su vida funcional.

Tabla 10 Descripción de Requerimientos No Funcionales

3.2 Análisis de la Solución

Una vez reconocido los requerimientos para el desarrollo de nuestro sistema continuamos con el siguiente punto que es el diseño de la solución, buscamos diferentes formas, métodos, herramientas, plataformas, diseños, etcétera, para el desarrollo de nuestro sistema, pero antes justificamos las diferentes factibilidades y las describimos a continuación.

○ Factibilidad Operativa

El personal docente de la Universidad Politécnica Salesiana, al cuál va dirigido nuestro sistema tiene suficientes conocimientos en el área de sistemas, los mismos que confirman que es necesario el sistema de detección anti plagio tanto de texto como de imágenes. Sin embargo al ser una herramienta completamente nueva, será absolutamente necesario capacitar a ciertos usuarios y éstos a la vez a otros que lo deseen. Por todo lo mencionado podemos asegurar que es factible Operativamente.

○ Factibilidad Técnica

La Universidad Politécnica Salesiana cuenta con equipo de hardware y software necesarios, y así que el sistema se puede ejecutar con el equipo y personal necesario para su eficiente funcionamiento, sin embargo en lo que se refiere a la elección de las herramientas de soporte y el lenguaje de programación en el cual se va a desarrollar el código del sistema, de deberá experimentar con algunas herramientas y diferentes lenguajes, ya que deseamos escoger la que proporcione un tiempo de respuesta vertiginoso, éstas herramientas las describiremos más adelante con mayor detalle, por lo tanto se puede decir que nuestro sistema de detección de plagio de imágenes es factible técnicamente.

○ Factibilidad de Fechas

Sabiendo a ciencia cierta que éste Sistema es una Tesis previa a la obtención del título en Sistemas, se había programado que el proyecto del sistema de detección de plagio de documentos se desarrollara en un estimado de 6 meses. Por lo tanto consideramos que es factible temporalmente.

○ Factibilidad Económica

La institución cuenta con los suficientes recursos económicos, que permita compensar los requerimientos que solicita el sistema anti plagio, tomándolo desde un punto de vista que es una inversión, sobretodo en equipo, ya que los conocimientos y la investigación es parte de los estudiantes que desarrollan el software. Por lo tanto es factible económicamente.

Una vez que percibiéramos que el desarrollo de nuestro sistema era factible para su desarrollo, comenzamos con el análisis de diferentes soluciones, entre éstas soluciones vamos a detallar la primera para seguir adelante con nuestro proyecto.

3.2.1 Primera Solución

Cómo ya sabíamos, pretendemos desarrollar es un sistema de detección de plagio de imágenes, que sea capaz de identificar imágenes plagiadas y que a partir de un documento extraigamos las imágenes que se encuentran en el mismo y compararlas con imágenes que se encuentran en la web, de ésta manera comprobar si cierta imagen es plagiada o no, independientemente si éstas imágenes hayan sido modificadas de alguna manera, ya sea cambiando su tamaño, recortándola o cambiando de color, en fin.

Como primera solución analizamos la posibilidad de intentar comparar imágenes utilizando el método de comparación de imágenes como son los metadatos, que como ya lo habíamos mencionado en la sección 1.7.4., servían para comparar imágenes entre sí, obteniendo los datos ocultos de cada imagen a través de bits.

3.2.2 Segunda Solución

Analizamos como segunda solución a nuestros requerimientos, la misma que la describimos a continuación.

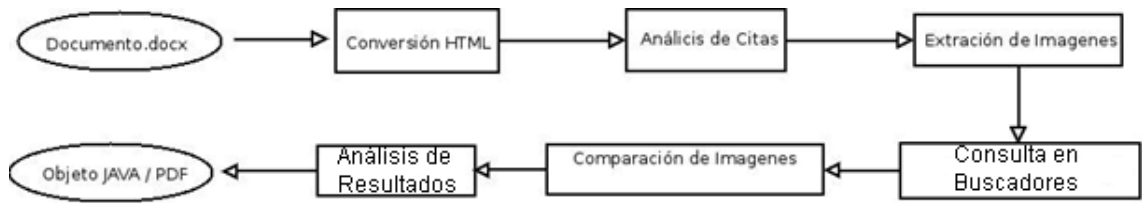
Como primer punto y una vez que poseamos el documento sospechoso a analizar; continuamos con ver que imágenes están referenciadas y cuales no lo están, ya que en caso de tener referencia no se considera que cierta imagen haya sido plagiada.

También pretendemos restringir los formatos de archivo que el sistema reconocerá, por el momento consideramos DOC y DOCX ya que esto son los que más se utilizan para la elaboración de trabajos escritos.

Una vez que tengamos el documento en formato DOC o DOCX; se intenta convertir el documento en HTML, para tratar en encontrar las citas a través de expresiones regulares, de ésta forma encontraremos y extraeremos las respectivas imágenes que existan en el documento a analizar para su posterior consulta en buscadores como es GOOGLE.

Ya teniendo los resultados de búsqueda y teniendo las imágenes extraídas del documento DOC o DOCX, realizamos la comparación de las mismas, utilizando algoritmos de comparación ya sea cualquiera de los siguientes algoritmos SIFT, ASIFT, SURF o RANSAC.

A continuación presentamos el siguiente diagrama de bloques donde se indica el funcionamiento general del sistema de basándose en la segunda solución planteada.



En éste diagrama general de los componentes expuesta en módulos, podemos ver como interactuarían cada una de los componentes de nuestro sistema.

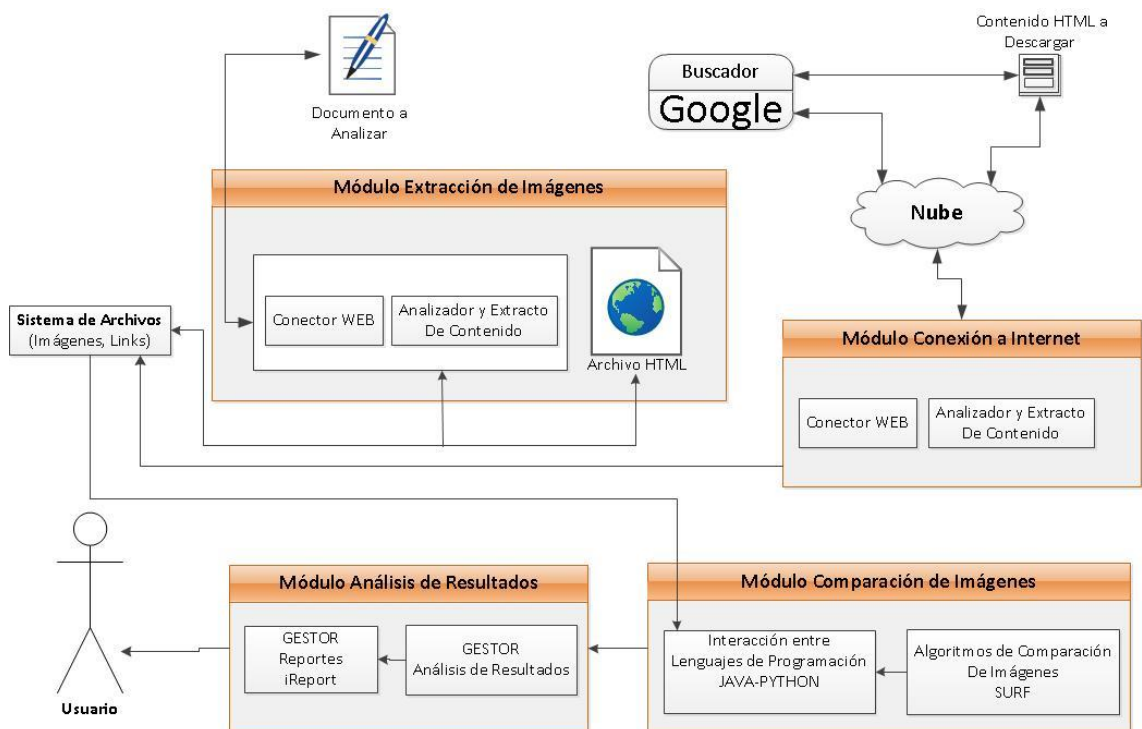


Ilustración 7 Componentes de los Módulos de administración y gestión del Sistema Anti plagio De la solución 2

3.2.3 Tercera Solución

Nuestra tercera posible solución analizada es semejante a la Solución 2, que generalmente va a seguir el mismo proceso, la diferencia radica en que ya no se convertirá el documento a analizar a HTML. El análisis que se realizó para llegar a ésta decisión es que si bien es más fácil obtener las referencias a partir de un archivo HTML, no es necesariamente un punto que tengamos necesariamente tenerlo presente, ya que a diferencia del proceso de análisis de citas, extracción, búsqueda, comparación, y resultados de imágenes, éstos son indispensables en el proceso en general, para llevar a cabo el desarrollo del sistema de detección de plagio.

A continuación presentamos el siguiente diagrama de bloques donde se indica el funcionamiento general del sistema de acuerdo a ésta solución, al igual que un diagrama de los componentes de nuestro sistema que en los próximos capítulos se detallara con más cuidado.

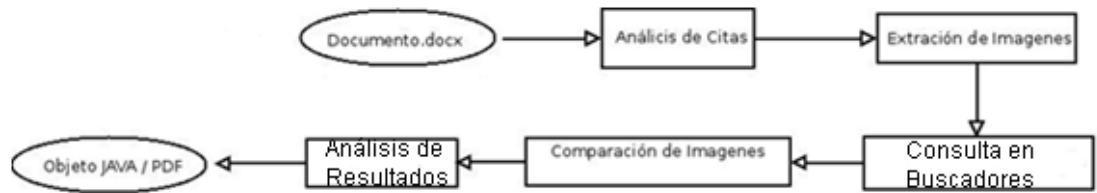


Ilustración 8 Funcionamiento General del Sistema Solución 4

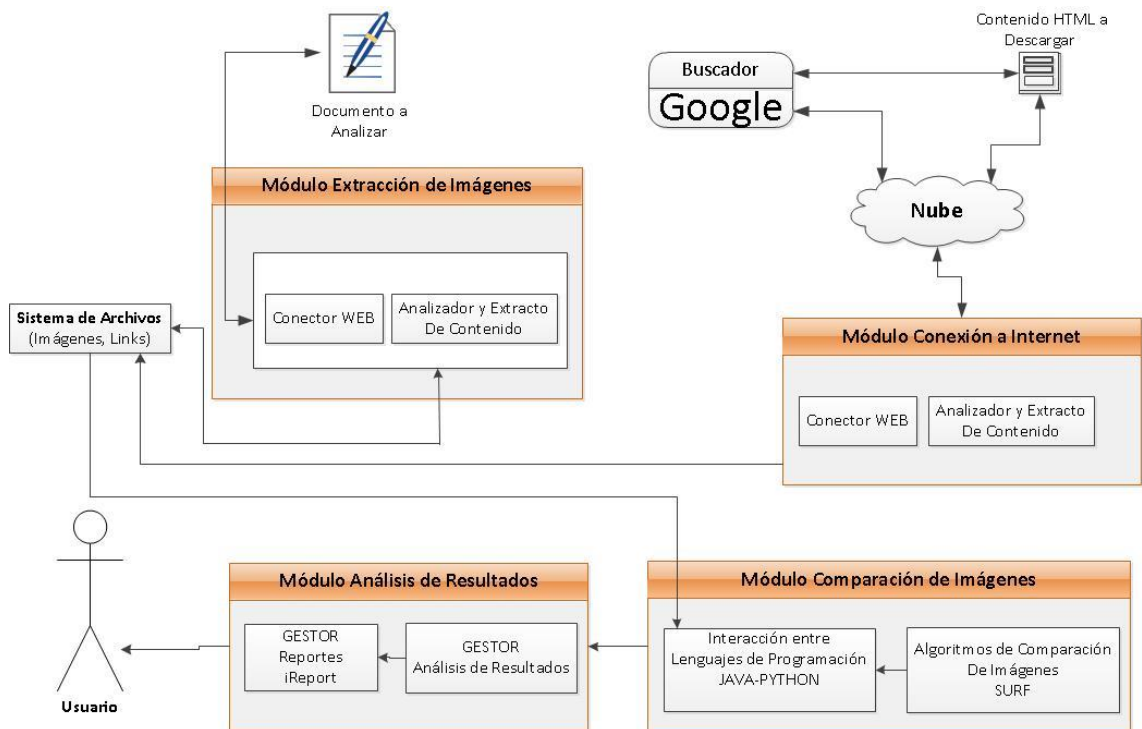


Ilustración 9 Componentes de los Módulos de administración y gestión del Sistema Anti plagio De la solución 3

3.3 Especificación de los Módulos de Trabajo

Para el desarrollo de nuestro sistema de detección de plagio de imágenes, nos hemos planteado constituir por varios módulos indispensables, independientemente de la solución que se elija, de esta manera podemos segmentar adecuadamente el código a desarrollarse. Con el fin de dar a conocer y de entender de mejor manera el sistema mencionamos y detallamos a continuación los módulos que conformaría nuestro sistema de detección de plagio.

3.3.1 Módulo Conexión a Internet

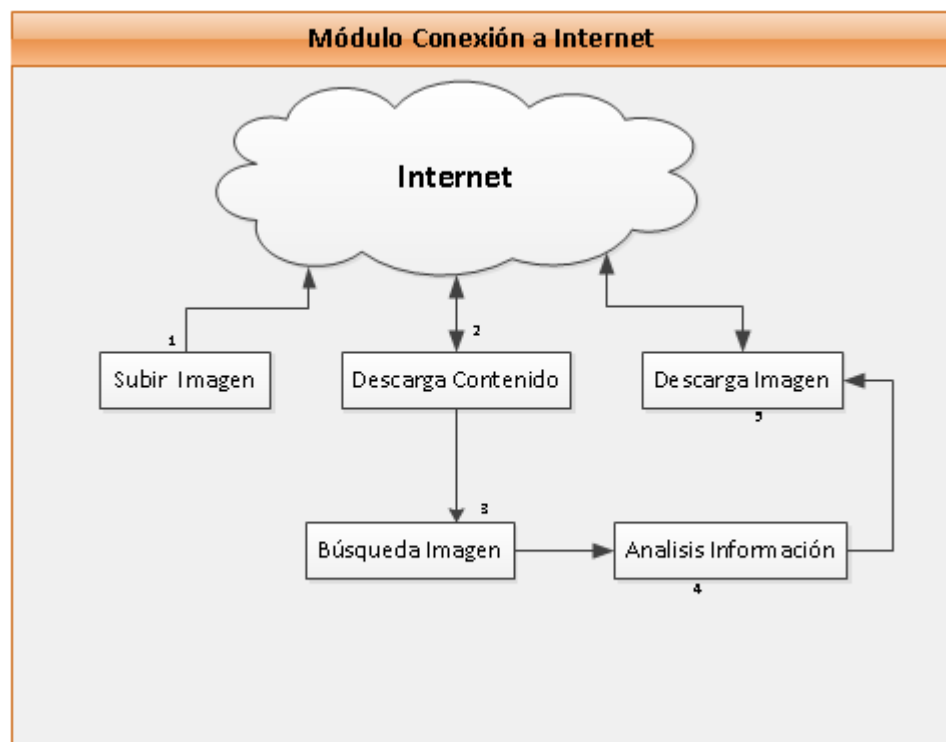


Ilustración 10 Módulo Conexión a Internet

Este módulo se encargaría de realizar las respectivas conexiones a internet, esto comprende los siguientes puntos:

En primer lugar se sube la imagen a los repositorios de internet (1), para posteriormente realizar la búsqueda (2) a través del motor de búsqueda de Google, para que consecutivamente descargar el contenido de la página con los resultados de la búsqueda realizada (3), luego se procede a realizar un análisis de los resultados a partir del contenido HTML(4), para obtener los link de referencia de la imagen que se tiene como objetivo (5) y por último se procede a descargar las imágenes a partir del link de referencia (6).

3.3.2 Módulo de Extracción de Imágenes

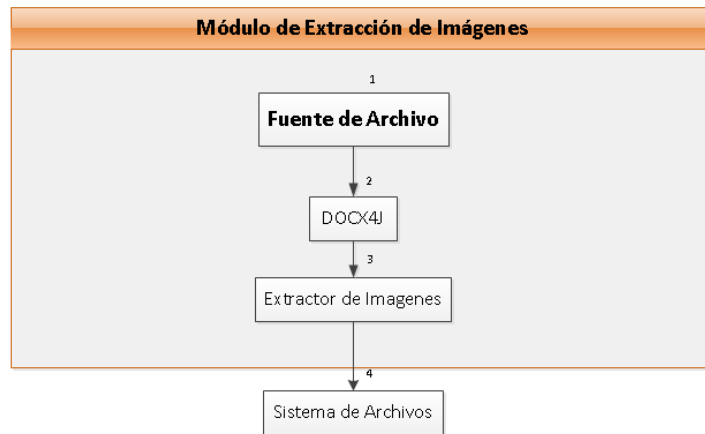


Ilustración 11 Módulo de Extracción de Imágenes

Este módulo tiene por objetivo realizar la extracción de imágenes de un documento dependiendo de ciertos parámetros uno de ellos es que las imágenes que tengan sus respectivas citas no serán analizadas ya que como lo habíamos mencionado no se considera plagio, pero caso contrario que una cierta imagen no contenga su respectiva cita será seleccionada, extraída y analizada, para comprobar si es de autoría propio o plagiada.

Realizando de la siguiente manera:

A partir de un documento a analizar (1), usando librerías de apoyo como es en nuestro caso la librería DOCX4J (2), se procede a realizar la extracción de las imágenes contenidas en el documento, y también se verifica si es que esta o no referenciado o citado correctamente (3), todas las imágenes extraídas del documento analizado se almacena en un sistema de archivos (4), para posteriormente realizar el respectivo análisis.

3.3.3 Módulo de Comparación de Imágenes

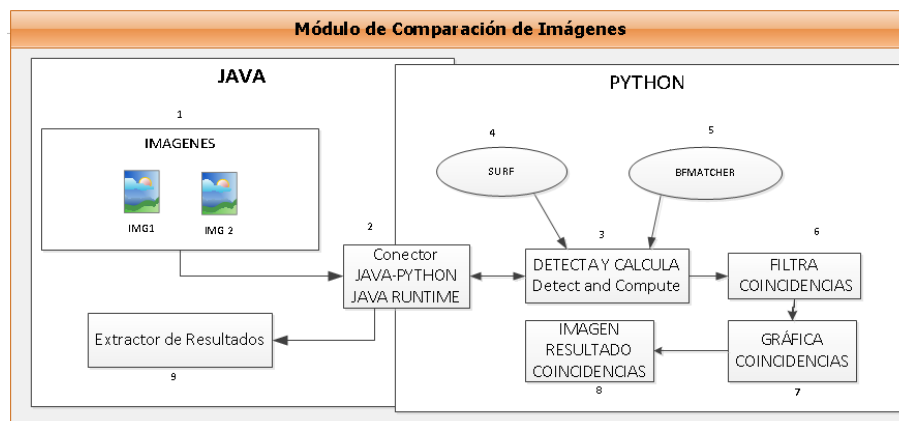


Ilustración 12 Módulo de Comparación de Imágenes

Este módulo tiene como objetivo realizar la comparación entre imágenes locales e imágenes descargadas de Internet, interactuando con el lenguaje de programación Python.

La comparación de imágenes se realiza mediante el algoritmo surf con Opencv, dándonos como resultado puntos similares, y el porcentaje de puntos similares entre las imágenes. A la vez que guarda una imagen comparativa, que refleja gráficamente la diferencia entre las dos imágenes.

La secuencia seria la siguiente:

A partir de dos imágenes (1), en nuestro caso seria las imagen descargada del internet y la del documento, a través de una clase que hace de conector (2), ejecutamos un script realizado en el lenguaje Python, dentro del cual se realiza el siguiente proceso, se detecta y calcula (3) los puntos más característicos de una imagen y similitud a través del algoritmo SURF (3), posteriormente se filtra coincidencias entre las imágenes(6), y se grafica las líneas entre los puntos similares (7), obteniendo una imagen que muestra la similitud entre imágenes.

Desde JAVA se lee los resultados obtenidos por el script (8), para agregar información a los objetos correspondientes, y guardar temporalmente en el sistema de archivos.

3.3.4 Módulo de Análisis de Resultados

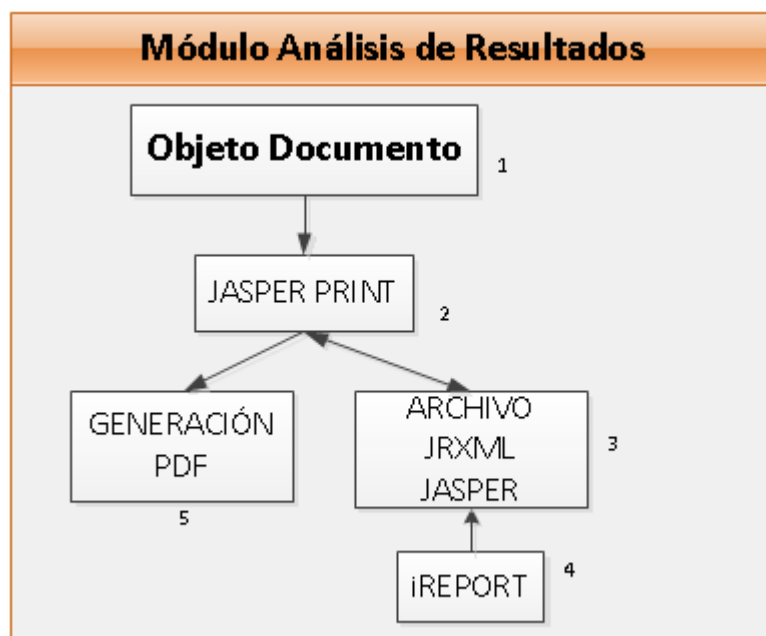


Ilustración 13 Módulo de Análisis de Resultados

Este módulo como su nombre lo dice, se encarga de arrojar los resultados correspondientes al análisis del documento analizado, es decir se encarga de generar un reporte en formato PDF, a través de Jasper Print. Obtendremos el archivo .jrxml

que se genera con la herramienta IReport, para transformarlo a un archivo PDF con datos del Objeto Documento.

Realizándose de la siguiente manera:

A partir de un objeto (1), con los datos a ser mostrados en el reporte, mediante Jasper Print (2), se accede al archivo del diseño del reporte (3), el mismo que está realizado con la herramienta iReport (4), se genera un archivo PDF (5), con los resultados obtenidos, una vez terminado el respectivo proceso de búsqueda y análisis de imágenes.

3.4 Selección de las Herramientas de Soporte

Para que el desarrollo de un sistema se considere un sistema flexible, se debe seleccionar las mejores condiciones para que el proyecto se realice con éxito, las decisiones tomadas deben cubrir las necesidades del proyecto, evitando la realización de actividades innecesarias, por esto la selección incorrecta de herramientas de trabajo podría ocasionar trastornos durante el proyecto e incluso entorpecer el logro de sus objetivos; se consideraron ciertos criterios para la elección de las herramientas que utilizaremos, algunos de estos razones lo describimos a continuación.

- **Su Uso:** Nos referimos al incremento del uso que la herramienta está teniendo en el desarrollo de aplicaciones, o la frecuencia con la que se usa en el desarrollo de diferente software.
- **Su Dominio:** Se refiere al grado de dominio que tiene los programadores hacia la herramienta seleccionada.
- **Su Documentación:** Se refiere a la cantidad de información existente de dicha herramienta, ya que de existir algún inconveniente se pueda recurrir a manuales o consultas en la comunidad informática.

Por éstas razones describiremos a continuación algunas de las herramientas escogidas para el desarrollo y ejecución de nuestro sistema.

3.4.1 Java

Java un lenguaje de programación mundialmente conocido por los desarrolladores, se caracteriza por ser un lenguaje simple, orientado a objetos, distribuido, consistente, seguro, portable, gratuito, de alto desempeño, dinámico, documentado y soporte [44].

Dicho de otra forma, java permite manejar de una manera más simple el código y nos salva de copiar y pegar varias veces el mismo código para realizar un procedimiento similar, además de ésta gran característica java es multiplataforma, por lo que podemos cambiarnos de Sistema Operativo e igualmente podremos trabajar con Java.



Ilustración 14. Logo de Java [44]

Al igual que su dominio, documentación y sobretodo que no depende de licencias se escogió a java como el lenguaje de programación que utilizaremos para el desarrollo de nuestro sistema.

3.4.2 Python

Python es un lenguaje de programación joven a comparación con Java, es una herramienta que se está apoderando dentro de los desarrolladores jóvenes, ya que a comparación con Java es mucho más fácil desarrollar el código, especialmente para los programadores principiantes, ya que cuando no se tiene conocimientos de programación, es recomendable iniciar a practicar con un lenguaje de programación un poco más sencillo, y una de las mejores opciones es Python, que además tiene sus características propias, como es orientado a objetos, multiplataforma, programación imperativa, gratuito, múltiples librerías y más rápido que Java.



Ilustración 15. Logo Python

Java	Python
Tiene una curva de aprendizaje más marcada, más complicada.	La sintaxis perfilada para ser intuitiva y simple.
Tipado ⁶ estático, obliga a indicar el tipo de una variable cuando la declara por primera vez, por lo que no le permitirá cambiar el tipo en un futuro dentro del programa.	Tipado dinámico, permite cambiar el tipo de una variable.
Utiliza las llaves para definir el principio y el final de cada función o clase.	Utiliza indentación ⁷ para separar el código en bloques.
Es un lenguaje compilado a bytecode es necesario compilar una clase en código que puede ser entendido por la Máquina Virtual de Java	Es un lenguaje interpretado, lo que significa que no se necesita compilar el código fuente para poder ejecutarlo. Por lo que lo hace más rápido.

Tabla 8. Diferencias entre Java y Python.

⁶ Tipos de datos que se manejan en un lenguaje de programación, y la forma en que deben usarse.

⁷ Mover hacia la derecha un bloque de texto con espacios o tabuladores.

3.4.3 Algoritmo Surf

El algoritmo SURF en inglés "Speeded Up Robust Features", es un algoritmo para la detección de características específicas de imágenes; es uno de los algoritmos más utilizado para la extracción de puntos de interés y la ubicación de los mismos en una imagen, este algoritmo es invariante a rotaciones, cambios de escala y cambios en los colores de las mismas. Este algoritmo computacionalmente es más rápido que otros algoritmos de detección de imágenes, sin afectar a su rendimiento.



Ilustración 16 Comparación de dos imágenes aplicando el algoritmo Surf [52]

A continuación presentamos los ciclos de este algoritmo son:

- Reproducción de imagen integral para apresurar los cálculos.
- Creación del espacio escala mediante aproximaciones a la segunda derivada de la gaussiana.

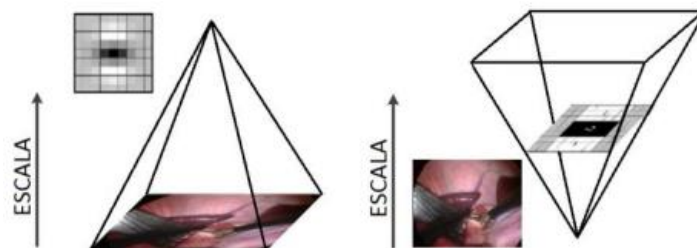


Ilustración 17 Espacio escala mediante SIFT (Izquierda) y SURF (Derecha) [46]

- Orientación de puntos máximos y mínimos del espacio escala;
- Asignación de orientación (con wavelets Haar)
- Generación del descriptor [46].

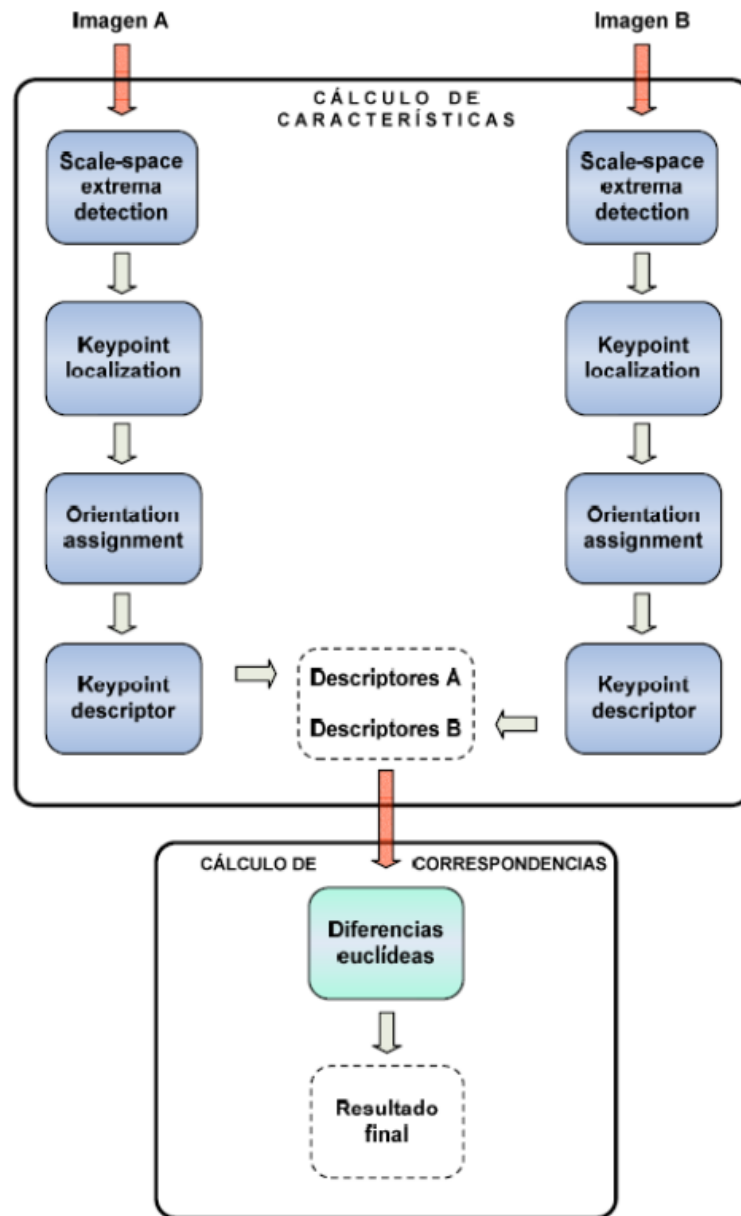


Ilustración 18 Esquema del proceso para realizar nuestro proyecto de clasificación SURF [47]

3.4.4 OpenCV

OpenCV, Visión de computadora de código abierto⁸, es una biblioteca de visión artificial, dirigidas principalmente a la visión en tiempo real por el ordenador [48]. OpenCV es gratuito y multiplataforma (Windows, Linux, Mac OS, iOS y Android). Acogido mundialmente, OpenCV tiene más de 40 mil personas de la comunidad de usuarios y el número estimado de descargas superior a 7 millones.

⁸ Open Source Computer Vision



Ilustración 19 Logo de OpenCV [53]

En nuestro sistema utilizamos OpenCV-Python que es el API de Python para OpenCV, el mismo que combina las mejores cualidades del API OpenCV y el lenguaje Python. La biblioteca OpenCV está lleno de algoritmos y funciones que ayudan a crear software que proporcionará visión a tu computadora, es decir que una computadora, procesará e interpretará las imágenes analizadas enviando los respectivos resultados de acuerdo a la programación realizada por el usuario.

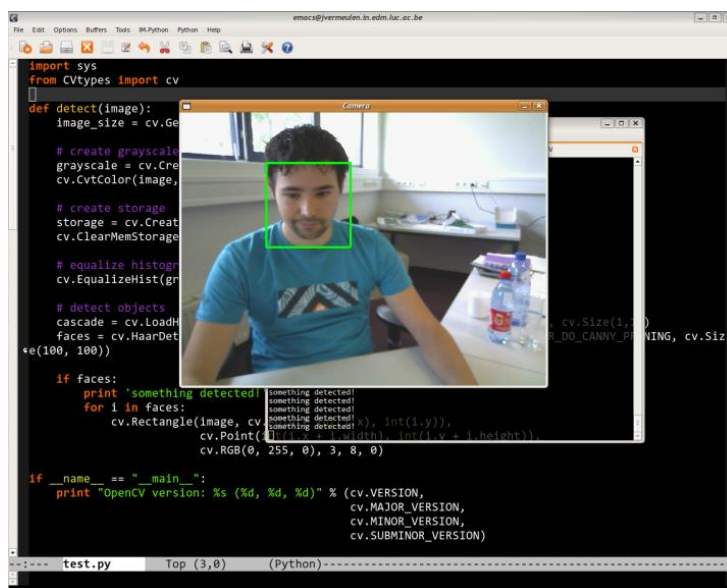


Ilustración 20 Ejemplo de funcionamiento de OpenCV con Python [49]

3.4.5 JASPERREPORT - iReport

JasperReport es una librería que nos proveen la generación de informes de los sistemas en general, y que, junto a iReport que es una herramienta que nos servirá para mostrar los resultados finales en forma gráfica, hacen que la presencia de la mismas sea indispensables en el sistema de detección de plagio, ya que nos servirá para diseñar o generar informes visuales. Lo que iReport permite es corregir visualmente informes ya sea en forma de escritos, imágenes, etc., por lo que es muy fácil de usar; para JasperReports en Java iReport también está integrado con JFreeChart, que es una de la biblioteca gráficas OpenSource más circulada para Java. Los reportes finales se pueden obtener en formatos PDF, XML, XLS, CSV, HTML, XHTML, DOCX, etc., [54].

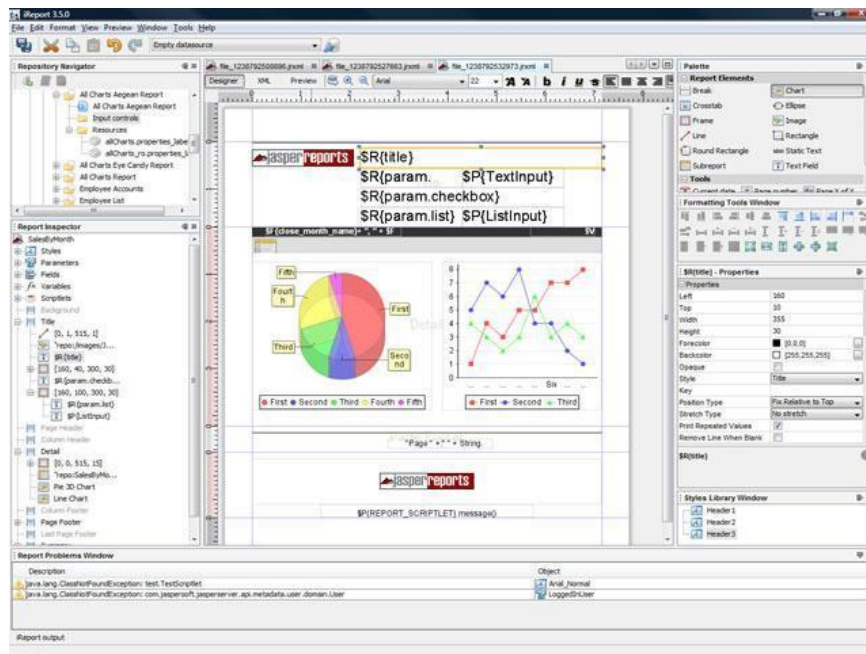


Ilustración 21 iReport 3.5.0 [55]

3.4.5.1 Características de iReport

A continuación mencionaremos algunas características fundamentales propias de iReport.

- “Totalmente escrito en JAVA y además OPENSOURCE y gratuito.
- Maneja el 98% de las etiquetas de JasperReports
- Permite diseñar con sus propias herramientas: rectángulos, líneas, elipses, campos de los textfields, cartas, subreports (subreportes).
- Soporta internacionalización nativamente.
- Browser de la estructura del documento.
- Recopilador y exportador integrados.
- Soporta JDBC.
- Soporta JavaBeans como orígenes de datos (éstos deben implementar la interface JRDataSource).
- Incluye Wizard’s (asistentes) para crear automáticamente informes.
- Tiene asistentes para generar los subreportes
- Tiene asistentes para las plantillas.
- Facilidad de instalación” [54].
- Vista previa integrada es decir, permite visualizar el reporte en diferentes formatos a través del Preview.
- Soporte para construcción de tablas anidadas.

CAPITULO IV

4. DESARROLLO DEL DISEÑO DEL SISTEMA PROTOTIPO

4.1 Análisis de la Arquitectura

Considerando las características del problema de desarrollar un sistema anti plagio de documentos, la identificación de los requerimientos y el empleo posterior de la herramienta de desarrollo mencionados anteriormente, la arquitectura del sistema quedaría establecida de tal manera que se pueda precisar la estructura Entidades, Atributos y Dominios. Una idea planteada por el equipo de desarrollo del sistema anti plagio consiste en dar una ilustración de la arquitectura del sistema en general ya que la arquitectura no puede ser explicado solo en palabras; más bien está definida por algunos parámetros.

A continuación mencionaremos algunos de éstos parámetros para el desarrollo del sistema bajo la respectiva arquitectura dispuesta.

- ✓ Identificación de los estándares y marco de trabajo.
- ✓ Organización lógica.
- ✓ Identificar la funcionalidad que debería tener el sistema a desarrollarse.
- ✓ Identificación de los respectivos paquetes y clases que conformaran el sistema.
- ✓ Organización de los paquetes, clases y las respectivas relaciones entre los mismos.
- ✓ Organización o distribución del sistema en la plataforma.

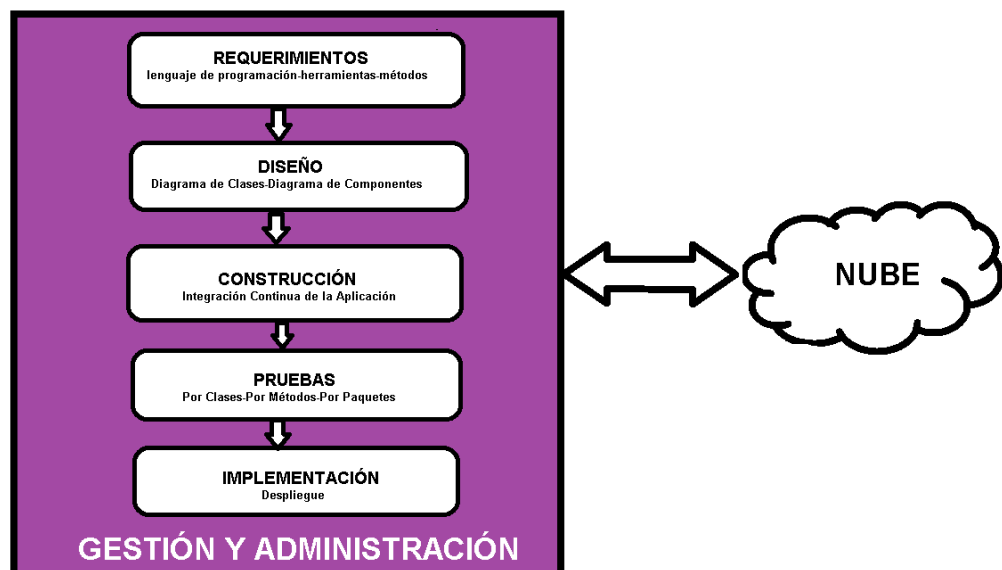


Ilustración 22 Arquitectura integrada por fases en general

La solución que decidimos acoger para resolver nuestros requerimientos fue la Tercera Solución, detallada en la sección 3.2., descartando la primera solución ya que los metadatos no son únicos en una imagen, y pueden haber imágenes iguales con metadatos completamente diferentes, al igual que se descartó también la segunda

solución ya que no la obtención de citas del archivo HTML se dificultó en gran magnitud al momento de la programación.

Para llevar a cabo el desarrollo del tema planteado se realizó un análisis mediante el cual se determinó dividir en dos etapas:

4.1.1 Primera etapa:

La primera etapa estuvo estrechamente relacionada a la teoría, ya que se dio inicio investigando e indagando temas relacionados con técnicas para análisis de imágenes y tratamiento de las mismas, así como también se revisó algunas librerías que serán necesarias para llevar a cabo el desarrollo del tema planteado.

Como resultado de la investigación previa y análisis se tomó la decisión de fragmentar en dos sub módulos o etapas.

4.1.1.1 Análisis de técnica para comparar imágenes

El primer sub módulo comprendería el análisis comparativo de la o las imágenes, utilizando para ello el método o técnica denominado SURF. Cabe recalcar que fueron varias las razones por las que se optó por esta técnica, ya que es eficaz y eficiente, pero el resultado determinante para la utilización de la misma fueron los tiempos para la entrega de resultados y de la alta precisión de los mismos.

4.1.1.2 Proceso de búsqueda en Internet

El segundo sub módulo contemplaría la investigación relacionada a la implementación del proceso o algoritmo necesario para realizar las búsquedas de las imágenes en el internet, así como también investigar expresiones regulares necesarias para la extracción de información estrictamente necesaria y útil para la búsqueda.

4.1.2 Segunda etapa:

En esta segunda etapa se estableció realizar una investigación referente a las herramientas que se utilizaran para el desarrollo del sistema.

Para la implementación del punto 4.1.1.1 se decidió utilizar la librería OpenCV para tratamiento de las imágenes, y como lenguaje de programación se probarían Java en primera instancia y como alternativa Python, ya que según las investigaciones realizadas este último es muy bueno para procesamiento de grandes cantidades de datos, que en nuestro caso vendrían siendo las imágenes.

Para el desarrollo del punto 4.1.1.2 se investigó y se concluyó que se utilizaría la librería JavaNet para realizar las consultas de las imágenes en el internet y para descargar el código HTML obtenido como resultado de la búsqueda.

4.2 Arquitectura Lógica y Física

4.2.1 Arquitectura Lógica

Se optó ésta topología lógica ya que se adapta de manera general a los requerimientos establecidos en un principio, ya que contiene, se distribuye y se acopla justo y de manera eficaz a nuestras necesidades.

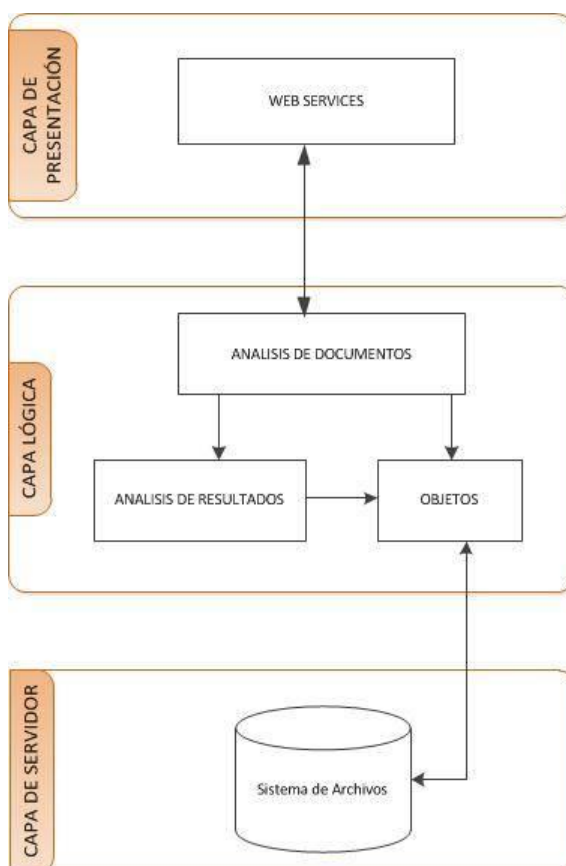


Ilustración 23. Arquitectura Lógica del Sistema de Anti plagio de Imágenes

A continuación mencionaremos una breve descripción, de la Ilustración anterior.

En la capa de presentación, el cliente accede al sistema de análisis mediante servicios web SOAP, estos servicios ejecutan en el core, una serie de procesos tales como análisis del documentos, la extracción de imágenes del documento, análisis de citas, etc., a la vez también realizan el análisis de las imágenes como son la búsqueda en internet, descargas de imágenes y posteriormente el análisis de resultados y generación del reporte, cuando se realiza el análisis se guarda los datos como objetos, los mismo que tiene referencias de path's o referencias de cada imagen, y que se encuentran almacenando en un sistema de archivos, este sistema de archivos tiene la siguiente estructura: carpeta con nombre del documento dentro de la cual se encuentra las imágenes del documento al igual de una carpeta con nombre de cada imagen, en la cual se almacena las imágenes descargadas del internet, y un archivo con información de cada imagen descarga, para su posterior análisis.

4.2.2 Arquitectura Física

El diagrama físico es una representación de cómo se distribuirá la aplicación, tomando en cuenta que la representación del diagrama no siempre coincidirá con la parte física sean están por recursos, tecnología o factores externos y no controlables por el o los programadores.

Se escogió esta topología Física en nuestra aplicación ya que cumple con todos los requerimientos que necesitamos para implementar la parte lógica del sistema a realizarse. Los procesos se ejecutan en máquinas dedicadas que se configuran de la manera más adecuada para ofrecer los servicios que requiere cada parte de la aplicación.

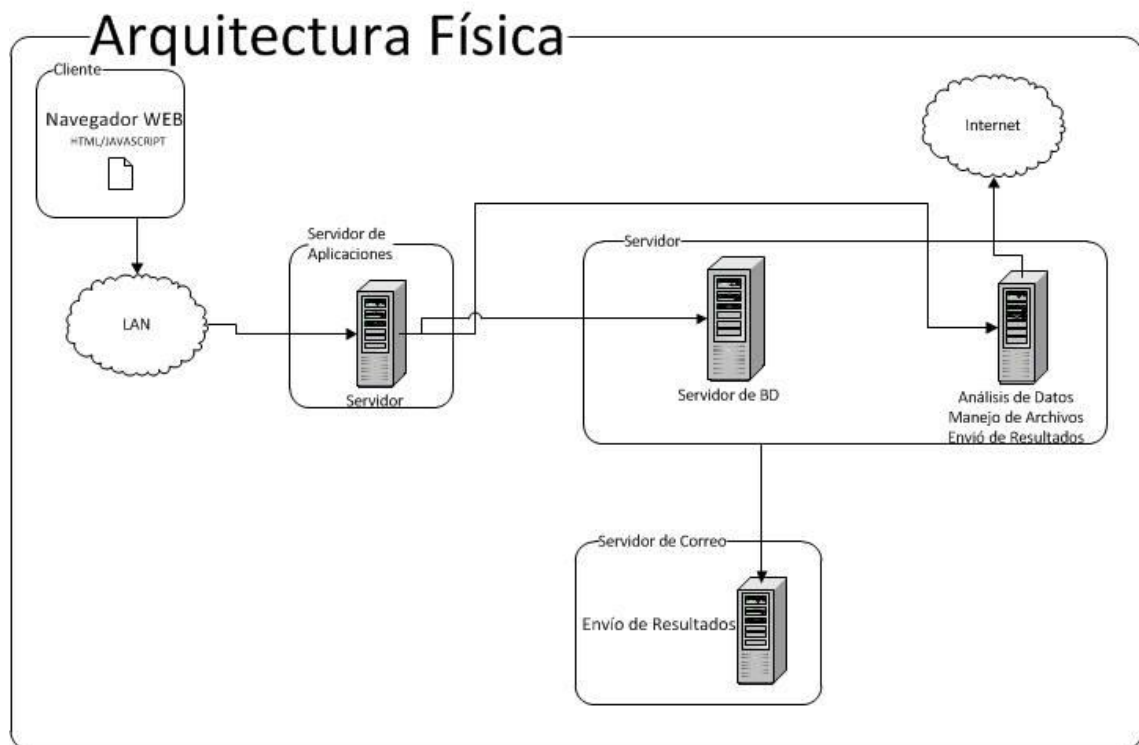


Ilustración 24. Arquitectura Lógica del Sistema de Anti plagio de Imágenes

El cliente mediante un navegador accede al sistema interconectado a través de una red TCP/IP a un servidor de aplicaciones, este servidor realiza la respectiva autenticación, para luego enviar a ejecutar el análisis en un servidor dedicado a este proceso. El servidor donde se realiza el análisis está conectado a internet para realizar las búsquedas de imágenes para posteriormente realizar el respectivo tratamiento de cada una de ellas.

4.3 Documentación UML

Teniendo en cuenta que la documentación es de vital importancia en los diferentes proyectos de desarrollo de sistemas ya que nos permite evaluar el progreso de la aplicación que se está o se vaya a desarrollar, también permite informar de las diferentes funciones que deben cumplir los miembros del equipo y los desarrolladores del sistema.

Asimismo, los diferentes diagramas nos ayudaran a marcar o precisar una ruta que debemos seguir, y que tal vez y con mayor probabilidad más adelante sea necesario actualizar la aplicación.

A continuación ilustraremos algunos de los diagramas esenciales para el desarrollo de nuestro sistema anti plagio.

4.3.1 Diagramas de Casos de Uso

4.3.1.1 Caso de Uso de la Aplicación en General

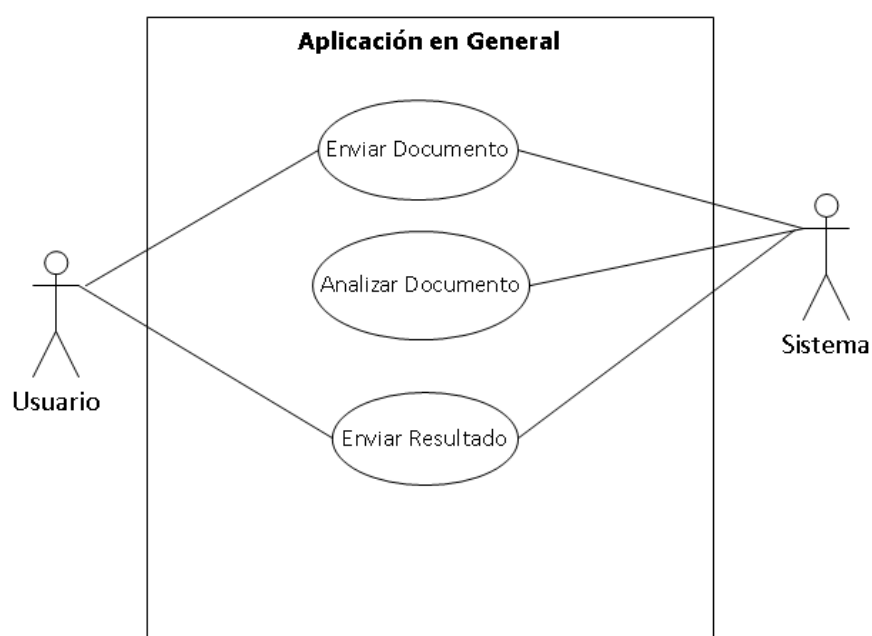


Ilustración 25. Caso de Uso de la Aplicación en General

1	Aplicación en General del Sistema de Detección de Plagio.
Requerimiento	El sistema debe analizar un documento entregado por el usuario, y devolver al mismo los respectivos resultados en formato PDF que el usuario final sea capaz de entender.
Nivel del caso de Uso: ☐ Negocio ☐ <u>Sistema de Análisis e Información</u>	
Actores	Usuario y Sistema.
Requisitos asociados	Cuando tengamos el documento a analizar; se procederá a comparar las imágenes que no se encuentren referenciadas, ya que en caso de tener referencia no se considera que cierta imagen haya sido plagiada.

Curso	1. Iniciar Aplicación 2. Cargar documento a analizar. 3. Esperar respuesta por parte del sistema. 4. Obtener los resultados en formato PDF. 5. Análisis visual por parte del usuario.
Post- estado	Éxito: Carga y analiza el documento en el sistema.
	Fracaso: La aplicación no puede cargar o analizar el documento.
Excepciones	La aplicación no analiza el documento en caso de no existir conexión a Internet.
Frecuencia Esperada	Dependerá del criterio del docente o del usuario final.
Comentarios	La decisión de la existencia o no de plagio dependerá única y exclusivamente del docente o usuario final, el sistema de detección de plagio dará un reporte del análisis del documento de las imágenes parecidas en la web.

4.3.1.2 Caso de Uso – Envío Del Documento

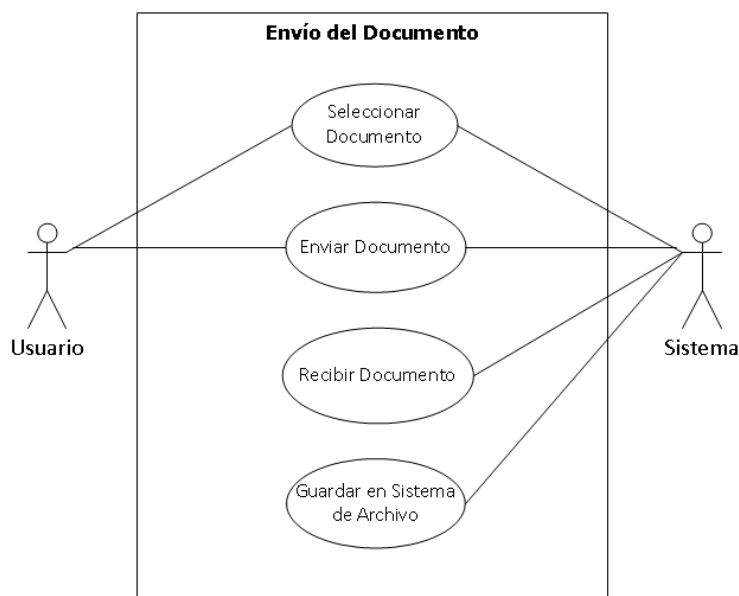


Ilustración 26. Diagrama Caso de Uso Envío Del Documento

2	Envío del Documento
Requerimiento	El usuario debe seleccionar y enviar el documento y el sistema tendrá que recibir el documento, y guardarlo en el sistema de archivo.
Nivel del caso de Uso: ☐ Negocio ☐ <u>Sistema de Análisis e Información</u>	

Actores	Usuario y Sistema.
Curso	1. Selección del documento a analizar por parte del usuario final. 2. El usuario cargará en el sistema el documento que desea analizar. 3. El sistema receptara el documento. 4. El sistema se encarga de guardar en el sistema de archivo.
Post- estado	Éxito: El sistema recibe y guarda el documento enviado por el usuario final.
	Fracaso: No exista conectividad con el servidor por lo tanto no recibe ni guarda el documento a analizar.
Excepciones	El sistema no recibe documento en caso de no existir conexión con el servidor.
Frecuencia Esperada	Dependerá del criterio del docente o del usuario final.
Comentarios	No existe mayor dificultad en el proceso de envío del documento.

4.3.1.3 Caso de Uso – Resultados

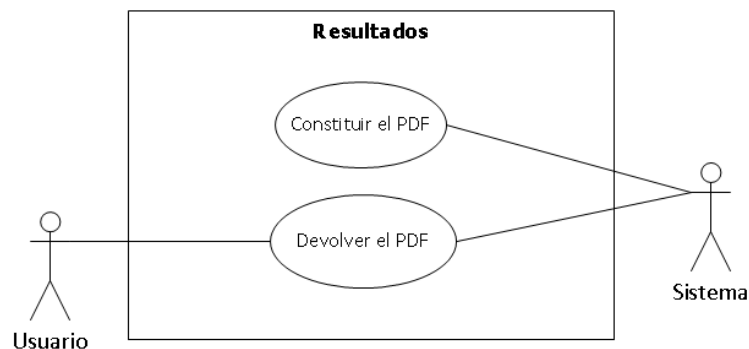


Ilustración 27 Caso de Uso General

3	Resultados
Requerimiento	El sistema construirá resultados del análisis del documento en formato PDF para luego devolver al usuario dicho reporte PDF.
Nivel del caso de Uso: ☐ Negocio ☐ <u>Sistema de Análisis e Información</u>	
Actores	Usuario y Sistema.

Curso	1. Construir el PDF luego del respectivo análisis. 2. El sistema devuelve al usuario el archivo PDF. 3. El usuario recibe archivo PDF con resultados respectivos del documento analizado.
Post- estado	Éxito: Construir y devolver resultados en archivos PDF.
	Fracaso: No devolver ningún archivo, o devolver datos erróneos.
Excepciones	El sistema no recibe documento en caso de no existir conexión con el servidor.
Frecuencia Esperada	Dependerá del criterio del docente o del usuario final.
Comentarios	El sistema brindara bastante estable, como para que el usuario final sienta confianza en tomar una decisión correcta.

4.3.1.4 Caso de Uso – Análisis del Documento

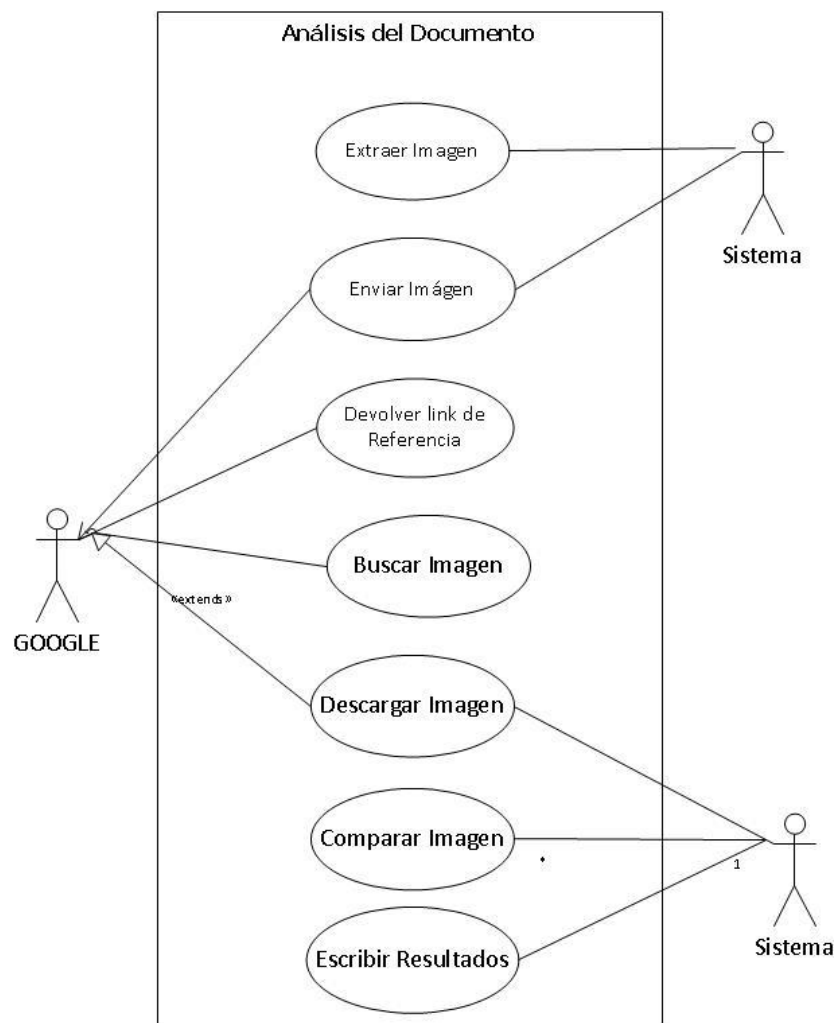


Ilustración 28. Diagrama Caso de Uso Análisis del Documento

4	Análisis del Documento
Requerimiento	Realizar el respectivo análisis de las imágenes que se encuentran en un documento.
Nivel del caso de Uso: ☐ Negocio ☐ <u>Sistema de Análisis e Información</u>	
Actores	Usuario y Sistema.
Curso	1. El sistema extraerá la imagen y la enviara a un buscador (Google). 2. Google revolverá el link de referencia, buscará y descargará la imagen respectiva. 3. El sistema comparará la imagen y escribirá los resultados.
Post- estado	Éxito: Extracción de las imágenes sin referenciar del documento y compararlo en la web y determinar si existe un grado de igualdad o no.
	Fracaso: No exista conectividad con internet, por lo tanto o se de la respectiva búsqueda en la web.
Excepciones	El sistema no recibe documento en caso de no existir conexión con el Internet.
Frecuencia Esperada	Dependerá del criterio del docente o del usuario final.
Comentarios	El proceso de análisis es uno de los puntos más críticos ya que aquí se integra las búsquedas en internet y comparación de imágenes de esto dependerá los resultados que se mostrara en el reporte final.

4.3.2 Diagrama de Secuencia

4.3.2.1 Diagrama de Secuencia – General

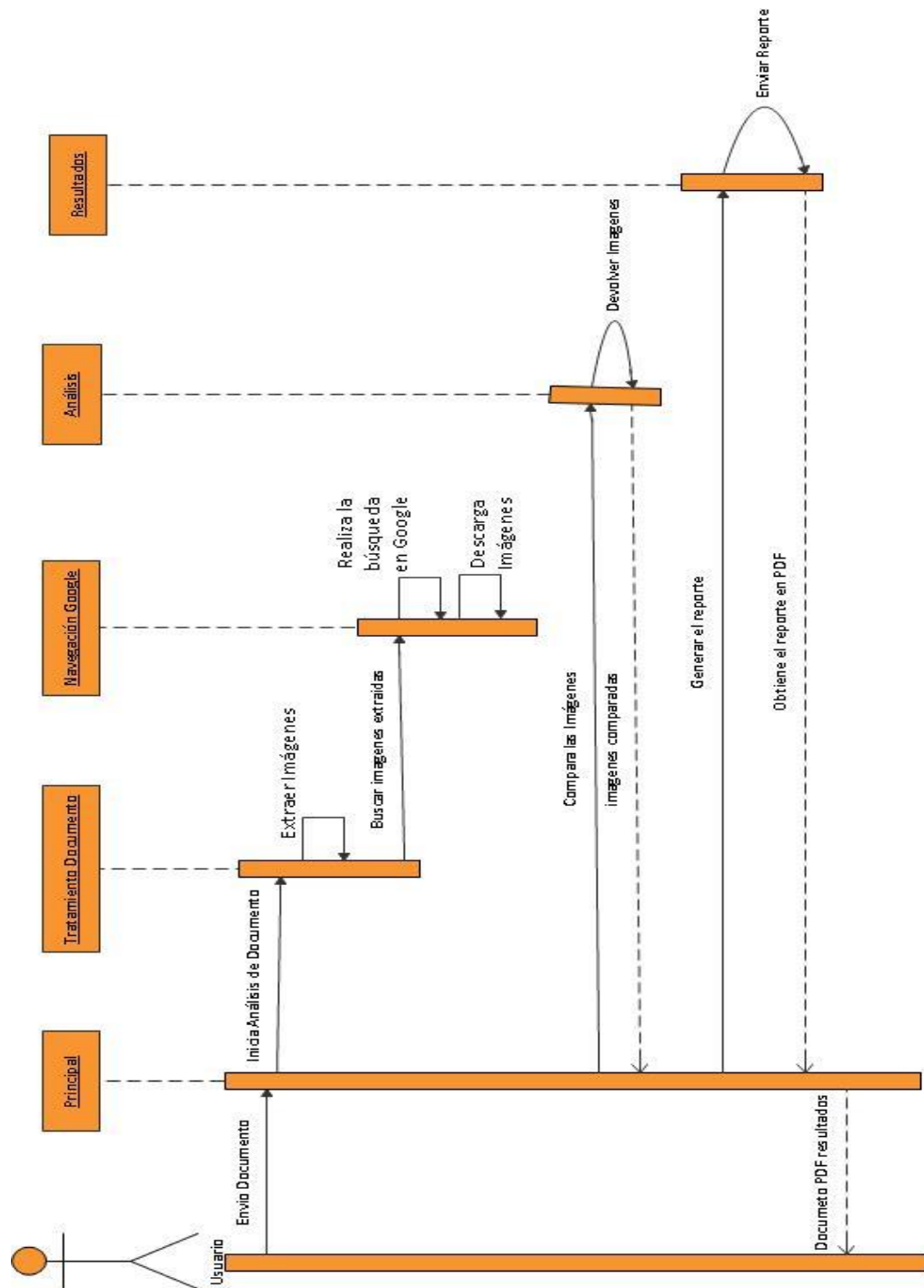


Ilustración 29. Diagrama de Secuencia **General**

4.3.2.2 Diagrama de Secuencia - Envió del Documento

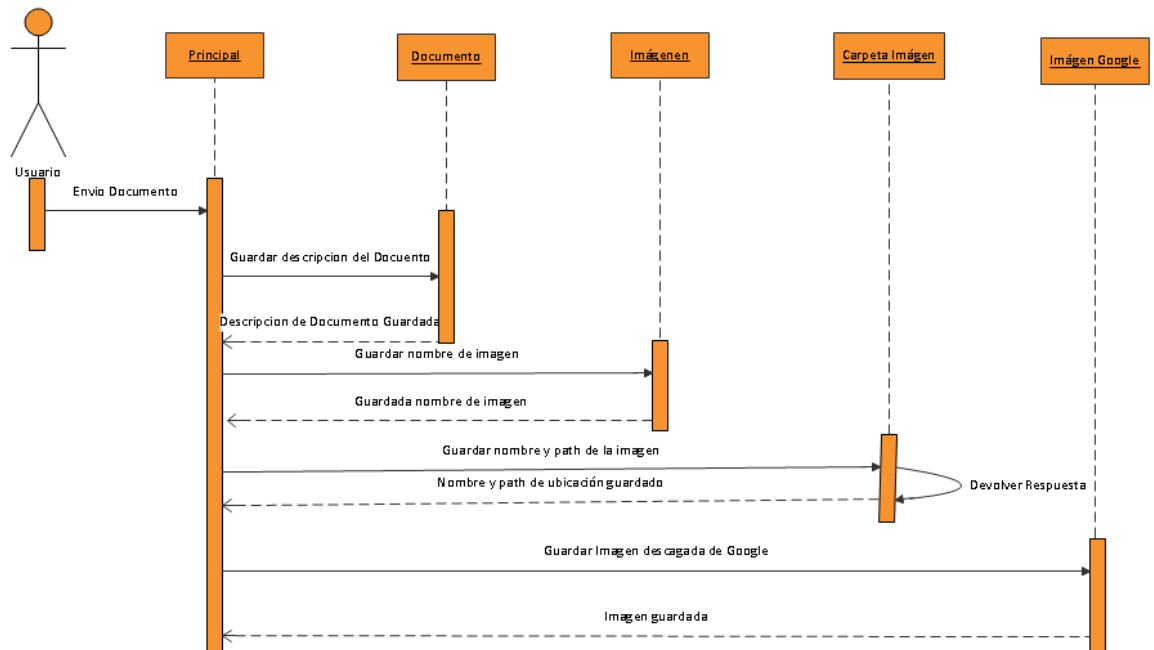


Ilustración 30 Diagrama de Secuencia Envío del Documento

4.3.2.3 Diagrama de Secuencia – Resultados

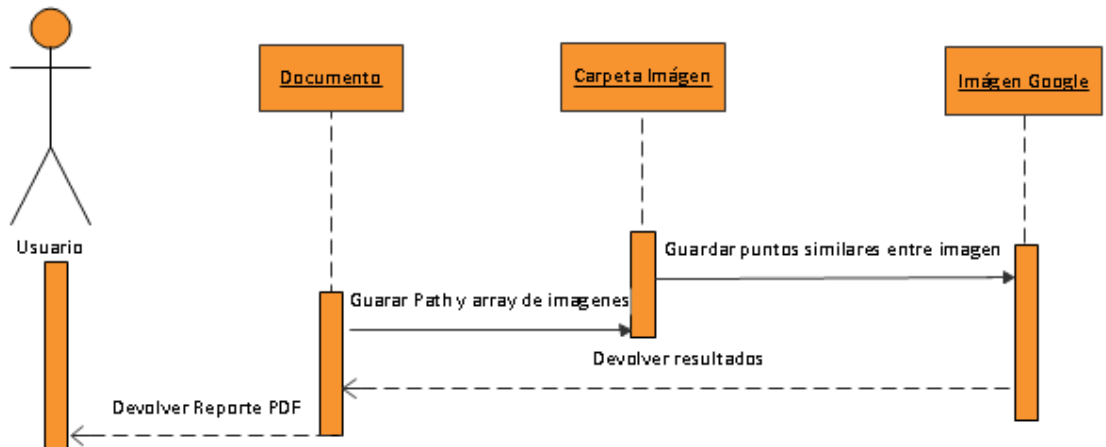


Ilustración 31 Diagrama de Secuencia Resultados

4.3.2.4 Diagrama de Secuencia - Análisis del Documento

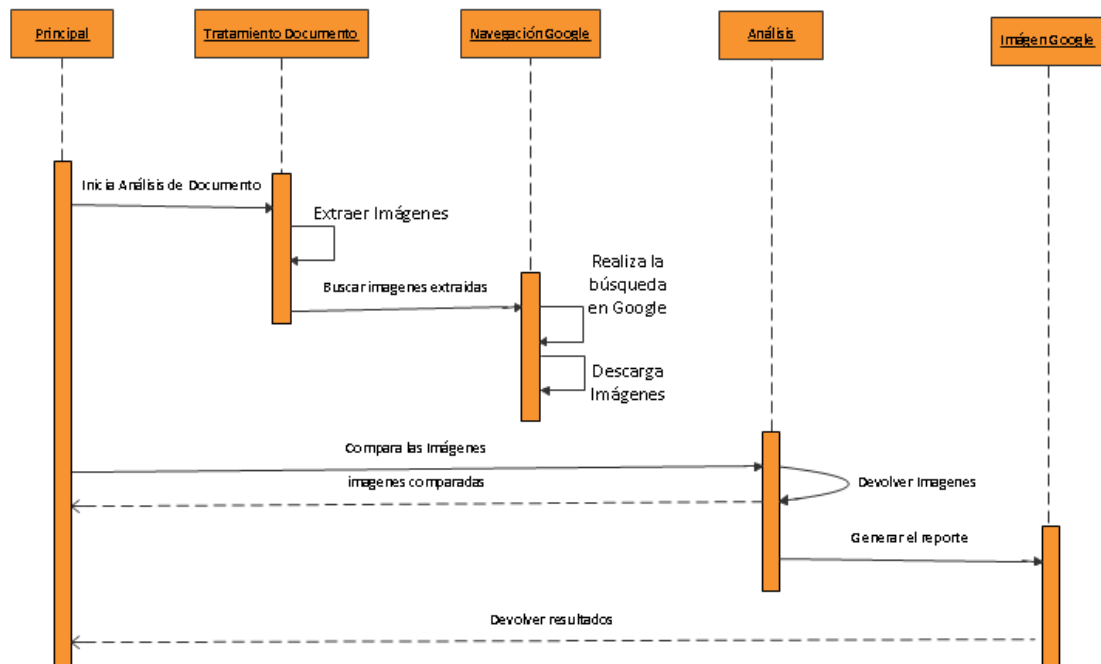


Ilustración 32 Diagrama de Secuencia de Análisis del Documento

4.3.3 Diagrama de Clase

4.3.3.1 Diagrama de clase interacción entre módulos.

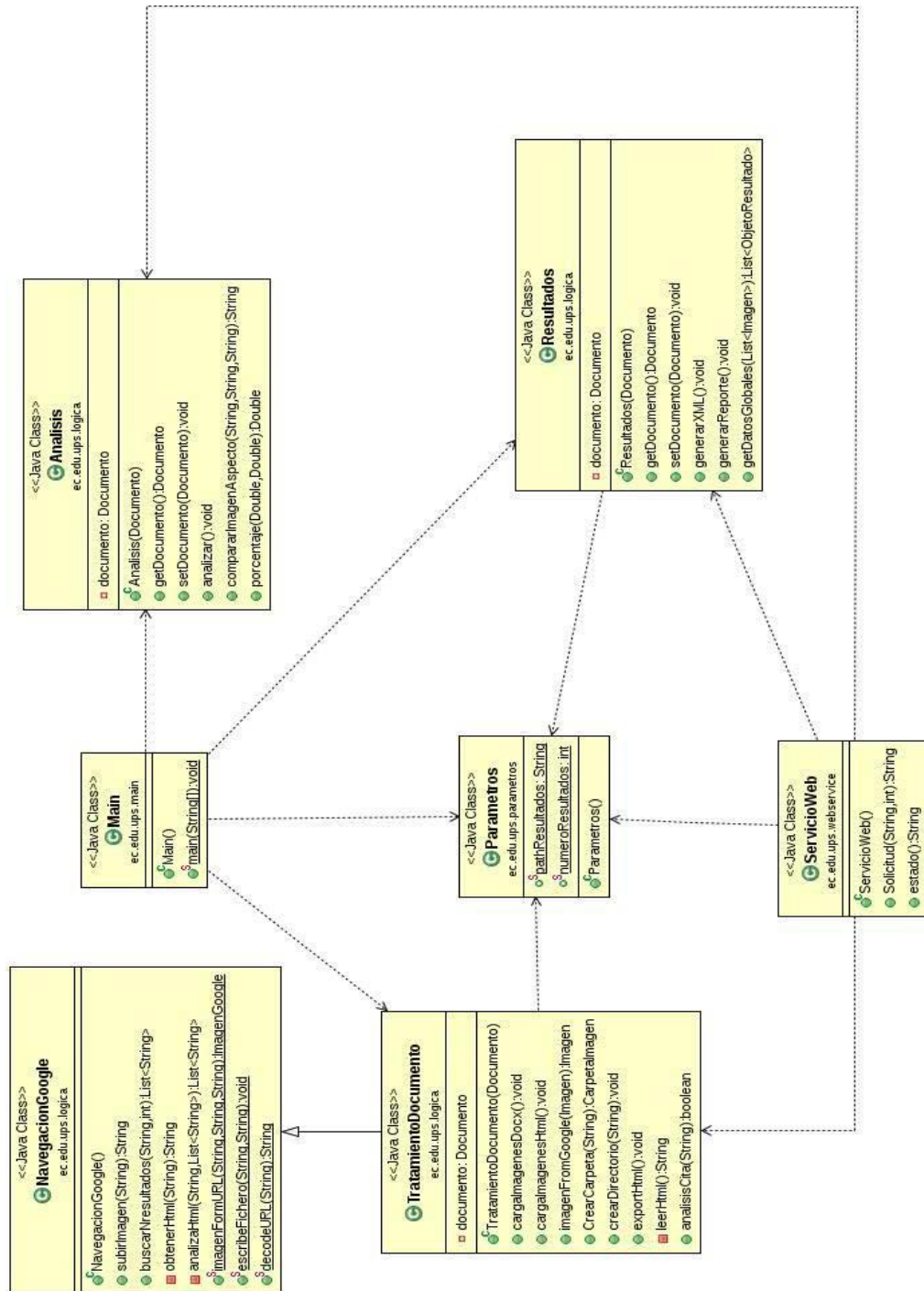


Ilustración 33 Diagrama de Clase de Sistema Anti Plagio de Imágenes

4.3.3.2 Diagrama de Clase del Módulo de Análisis de Resultados.



Ilustración 34 Tabla Resultados

4.3.3.3 Diagrama de Clase del Módulo de Comparación de Imágenes.

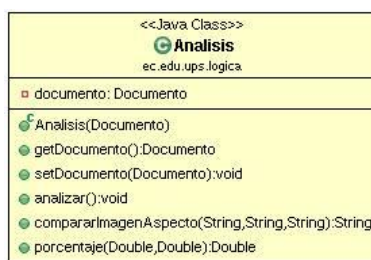


Ilustración 35 Tabla Análisis

4.3.3.4 Diagrama de Clase del Módulo de Extracción de Imágenes.



Ilustración 36 Tabla TratamientoDocumento

4.3.3.5 Diagrama de Clase del Módulo de Conexión a Internet.

<<Java Class>>	
 NavegacionGoogle ec.edu.upv.logica	
	NavegacionGoogle()
	subirImagen(String):String
	buscarResultados(String,int):List<String>
	obtenerHtml(String):String
	analizaHtml(String,List<String>):List<String>
	<u>imagenFormURL(String,String,String):ImagenGoogle</u>
	<u>escribeFichero(String,String):void</u>
	<u>decodeURL(String):String</u>

Ilustración 37 Tabla NavegacionGoogle

4.3.4 Diagrama de Paquetes

En el siguiente diagrama ilustramos la interacción que tienen los paquetes en sí dentro del sistema de detección de plagio, y la dependencia que tienen entre sí, la explicación de cada uno de ellos, se detalla más cuidadosamente en la siguiente sección 4.4.

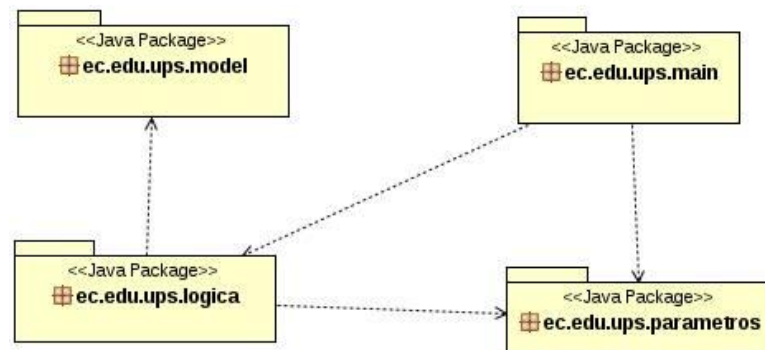


Ilustración 38 Diagrama Relación entre Paquetes del Sistema Anti Plagio de Imágenes

4.4 Diseño de la Arquitectura de la Solución

Comenzaremos a realizar un análisis detallado sobre nuestra arquitectura, el mismo que será reflejado en respectivos módulos que conformaran la misma, donde se muestra cómo y donde será desarrollado el sistema anti plagio, aquí se muestran los diferentes componentes que corren en cada uno. El diagrama puede tener distintas implementaciones de acuerdo al grado de detalle al que queremos llegar, es decir podemos agregar en un futuro más módulos con nuevas funcionalidades sin necesidad que esto afecte al funcionamiento general del sistema. A continuación presentamos un esquema general del sistema con los siguientes módulos: Módulo de Conexión a Internet, Módulo de Extracción de Imágenes, Módulo de Comparación de Imágenes y Módulo de Análisis de Resultados.

4.4.1 Paquete Lógica

Como su nombre lo indica, este paquete se encarga de toda la lógica del sistema, es el paquete más importante de nuestro sistema, ya que aquí se realiza las comparaciones de imágenes, la búsqueda en Google, las conversiones y los resultados, a continuación detallamos con más precisión las clases existentes dentro de éste paquete.

4.4.1.1 Clase Análisis

Se encarga de comparar las imágenes (A con B), con el algoritmo Surf, a través de Open CV, interactuando con el Lenguaje de Programación Python, el mismo que arrojan los puntos de coincidencia entre las dos imágenes y los puntos más

destacados de cada imagen, a la vez se genera otra imagen a partir de las dos imágenes (A y B), haciendo una comparación visual o gráfica de los puntos comparados.



Ilustración 39 Imagen A [50]



Ilustración 40 Imagen B [51]



Ilustración 41 Comparación de los Puntos de coincidencia entre Imagen A y B

```
public String compararImagenAspecto(String img1, String img2,
    String imgResultado) throws IOException {
    Process p = Runtime.getRuntime().exec(
        "python ./files/surf.py " + img1 + " " + img2 + " "
        + imgResultado);
    BufferedReader bri = new BufferedReader(new InputStreamReader(
        p.getInputStream()));
    String tmpLine = "";
    while ((tmpLine = bri.readLine()) != null) {
        return tmpLine;
    }
    return "";
}
```

Ilustración 42 Código Fuente Java, Clase Análisis

4.4.1.2 Clase Navegación Google

Esta clase se encarga básicamente de la conexión con el servidor de Google, está compuesta por varios métodos, los mismos que se encargan de:

- Subir la imagen al servidor de Google a través de código java.
- Obtener a partir del link de Google el código fuente del mismo.
- Permite a través de expresiones regulares obtener los Tags que contiene las imágenes y el link respectivo de cada imagen analizada.
- Permite escribir en un fichero el link en un directorio específico.
- Por último descargar la imagen encontrada a través del link.

```
private List<String> analizaHtml(String cadena,List<String> arrayIngreso){
    String expr1 = "<li\\s+class=\"g\"[^>]*>"
        + ".*?<div\\s+class=\"rc\"\\s+data-hveid=\".*?\"><h3 class=\"r\">"
        + "<a\\s+href=\""
        + "([^\"]+)" //link referencial group1
        + ".*?<div class=\"s\">.*?"
        + "<a\\s+href=\".*?"
        + "imgurl=http://"
        + "([^\"]+)&amp;imgrefurl" //link referencial group2
        + ".*?</li>";

    Pattern patt1 = Pattern.compile(expr1,
        Pattern.DOTALL | Pattern.UNIX_LINES);
    Matcher m1 = patt1.matcher(cadena);
    while (m1.find()) {
        String newString = m1.group(2);
        arrayIngreso.add(newString);
    }
    return arrayIngreso;
}
```

Ilustración 43 Código Fuente Java, Clase Navegación Google

4.4.1.3 Clase TratamientoDocumento

Esta clase es también una de las clases más importantes de nuestro sistema, se encarga de extraer las imágenes del respectivo documento a analizar, a la vez ésta clase hereda los procedimientos de la clase Navegación Google, es decir se encarga de unir y llamar a los métodos de ésta clase.

```
public void cargaImagenesDocx() throws IOException {

    List<Imagen> imagenes = new ArrayList<Imagen>();
    InputStream in = new FileInputStream(new File(documento.getPath()));
    XWPFDocument document = new XWPFDocument(in);
    List<XWPFPictureData> allpictures = document.getAllPictures();
    String contenido = document.getDocument().toString().trim();
    System.out.println("Total de Imagenes contenidas en el documento:"
        + allpictures.size());
    for (int i = 0; i < allpictures.size(); i++) {
        Imagen img = new Imagen();
        XWPFPictureData pic = allpictures.get(i);
        img.setNombre(pic.getFileName());
        String expr1 = "embed=\"" + pic.getPackageRelationship().getId()
            + "\".*?<w:t>(.*?)</w:t>";
        Pattern patt1 = Pattern.compile(expr1, Pattern.DOTALL
            | Pattern.UNIX_LINES);
        Matcher m1 = patt1.matcher(contenido);
        String cita = "";
        while (m1.find()) {
            cita = m1.group(1);
        }
    }
}
```

Ilustración 44 Código Fuente Java, Clase Tratamiento Documento

4.4.2 Paquete Principal

Paquete principal también conocido como paquete Main, ésta clase se encarga de configurar el Path de donde se encuentra el documento, así mismo el número de resultados que se van a buscar en la web, y el Path donde se van a guardar los resultados, también es el que interactúa con el Paquete Lógica, para realizar los respectivos procesos de éste paquete, es decir se encarga recibir y almacenar los parámetros que envía el usuario para configurar el programa.

4.4.3 Paquete Modelo

Aquí se encuentra todo lo que se refiere al modelo de objetos del sistema, a continuación describiremos brevemente las clases que la conforman.

4.4.3.1 Clase Documento

Es un objeto de tipo documentos, aquí se va a guardar la descripción del documento, el path del documento y un array de imágenes de tipo Imagen.

4.4.3.2 Clase Imagen

También es un objeto de tipo Imagen, el cual guarda el nombre de la imagen, el path y la carpeta a la que hace referencia.



Ilustración 45 Descripción de la Clase Imagen

4.4.3.3 Clase CarpetaImagen

Es un objeto en el cual se guarda el Path en donde está la carpeta y el nombre de dicha carpeta, al igual que se da un array de las imágenes que se encuentra en Google.

4.4.3.4 Clase ImagenGoogle

Es en la que se guarda como objeto las referencias de las imágenes descargadas de Google, a la vez en ésta clase se guarda también las referencias de comparación entre las imágenes, y el porcentaje respectivo de puntos coincidentes, entre las respectivas imágenes comparadas, es decir mantiene en la memoria del ordenador las respectivas propiedades como son los atributos o datos sobre sí mismo almacenados por el objeto, al igual que ciertas operaciones disponibles específicas como son los métodos.

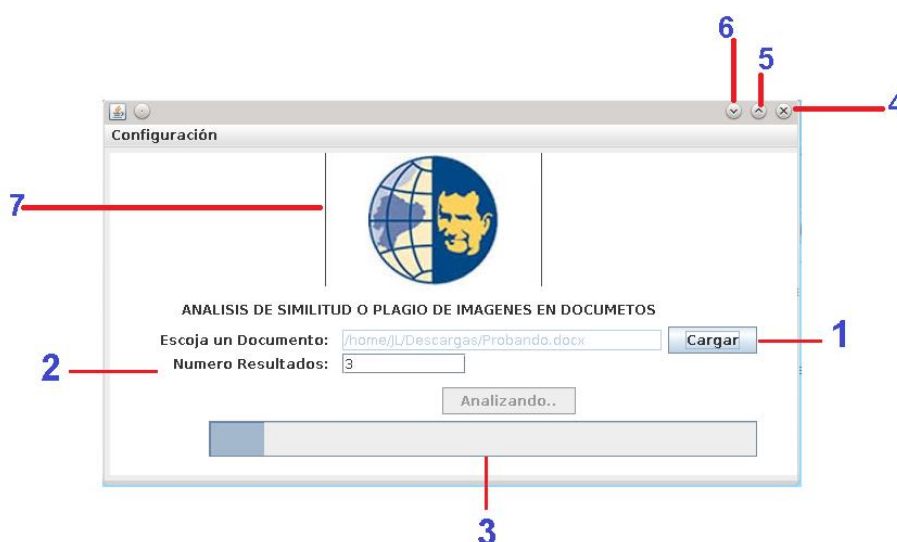
4.4.4 Paquete Parámetros

Este paquete contiene una sola clase, se encarga de guardar los parámetros que van a ser utilizado a nivel de toda la aplicación.

4.1 Diseño de la Interfaz del Usuario

Como ya sabemos todo programa necesita tener alguna forma de interactuar con el usuario, para recibir la petición y entregar los resultados al mismo, es por esta razón que hemos desarrollado o diseñado una interfaz gráfica sencilla, rápida, comprensible y fácil de manejar.

Describiremos brevemente el modo de operar por parte del usuario la interfaz gráfica. Como primer punto la siguiente ventana es la que va a visualizar el usuario para el análisis del documento.



1. **Escoja un Documento (Cargar):** Como su nombre lo indica, en ésta opción nos permite buscar el documento que deseamos analizar y cargarlo al sistema para su posterior análisis.

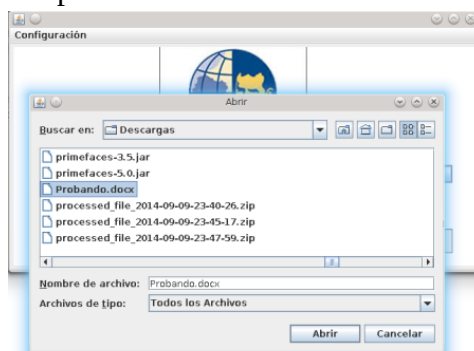


Ilustración 46 Interfaz Gráfica Carga de Documento

2. **Número Resultados:** Números de resultados que se analizara en el buscador de Google.

3. **Analizando:** Barra de estado, aquí se muestra el estado de análisis en el que se encuentra el documento, sirve que él usuario tenga una pequeña o aproximada idea de cuánto tiempo va a demorarse el análisis de todo el documento, ya que dependiendo del tamaño del documento dependerá el tiempo de respuesta final, y por último el sistema enviará el siguiente cuando de mensaje en el cual indica la ubicación del archivo que contiene el reporte final.



4. Cerrar.
5. Maximizar.
6. Minimizar
7. Logo Universidad Politécnica Salesiana.

A continuación mostraremos los resultados que botan después del análisis de cierto documento, en éste caso los resultados mostrados se refieren al documento Tesis1 de la sección de 5.3.1 en la ejecución de pruebas.

Como primer punto tenemos los resultados Globales o generales de todo el documento, aquí se muestra el nombre del documento (tesis1), el número de imágenes analizadas (9), un gráfico en barras que muestra que cierta imagen del documento tiene un parecido del cero al cien por ciento (0% - 100%) que otra imagen en la web y por último la tabla de criterios de valoración, para que el docente o quien corresponda disponga o decida la existencia de plagio.



Ilustración 47 Resultados Generales Tesis 1

Una vez revisado los resultados globales, también contamos con resultados específicos de cada imagen analizada, como primera imagen tenemos la imagen “**Junin**”, llamada de esa forma en el documento, a lado izquierdo tenemos la **imagen de la web**, que como su nombre lo indica es la imagen que el sistema encontró en la web y con la que va a realizar la respectiva comparación, que se encuentra en la imagen de la derecha, **comparación gráfica**.

También este reporte nos indica los Puntos Similares, Puntos analizados y el porcentaje de parecido entre las dos imágenes, (**Imagen de la web – comparación grafica**), representado también de forma gráfica, al igual que el link de donde se encuentra la imagen descargada de la web.

Para no caer en el empalago solamente resumiremos 3 de las 9 imágenes ya que en el procedimiento es el mismo para cada una de ellas.

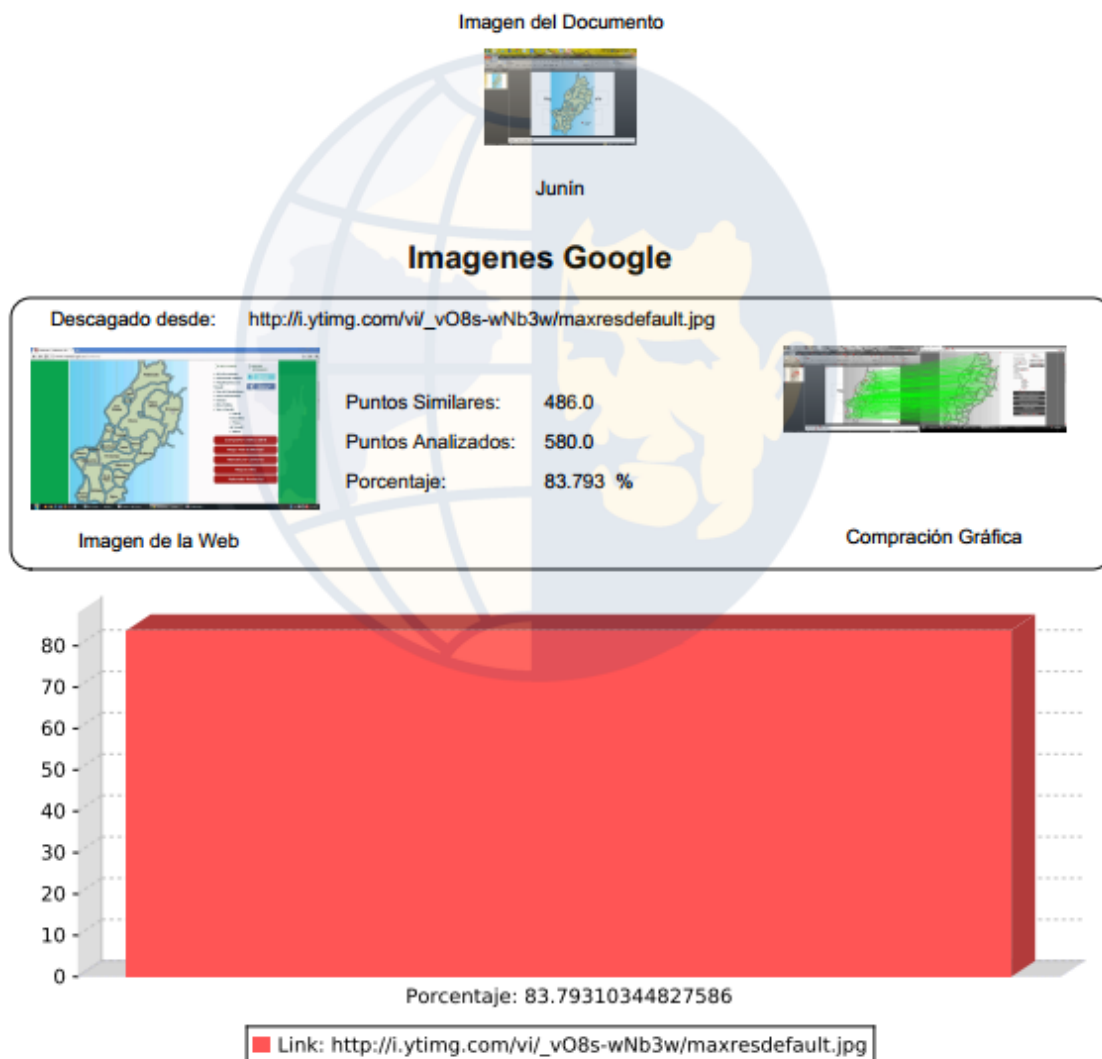
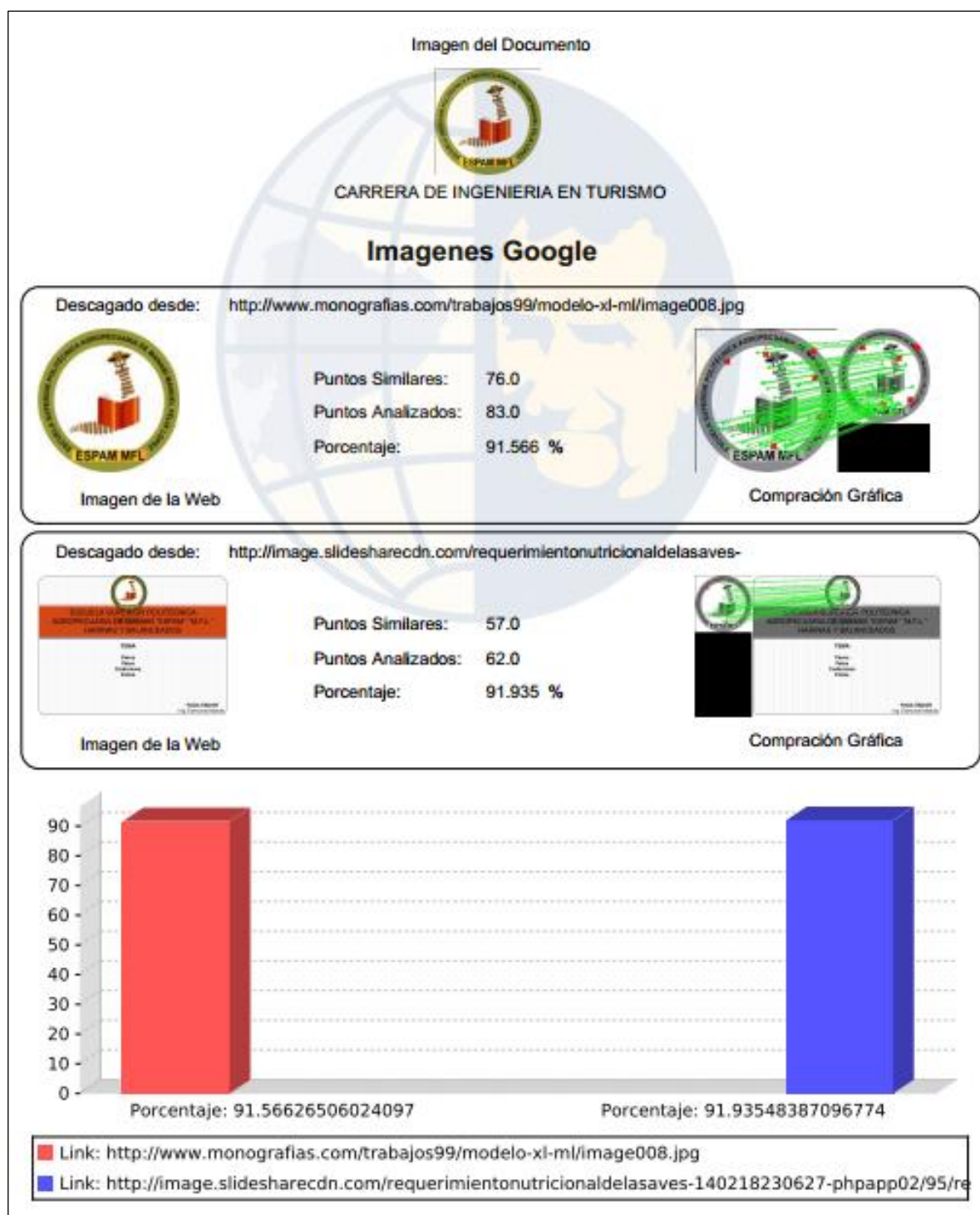


Ilustración 48 Anexo Detalle del Análisis de la Imagen Junín del Documento tesis1

En el análisis de la imagen “CARRERA DE INGENIERIA EN TURISMO”, se obtuvo dos imágenes parecidas en la web, de las mismas que se hizo el mismo procedimiento de comparación anterior con cada una de las imágenes encontradas en la web.



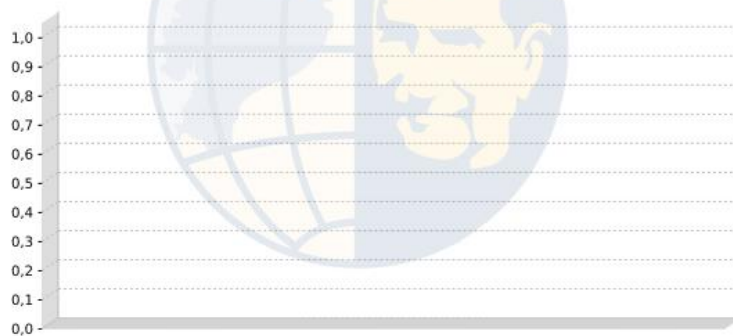
En este caso la “Figura 5.1 Vista aérea del taller comercial”, no tiene ninguna coincidencia en la web, por lo que no se muestra ningún valor de comparación.

Imagen del Documento



FIGURA 5.1. Vista aérea del área del taller comercial

Imágenes Google



CAPITULO V

5. DESARROLLO DEL PROTOTIPO Y ANÁLISIS DE RESULTADOS

5.1 Implementación del prototipo.

Para llegar a la etapa de implementación se tuvo que pasar por etapas previas de análisis y desarrollo, en las cuales se tuvieron varios escenarios tanto contemplados como no contemplados, a continuación se detallan las dos siguientes etapas que sería una continuación de la sección 4.1.

El fundamento de la etapa, que nosotros lo consideraríamos la tercera etapa, está centrado en el análisis y pruebas que se tienen que realizar sobre las mismas para establecer cuál es la mejor desde el punto de la eficiencia así como también factores relacionados a precisión y consumo de recursos.

En etapas previas y en relación al punto 4.1, se determinaron utilizar la librería OpenCV para tratamiento de las imágenes, y como lenguajes de programación Java y Python, siendo Java el primer lenguaje sobre el cual se elaboró un algoritmo para el análisis comparativo de las imágenes.

Como resultado de este primer algoritmo se obtuvo que los tiempos de respuesta son demasiado prolongados, así como también demandaba de un alto consumo de recursos de la máquina sobre la cual se ejecutaba la aplicación.

Posteriormente como alternativa a Java se procedió a elaborar un algoritmo en Python y una vez terminado se pasó a la etapa de análisis o pruebas con las imágenes.

Finalmente se concluyó que Python era la mejor opción por dos motivos determinantes, el primero por el corto tiempo de respuesta en comparación con Java, y el segundo que la ejecución del mismo no demandaba de un alto consumo de recursos.

Una vez que se ha llegado a este proceso y el algoritmo entra en ejecución procede a analizar el documento para la extracción de las fotografías, que no estén referenciadas bajo el estándar, mediante el uso de expresiones regulares extrae la información necesaria para posteriormente realizar una búsqueda de la imagen en el internet, todo esto haciendo uso de la librería JavaNet.

Una vez que se ha realizado la búsqueda de igual manera con la ayuda de expresiones regulares se extrae el path de la imagen, procediendo finalmente a descargar y analizar niveles de coincidencia mediante la técnica de análisis SURF.

Esta es la última etapa y comprende en si la implementación del prototipo lo cual permitirá llevar a cabo la etapa de plan de pruebas y un análisis de las mismas.

Cabe destacar que en esta etapa fue necesario tomar en cuenta varios elementos, necesarios para la implementación tales como distribución física y lógica de la aplicación, así como también los requerimientos exigidos para el correcto funcionamiento.

5.1.1 Requerimientos de la Aplicación

En cuanto a los requerimientos de la aplicación se pueden clasificar en dos:

➤ Requerimientos a nivel de Hardware

En lo referente a requerimientos de hardware ese no es necesariamente imposible o difícil de adquirir, tomando en cuenta el diagrama físico se requerirá de los siguientes equipos:

- Servidor con mínimo, de 4Gb de RAM.
- Procesador de 4 núcleos.
- 256 GB de Disco Duro.
- Tarjeta Ethernet.

Se debe aclarar que se debe disponer de una conexión a internet para que la aplicación se desempeñe correctamente.

➤ Requerimientos a nivel de Software

En cuanto a los requerimientos de software este no exige ninguna dependencia fuera de lo común, el único requerimiento y obligatorio es tener instalado el JRE perteneciente a Java para su correcta ejecución.

5.2 Preparación y Definición del Plan de pruebas.

Para definir el plan de pruebas se estableció etapas, mismas que al ejecutarlas de forma coordinada y sucesivamente permitieron realizar una evaluación integra y completa del sistema.

Cabe acotar que definir y ejecutar la etapa de pruebas permitieron realizar ajustes, así como también definir el alcance del sistema. Estas bases, estructuras y alcances permitirán dar una robusta continuidad a etapas posteriores que se deberán complementar con esta.

Las etapas que se definieron son las detalladas a continuación:

5.2.1 Etapa de Socialización

Para que una aplicación sea eficiente y cumpla un buen nivel de expectativas es necesario, socializar la aplicación directamente con los involucrados, es por ello que

en una primera instancia, se mostró la aplicación al tutor encargado del presente trabajo, siendo la primera persona en sugerir cambios y adecuaciones que permitan un correcto funcionamiento del mismo.

Entre los puntos que se trataron para la definición de las métricas opciones y parámetros con los que deberá contar, estuvieron presentes los actores involucrados al proyecto (Director de Tesis, desarrolladores).

Entre los puntos que se trataron en dicha reunión destacan los siguientes:

- Socialización de estructura para una mejor comprensión del código desarrollado.
- Definición de requerimientos para acoplamiento de módulos desarrollados posteriormente.
- Definición de la herramienta iReport para el diseño y elaboración de los reportes con los que contara el sistema.
- Especificar los reportes con los que deberá contar, y la información que deberán visualizar los mismos.
- Definición de reportes que se integrarán en etapas posteriores.
- Especificación de los estándares que se manejarán para referenciar una imagen dentro de un documento de tesis.

Cabe destacar que una de las sugerencias más estrictas fue el hecho de que los reportes no debe de estar ligados a ninguna base de datos lo cual permitiría total confianza de la aplicación hacia una motor de datos.

5.2.2 Etapa de Aprobación

En la etapa de aprobación se esclareció algunas ambigüedades, también se realizaron algunas pruebas de demostración que permitan manifestar y exponer las funcionalidades del sistema así como también establecer algunos parámetros y opciones necesarios para mejorar la funcionalidad del mismo. Cabe recalcar que estas opciones están relacionadas con la optimización del sistema, entre estos parámetros están:

- Parametrización del número de imágenes a buscar en la internet.
- Configuración del path del documento.
- Configuración del path para guardar resultados del análisis.
- Definición de estándar para búsqueda de imágenes.

Esta última acotación establece que el algoritmo solo deberá realizar la búsqueda de imágenes dentro del documento, que no estén citadas bajo los estándares establecidos, que en nuestro caso el sistema está basado en los estilos más sobresalientes y utilizados a la hora de referenciar y citar, como es APA e IEEE, que lo mencionamos en la sección 2.4.1.

5.3 Ejecución de Pruebas y Recopilación de Datos

5.3.1 Definición del Corpus de Pruebas

Como es de conocimiento la importancia de realizar pruebas con diferentes corpus, para que se ajuste de mejor manera a un ambiente real de procedimientos, razón por la cual hemos desarrollado un pequeño corpus que acoge todas las características que necesitamos para considerando cualquier tipo de eventualidad que se pueda dar, al ejecutar las pruebas de funcionalidad del sistema desarrollado, conteniendo de esta manera, imágenes con referencia e imágenes sin referencia. Nuestro corpus está conformado por 13 documentos detallados en la sección siguiente.

5.3.2 Ejecución de Pruebas

Con el objetivo de medir la eficiencia y el nivel o porcentaje de aproximación de resultados se elaborara un plan de pruebas que permitirá evaluar la aplicación.

El plan de pruebas contempla analizar variables que permitan obtener métricas y en base a esos resultados hasta cierto punto medible poder concluir algo, también es importante aclarar que habrá parámetros que se tendrán que variar o modificar con el afán de experimentar resultados.

Cabe destacar que para este proceso será necesario recopilar varios documentos y mandarlos a analizar con el sistema, mismos que de antemano tendrá imágenes referenciadas bajo el estándar establecido así como imágenes sin referenciar.

Entre las variables que se consideran de interés y pueden ser medidas están las siguientes:

- Tiempo de análisis (retorno de resultados en segundos)
- Porcentaje de precisión de resultados.
- Tiempos de búsqueda de imágenes
- Tiempos de descarga de imágenes desde la web

➤ Agente Externos

Dentro de estos parámetros también hay que considerar algunas variables que no dependen estrictamente de la aplicación pero que pueden afectar directamente la funcionalidad o tiempos de desempeño de la misma.

Estas son:

- Velocidad de conexión de internet.
- Interrupción de conexión.

También se evaluarán los reportes establecidos, ya que estos deberán ser lo suficientemente comprensivos, legibles e intuitivos.

5.3.3 Recopilación de Datos

Las etapas previas de definición de entregables y socialización con los involucrados permitió obtener datos relevantes, y que permitieron realizar ciertos ajustes a la aplicación.

También se obtuvieron datos informativos en base a las variables que se propuso medir en la etapa de ejecución de pruebas.

Es importante estipular que para realizar esta recopilación de datos informativos se manejó un conjunto de parámetros que estaban sujetos a variaciones y son los siguientes:

✓ **Imágenes en documento que serán analizadas.**

Este indicador permitió manipular los números de imagen presentes o existentes en el documento a analizar.

✓ **Número de páginas.**

Este indicador permite ver la magnitud de páginas que existe en el documento a analizar.

✓ **Tiempo promedio en extraer una imagen.**

Esta variable me indica el tiempo promedio en segundos que se demora el sistema en extraer una imagen.

✓ **Tiempo promedio en buscar una imagen.**

Esta variable me indica el tiempo promedio en segundos que se demora el buscador Google en buscar una imagen.

✓ **Tiempo total en la búsqueda de una imagen.**

Esta variable me indica el tiempo total en segundos que se demora el sistema en extraer una imagen.

✓ **Total de imágenes descargadas.**

Este indicador me muestra el total de imágenes descargadas de la web.

✓ **Tiempo promedio y total en descargar una imagen.**

Me indica el tiempo promedio en segundos que se demora en descargar la imagen de la web, cabe recalcar que éste tiempo dependerá del ancho de banda que disponga el servidor al momento de realizar el análisis del documento.

✓ **Tiempo promedio y total en comparar imagen.**

Esta variable me indica el tiempo total en segundos que se demora el sistema en comparar la imagen extraída del documento con la imagen descargada de Google.

✓ **Tiempo en generar un reporte.**

Este indicador me muestra el tiempo en segundos que se demora el sistema en generar un reporte es decir el resultado final en formato PDF.

✓ **Tiempo total.**

Y por último tenemos el indicador del tiempo total que se demora todo el análisis del documento.

Con cada uno de las opciones o parámetros (variables) se obtuvo las siguientes variaciones en los resultados, mismos que se exponen en la siguiente tabla según su rendimiento.

	Tesis 1	Tesis 2	Tesis 3	Tesis 4
Titulo	ESCUELA SUPERIOR POLITÉCNICA AGROPECUARIA DE MANABÍ MANUEL FÉLIX LÓPEZ	UNIVERSIDAD TÉCNICA DEL NORTE FACULTAD DE EDUCACIÓN, CIENCIA Y TECNOLOGÍA	ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL INSTITUTO DE CIENCIAS MATEMÁTICAS	UNIVERSIDAD ESTATAL DE BOLÍVAR FACULTAD DE CIENCIAS AGROPECUARIAS RECURSOS NATURALES Y DEL AMBIENTE
Número de Imágenes	9	97	29	56
Número de Paginas	76	187	86	163
Tiempo Promedio en Extraer Imagen (Seg)	1	2	3	2
Tiempo Promedio en Buscar Imagen (Seg)	2.44	2.61	1,63	1,42
Tiempo Total búsqueda de Imagen (Seg)	22	254	47	80

Total Imágenes Descargadas	14	365	58	5
Tiempo Promedio Descargar Imagen (Seg)	3.64	1.6	1,9	37
Tiempo Total Descargar Imágenes (Seg)	51	586	115	184
Tiempo Promedio Comparar Imagen (Seg)	1.6	0.9	1,6	3.6
Tiempo Total Comparar Imagen (Seg)	14	303	97	18
Tiempo Generar Reporte (Seg)	4	18	7	13
Tiempo Total (Seg)	91	1163	269	297

Tabla 11 Recopilación de Datos del Sistema Anti Plagio de Imágenes

La ejecución de las pruebas se realizó bajo las siguientes circunstancias, parámetros o escenarios

- Computadora HP Pavilion dev5.
- Procesador AMD.
- Sistema Operativo de 64bits.
- Memoria RAM 4,00GB
- Velocidad de internet 225kbs de bajada y 94kbs de subida.

Un tiempo promedio podemos decir que la comparación entre imágenes está en un promedio de 1 segundo, la descarga de cada imagen tomando en cuenta las velocidades que se disponen al momento de las pruebas ya antes mencionadas es de un promedio de 2s. El análisis completo de una imagen dependerá del número de imágenes que se descargue de la web.

A continuación realizamos el análisis de los demás documentos para recopilar información y medir el rendimiento del sistema.

SISTEMA SOLAR		
	ANALISIS MANUAL	ANALISIS POR SISTEMA
Número de Imágenes	7	7
Número de Imágenes no plagiadas	2	2
Páginas	7	
Imágenes Descargadas		20
Porcentaje Global		91.4%
IMÁGENES ANALIZADAS		
	PLAGIO	
1	100%	99%
2	100%	96%
3	100%	100%
4	100%	100%
5	100%	62%
Efectividad	95.7	

Tabla 12 Análisis de Efectividad del Documento Sistema Solar

BIOLOGIA ANIMAL		
	ANALISIS MANUAL	ANALISIS POR SISTEMA
Número de Imágenes	7	7
Número de Imágenes no plagiadas	0	0
Páginas	13	
Imágenes Descargadas		27
Porcentaje Global		83.2%
IMÁGENES ANALIZADAS		
	PLAGIO	
1	100%	100%
2	100%	100%
3	100%	98%
4	100%	88%
5	0%	0%
6	100%	98%
7	100%	99%
Efectividad	98.78%	

Tabla 13 Análisis de Efectividad del Documento Biología Animal

ENLACES DE COMUNICACIONES		
	ANALISIS MANUAL	ANALISIS POR SISTEMA
Número de Imágenes	12	12
Número de Imágenes no plagiadas	12	12
Páginas	5	
Imágenes Descargadas		5
Porcentaje Global		28.66%
IMÁGENES ANALIZADAS		
	PLAGIO	
1	0%	0%
2	0%	0%
3	0%	0%
4	0%	0%
5	0%	44%
6	0%	0%
7	0%	0%
8	0%	0%
9	0%	0%
10	0%	100%
11	0%	100%
12	0%	100%
Efectividad	70%	

Tabla 14 Análisis de Efectividad del Documento Enlaces de Comunicaciones

RED INTEGRAL DE SERVICIOS DIGITALES (RISD)		
	ANALISIS MANUAL	ANALISIS POR SISTEMA
Número de Imágenes	4	4
Número de Imágenes no plagiadas	0	4
Páginas	3	
Imágenes Descargadas		0
Porcentaje Global		0%
IMÁGENES ANALIZADAS		
	PLAGIO	
1	100%	0%
2	100%	0%
3	100%	0%
4	100%	0%
Efectividad	0%	

Tabla 15 Análisis de Efectividad del Documento Red Integral de Servicios Digitales

Sample IEEE Paper for A4 Page Size		
	ANALISIS MANUAL	ANALISIS POR SISTEMA
Número de Imágenes	1	1
Número de Imágenes no plagiadas	0	0
Páginas	3	
Imágenes Descargadas		0
Porcentaje Global		100%
IMÁGENES ANALIZADAS		
PLAGIO		
1	0%	0%
Efectividad	100%	

Tabla 16 Análisis de Efectividad del Documento Sample IEEE Paper for A4 Page Size

Implementación de un portal web para la automatización del proceso de consultorías de mentores GOLD de la Región Latinoamericana del IEEE (R9), utilizando arquitectura Java 2 Enterprise Edition - J2EE y tecnología Ajax		
	ANALISIS MANUAL	ANALISIS POR SISTEMA
Número de Imágenes	7	7
Número de Imágenes no plagiadas	0	4
Páginas	5	
Imágenes Descargadas		20
Porcentaje Global		3.24%
IMÁGENES ANALIZADAS		
PLAGIO		
1	0%	0%
2	0%	22.72%
3	0%	0%
4	0%	0%
5	0%	0%
6	0%	0%
7	0%	0%
Efectividad	98.37%	

Tabla 17 Análisis de Efectividad del Documento Implementación de un portal web para la automatización del proceso de consultorías de mentores GOLD de la Región Latinoamericana del IEEE (R9), utilizando arquitectura Java 2 Enterprise Edition - J2EE y tecnología Ajax

Formato y Guía para presentación de artículos académicos y científicos		
	ANALISIS MANUAL	ANALISIS POR SISTEMA
Número de Imágenes	3	3
Número de Imágenes no plagiadas	1	1
Páginas	4	
Imágenes Descargadas		5
Porcentaje Global		97.67%
IMÁGENES ANALIZADAS		
PLAGIO		
1	0%	0%
2	100%	97.67%
Efectividad	99.41%	

Tabla 18 Análisis de Efectividad del Documento Formato y Guía para presentación de artículos académicos y científicos

Magnetron		
	ANALISIS MANUAL	ANALISIS POR SISTEMA
Número de Imágenes	4	4
Número de Imágenes no plagiadas	0	0
Páginas	3	
Imágenes Descargadas		19
Porcentaje Global		98.39
IMÁGENES ANALIZADAS		
PLAGIO		
1	100%	99.4%
2	100%	94.5%
3	100%	100%
4	100%	99.67%
Efectividad	99.19%	

Tabla 19 Análisis de Efectividad del Documento Magnetron

ESTADO DEL ARTE DEL GRID		
	ANALISIS MANUAL	ANALISIS POR SISTEMA
Número de Imágenes	4	4
Número de Imágenes no plagiadas	2	2
Páginas	25	

Imágenes Descargadas		20
Porcentaje Global		45.2%
IMÁGENES ANALIZADAS		
PLAGIO		
1	0%	0%
2	0%	31.86%
3	100%	83.64%
4	100%	65.30%
Efectividad	89.63	

Tabla 20 Análisis de Efectividad del Documento Estado del Arte del Grid

A continuación presentamos una tabla que resumen la efectividad del sistema de detección de plagio.

EFFECTIVIDAD	
DOCUMENTO	PORCENTAJE
SISTEMA SOLAR	95,7
BIOLOGIA ANIMAL	98,78
ENLACES DE COMUNICACIONES	70
RED INTEGRAL DE SERVICIOS DIGITALES (RISD)	0
SAMPLE IEEE PAPER FOR A4 PAGE SIZE	100
IMPLEMENTACIÓN DE UN PORTAL WEB PARA LA AUTOMATIZACIÓN DEL PROCESO DE CONSULTORÍAS DE MENTORES GOLD DE LA REGIÓN LATINOAMERICANA DEL IEEE (R9), UTILIZANDO ARQUITECTURA JAVA 2 ENTERPRISE EDITION - J2EE Y TECNOLOGÍA AJAX	98,37
FORMATO Y GUÍA PARA PRESENTACIÓN DE ARTÍCULOS ACADÉMICOS Y CIENTÍFICOS	99,41
MAGNETRÓN	99,19
ESTADO DEL ARTE DEL GRID	89,63
TOTAL	83.45

Tabla 21`Porcentaje de Efectividad del Sistema de Detección de Plagio

En el caso del porcentaje 0% es porque las imágenes son recortadas en el mismo documento, esto hace que no se pueda realizar una búsqueda efectiva.

De acuerdo a la ejecución y recopilación de datos expuestos anteriormente podemos llegar a la conclusión de que el tiempo de respuesta dependerá de gran parte de la extensión que tenga el documento, así como también el número de imágenes que contenga el mismo, de igual manera todo lo que se refiere a la búsqueda y descarga de imágenes el tiempo de respuesta dependerá del ancho de banda que disponga el servidor al momento de realizar el análisis. Al igual es el número de imágenes que se manda como parámetro, por parte del usuario, para que Google realice la búsqueda de las imágenes, ya que existe una gran diferencia en mandar a buscar solamente 3 o 5 resultados que mandar a buscar unos 100 resultados.

Otra conclusión a la que llegamos, es la escasa influencia del tamaño de imagen al momento de la comparación, sin embargo tiene cierta influencia a la hora de descargar la imagen de Internet, y por ultimo a lo que se refiere a términos de eficiencia podemos decir que el sistema de detección de plagio tiene un promedio de 95% de eficiencia y confiabilidad, el otro 5% de vulnerabilidad se debe a la existencia de diferentes métodos de disimular el plagio y que el sistema no lo puede detectar.

5.4 Despliegue de la Aplicación.

En general los procesos de despliegue de las aplicaciones, sistemas o software, tienden a ser costosos ya que conlleva varios aspectos como es el personal que se necesita para la ejecución del mismo, igual que el cuantioso cantidad de actividades que se lleva a cabo, el cual incluye el **dar soporte del software**, ya que este es el paso que se da con la participación conjunta y directa de los usuarios finales del software creado.

Para desplegar la aplicación es necesario planificar y definir como se realizará transitoriamente este proceso, y también debemos predecir cualquier situación que se presente a lo largo del despliegue, y definir estrategias para mitigar con los mismos.

A continuación presentamos el diagrama correspondiente al despliegue de nuestra aplicación. Y algunos de los aspectos más importantes que debemos tener en cuenta para el proceso de despliegue.

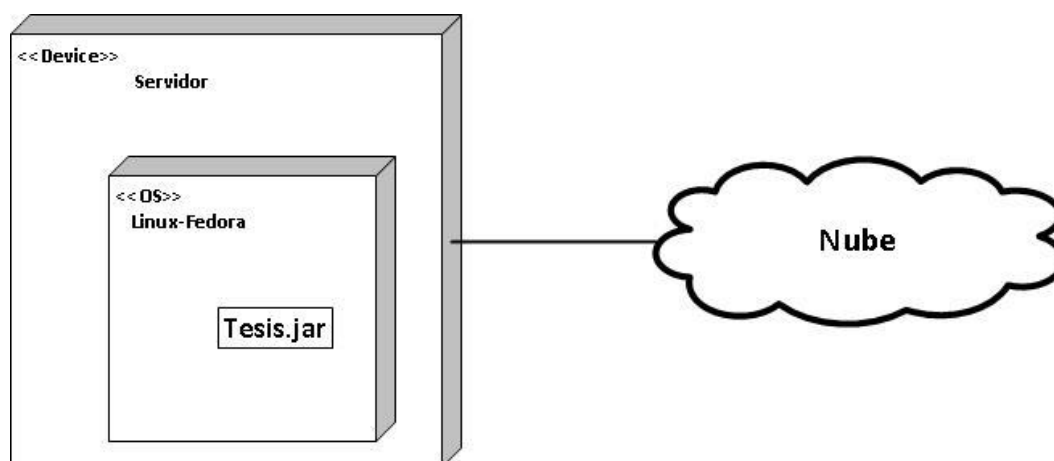


Tabla 22 Diagrama de Despliegue del Software Anti Plagio de Imágenes

✓ Personal disponible.

En lo que se refiere al personal involucrado en el despliegue de la aplicación, podemos mencionar que contamos con el adecuado y capacitado grupo de personas para esta función, entre ellos se encuentra nuestro director de tesis, profesores especializados en el área de la informática y el grupo que desarrollo el sistema anti Plagio.

✓ **Equipo a desplegar.**

Como ya lo habíamos mencionado en la sección 5.1, en lo que se refiere al equipo físico o hardware no se necesita mayor cosa, sino que contar con mínimo de los siguientes requerimientos

- Servidor con mínimo, de 4Gb de RAM.
- Procesador de 4 núcleos.
- 256 GB de Disco Duro.
- Tarjeta Ethernet

✓ **Características de los clientes.**

Lo que pretende conseguir con este punto, es identificar las características específicas que cuenta el cliente, ya sea el nivel profesional de los clientes de la aplicación, o el nivel de conocimiento en el área de la informática, ya que el programa anti plagio de imágenes desarrollado va enfocado a un usuario común y corriente, y debe ser fácil de manejar, obviamente otorgándoles una breve introducción de su funcionamiento.

5.5 Pruebas de Funcionalidad.

En la sección 5.3.1 se realizó la ejecución de pruebas para comprobar la funcionalidad y eficiencia del sistema anti plagio de imágenes, para lo cual se realizó varios experimentos basados varios parámetros.

Ésta sección básicamente respalda lo ocurrido en la sección 5.3.1 y 5.3.2, ya que las pruebas de funcionalidad están estrechamente relacionadas también con la etapa de recopilación de datos lo cual permitió probar las siguientes funciones y rectificar su correcto funcionamiento:

- Identificación de imágenes referenciadas y no referenciadas con total éxito.
- Extracción de información necesaria para realizar búsqueda en internet con un resultado exitoso.
- Búsqueda de información (imágenes) en el internet con un resultado exitoso.
- Descarga de resultados con un resultado exitoso.
- Análisis en la comparación de imágenes utilizando el algoritmo SURF con un resultado exitoso.
- Visualización de reportes en documento PDF, representado con un cuadro de estadística con un resultado exitoso.

Dentro de esta fase también se puede concluir que muchas de las pruebas de funcionalidad llegan o culmina con un resultado exitoso siempre y cuando la conexión a internet no se vea interrumpida, caso contrario el sistema esperará aproximadamente por 20 segundos para enviar un mensaje de excepción, es decir un time out.

También hay que tomar en cuenta que la velocidad de conexión a internet es un factor muy influyente en los tiempos de respuesta.

CONCLUSIONES Y RECOMENDACIONES

A lo largo de la existencia humana se ha ido desarrollando y evolucionando diferentes técnicas, herramientas, etc., que ayuden a realizar y facilitar diferentes tareas de diversas índoles y que en los últimos tiempos y con la creación de los ordenadores, el hombre ha llevado a cabo sus proyectos a una nueva escala. En la actualidad, se dispone de herramientas que permiten progresar en la creación de nuevas aplicaciones.

A partir de la investigación y del análisis efectuado a lo largo del desarrollo del sistema anti plagio de imágenes, se puede destacar la importancia de desarrollar software anti plagio aplicable en cualquier entorno, es así que ya por concluir ésta respectiva tesis, nos hemos dado cuenta de la importancia de colaborar con el desarrollo y correcto funcionamiento un sistema anti plagio, sobretodo orientado a imágenes, y que éste proceso de desarrollo sea completamente transparente y fácil de manejar para el usuario final, permitiendo lograr únicamente un reporte final con resultados entendibles para dar una perspectiva o sugerencia de cuando exista o no plagio.

Pues dicho de ésta manera y de acuerdo al trabajo realizado en la presente tesis, podemos mencionar con mayor especificación las siguientes conclusiones.

Se indagó todo lo que se refiere a temas de plagio, ya sea de texto como de imágenes, y llegando a la conclusión que existe un gran impacto o situándose mayormente en el entorno académico, por lo que se ve la necesidad de la creación de sistemas que ayuden a la detección de plagio.

Se investigó la existencia de metodologías, recursos y herramientas disponibles, llegando a la conclusión que existen diferentes metodologías ya establecidas y utilizadas para afrontar el problema de plagio, muchas de éstas metodologías y herramientas fueron utilizadas como base para el desarrollo de varios sistemas anti plagio, sin excepción nosotros que también nos ayudamos de algunas de estas herramientas.

Un punto muy importante son las técnicas utilizadas en la detección de plagio, en lo que se refiere a nuestro caso, nos interesó en gran magnitud todo lo que se refiere a citar y referenciar correctamente una bibliografía ya que dependiendo de éste parámetro, nuestro sistema analizará o no la imagen sospechosa.

También queremos resaltar la importancia que tiene realizar el análisis de los requerimientos, ya que basados en éstos nos dimos una idea más acertada de lo que pretendíamos que nuestro sistema haga, además de mostrarnos las fortalezas y debilidades que teníamos, también podíamos ver otros puntos de vista de acuerdo a otras soluciones, en el caso de que una solución no fuera la correcta.

En cuanto a lo que se refiere a la selección herramientas para el desarrollo del sistema, tomamos en consideración interactuar con dos de los lenguajes esta está acaparando mucha demanda en los últimos tiempos, asumiendo como lenguaje de programación principal Java, y posteriormente y para complementar Python, éste último lo utilizamos para agilizar el proceso de análisis de imágenes, ya que al ser un lenguaje menos sistemático es más rápido.

Para la comparación de imágenes, decidimos usar el algoritmo Surf puesto que éste algoritmo nos muestra varios puntos de interés y la ubicación de los mismos, en un rápido tiempo de respuesta.

La mayor limitación de nuestro sistema anti plagio, es la dependencia que se tiene con el buscador de Google, cabe mencionar que la elección de Google como buscador, fue por su gran acogida que tiene al ofrecer una herramienta para realizar búsquedas por imágenes.

El sistema es bastante eficiente y confiable a pesar de que en una de las pruebas fue poco efectivo, ya que en este documento existieron imágenes que fueron recortadas usando herramientas propias de Word y esto impidió que se esquivase de alguna forma su detección.

Para la interoperabilidad del sistema se optó por implementar web service, esto nos brinda la oportunidad, de acceder al sistema desde otros sistemas independientemente del lenguaje o Sistema operativo.

Una recomendación que daríamos para obtener un óptimo tiempo de respuesta en descarga y búsqueda de imágenes es de tener un ancho de banda mínimo 225kbps de conexión a internet.

Para futuros desarrolladores que deseen implementar mejoras en el sistema actual se recomienda realizar un repositorio para búsquedas de imágenes, para no depender de un determinado motor de búsqueda.

ANEXOS

ANEXO 1 REPORTE DE DOCUMENTO TESIS 1



Ilustración 49 Anexo Reporte General Tesis 1, pág. 1

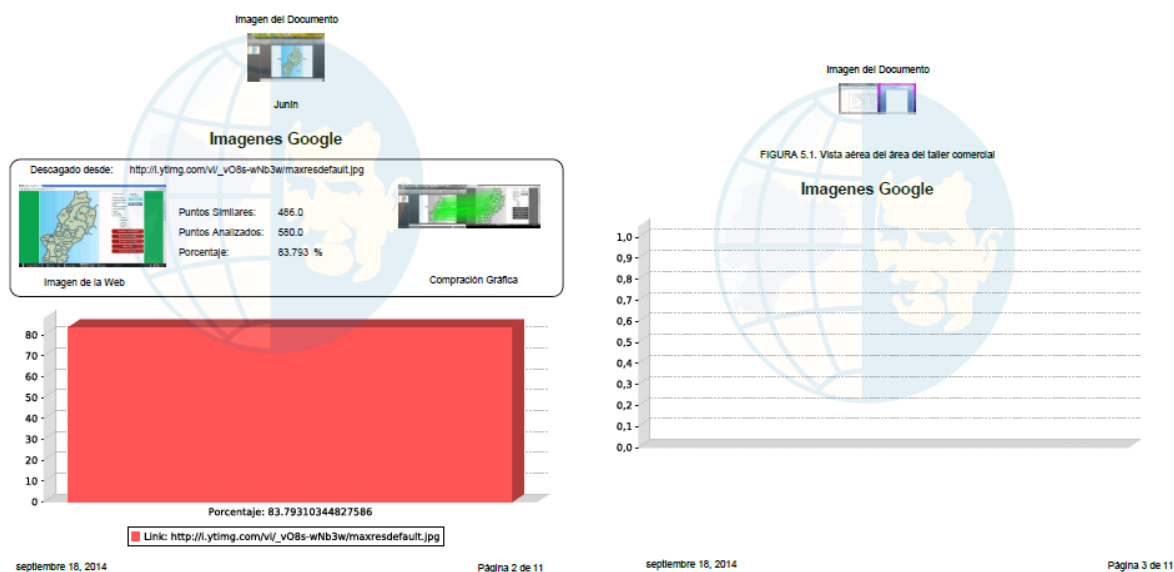


Ilustración 50 Anexo Reporte específico tesis Pag.2 y 3

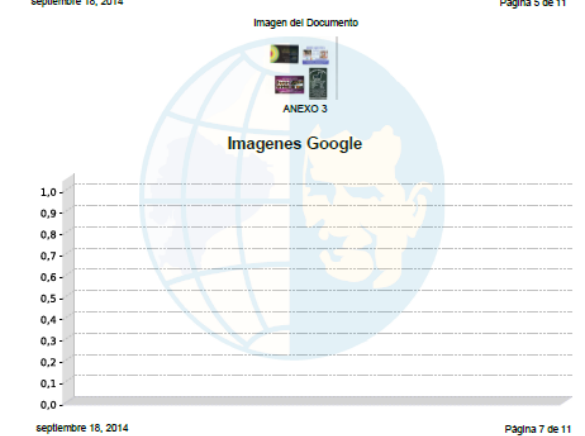
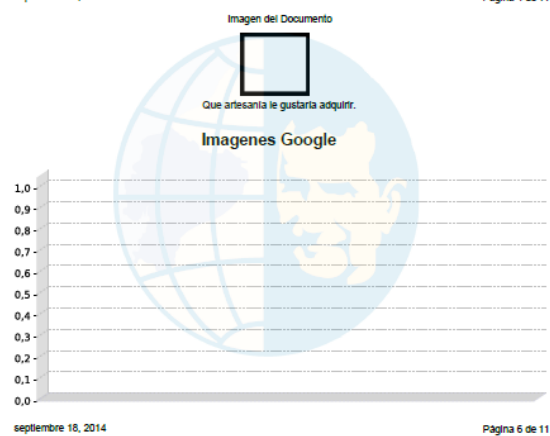
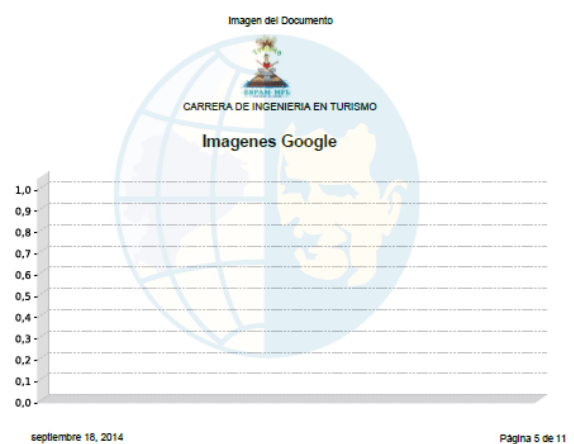
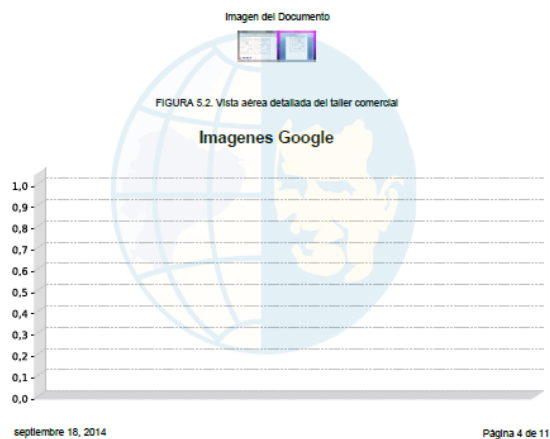


Ilustración 51 Anexo Reporte específico tesis Págs. 4/5/6/7

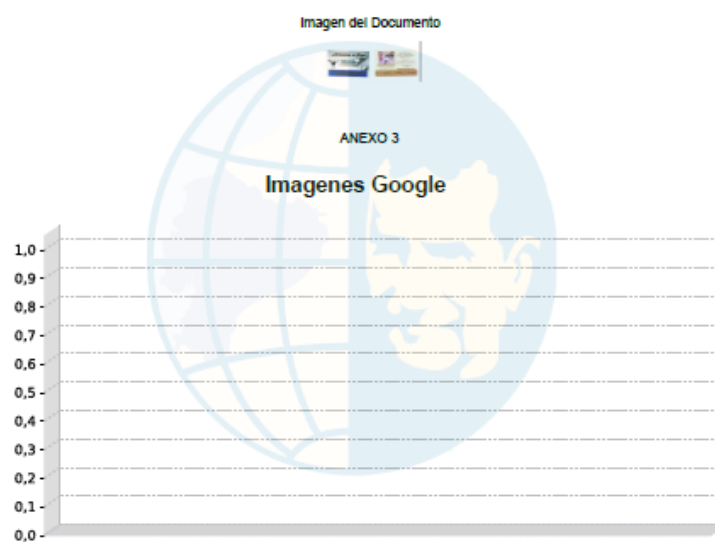


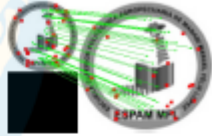
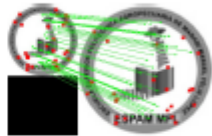
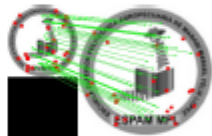
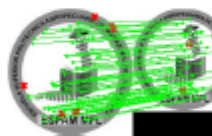
Ilustración 52 Anexo Reporte específico tesis Pag.8

Imagen del Documento



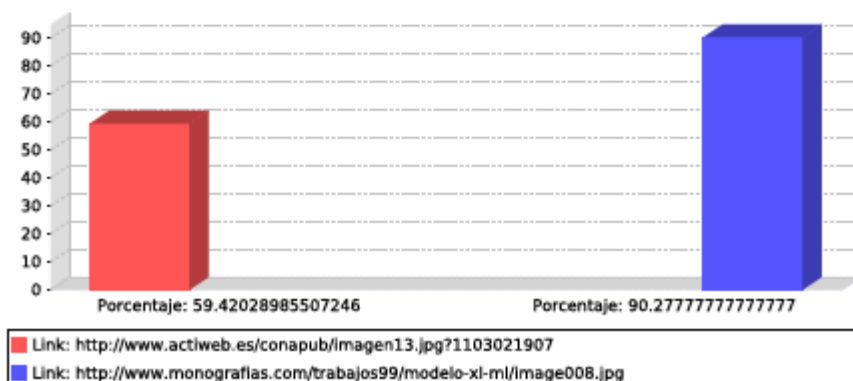
ESCUELA SUPERIOR POLITÉCNICA AGROPECUARIA DE

Imágenes Google

Descargado desde: http://www.actiweb.es/conapub/imagen13.jpg?1103021907  Imagen de la Web	Puntos Similares: 41.0 Puntos Analizados: 69.0 Porcentaje: 59.420 %	 Comparación Gráfica
Descargado desde: http://www.actiweb.es/conapub/imagen13.jpg?1103021907  Imagen de la Web	Puntos Similares: 41.0 Puntos Analizados: 69.0 Porcentaje: 59.420 %	 Comparación Gráfica
Descargado desde: http://www.actiweb.es/conapub/imagen13.jpg?1103021907  Imagen de la Web	Puntos Similares: 41.0 Puntos Analizados: 69.0 Porcentaje: 59.420 %	 Comparación Gráfica
Descargado desde: http://www.monografias.com/trabajos99/modelo-xi-ml/image008.jpg  Imagen de la Web	Puntos Similares: 65.0 Puntos Analizados: 72.0 Porcentaje: 90.277 %	 Comparación Gráfica

septiembre 18, 2014

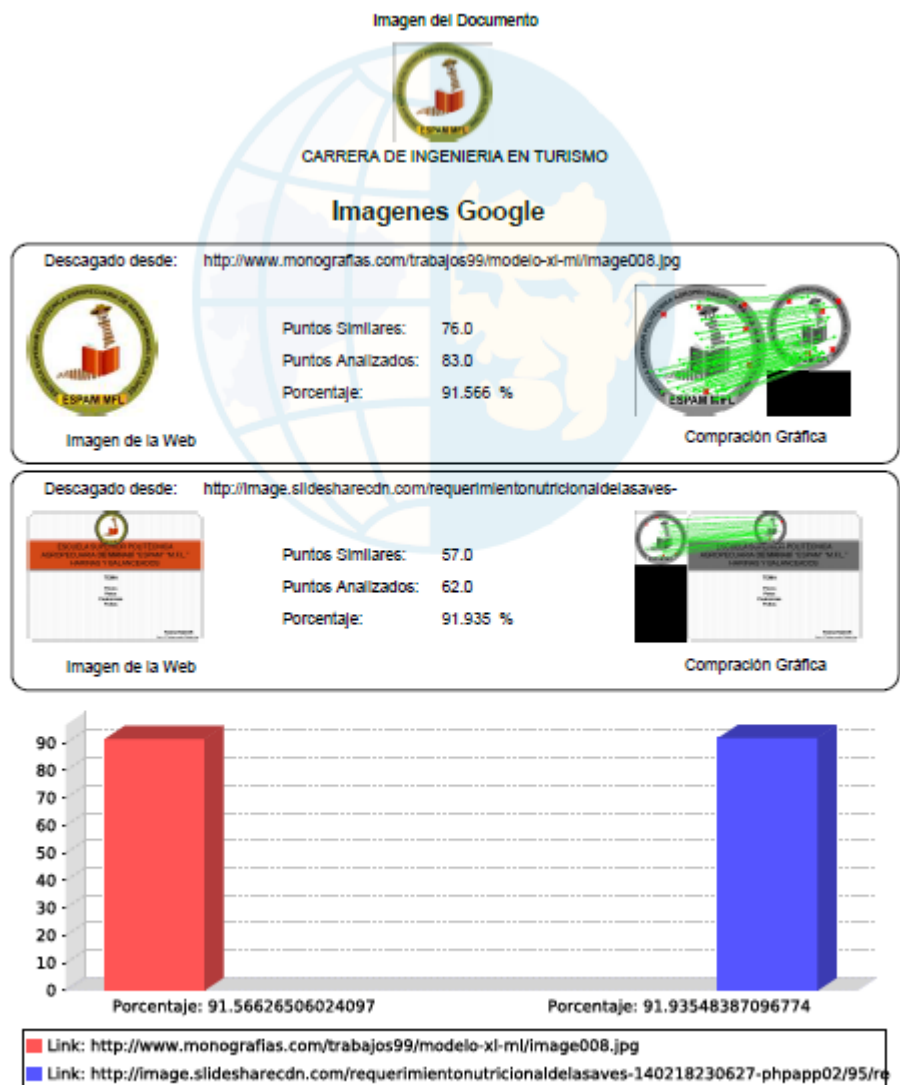
Página 9 de 11



septiembre 18, 2014

Página 10 de 11

Ilustración 53 Anexo Reporte específico tesis Pag.9 y 10



septiembre 18, 2014

Página 11 de 11

Ilustración 54 Anexo Reporte específico tesis Pág. 11

ANEXO 2 JavaDoc

ec.edu.ups.webservice	
Class AnalisisPlagioWS	
java.lang.Object ec.edu.ups.webservice.AnalisisPlagioWS	
<pre>public class AnalisisPlagioWS extends java.lang.Object</pre>	
Clase que se encarga de iniciar servicios web, para interaccion con otras plataforma	
Author: Juan Lozano & Jessica Zhunio	
Method Summary	
Methods	
Modifier and Type	Method and Description
byte[]	convertirBytes (java.lang.String path) Convierte a bytes a partir de un path específico
java.lang.String	estado () Consulta el estado en el que se encuentra el analisis
java.io.File	fromBytes (byte[] bytes, java.lang.String path) Convierte a file una cadena de bytes
java.lang.String	Solicitud (java.lang.String path, int numResultados) Inicia servicios para analisis de documentos
Methods inherited from class java.lang.Object	
clone, equals, finalize, getClass, hashCode, notify, notifyAll, toString, wait, wait, wait	

Ilustración 55 Anexo JavaDoc Class AnalisisPlagioWS

ec.edu.ups.webservice

Class Cola

java.lang.Object
ec.edu.ups.webservice.Cola

All Implemented Interfaces:

java.lang.Runnable

```
public class Cola
extends java.lang.Object
implements java.lang.Runnable
```

Esta clase ejecuta el proceso de deteccion en modo background para que el servicio no sea sincrónico

Author:

Juan Lozano & Jessica Zhunio

Method Summary

Methods	
Modifier and Type	Method and Description
java.lang.String	getEstado() Retorna el valor del estado en que se encuentra el analisis
int	getNumResultados() Retorna el numero de resultados
java.lang.String	getPath() Retorna la direccion de donde se encuentra el documento
void	run() Ejecuta el hilo, y ejecuta el metodo solicitud para iniciar el analisis del documento en modo background
void	setEstado (java.lang.String estado) Aisgna un valor al estado en el que se encuentra el analisis
void	setNumResultados (int numResultados) Asigna un valor a numero de resultados
void	setPath (java.lang.String path) Asigna un vaor al atributo path
java.lang.String	Solicitud (java.lang.String path, int numResultados) Solicitud de ejecucion de analisis

Methods inherited from class java.lang.Object

clone, equals, finalize, getClass, hashCode, notify, notifyAll, toString, wait, wait, wait

Ilustración 56 Anexo JavaDoc Class Cola

ec.edu.ups.webservice

Class Proceso

java.lang.Object
ec.edu.ups.webservice.Proceso

```
public class Proceso
extends java.lang.Object
```

Clase para inspeccionar tiempos de ejecucion.

Method Summary

Methods

Modifier and Type	Method and Description
void	comenzarProceso () Inicio del proceso
void	finalizaProceso (java.lang.String descripcion) Finalizacion del proces

Methods inherited from class java.lang.Object

clone, equals, finalize, getClass, hashCode, notify, notifyAll, toString, wait, wait, wait

Ilustración 57 Anexo JavaDocClass Proceso

ANEXO 3 Manual de Uso



1. **Cargar:** Nos permite buscar el documento que deseamos analizar y cargarlo al sistema para su posterior análisis.

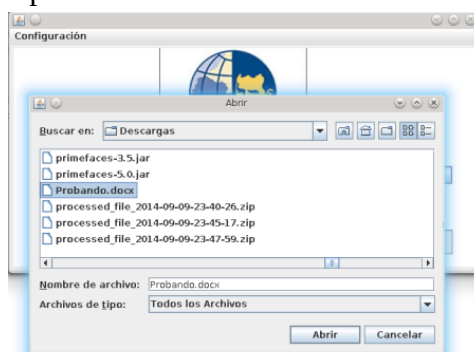


Ilustración 58 Interfaz Gráfica Carga de Documento

2. **Número Resultados:** Números de resultados que se analizara en el buscador de Google, y a esperar un tiempo prudencial de acuerdo a la magnitud del documento y al número de imágenes para que me aparezca la siguiente ventana con la dirección donde se encuentra el reporte final con los respectivos resultados.



ANEXO 4 Manual de Instalación

Manual de Ejecución del JAR:

1. Dependencias del sistema:
 - PYTHON
 - OPENCV
 - JAVA JDK
2. UNA VES INSTALADO TODAS LAS DEPENDENCIAS
Modifique el archivo config.txt con el path de donde se guardara los resultados y las imágenes descargadas de internet
cd %DetectorPlagiolImagenes
3. EJECUTAR EL JAR analisis-plagio-imagenes.jar
en consola java -jar analisis-plagio-imagenes.jar
El sistema detecta las ip de todas las interfaces disponible por defecto en el puerto 80.
4. Y por último se mostrará las dirección en las cuales está disponible los wsdl.

REFERENCIAS BIBLIOGRAFICAS

- [1] Real Academia Española (2001). Diccionario de la lengua española de la Real Academia Española.
- [2] Bordigon, F., Tolosa, R.A Rodriguez y Peri, J. Primeras Experiencias en la detección de Plagio en el Ambiente Educativo. *Actas, Primera Jornada de Educación en Informática y TICS en Argentina*, pp. 97-104, 2003.
- [3] SENESCYT anuncia nuevas políticas para el desarrollo de la educación superior (Noviembre 13,2013). Recuperado de:
http://ecuadoruniversitario.com/noticias_destacadas/senescyt-anuncia-nuevas-politicas-para-el-desarrollo-de-la-educacion-superior/
- [4] Renuncia ministro de Defensa en Alemania, (Marzo 2, 2011). Recuperado de:
http://www.nacion.com/mundo/Renuncia-ministro-Defensa-Alemania_0_1180681956.html
- [5] Armando Soto Rodríguez, (Enero 1, 2012), El plagio y su impacto a nivel académico y profesional, Recuperado de: <http://eprints.rclis.org/19890/1/2-1-2.pdf>
- [6] Revista Ingeniería de sistemas, Volumen XXV, Septiembre 2011, El Fenómeno del Plagio en Documentos Digitales: Un Análisis de la situación Actual en el Sistema Educacional Chileno, Francisco Molina, Juan D. Velásquez, Sebastián Ríos, Paulina A. Calfucoy, Matías Cociña.
- [7] El Universal, Bryce Echenique recibió en su casa de Perú el Premio FIL 2012, (Octubre 27,2012). Recuperado de:
<http://www.eluniversal.com.mx/cultura/70180.html>
- [8] La ministra alemana de Educación dimite por el presunto plagio de una tesis, (Febrero 9, 2013). Recuperado de:
<http://www.lavanguardia.com/internacional/20130209/54365225827/ministra-alemana-educacion-dimite-plagio-tesis.html>
- [9] Grupo Difusión Científica, (marzo 21, 2013), URKUND: El mejor recurso para la prevención del plagio en documentos. Recuperado de
<http://grupodifusioncientifica.wordpress.com/2013/03/21/urkund-el-mejor-recurso-para-la-prevencion-del-plagio-en-documentos/>
- [10] The Sherlock Plagiarism Detector. Recuperado de:
<http://sydney.edu.au/engineering/it/~scilect/sherlock/>
- [11] JPlag. Recuperado de: <http://jplag.ipd.kit.edu/>

- [12] SENESCYT anuncia nuevas políticas para el desarrollo de la educación superior (Noviembre 13,2013). Recuperado de: http://ecuadoruniversitario.com/noticias_destacadas/senescyt-anuncia-nuevas-politicas-para-el-desarrollo-de-la-educacion-superior/
- [13] TinEye, Recuperado de: <http://www.tineye.com/>
- [14] S. Eissen y B. Stein. Intrinsic Plagiarism Detection. En Proceedings of the 28th European Conference on IR Research, ECIR 2006 Londres, pp. 565-569, Springer 2006
- [15] Word Sequences. Lecture Notes in Computer Science, 2006, Volumen 4225/2006, pp 844-853, Springer
- [16] Hector Garcia Molina. SCAM: Un mecanismo de detección de copia de los documentos digitales, 1995
- [17] M. Potthast, A. Barrón-Cedeño, A. Eiselt, B. Stein y P. Rosso. Vista general del segundo Concurso Internacional de detección de plagio. Italia, 22-23 de Septiembre 2010.
- [18] B. Stein, S. Meyer zu Eissen y M. Potthast. Estrategias para la recuperación de documentos plagiados.
- [19] Victoria Elizalde, Estudio y Desarrollo de nuevos algoritmos de detección de plagio, Junio 2011
- [20] J. Kasprzak, M. Brandejs, y M. Kripac. Encontrar plagio evaluando Similitudes del documento. En Actas del Taller SEPLN'09 en descubrir el plagio, Autoría y Software Social El mal uso.
- [21] B. Stein y S. Meyer zu Eissen. La zona de Búsqueda de Similitud y Análisis de plagio, páginas 430 a 437, Springer 2006.
- [22] Jose Fernando Sanchez Vega, Detección automática de plagio basada en la distinción y fragmentación del texto reutilizado.
- [23] R. Rehúrek. Plagiarism Detection through Vector Space Models Applied to a Digital, Masarykova Univerzita, 2008, pp. 75-83.
- [24] Manning, Christopher D.: Introduction to Information Retrieval, Cambridge University Press, Cambridge, England, 1 Abril, 2009, p. 581.
- [25] J. Helfman. Dotplot patterns: a literal look at pattern languages. En: Journal Theory and Practice of Object Systems - Special issue on patterns archive Volume 2 Issue1, 1996.
- [26] ARAUJO Lourdes, Procesamiento de Lenguaje Natural, Noviembre 2011. Recuperado de: <http://tabasco.torreingenieria.unam.mx/gch/PLN/cap1.pdf>

- [27] Fellbaum Christiane, (2005). Recuperado de: <http://wordnet.princeton.edu/>
- [28] G. Miller. WordNet: A Lexical Database for English. En: Magazine Communications of the ACM CACM Homepage archive Volume 38 Issue 11, Nov. 1995.
- [29] Sistemas De Detección De Plagio En Textos. Recuperado de: http://discursoforense.unam.mx/sites/default/files/archivos/publicaciones/s_plagio.pdf
- [30] Nonis, S. & Owens, C. (2001). An Examination of the Relationship Between Academic Dishonesty and Workplace Dishonesty. A Multicampus Investigation. Journal of Education for Business.
- [31] Jaramillo Fernando, Recuperado de: http://www.virtualepn.edu.ec/moodleday2014/descargar/p_fjaramillo.pdf
- [32] Software Antiplagio En La Uleam-Capacitan A Docentes Para El Manejo De Urkund. Recuperado de: <http://www.uleam.edu.ec/software-antiplagio-en-la-uleam-capacitan-a-docentes-para-el-manejo-de-rkund/>
- [33] Marck Franco Salvador, Detección de Plagio Traslingüe Utilizando una Red Semántica multilingüe, Departamento de Sistemas Informáticos y Computación.
- [34] Alberto Barrón Cedeño, Paolo Rosso, David Pinto, and Alfons Juan. On cross - lingual plagiarism analysis using a statistical model. In Proceedings of the ECAI'08 PAN Workshop Uncovering Plagiarism, Authorship and Social Software Misuse, pages 9–13, Patras, Greece, 2008.
- [35] Paul Clough, Robert Gaizauskas, and Scott Piao. Building and annotating a corpus for the study of journalistic text reuse. In Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC-02), volume V, pages 1678–1691, Las Palmas de Gran Canaria, Spain, 2002.
- [36] Paul Clough, (February,2005). The METER Corpus Download page. Retrieved from: <http://nlp.shef.ac.uk/meter/>
- [37] Elizalde Victoria, (Junio, 2011). Estudio y desarrollo de nuevos algoritmos de detección de plagio. Recuperado de: <http://www.dc.uba.ar/inv/tesis/licenciatura/2011/elizalde>
- [38]M. Errami, Z. Sun, T. Long, A. George y H. Garner. Deja vu: a database of highly similar citations in the scientific literature. En Nucleic Acids Research, Vol.37, No. suppl 1. (1 January 2009), pp. D921-D924.
- [39] Elizalde Victoria, (Junio, 2011). Estudio y desarrollo de nuevos algoritmos de detección de plagio. Recuperado de: <http://www.dc.uba.ar/inv/tesis/licenciatura/2011/elizalde>

- [40] Lenguajes Documentales. Los Tesauros: Creación y Mantenimiento. Norma ISO 2788. Recuperado de: <http://www.bibliopos.es/Biblion-A2-Bibliografia-Documentacion/11Lenguajes-documentales-Tesauros-ISO-2788.pdf>
- [41] Tesoro Elementos y Funciones. Recuperado de: <http://www.oocities.org/es/tesauros/elementosyfunciones.htm>
- [42] Gil Urdiciain Blanca, 1997, Evolución histórica de los tesauros españoles y análisis de su rendimiento en el proceso de recuperación de información. Recuperado de: <http://biblioteca.ucm.es/tesis/19972000/S/3/S3029701.pdf>
- [43] Requisito, Recuperado de: [http://es.wikipedia.org/wiki/Requisito\(sistemas\)](http://es.wikipedia.org/wiki/Requisito(sistemas))
- [44] JAVA Características de Java como Lenguaje de programación. Recuperado de: <http://www.infor.uva.es/~jmrr/tgp/java/JAVA.html>
- [45] Eclipse IDE, (Enero 10, 2014). Recuperado de: <http://www.genbetadev.com/herramientas/eclipse-ide>
- [46] García Barquero, P. Sánchez González, M. Luna Serrano, E.J. Gómez Aguilera. Comparación de algoritmos detectores de puntos singulares para reconocimiento de objetos en vídeo quirúrgico. Recuperado de: [http://oa.upm.es/20480/1/INVE MEM 2012 135438.pdf](http://oa.upm.es/20480/1/INVE_MEM_2012_135438.pdf)
- [47] Rafael Aracil López, Desarrollo de un sistema cognitivo de visión para la navegación robótica, (Julio 28, 2012)
Recuperado de: <https://riunet.upv.es/bitstream/handle/10251/17010/memoria.pdf?sequence=1>
- [48] OpenCv. Recuperado de: <http://es.wikipedia.org/wiki/OpenCV>
- [49] Fun with Python, OpenCV and face detection, (June 27, 2008). Retrieved from: <http://blog.jozilla.net/2008/06/27/fun-with-python-opencv-and-face-detection/>
- [50] Artículo sobre las virtudes necesarias en el uso de los nuevos medios, (Mayo 19, 2014). Recuperado de: <http://sacerdotesyseminaristas.org/enlaces.php?listEntrada=2378>
- [51] Actividades de Clases, (Agosto 16, 2014) Recuperado de: <http://actividadesdeclases.blogspot.com/>
- [52] Búsqueda, reconocimiento y seguimiento de imágenes mediante el uso del algoritmo SIFT. Recuperado de: <http://iie.fing.edu.uy/investigacion/grupos/gti/timag/trabajos/2011/keypoints/PaginaWeb.html>
- [53] Python, (Diciembre 27, 2007). Recuperado de: <http://plagatux.es/2007/12/python-qt-pyqt-desarrollo-facil-y-rapido-de-aplicaciones-con-gui/>

- [54] Cristhian Herrera,(Abril 9, 2005), Introducción a iReport Recuperado de:
<http://www.adictosaltrabajo.com/tutoriales/tutoriales.php?pagina=ireport>
- [55] iReport-Designer for JasperReports. Recuperado de:
<http://sourceforge.net/projects/ireport/>
- [56] Google Images. Recuperado de: http://en.wikipedia.org/wiki/Google_Images
- [57] List of CBIR engines Recuperado de:
http://en.wikipedia.org/wiki/List_of_CBIR_engines
- [58] Metadatos, (2003.) Recuperado de:
<http://www.gitsinformatica.com/metadatos.html>
- [58] Juan Luis Quirós Cassillas. Marcelo Simón Penas. Recuperado de:
http://cala.unex.es/cala/epistemowikia/index.php?title=Tratamiento_y_procesado_de_imagenes._Tecnicas_CBIR
- [59] Xiawi, Recuperado de:
http://commons.wikimedia.org/wiki/File:Principe_cbir.png
- [59] Universidad de Piura, Facultad de Comunicaci[o , Guia para citas y referencias bibliográficas. Recuperado de:
http://www4.ujaen.es/~emilioml/doctorado/guia_rapida_de_citas_apa.pdf
- [60] MLA, Recuperado de: <http://www.mla.org/style>
- [61] Daily Miranda Pardo¹, Juniel Tamayo Hernández¹, Procedimiento para el despliegue de soluciones de software desarrolladas en la Universidad de las Ciencias Informáticas, basado en casos de estudio.
- [62] Purdue University, (abril 23, 2011), Las referencias bibliográficas en el estilo MLA, Recuperado de:
http://www.bbtck.ull.es/view/institucional/bbtck/Referencias_MLA/es
- [63] Normas ISO, Referencias Bibliográficas. Recuperado de
<http://www.sibum.cl/descargables/normas%20ISO%20690.pdf>
- [64] Sara PÉREZ ÁLVAREZ, Análisis de usabilidad de sistemas CBIR, Departamento de Biblioteconomía y Documentación Facultad de Ciencias de la Información Universidad Complutense de Madrid.
- [65] Universidad de Chile, (abril 26, 2013), INGENIERÍA INDUSTRIAL LANZÓ SOFTWARE DOCUMENT COPY DETECTOR (DOCODE). Recuperado de
<http://www.dii.uchile.cl/2012/05/03/ingenieria-industrial-lanzo-software-document-copy-detector-docode/>