Universidad de las Ciencias Informáticas

Facultad 3

HERRAMIENTA PARA LA DETECCIÓN DE INFORMACIÓN ÚTIL EN INFORMES DE CONTROLES A CLASES

Trabajo final presentado en opción al título de Ingeniero en Ciencias Informáticas

Autora: Claudia Salcedo García

Tutores: MsC. Elizabeth Rodríguez Stiven

Ing. Vladimir Milián Núñez

Ciudad de La Habana, junio de 2019

DEDICATORIA

A mi mamá porque lo es todo en mi vida, sin ella no sé qué sería de mí y todo lo que soy se lo agradezco a ella. Siempre me apoyó para que estudiara, me superara y me enseñó que con esfuerzo todo se puede alcanzar. Gracias por haber sido mi motor impulsor en los años de estudios.

A mi papá, mi hermana, mi tío Jorgito y mi sobrino por su amor y por regalarme todo ese apoyo incondicional.

A toda mi familia en general que son el tesoro más preciado que tengo.

A mi Juan por lo especial y maravillo que ha sabido ser conmigo, por ser paciente y ayudarme y darme todo su amor.

AGRADECIMIENTOS

Aunque no lo parezca, este es el fragmento del documento que me es más difícil de escribir. Son muchos los que estuvieron cerca a lo largo de este trabajo, y que me tendieron su mano de una forma u otra. A todos, muchas gracias, en especial:

A todos los profesores que he tenido desde que comencé en la primaria hasta terminar en la UCI, en especial a la jefa de año que tanto se preocupaba y corría por todos.

A mis tutores, Vladimir y Elizabeth por guiarme en todo este proceso que parecía interminable.

A mi oponente Dailién por sus correcciones, apoyo y regaños, espero que el lenguaje haitiano haya mejorado, A Osiel por su ayuda con todas las correcciones.

Al tribunal en general por todas sus críticas para bien.

A mi mamá, por estar siempre al tanto de TODO, de todo lo que me hacía falta, de preguntarme cómo va la tesis sin ni siquiera entender de que iba.

A mi papa, tú que me diste la vida, eres un hombre maravilloso y aunque nunca diga cuanto te quiero eres mi hombre favorito.

A mima, donde quiera que estés, sé que te sentirías muy orgullosa.

A mi hermana, mi tío Jorgito y mi sobrino son mi gran familia, gracias por ser parte de mí.

A mis tías, Espe, Jenny, mechy por estar siempre al pendiente de mí en todo momento.

A mi Doris y Vertico, el día que existan las palabras que lleva mi agradecimiento, les prometo que vuelvo a hacer otra tesis (que dure menos, claro) para ponerlas.

A eli, y yordan, por acogerme en su familia, todos los regaños consejos y ayuda que me brindaron, eli más que mi tutora fue mi guía mi amiga en todos estos años, y no te preocupes ya no te darán más quejas.

A Juan, agradecida estoy a dios de haberte puesto en mi camino, gracias por tu paciencia, comprensión y amor en todo este proceso.

A mis amigas, las mejores y más locas que existen con ellas he pasado los mejores momentos desde risa hasta llantos, de todas me llevo lo mejor, Daylin, Daima, Angélica y Dayana e Ingrid. ¡Las voy extrañar mucho!!!

A mi Ede, aunque no está, sabe cuánto trabajo he pasado para llegar a escribir estas líneas y que lo extraño mucho.

A Claudita, Rosi, Yanci y Raime mis amigas desde la primaria, aguantándome.

A todas las buenas amistades que hice en estos años Harold y Pablo, Pedro, el gordo, Oslen, el pucho, norbe, chichi, Raykof y hasta el puti.

A todos mil GRACIAS, los que están y los que no, por creer en mí, cuando yo no lo hacía.

Declaración Jurada de Autoría.

Declaro ser autora de la presente tesis y reconozco a la Universidad de las Ciencias Informáticas los derechos patrimoniales de la misma, con carácter exclusivo.

Para que así conste firmo la presente a los __ días del mes de ___del año ___.

Nombre y firma autor **<Nombre y firma tutor**

RESUMEN

Dada la necesidad de elevar la calidad educacional en las instituciones educativas es importante expresar el papel que juega el trabajo metodológico en las mismas, pues va dirigido a cambiar puntos de vista, estilos de trabajo y modos de actuación de los implicados, con el fin de obtener mayor eficiencia en su práctica pedagógica. Sin embargo, la extracción de información del creciente volumen de datos almacenados en los repositorios de informes de controles a clases es un aspecto pobremente tratado en el ámbito educacional, resultando escasos los trabajos donde se modele la extracción de información de informes de controles a clases desde un enfoque de agrupamiento de temas.

En el presente trabajo se propone una herramienta para la extracción de información útil en informes de controles a clases basada en algoritmos de minería de texto para el agrupamiento por temáticas. La propuesta está constituida por el método de agrupamiento de temas en informes de controles a clases encargado de generar tres modelos de temas (Logros, Señalamientos y Recomendaciones) que representen los principales temas tratados en un conjunto de informes de controles a clases; y la Herramienta Info_Controles a Clases que soporta al método. Además, se evalúa la calidad del método a partir del valor de coherencia, que certificará la interpretabilidad del modelo teniendo en cuenta el valor óptimo de la cantidad de temas. Finalmente se valora la viabilidad del uso de la herramienta mediante dos casos de estudio, uno para analizar el tiempo requerido para extraer información y otro para evaluar la calidad de información extraída a través de la precisión.

Palabras Claves: modelo de temas, agrupamiento de temas, informes de controles a clases.

RESUMEN

ABSTRACT

Given the need to improve the quality of education in educational institutions, it is important to express the role that methodological work plays in them, since it is aimed at changing the points of view, work styles, and modes of action of those involved, in order to obtain greater efficiency in their pedagogical practice. However, the extraction of information from the growing volume of data stored in the repositories of control reports to classes is an aspect that is poorly addressed in the educational sphere.

This paper proposes a tool for the extraction of useful information in control reports to classes based on text mining algorithms for thematic grouping. The proposal is constituted by the method of grouping topics in reports of controls to classes in charge of generating three models of topics (Achievements, Signals and Recommendations) that represent the main topics treated in a set of reports of controls to classes; and the Tool Info_Controles a Clases that supports the method. In addition, the quality of the method is evaluated based on the coherence value, which will certify the interpretability of the model taking into account the optimal value of the number of topics. Finally, the feasibility of using the tool is evaluated through two case studies, one to analyze the time required to extract information and the other to evaluate the quality of information extracted through precision and accuracy.

Keywords: theme model, grouping of topics, reports of controls to classes.

Tabla de contenido

Índice de Figuras

Índice de Tablas

INTRODUCCIÓN

El rápido desarrollo del conocimiento y la información y el constante proceso de globalización en todas las ramas del saber y sus consecuencias, repercuten en todas las esferas de la sociedad. Para poder asumir tales retos las universidades perfeccionan sus actividades sustantivas, con el fin de formar profesionales más preparados que den respuestas a los requerimientos sociales. Organizaciones internacionales como la Organización de Naciones Unidas para la Educación, la Ciencia y la Cultura (UNESCO) y la Organización Regional de Educación Superior en América Latina (CRESAL), proponen grandes esfuerzos en función del estudio de las particularidades y retos de la educación superior en la actualidad, encontrándose entre ellos la formación del personal académico, para que estén a tono con las exigencias actuales en la formación de profesionales (1).

Nuestro Comandante en Jefe Fidel Castro Ruz expresó(2):

> *"…Hay que trabajar para enriquecer los conocimientos adquiridos durante los estudios, para saberlos aplicar en la práctica de manera creadora y recordar que la realidad es siempre mucho más rica que la teoría, pero que la teoría es imprescindible para desarrollar el trabajo profesional de un modo científico…"*

Sobre esta base se perfecciona la gestión con el objetivo de lograr niveles superiores de calidad en cada uno de los procesos universitarios, que, unido a la evaluación periódica, a la elaboración y cumplimiento de planes de mejora que garantizan la atención a las debilidades, dan continuidad al proceso de perfeccionamiento y conciben de forma organizada, en diferentes etapas, el sistema de trabajo metodológico.

Dada la necesidad de elevar la calidad educacional en las instituciones educativas es importante expresar el papel que juega el trabajo metodológico en las mismas, pues va dirigido a cambiar puntos de vista, estilos de trabajo y modos de actuación de los implicados, con el fin de obtener mayor eficiencia en su práctica pedagógica. Tiene un efecto multiplicador de las experiencias y permite resolver problemas que se presentan en el proceso docente educativo, ofreciendo el modo de proceder.

El Trabajo Metodológico es la vía fundamental de superación, el cual garantiza la preparación profesional, logrando una adecuada integración de las clases con la actividad investigativa y laboral. Las funciones principales son la planificación, la organización, la regulación y el control del proceso docente educativo. El adecuado desempeño de estas funciones, garantiza el eficiente desarrollo del proceso docente educativo.

Desarrollar la gestión del trabajo metodológico, de manera que tenga un impacto en la formación integral del profesional que requiere la Sociedad, es una de las vías más pertinentes que tiene la universidad actual y en particular la universidad cubana, en la búsqueda de un modelo de gestión propio. En este contexto, es esencial el rol que desempeña el jefe de departamento docente y/o el coordinador (Director) de la carrera, en la gestión de las actividades sustantivas de la universidad, ya que es un metodólogo que contribuye al desarrollo exitoso del trabajo en el proceso pedagógico de la universidad, desde la gestión, a nivel departamental o de la carrera, según corresponda.

Una superación integral que garantice la elevación del saber profesional de los jefes de departamentos docentes en lo político, científico y pedagógico, desde una perspectiva gerencial que les permitirá trazar estrategias de dirección efectivas. Para lograrlo es de vital importancia realizar un adecuado control de la actividad docente basado en la revisión y análisis de informes y actas generadas como evidencias de la ejecución del trabajo metodológico (3).

El control de la actividad docente es aquel que se realiza a una de las formas organizativas del proceso docente educativo, previsto en el horario de clases de los estudiantes, y estará dirigido a comprobar el logro de los objetivos propuestos para dicha actividad(4).Debido a la necesidad de mejorar el proceso docente educativo, los controles a clases se realizan para contribuir a la mejora continua del trabajo metodológico como acción correctiva del proceso.

La Universidad de las Ciencias Informáticas (UCI), cuenta con un sistema de gestión académica (AKADEMOS), que incluye entre sus funcionalidades el registro de los controles a clases realizados en las siete facultades que componen la universidad. Esta información, resulta de gran importancia para los directivos docente, pues una adecuada revisión y análisis de la misma les permitirá tener una visión de la efectividad del trabajo metodológico, a partir del cual se pueden corregir los planes de acción. Sin embargo, una entrevista realizada a directivos

docentes refleja insuficiencias en el proceso de revisión y análisis de dicha información como:

- La revisión de informes se realiza de forma manual, que unido al alto número de informes a revisar, trae consigo que los directivos docentes tengan que invertir mucho tiempo para extraer información que facilite la toma de decisiones.
- Los análisis realizados se basan en las experiencias de los directivos y en la revisión de una pequeña muestra de los informes que no excede el 10% de la población.

Partiendo de la situación problemática expuesta anteriormente se plantea el siguiente **problema a resolver**: el modo en que se procesan los informes de controles a clases en la UCI limita la obtención de información útil para el trabajo de los directivos docentes.

Se plantea como **objeto de estudio**: la minería de texto.

Para dar solución al problema se define como **objetivo general**: desarrollar una herramienta para el agrupamiento de los elementos que componen los informes de controles a clases que permita la detección de información útil para el trabajo de los directivos docentes.

Para dar cumplimiento al objetivo general se definen los siguientes objetivos específicos:

- Seleccionar el enfoque utilizado por los algoritmos en la minería de texto para el agrupamiento por temáticas de los elementos que componen los informes de controles a clases.
- Diseñar un método que posibilite el agrupamiento de texto por temáticas a partir de los informes de controles a clases.
- Implementar una herramienta que soporte el método diseñado que permita la detección de información útil para el trabajo de los directivos docentes.
- Valorar la viabilidad del uso de la herramienta mediante un caso de estudio.

Definiéndose como **campo de acción**: minería de texto para el análisis de datos educacionales, específicamente los informes de controles a clases.

Idea a defender: si se desarrolla una herramienta que soporte el método diseñado para el agrupamiento de texto por temáticas a partir de los informes de controles

a clases, se logrará la obtención de información útil para el trabajo de los directivos docentes.

Métodos Teóricos

Análisis-síntesis: el empleo de este método permitió analizar individualmente los principales conceptos relacionados con la minería de texto que es el área de estudio a tratar, posibilitando un análisis profundo de cada uno.

Hipotético-deductivo: se utilizó para guiar la investigación desde el planteamiento de la necesidad de agrupar los elementos que componen los informes de controles a clases hasta desarrollar una herramienta capaz de realizar esta función; de manera tal que determine la información útil para el trabajo de los directivos docentes.

Histórico-Lógico: se empleó para estudiar la trayectoria y el desarrollo histórico de los sistemas de gestión universitaria y comprender la lógica de sus aportes, así como las tendencias actuales.

Métodos Empíricos

Entrevista: se realizó una entrevista a los directivos docentes de la facultad 3 para:

- Identificar las principales deficiencias existentes en la gestión del trabajo metodológico y la necesidad de contar con una herramienta informática para solucionar el problema.

Experimentación: se empleó para verificar la comparación de los resultados obtenidos con los distintos algoritmos utilizados para el diseño del método.

El presente documento está estructurado en tres capítulos cuyos contenidos son:

- Capítulo 1: se describen los conceptos asociados al objeto de estudio: la minería de texto, además de las técnicas de la minería de texto. Se definen las herramientas necesarias para realizar el trabajo, orientadas a CRISP-DM como metodología seleccionada adaptándose sus fases a la investigación.
- Capítulo 2: se describen los pasos a seguir para desarrollar la propuesta de solución guiándose por la metodología definida. Se realiza una guía para el uso de la herramienta, de manera tal que se logre una mejor interpretación por parte de los directivos docentes.

- Capítulo 3: se evalúan los modelos obtenidos aplicando la medida valor de coherencia y se valora la viabilidad del uso de la herramienta propuesta mediante la estimación del tiempo requerido por los especialistas para extraer información útil, además de la calidad de la información obtenida por la herramienta con la medida de precisión en un caso de estudio.

CAPÍTULO 1. Fundamentación teórica

Introducción

En este capítulo se realiza un estudio de las características orientadas al dominio de los controles a clases y su importancia en el trabajo metodológico. Se analizan los elementos de la minería de texto y algunas de sus técnicas enfocadas a la extracción de información de un conjunto de informes. Se define CRISP-DM como metodología adoptada para guiar el proceso de desarrollo y se describe el porqué de la selección.

1.1- Informes de controles a clases

A fin de lograr la pertinencia de la labor didáctica, en correspondencia con el modelo pedagógico que instaura el perfeccionamiento de los planes y programas para la formación, en la actual etapa del desarrollo educacional cubano, se ha planteado el propósito de mejorar el diseño y ejecución del trabajo metodológico, desde diferentes direcciones(5).

El trabajo metodológico es la labor que, apoyados en la didáctica, realizan los sujetos que intervienen en el proceso docente educativo, con el propósito de alcanzar óptimos resultados en dicho proceso, jerarquizando la labor educativa desde la instrucción, para satisfacer plenamente los objetivos formulados en los planes de estudio. Se orienta básicamente hacia la preparación de los directivos académicos, profesores y personal de apoyo, a fin de ponerlos en condiciones de dirigir con eficiencia y eficacia el proceso de formación(4).

Una de las funciones del Trabajo Metodológico es el control al proceso docente educativo. Siendo este el medio fundamental para conocer la calidad de dicho proceso, evaluar sus resultados y dirigirlo hacia el cumplimiento de sus objetivos, facilitando así la compresión del control a la calidad de la ejecución del proceso docente educativo. Para controlar la calidad con que se ejecuta el proceso docente educativo uno de los aspectos que se toman en consideración es el control a la actividad docente el cual se define como:

El control que se realiza a una de las formas organizativas del proceso docente educativo, previsto en el horario de clases de los estudiantes, y estará dirigido a comprobar el logro de los objetivos propuestos para dicha actividad. Al finalizar el control, sin la presencia de los estudiantes, el responsable dirigirá el análisis, dará las conclusiones al controlado, señalará los principales logros, los señalamientos y las recomendaciones. Los resultados del control se recogerán en un documento

que será firmado por el docente controlado, como constancia de que fue informado de sus resultados. Este documento que recoge todos los datos del control de la clase se le conoce como: informe de control a clases(ICC)(4).

Los directivos docentes utilizan la información contenida en los informes de control a clases para evaluar la calidad de la ejecución del proceso docente educativo. El análisis de esta información es de gran importancia para corregir estrategias metodológicas. De ahí que, la conjugación de métodos teóricos y empíricos como el análisis estadístico y el análisis de documentos de texto en los informes de controles, constituyen hoy un instrumento de trabajo fundamental para el desempeño exitoso de las funciones de los directivos docentes. Sin embargo, se consideran insuficientes los esfuerzos realizados en los sistemas de gestión académica dirigidos a extraer la valiosa información que contienen los informes de controles a clases para el trabajo de los directivos docentes.

1.2- Herramientas de Sistemas de Gestión Universitaria

En la bibliografía consultada se analizaron cinco (5) Sistemas de Gestión Universitaria (SGU), dos en el ámbito internacional y tres en el ámbito nacional. El estudio de estos sistemas está basado en seis principales indicadores: universidades que emplean estos sistemas para conocer el comportamiento geográfico de los mismos, el objetivo con el que fue desarrollado y puesto en funcionamiento en la universidad, si realiza análisis de los datos almacenados en el sistema de manera tal que apoye la toma de decisiones, si emplean gestión del trabajo metodológico como unas de las necesidades de nuestra investigación, el uso de técnicas de minería de texto y tratamiento a controles a clases como principal función definida en este investigación.

1.2.1- SIU-guaraní: Sistema de Gestión Académica

Es un sistema de gestión académica de libre acceso desarrollado en Argentina, por la Facultad de Informática de la Universidad Nacional de La Plata(6). Permite la gestión de los alumnos de forma segura con la finalidad de obtener información consistente para los niveles operativos y directivos (decano y secretario académico). Incorpora la planificación anual del calendario académico incluyendo comisiones y turnos de exámenes. Brinda la posibilidad de que el alumno realice por sí mismo operaciones tales como la inscripción a cursos y materias y la consulta de su situación académica. Permite a los alumnos de postgrado la inscripción a materias, consulta de notas y solicitar certificados. Realiza análisis

estadísticos de la información almacenada. Algunos de los módulos que ofrece este sistema son:

- Gestión de Matrículas: permite la inscripción y admisión, la reinscripción del alumno, el tratamiento de sanciones y su cambio de plan.
- Gestión de Cursado: administra los actos por los cuales un alumno selecciona las materias a cursar, el seguimiento de las actuaciones académicas de los alumnos durante el cursado de la materia y el registro del resultado de dicha cursada en las actas correspondientes.
- Gestión de Exámenes: administra los actos por los cuales un alumno selecciona las materias a rendir y el registro del resultado de dichos exámenes en las actas correspondientes.
- Estadísticas Generales: permite generar información estadística asociada al alumno, censados, ingresantes y egresados. Debe generar, también, un archivo de interface que actualice los datos del sistema.
- Información General: permitirá a los niveles directivos de la universidad explorar la información de manera tal que ayude en la toma de decisiones.

A pesar de la información que brindan los módulos que tiene este sistema y de realizar análisis estadístico a los datos almacenados, no se ajusta a los intereses de la investigación; ya que no realiza gestión del trabajo metodológico, no incluye la gestión de los controles a clases en sus módulos, ni aplica técnicas de minería de textos en los datos.

1.2.2- SIGA: Sistema Integrado de Gestión Académica

Sistema comercial desarrollado en España, utilizado en la Universidad de Piura. Es de libre acceso que puede ser utilizado en muchos centros educacionales, no solo universitarios, sino también en conservatorios, academias, colegios, centro de formación de empresas, maestrías, postgrados, entre otros. Además, realiza análisis estadísticos de la información almacenada(7). Está compuesto por módulos que cubren en su gran mayoría las necesidades de cualquier institución. Entre sus principales funcionalidades están:

- Alumno: gestión integral de alumnos matriculados en el centro, de tutorías, asistencias o faltas al aula, calificaciones y mensajerías.
- Generador de diplomas: para la realización propia y personalizada de diplomas y certificados del centro. Diversidad de formatos a definir por el usuario.

- Horarios: Se realiza la programación de los horarios de las clases tanto como el local donde se imparten.

Este sistema brinda varias funciones que mejoran las necesidades de la institución, pero no se ajusta a la tarea de investigación porque no gestiona el trabajo metodológico, pues no realiza la gestión de los informes de controles a clases, ni aplica técnicas de minería de textos.

1.2.3- GestAcad: Sistema para la Gestión Académica

Sistema creado en Cuba por un grupo de jóvenes desarrolladores de la Universidad de Matanzas Camilo Cienfuegos en un acercamiento a la solución del problema de la gestión de la información docente en las instituciones de educación superior cubana. Su principal característica es que permite llevar el control de la academia de enseñanza o centro de estudios de manera fácil y fiable. También facilita la actualización y el procesamiento de informaciones docentes de postgrado. Sus datos se muestran en la INTRANET de la Universidad de Matanzas en formato Web(8). Consta de los siguientes componentes:

- Administración: para la gestión de las tablas del sistema vía Web, así como agregar nuevas consultas al sitio oficial y establecer los distintos niveles de acceso a estas.
- Web para las Secretarias Docentes: para la Gestión de Estudiantes que permite hasta el momento la realización de acciones generales comunes en una Secretaría Docente, así como la obtención de reportes oficiales.
- Web para los Jefes de Departamentos Docentes: donde se incluyen acciones relativas como la asignación de la carga docente y el control sobre los profesores del Dpto.
- Web para los Profesores: donde estos pueden llevar el control docente de sus estudiantes, el control de las evaluaciones, así como reportes relativos a su carga docente.

El sistema resuelve los problemas de la Universidad de Matanzas, sin embargo, para la presente investigación no cumple con los indicadores de interés ya que no realiza la gestión de trabajo metodológico, ni de los informes de controles a clases.

1.2.4- SIGENU: Sistema de Gestión de la Nueva Universidad

Es un sistema de libre acceso desarrollado en Cuba por la Universidad Tecnológica de la Habana “José Antonio Echeverría” (CUJAE) con el propósito de automatizar los procesos vinculados a la gestión docente de todos sus centros

adscritos, el cual pretende controlar y motorizar el proceso de gestión académica a nivel nacional. Ha sido implementado en los Centros de Educación Superior del país. Permite el apoyo a la toma de decisiones acorde a los principales procesos docentes como matrícula, bajas y graduados(9).

Algunas de sus funcionalidades son:

- Codificadores: contiene toda la información con que debe contar el sistema y que es provista por el Ministerio de Educación Superior (MES).
- Matrícula: permite realizar el proceso de matrícula a través del cual los estudiantes pasarán a ser registrados en el sistema como estudiantes de Educación Superior.
- Control de estudiantes: permite buscar un estudiante registrado en el sistema, modificar los datos de un estudiante tanto personales como docentes, ubicar a un estudiante en un grupo o cambiarlo de grupo, realizar el pase de estudiantes a otros años de estudio y definir los que serán repitentes, así como dar baja a un estudiante del centro ya sea por licencia de matrícula, resolución o traslado.
- Plan de Estudio: permite la creación de los planes de estudio para las diferentes carreras del centro, así como realizar ajustes y modificaciones a un plan de estudio específico.
- Evaluaciones: permite registrar, modificar o eliminar las evaluaciones de los estudiantes registrados en el sistema, así como premios y bonificaciones.
- Reportes: permite obtener diversos reportes con los que se puede recuperar toda la información necesaria del sistema.

El sistema, a pesar de que pretende controlar y motorizar el proceso de gestión académica a nivel nacional, no aplica análisis sobre la información almacenada. No existen evidencias de la gestión de los controles a clases por lo que no se ajusta a la investigación.

1.2.5- AKADEMOS: Sistema de Gestión de la Universidad

Es un sistema Web cubano desarrollado en la Universidad de la Ciencias Informáticas(UCI). Brinda una interfaz común para todos sus usuarios. Además, realiza la gestión de toda la información referente a la formación de pregrado de un estudiante. Su principal misión es el control de procesos que intervienen en la gestión académica de un centro de estudios universitarios. Surge como respuesta a la necesidad de sustentar y dar soporte en la UCI a la labor del personal de

secretaría. Asimismo, tiene como visión obtener un producto genérico, capaz de ser aplicable y adaptable en cualquier centro que implemente el control docente universitario. Los primeros módulos liberados fueron: Matrícula y Plan de Estudio, a pocos meses de haber comenzado su desarrollo y, posteriormente, se fueron incorporando los demás módulos que conforman el sistema: Expediente, Profesor, Estudiante, Control Docente y Reportes (10).

A continuación, se nombra y describen algunos de estos módulos:

- Matrícula: permite el control de los datos de los estudiantes que van a comenzar a ser parte de la Universidad, así como la gestión de los movimientos a que estos son sometidos en su paso por la misma.
- Profesor: da la posibilidad de planificar la carga docente de los profesores pues es el encargado de asignar al mismo a un departamento y además asignarle grupos. Los profesores son la materia fundamental con que cuenta un centro de estudios para desarrollar con éxito su labor formativa.
- Estudiantes: brinda a los estudiantes, principales protagonistas del proceso docente, un espacio destinado a brindarles información referida a su desempeño académico en su paso por la universidad.

Cuenta con áreas de procesos identificadas (Pregrado, Postgrado, Cooperación, Residencia, Biblioteca, Desarrollo, Tecnologías, Investigación, Extensión, Organizaciones y Egreso). Los avances obtenidos con su despliegue en la UCI son palpables, por ejemplo, en todo momento, el sistema está disponible para la consulta y actualización de la información, ya sea por directivos, profesores y estudiantes, lo que agilizó la gestión, y eliminó la necesidad de intermediarios entre la información y los que la generan o necesitan. De igual forma, incluye la gestión de horarios docentes.

Del punto de vista metodológico el sistema solo se limita a la gestión del plan de controles a clases por áreas docentes y el análisis estadístico del cumplimiento del mismo. Sin embargo, no se utiliza ninguna técnica de minería de datos textuales que permita realizar el análisis de información que contienen estos informes.

Después del estudio y análisis realizado a los distintos sistemas de gestión universitaria, se concluye que resuelven parte de los problemas que existen en las universidades, ya que están concebidos para una mejor gestión, organización y control de los procesos docentes. Sin embargo, ninguno utiliza técnicas de análisis

de datos textuales que posibiliten la extracción de información útil para el trabajo de los directivos docentes en los informes de controles a clases. Para un mayor entendimiento del análisis de los indicadores por sistemas estudiados, se muestra la Tabla 1.

Tabla 1 : Análisis de Sistemas de Gestión Universitaria (SGU).

IND. / SGU.	UNIV.	OBJ.	ADatAlm	TM	TCC.	MTEXTO.
SIU-Guaraní	Universidad de Córdoba	Gestión de los alumnos	No	No	No	No
SIGA	Universidad de Piura	La gestión de los alumnos(no solo para universidad)	No	No	No	No
GestAcad	Universidad de Matanzas	Llevar el control del centro de estudios	No	No	No	No
SIGENU	CUJAE	Automatizar los procesos de la gestión docente	No	No	No	No
AKADEMOS	UCI	Control de procesos que intervienen en la gestión académica	No	Si	Si	No

Nota: IND: indicadores, **UNIV**: Universidad, **OBJ**: Objetivo, **TCC**: Tratamiento a los Controles a Clases, **MTEXTO**: Uso de técnicas de análisis de datos textuales, **ADatAlm**: Análisis de datos almacenados (Minería de Datos), **TM**: Trabajo Metodológico. Fuente: elaboración propia.

1.3- Extracción del conocimiento

En la actualidad la gran cantidad de información, textual y no estructurada, que se encuentra almacenada y que continuamente se genera, demanda de iniciativas que propicien un mayor aprovechamiento de ese recurso - información - para el descubrimiento de conocimiento y la toma de decisiones, por tal motivo la sociedad se enfrenta al reto de trabajar con volúmenes de información cada vez mayores, pero la mayoría de la información dentro de una organización se encuentra en formato de texto, ya sea en documentos, informes de trabajo o publicaciones. De este modo da lugar a la extracción de conocimiento de documentos y surge la necesidad de la llamada minería de textos. La Minería de Texto (MT) constituye el área de conocimiento dentro de la que se estudia esta problemática, y desde donde se generan soluciones para el descubrimiento de conocimientos potencialmente útiles, y no explícito, en una colección de textos, a partir de la identificación y exploración de patrones interesantes(11). La minería de texto difiere de la minería de datos en el tratamiento de la información, donde la información textual carece de una estructura. De esta forma, se hace necesario buscar alguna representación intermedia del texto que pueda ayudar a la aplicación de técnicas de descubrimiento, que nos permitan extraer información útil que se encuentra oculta por decirlo de un modo(12).

La minería de texto es un proceso que consiste en extraer información útil de conjuntos de documentos no estructurados de texto, e identificar automáticamente patrones interesantes no triviales o conocimiento(13).

Para llegar a descubrir conocimiento en texto, se debe pasar por algunas etapas importantes en este proceso, como es la etapa del preprocesamiento que le brinda al texto una forma intermedia que permita ser tratada computacionalmente, luego aplicar alguna técnica de minería de texto y finalmente la visualización de los resultados.

1.3.1- Procesamiento de lenguaje natural

El lenguaje es una de las herramientas centrales en nuestra vida social y profesional. El ser humano no domina con exactitud las reglas que definen y describen formalmente el lenguaje. Por este motivo, entender y producir el lenguaje por medio de una computadora es un problema muy difícil de resolver. Este problema, es el campo de estudio que en la inteligencia artificial se conoce como Procesamiento del Lenguaje Natural.

El lenguaje natural se entiende como el lenguaje hablado y escrito con el propósito que exista comunicación entre una o varias personas, es más directo para expresar lo que se quiere comunicar(14).

Aunque las diferencias en los lenguajes humanos y de computadora son expansivas, han sido los avances tecnológicos los que han comenzado a cerrar la brecha. El campo del procesamiento del lenguaje natural ha producido tecnologías que enseñan lenguaje a las computadoras para que puedan analizar, entender e incluso generar texto. Algunas de las tecnologías que se han desarrollado y se pueden utilizar en el proceso de minería de texto son: extracción de información, seguimiento de temas, resumen, categorización, agrupación, concepto, vinculación y visualización de información.

Preprocesamiento de los datos

Los textos se transforman en algún tipo de representación estructurada o semi-estructurada que facilite su posterior análisis. Es decir, el primer paso dentro de la minería de texto sería definir el conjunto (corpus) de documentos. Se debe evitar en esta etapa la duplicación de documentos dentro del corpus. Incluye:

- Tokenización: Dada una secuencia de caracteres y una unidad de documento definida, se entiende por tokenización al proceso de segmentación de una sentencia en unidades más simples denominadas tokens, al mismo tiempo se busca remover ciertos caracteres especiales y signos de puntuación. Un ejemplo de tokenización es:

Tabla 2: Tokenización.

Entrada	En un lugar de la Mancha, de cuyo nombre no quiero acordarme	
Salida	En un lugar de la Mancha	de cuyo nombre no quiero acordarme

- Eliminar todos los caracteres no alfanuméricos del texto (como signos de puntuación, etc.), se convierten todas las palabras en minúsculas y se eliminan los acentos. El resultado de este paso es un texto en el que las palabras no contienen ningún carácter no alfanumérico, están en minúsculas, sin acentos, y separadas entre sí por un solo espacio.
- Eliminación de palabras de función (stopwords: preposiciones, artículos, conjunciones, pronombres, adverbios, etc.) y otras dependientes del dominio. Esta eliminación está basada en un diccionario.

- Radicalizar: Es un proceso que permite la reducción de los tokens a sus “raíces”, eliminando de esta manera los sufijos. La complejidad de un proceso de radicalización es alta, debido a las características propias de cada lenguaje. A manera de resumen, en este análisis morfológico, se busca catalogar cada token de una sentencia para extraer sus “morfemas” y “raíces” para su posterior análisis.

1.3.2- Aprendizaje automático

En el campo del aprendizaje automático, los métodos de predicción se conocen como aprendizaje supervisado y los métodos de descripción como aprendizaje no supervisado. El aprendizaje supervisado es una técnica para deducir una función a partir de datos de entrenamiento, se basa en entrenar un set de datos de un modelo que, por medio de diferentes datos, se pueda predecir el comportamiento de una variable. En el aprendizaje no supervisado el modelo se ajusta a las observaciones y al contrario que en el aprendizaje supervisado no existe un conocimiento, a priori, de las clases y no se provee al modelo de datos de entrenamiento. El método permite agrupar datos de forma rápida y también son llamados métodos simétricos o indirectos(15).

La característica general de los algoritmos de aprendizaje no supervisado es que no requieren ninguna información previa sobre los documentos y que pueden ser aplicados a cualquier documento nuevo. Los principales métodos no supervisados son el agrupamiento y el modelado de temas.

El agrupamiento divide un conjunto de objetos en grupos que presentan características similares. El objetivo del agrupamiento es ubicar los objetos similares en el mismo grupo y así, asignar objetos distintos a diferentes grupos. Para buscar similitud entre las palabras se toman en cuenta las palabras vecinas inmediatas. De esta manera, cada palabra forma su propio cúmulo.

En el modelado de temas se usa un modelo probabilístico para determinar la probabilidad de membresía de los documentos en grupos determinados. El modelado de temas se considera como un proceso de agrupamiento con un modelo generativo probabilístico. Cada documento puede ser expresado como una combinación probabilística de diferentes temas, así, los temas se pueden considerar como una especie de cúmulo y la membresía del documento en ese tema tiene naturaleza probabilística.

Otra de las formas más habituales en el aprendizaje no supervisado suelen ser basados en redes neuronales que consiste en identificar relaciones de asociación o correlación entre un conjunto extenso de datos. Originalmente, las reglas de asociación surgen de la necesidad de muchas industrias de encontrar relaciones entre los registros o transacciones almacenadas en sus bases de datos.

1.3.3- Modelos de temas

Como se había dicho en el aprendizaje automático y el procesamiento del lenguaje natural, los modelos probabilísticos de temas utilizan la teoría de probabilidad para definir la distribución que mejor se ajusta a los datos observados y cuyo propósito básico es estudiar la condición de similitud que entre sí guarda un grupo grande de documentos, es decir, un corpus.

Un modelo temático es un tipo de modelo estadístico para descubrir los "temas" abstractos que ocurren en una colección de documentos. El modelado de temas es una herramienta de minería de texto de uso frecuente para el descubrimiento de estructuras semánticas ocultas en un cuerpo de texto. Los "temas" producidos por las técnicas de modelado temático son grupos de palabras similares. Un modelo de temas capta esta intuición en un marco matemático, que permite examinar un conjunto de documentos y descubrir, sobre la base de las estadísticas de las palabras en cada uno, cuáles son los temas y cuál es el balance de los temas de cada documento.

En la era de la información, la cantidad de material escrito que se encuentra cada día está simplemente más allá de la capacidad de procesamiento. Los modelos temáticos pueden ayudar a organizar y ofrecer conocimientos para comprender grandes colecciones de cuerpos de texto no estructurados. Originalmente desarrollado como una herramienta de minería de texto, se han utilizado modelos temáticos para detectar estructuras instructivas en datos tales como información genética, imágenes y redes. (16).

1.3.3.1- Modelo TF-IDF

El modelo de Frecuencia del Término - Frecuencia Invertida del Documento (TF-IDF) se propuso en 1983(17). Su enfoque es establecer una matriz $V \times D$, donde V representa un vocabulario que contiene todas las palabras posibles, y $|V|$ es el tamaño del vocabulario. D representa un conjunto de texto, y $|D|$ es el tamaño del conjunto de texto. Para cada palabra, la Frecuencia del Término (TF, por sus siglas en inglés) se calcula en todos los documentos, y también se calcula la inversa del

número de todos los documentos que contienen la palabra. Finalmente, el producto de TF y Frecuencia Invertida del Documento (IDF, por sus siglas en inglés) se almacena en la posición correspondiente de la matriz. Según la idea central del modelo TF-IDF, si la frecuencia de una palabra que aparece en el mismo documento es mayor, lo que se puede medir por TF, o la frecuencia de la palabra que aparece en todos los documentos es menor, lo que se puede medir por IDF, la palabra será más importante para el documento. Por lo tanto, la importancia de todo el documento se puede obtener calculando la importancia de todas las palabras (18)(19).

1.3.3.2- Modelo LSI

La idea del modelo de tema de probabilidad se originó a partir del Análisis de la Semántica Latente (LSA, por sus siglas en inglés). LSA también se conocía como Indexación Semántica Latente (LSI, por sus siglas en inglés), que fue propuesto por Scott Deerwester como un modelo espacial semántico en 1990 (20). La indexación semántica latente es un método de indexación y recuperación para identificar patrones en las relaciones entre los términos contenidos en una colección de textos no estructurados.

Es una extensión del modelo de espacio vectorial, que aborda los problemas de TF-IDF. LSI asume que las palabras en los documentos tienen alguna estructura semántica latente. Su principio es usar la técnica de Descomposición en Valores Singulares (SVD, por sus siglas en inglés) para convertir la matriz de TF-IDF en una matriz singular (20)(21).

Ventajas de LSI

- El LSI se puede ver como una mejora del modelo de espacio vectorial , que incorpora el concepto de semántica latente (22). Puede realizar la recuperación semántica en cierta medida y eliminar la influencia causada por los sinónimos y polisemas de algunas palabras.
- SVD en LSI solo ejecuta un tratamiento matemático a la matriz, que no necesita gramática, semántica y otros conocimientos básicos del procesamiento del lenguaje natural(23). Básicamente, es un método inflexible y fácil.
- SVD en LSI se usa para lograr el propósito de filtrar la información y eliminar el ruido. Mientras tanto, la matriz de rango reducido realiza una representación de alta dimensión del documento en el mapa del modelo de

espacio vectorial en la representación de baja dimensión del espacio semántico latente, reduciendo en gran medida la escala del problema (22).

- En la matriz original, una palabra solo aparece en unos pocos documentos, por lo que muchos valores de los elementos de la matriz serán cero. El problema de la escasez de datos hace que la manipulación de la matriz sea bastante difícil. Pero después de la SVD, la dimensión del espacio se reduce en gran medida, lo que hace que el problema de dispersión de datos mejore algo (21).

Desventajas de LSI

- El significado físico SVD no está claramente definido. Es difícil controlar el efecto de la clasificación y la agrupación de significados de las palabras (24).
- El proceso del algoritmo SVD no se puede controlar porque su tiempo, espacio y complejidad son demasiado grandes. Por lo tanto, es difícil tratar con datos a gran escala y aplicaciones reales bajo la capacidad operativa actual (25).
- La matriz actualizada después del algoritmo SVD tiene valores positivos y negativos, lo que significa que el valor de la similitud entre palabras y matriz puede ser negativo (24). Por lo tanto, es difícil presentar un significado físico definido de la matriz. Además, la incertidumbre del valor de la similitud traería algunas dificultades para otras aplicaciones.
- El efecto de resolver sinónimos y casi sinónimos no es obvio, aunque su efecto de reducción de dimensión es significativo. La razón es que el proceso de reducción de dimensión es inflexible sin la información previa. Por lo tanto, el rendimiento final de la clasificación de texto a menudo se verá muy dañado (20).

1.3.3.3- Modelo PLSI

El segundo avance importante de los modelos de probabilidad fue el modelo Probabilístico de Indexación Semántica Latente (PLSI, por sus siglas en inglés), que fue presentado por Hofmann en 1990 (24)(25). PLSI se utiliza para simular el proceso de generación de palabras, extendiendo LSI al marco de las estadísticas de probabilidad. Rediseña la idea de generar modelos. Abandona el método de transformación de matriz en LSI, pero hace uso del modelo generativo. Comparado con el modelo no generativo, el modelo generativo describe la razón de generar algunas funciones de densidad de probabilidad y el proceso de la interacción entre los factores.

PLSI es un modelo de probabilidad. Su idea principal es construir un espacio semántico donde la dimensión no sea alta. Y luego, todas las palabras y documentos se tratan por igual y se asignan al espacio semántico. De esta manera, no solo resuelve el problema de la alta dimensión, sino que también refleja la relación entre las palabras. Por ejemplo, si la semántica de las palabras está mucho más cerca, los puntos correspondientes a las palabras en el espacio semántico también estarán más cerca. En el proceso de construcción del mapeo, el modelo PLSI utiliza el algoritmo iterativo de Maximización de Expectativa (EM, por sus siglas en inglés), haciéndolo más eficiente (26)(27).

Ventajas del PLSI

- El espacio semántico latente de PLSI tiene un significado físico claro, que representa el tema latente. Además, otros valores de probabilidad también tienen sus significados físicos correspondientes (24).
- PLSI podría resolver el problema de los sinónimos y polisemas de manera efectiva, y utilizar el algoritmo de Maximización de Expectativas (EM) para entrenar las clases latentes (26). Comparado con LSI, tiene una base estadística sólida.
- PLSI utiliza el algoritmo EM para obtener soluciones por iteración mientras calcula el modelo, lo que reduce la complejidad del tiempo y aumenta la velocidad de la informática (27). Por lo tanto, es más fácil de lograr que el algoritmo SVD.

Desventajas de PLSI

- EM de PLSI es un algoritmo completamente sin supervisión por lo que su convergencia es lenta, mientras que el algoritmo itera (28).
- El espacio de parámetros de PLSI es proporcional a los datos de entrenamiento de PLSI, por lo que no es bueno para modelar un corpus de crecimiento dinámico o de gran escala (29).
- El modelo PLSI necesita obtener una probabilidad previa, que solo se basa en el conjunto de entrenamiento existente. Para el texto fuera del conjunto de entrenamiento, no existe una probabilidad previa adecuada (28).

1.3.3.4- Modelo Asignación Latente de Dirichlet

Con el objetivo de abordar los problemas anteriores en PLSI, Blei et al. presentaron el Modelo de Asignación Latente de Dirichlet (LDA, por sus siglas en inglés) en 2003 (30). Basado en el modelo PLSI, LDA usa una variable aleatoria latente de

dimensión K que obedece la distribución de Dirichlet para representar la proporción de mezcla de temas del documento, que simula el proceso de generación del documento. En la actualidad, el modelo LDA es uno de los modelos de temas de probabilidad más populares. Como se muestra en la (Figura 1) en el proceso de generación de texto, LDA utiliza el muestreo de la distribución de Dirichlet para generar un texto con el tema específico de la distribución multinomial, donde el texto suele estar compuesto por algunos temas latentes.

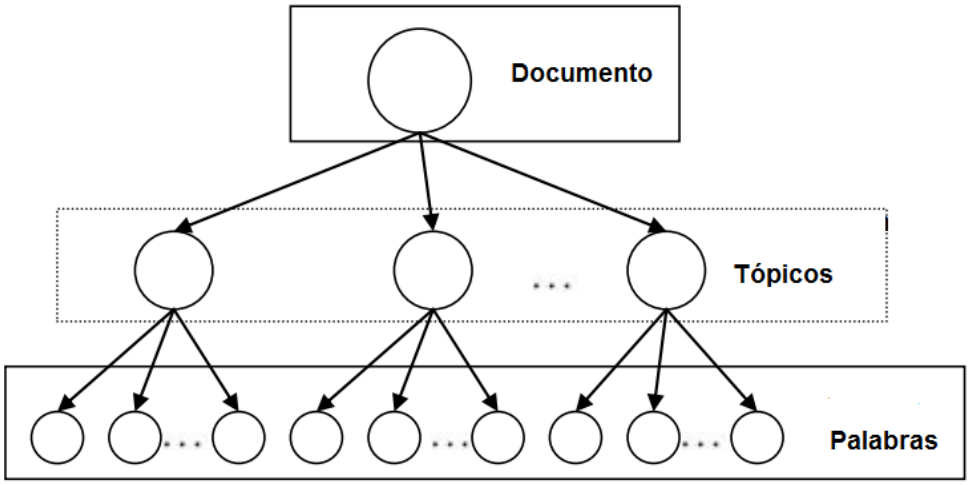


Figura 1: Topología del LDA.

El modelo LDA hereda todas las ventajas del modelo PLSI sin agregar cualquier condición idealizada, describiendo un modelo semántico de manera más objetiva. Por ejemplo, en la selección de temas, PLSI necesita determinar una etiqueta de clase de texto. De acuerdo con la etiqueta, se genera la distribución del tema, donde la probabilidad entre los diferentes temas es independiente entre sí. Pero, la distribución de Dirichlet se utiliza para generar un conjunto de temas en LDA. El conjunto de temas no es una serie de parámetros independientes de PLSI, pero está representado por las variables aleatorias latentes. Entonces, LDA coincide con las condiciones semánticas mejor que otros modelos en la realidad, y tiene un fuerte poder descriptivo. Comparado con el modelo PLSI, el espacio de parámetros de LDA es simple. Además, el tamaño del espacio de parámetros no tiene nada que ver con el número de documentos de capacitación en LDA para que no haya una situación de sobreajuste. Por lo tanto, el modelo LDA es un modelo generativo de probabilidad completa (28).

Ventajas del LDA

- El modelo LDA es el modelo generativo de probabilidad total, de modo que tiene una estructura interna clara y puede utilizar algoritmos de inferencia de probabilidad eficientes para calcular los parámetros del modelo (28).
- El tamaño del espacio de parámetros del modelo LDA no tiene nada que ver con el número de documentos de capacitación. Por lo tanto, es más adecuado para manejar corpus a gran escala(31).
- El modelo LDA introduce el híper-parámetro para el nivel de tema de documento, que es mejor que PLSI (31). Se agrega la información a priori, lo que significa que los parámetros se pueden ver como las variables aleatorias. Además, hace que LDA se convierta en un modelo jerárquico con una estructura más estable, evitando el sobreajuste (28).

El LDA es un modelo generativo de probabilidad que modela conjuntos de datos discretos, es un modelo bayesiano de tres niveles y es un método para modelar la información de temas de los documentos (30). Describe brevemente los documentos y mantiene la información estadística esencial, ayudando a procesar el conjunto de textos a gran escala de manera eficiente.

Es importante resaltar que, como dice Blei(16) los algoritmos no tienen información del tema sobre el cual los documentos están escritos y tampoco los documentos están etiquetados con los temas o palabras claves. La distribución de temas surge de analizar cuál es la estructura oculta más probable para generar la colección de documentos observada.

La utilidad de los modelos de temas se deriva de la propiedad que infiere la estructura oculta que se asemeja a la estructura temática de la colección. Esta estructura oculta interpretable clasifica cada documento en la colección mediante una minuciosa tarea y cada clasificación puede ser utilizada para ayudar a otras tareas como la recuperación de información, búsqueda, y la exploración de documentos. De esta forma, el modelo proporciona una solución algorítmica a la gestión, organización y anotación de grandes archivos de textos.

Producto de las desventajas que presentan los modelos de temas estudiados, se define el modelo LDA como el seleccionado para desarrollar en la investigación.

1.4- Visualización de los resultados

Para facilitar el descubrimiento de conocimiento la visualización de resultados en la minería de texto, tiene crucial importancia, ya que entrega un panorama general de una gran cantidad de datos. Métodos no supervisados de minería de texto,

como el agrupamiento, requieren un intenso trabajo que permita una correcta interpretación de los resultados con el fin de obtener información útil.

En esta etapa, una vez que han aplicado la técnica de minería de texto, se escoge la representación de los textos; la cual podría ser por medio de palabras, términos llaves, características, nube de palabras, conceptos, sugerencias de textos, etc. Esta representación debe ser fácil de manejar en tareas de minería de texto y debe ser lo más informativa posible, es decir, debe capturar los aspectos o características del espacio del problema. En esta etapa se proporciona la interfaz de exploración de los datos. Dicha interfaz debe ser lo más amigable para el usuario final.

Una nube de términos es un grupo de palabras clave etiquetadas en diferentes ubicaciones, formas, tamaños o colores, en forma de nube. Normalmente las de mayor tamaño y colores intensos, reflejan las temáticas de mayor importancia siendo las menos significativas aquellas más pequeñas y de colores más degradados. El objetivo principal de la nube de términos es facilitar al usuario la búsqueda de información relevante.

1.5-Metodología CRISP-DM

La metodología adoptada en este trabajo está basada en las etapas que caracterizan el modelo de proceso CRISP-DM: comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación y despliegue, ya que se desarrolla un proceso de minería de texto, el cual tiene un flujo de trabajo aceptado por la comunidad internacional.

Una metodología consiste en un conjunto de actividades organizadas que tienen como objetivo la realización de un trabajo. Para cada actividad se define, además de las entradas y salidas, la forma en la que debe llevarse a cabo (32).

La metodología CRISP-DM es introducida en una amplia gama de profesiones y está descrita en términos de un modelo de proceso jerárquico, estructurada en seis fases (33). Algunas de estas fases son bidireccionales, lo que significa que permiten revisar parcial o totalmente las fases anteriores (Figura 2).

El modelo consiste en:

Fase 1. Comprensión del negocio: esta fase se enfoca en la comprensión de los objetivos y requerimientos del proyecto desde una perspectiva no técnica, para luego convertir este conocimiento de los datos en la definición de un problema de

minería de datos y en un plan preliminar diseñado para alcanzar los objetivos. Sus principales tareas son:

- Establecimiento de los objetivos del negocio: cuáles son, en el contexto inicial, los objetivos que se tienen y los criterios de éxito.
- Evaluación de la situación: se realiza un inventario de los factores que deban ser considerados para alcanzar el objetivo del análisis de datos y conformar el plan del proyecto.
- Establecimiento de los objetivos de la minería de datos: determina los resultados esperados en términos técnicos que posibilitan alcanzar los objetivos propuestos.
- Generación del plan del proyecto, herramientas, tecnologías y técnicas: describe paso a paso cómo se pretenden alcanzar los objetivos de la minería y por tanto del negocio.

Fase 2. Comprensión de los datos: esta fase comienza con la recopilación inicial de datos y continúa con las actividades que permiten familiarizarse con los datos y verificar su calidad. Las principales tareas de esta fase son:

- Recopilación inicial de datos: se realiza la adquisición de los datos de las fuentes identificadas. Se identifican los problemas encontrados y las soluciones que se le dan a estos problemas.
- Descripción de los datos: caracterización general de los datos, ya sea su formato, cantidad, llaves y cualquier otra información descubierta.
- Exploración de los datos: realizar un análisis simple de los datos, usando preguntas, visualización o técnicas de reporte. Se analizan descubrimientos preliminares, hipótesis y su impacto en el resto del proyecto.
- Verificación de calidad de datos: se analizan posibles errores que tengan los datos (si están completos, faltantes, diferencia de formato, etc.). Se determina cuáles son las soluciones a seguir para el tratamiento de los errores.

Fase 3. Preparación de datos: esta fase cubre todas las actividades necesarias para construir el conjunto de datos que es utilizado en las herramientas de modelado a partir de los datos en bruto iniciales. Las tareas incluyen:

- Selección de atributos: se decide qué parte de los datos son incluidos o excluidos en el análisis y el porqué de esta selección (relevancia, calidad, restricciones, etc.).

- Limpieza de datos: su objetivo es elevar la calidad de los datos al nivel requerido por las técnicas de análisis seleccionadas (estimación de datos faltantes, decisiones tomadas para eliminar los problemas de la etapa anterior, transformaciones realizadas, etc.).
- Construcción de datos: se pueden crear atributos derivados a partir de atributos existentes o también generar registros, que no existen en la base datos almacenados, pero son válidos incluir en la modelación.
- Integración de datos: se mezclan datos de registros de diferentes tablas con información sobre el mismo objeto. Se pueden incluir datos agregados o dependientes de varios registros.
- Formateo de datos: se modifican los datos de forma tal que no cambie su significado pero que facilite la modelación.

Fase 4. Modelado: en esta fase varias técnicas de modelado son seleccionadas y aplicadas. También se realiza el diseño para la posterior evaluación del modelo. Las tareas principales con que cuenta esta fase son:

- Selección de la técnica de modelado: se escoge cuál técnica de modelación específica se va a utilizar (árboles de decisión, agrupamiento, redes neuronales, etc.). Se documenta la técnica escogida y el conjunto de requerimientos que deben cumplir los datos.
- Diseño de la evaluación: permite probar la validez y calidad del modelo. Se debe decidir cómo dividir el conjunto de datos, en conjunto de entrenamiento, prueba y validación.
- Construcción del modelo: se ejecuta la herramienta de modelación, se obtiene el modelo real que produce la herramienta de modelación y se realiza una interpretación del modelo resultante, describiendo las dificultades encontradas en la interpretación.
- Evaluación del modelo: se evalúa el modelo obtenido según el conocimiento del negocio y los criterios de éxito de minería de datos planteados, se realiza una comparación de la calidad de los modelos construidos. Según el resultado de la evaluación, se ajustan los parámetros y se vuelve a la etapa o fase anterior, hasta obtener los resultados esperados.

Fase 5. Evaluación: se evalúan los resultados finales obtenidos del modelo y se revisan los pasos del proceso realizado. Luego se compara el modelo obtenido con los objetivos inicialmente planteados. Las tareas de esta fase son:

- Evaluación de resultados: se evalúan los resultados del modelo desde el punto de vista de la exactitud técnica y del negocio. Los modelos aprobados son los que permiten alcanzar con éxito los objetivos propuestos.
- Revisar el proceso: se realiza la revisión del proceso resaltando si algo debe repetirse.
- Establecimiento de los siguientes pasos o acciones: de acuerdo con las pruebas realizadas hay que decidir cómo continuar, si finalizar o realizar nuevas iteraciones. Depende del cumplimiento de los objetivos y de los recursos disponibles.

Fase 6. Despliegue: en esta fase el conocimiento obtenido es organizado y presentado de modo que el usuario final pueda usarlo. Se establece una planificación de la monitorización y del mantenimiento del proceso, se generan los informes finales y se realiza la revisión del proyecto. Las tareas de esta fase son:

- Planificación de despliegue: según los resultados obtenidos en la evaluación, se decide una estrategia a seguir, sus pasos y cómo realizarlos.
- Planificación de la monitorización y del mantenimiento: una correcta estrategia de mantenimiento evita utilizar resultados incorrectos por largos períodos de tiempo.
- Generación de informe final: se debe organizar y resumir los resultados y experiencias del desarrollo del proyecto, los cuales son presentados en forma de reporte técnico, memoria, artículo, etc.
- Revisión del proyecto: definir qué estuvo bien, qué estuvo mal y qué se puede hacer mejor, es decir, resumir experiencias importantes en el desarrollo del proyecto.

En la Figura 2 se muestra el proceso CRISP-DM y la relación en cada una de sus fases.

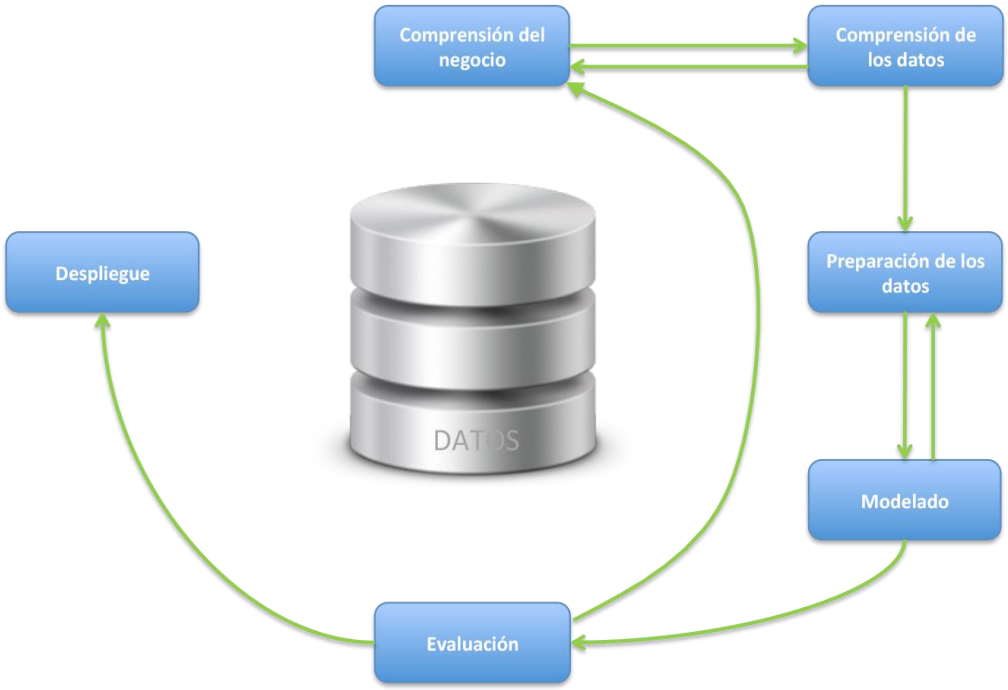


Figura 2: Metodología CRISP-DM

En el desarrollo del trabajo se adaptan las fases y tareas de la metodología CRISP-DM al problema de la investigación.

1.6- Herramientas tecnologías

1.6.1 Lenguajes de programación

Python 2.7

Lenguaje interpretado que tiene una sintaxis simple, clara y sencilla. Python usa tipado dinámico, pues no es necesario declarar el tipo de datos que va a contener una determinada variable. Es multiplataforma y tiene gran cantidad de bibliotecas disponibles para el trabajo con minería de texto.

Bibliotecas para la minería de datos y el procesamiento del idioma natural (35)

- **NLTK 3.2.5** (**Natural Language Toolkit**): conjunto de técnicas que permiten el análisis y manipulación del lenguaje natural. Se utiliza para la creación de programas en Python que interpretan el lenguaje humano. Permite realizar tareas de transformación y limpieza de documentos tales como: eliminar

caracteres especiales, signos de puntuación, convertir todo el texto en minúscula, eliminar palabras comunes o sin significado (conocidas como palabras de paradas) de la lengua en la que está escrito tales como: el, para, de, por, y, un, entre otras.

Bibliotecas básicas para la ciencia de los datos(34)

- **Sklearn 0.19.1**: es una biblioteca de código abierto para tareas de aprendizaje automático para el lenguaje de programación Python. Cuenta con varios algoritmos de clasificación, regresión y agrupación, y está diseñada para interoperar con las bibliotecas numéricas y científicas de Python, NumPy y SciPy (34).
- **Pandas 0.24.2**: es una biblioteca de Python diseñado para trabajar con datos "etiquetados" y "relacionales" de manera simple e intuitiva, está diseñado para una manipulación, agregación y visualización de datos rápida y fácil. Pandas agrega estructura de datos y herramientas que están diseñadas para el análisis de datos prácticos en finanzas, estadísticas e ingeniería. Pandas funciona bien con datos incompletos, desordenados y no etiquetados, es decir, el tipo de datos que es probable que encuentre en el mundo real, y proporciona herramientas para configurar, fusionar, remodelar y dividir conjuntos de datos (34).
- **NumPy 1.16.3**: se refiere a Numerical Python y es una biblioteca fundamental para la informática científica en Python ya que proporciona la vectorización de operaciones matemáticas en el tipo de matrices, mejorando el rendimiento y, en consecuencia, acelera la ejecución. Está orientada en administrar y tratar los datos como matrices, su propósito es proporcionar la capacidad de hacer operaciones complejas de matriz que son requeridas por redes neuronales y estadísticas complejas de manera fácil. NumPy es una biblioteca de administración de datos que normalmente está emparejado con TensorFlow, SciPy, Matplotlib y muchas otras bibliotecas de Python orientadas hacia Aprendizaje Automático y la ciencia de datos (34).
- **Gensim 3.7.2**: es una biblioteca de código abierto de Python para modelado de temas, indexación de documentos y recuperación de similitudes. Utiliza NumPy, SciPy y opcionalmente Cython para el rendimiento. Está específicamente diseñado para manejar grandes colecciones de texto y algoritmos incrementales eficaces, lo que lo diferencia de la mayoría de los paquetes de software científico que sólo se orientan al procesamiento por lotes y en memoria (34).

1.6.2- Entorno de desarrollo integrado

Una de las herramientas que desempeña un papel importante en el desarrollo de soluciones informáticas son los Entornos de Desarrollo Integrado (IDE). Es una aplicación informática que proporciona servicios integrales para facilitarle al programador desarrollo de software. Ofrecen facilidades al equipo de desarrollo cuando se implementan las aplicaciones debido a que permite la identificación de errores comunes que se comenten a diario.

- PyCharm 2019.1: PyCharm es un entorno de desarrollo integrado utilizado en la programación de computadoras, específicamente para el lenguaje Python. Está desarrollado por la empresa checa JetBrains (35).

Conclusiones parciales

Como conclusiones del capítulo:

- Tras el estudio bibliográfico de las tendencias actuales de las herramientas de gestión universitaria, se llega a la conclusión de que resuelven los problemas existentes en las universidades, pero no realizan análisis de datos textuales en los informes a clases que facilite el trabajo de los directivos docentes.
- El estudio de las tecnologías apropiadas para el desarrollo de la solución, determinó emplear Python como lenguaje de programación en su versión 2.7 por los beneficios que ofrece para el trabajo con minería de textos y PyCharm en su versión 2019.1 como entorno de desarrollo integrado para la implementación de la solución.
- Después de un análisis de los modelos de temas estudiados, se define LDA como modelo de tema a utilizar dada las ventajas que proporciona su uso para la investigación.

De esta forma se demuestra la necesidad de elaborar un método que obtenga información útil del conjunto de controles a clases y una herramienta que soporte dicho método para facilitar el trabajo de los directivos. Dando cumplimiento así, al primer objetivo específico definido en la investigación.

CAPÍTULO 2: Propuesta de solución

Introducción

En este capítulo se presenta la propuesta de solución, la cual está compuesta por el método para el agrupamiento por tema en informes de controles a clases y la Herramienta Info_Controles a Clases que soporta el método. Para su elaboración se siguió la metodología CRISP-DM adaptada a los requisitos de la investigación y las bibliotecas Gensim y Pandas de Python.

La Figura 3 describe de forma general la propuesta de solución. El método para el agrupamiento de temas es el encargado de construir un modelo para cada tópico de los ICC (señalamientos, logros y recomendaciones) que represente los principales temas tratados ellos. La herramienta fue diseñada para soportar el método, esta puede ser aplicable a cualquier conjunto de informe de controles a clases perteneciente a un departamento de la universidad. Su principal objetivo es facilitar el trabajo de los directivos en la extracción de la información útil para la toma de decisiones.

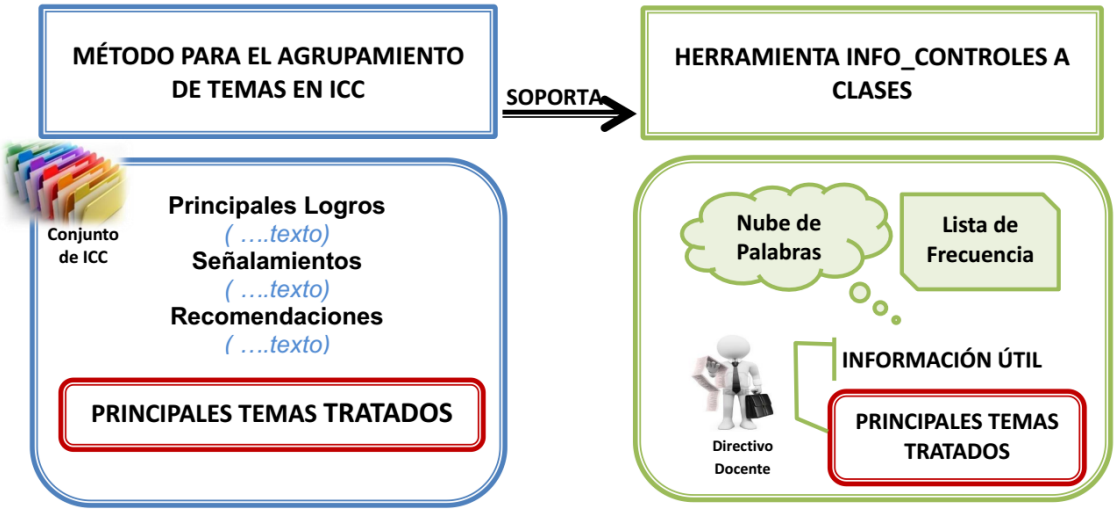


Figura 3: Propuesta de solución.

2.1- Método para la agrupación temática en informes de controles a clases

Siguiendo las etapas de la metodología CRISP_DM se describe el método para la agrupación temática en informes de controles a clases. La Figura 4 muestra un esquema de los pasos a seguir.

El método propuesto recibe como entrada un DATASET.xlsx que contiene todos los logros, señalamientos y recomendaciones del conjunto de ICC. Estos ficheros .xlsx representan el conjunto de datos a analizar. Luego se procede a realizar el

preprocesamiento al conjunto de datos. La tarea compuesta por los métodos llevar_Minúscula, eliminar_Puntuación y obtener_Tokens. Posteriormente se realiza el agrupamiento de temas mediante el algoritmo LDA. El método que se propone puede ser aplicable sobre un conjunto de informes a clases de cualquier departamento docente de la universidad. Como salida se obtienen tres modelos de temas (logros, señalamientos y recomendaciones) que representan los principales temas tratados en el conjunto de ICC.

Figura 4: Esquema del Método para el agrupamiento temático de controles a clases.

2.1.1-Conceptualización del problema

Detectar los temas en los tópicos[1] de ICC (logros, señalamientos y recomendaciones). Debido al alto volumen de información con el que se pudiera contar, obtener los principales temas tratados en el momento oportuno basada en los datos (información útil) requiere de un gran esfuerzo.

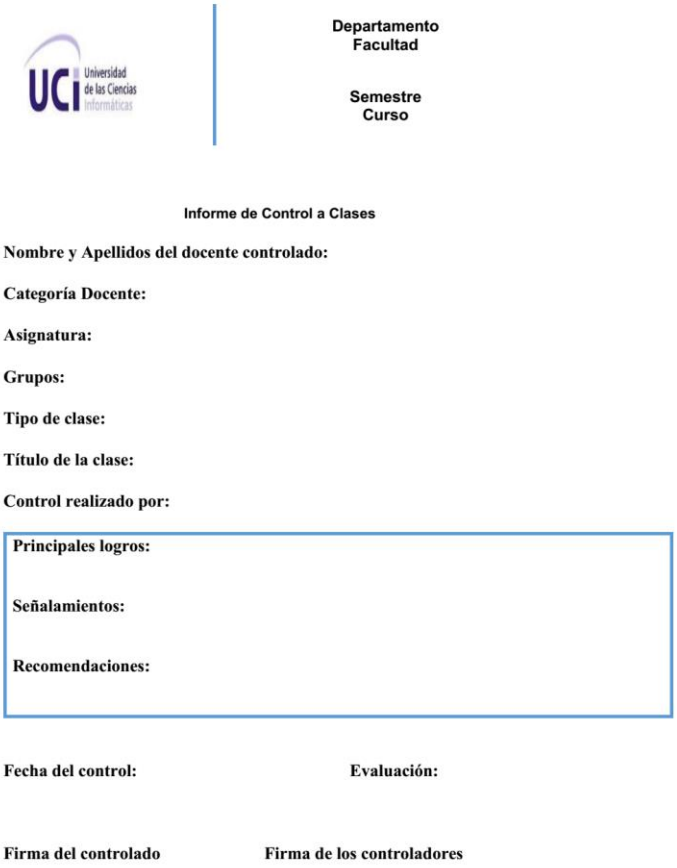
El objetivo que se persigue es detectar principales temas tratados en un conjunto de informes a clases, que sirva como información útil para el trabajo de los directivos. Como resultado se espera proporcionarles a los directivos docentes un conjunto de temas que caracterice la ejecución del proceso a partir de los comentarios escritos en los tópicos (logros, señalamientos o recomendaciones) de los informes de forma oportuna.

En la UCI los ICC se almacenan en AKADEMOS, contando así con todos los informes de controles a clases de cada departamento docente; lo cual constituye un gran volumen de datos. Se revisa la factibilidad de obtener los recursos necesarios para llevar a cabo este trabajo, como lo es la disponibilidad de un volumen considerable de ICC. En este contexto, se entiende que el estudio presenta ciertas restricciones, dado que no se cuenta hoy con un servicio que provea la información almacenada en AKADEMOS.

[1] En la propuesta de solución se le llamará tópicos a los logros, señalamientos y recomendaciones.

2.1.2-Comprensión de los datos

Los informes en su estructura contienen tres secciones, datos generales del control, datos generales de los profesores controladores y profesor controlado y las descripciones del control a partir de los logros, señalamientos y recomendaciones. Para la investigación los datos de interés están contenidos en la sección de descripciones del control. Los informes se almacenan siguiendo la estructura organizacional de la universidad (Facultad y Departamento), el período en el que se realiza (fecha, semestre y curso). Cada informe es generado siguiendo el siguiente formato (Figura 5).


Departamento
Facultad

Semestre
Curso

Informe de Control a Clases

Nombre y Apellidos del docente controlado:

Categoría Docente:

Asignatura:

Grupos:

Tipo de clase:

Título de la clase:

Control realizado por:

Principales logros:

Señalamientos:

Recomendaciones:

Fecha del control: Evaluación:

Firma del controlado Firma de los controladores

Figura 5: Formato de informes de controles a clases.

Los datos que contienen los ICC se transforman a un formato común, unificando toda la información recogida en un archivo DATASET.xlsx. (Ver Figura 6). El libro está compuesto por cuatro hojas: logros, señalamientos, recomendaciones y datos generales, donde cada fila de una hoja es la información referente a un informe de control a clases.

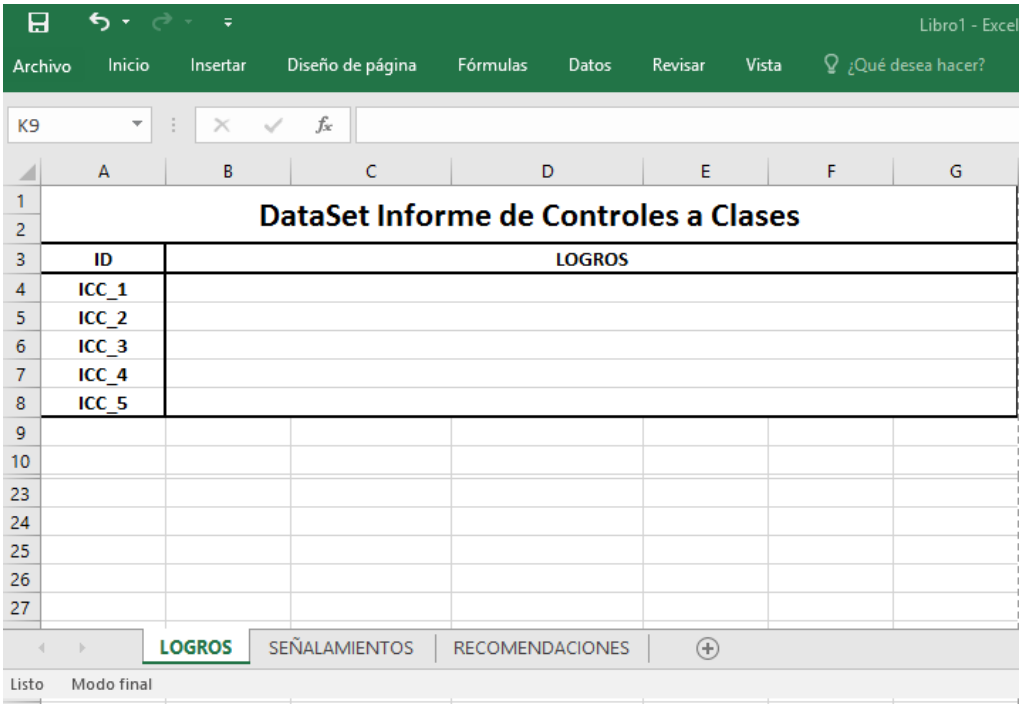


Figura 6: Excel para guardar la información.

Se cargan los datos del DATASET.xlsx y se transforman a codificación UTF-8[2](8-bit Unicode Transformation Format) utilizando la biblioteca pandas 0.24.2. Seguidamente se guarda toda la información de los logros, señalamientos y recomendaciones por separado en las siguientes variables: *Logros*, Senalamientos, *Recomenda* y *Datosg*. Quedando los datos listos para la próxima etapa.

Cargar Datos
Entrada: DATASET.xlsx
Salida: Logros, Senalamientos, Recomenda

```
#! -*- coding: utf-8 -*-
import pandas
Logros =pandas.read_excel("DATASET.xlsx",sheet_name="Logros")
Senalamientos=pandas.read_excel("DATASET.xlsx",sheet_name="S)
Recomenda =pandas.read_excel("DATASET.xlsx", sheet_name="R")
Datosg     =pandas.read_excel("DATASET.xlsx",     sheet_name="Datos
generales")
```

[2] UTF-8 (8-bit Unicode Transformation Format): Es un formato de codificación de caracteres Unicode e ISO 10646, que utiliza símbolos de longitud variable. Se recomienda usarlo para el latino que es más eficiente cuando aparecen tildes o “eñes”.

2.1.3-Preparación de los datos (Preprocesamiento)

En esta fase se procesan los ICC aplicándole las técnicas de procesamiento del lenguaje natural (ver epígrafe 1.3.1- Procesamiento de lenguaje natural).

- Se convierten todas las palabras en minúsculas.
- Se eliminan todos los caracteres no alfanuméricos del texto (los signos de puntuación).
- Se eliminan las palabras de paradas, (preposiciones, conjunciones, artículos). En los ICC existen palabras que son comúnmente utilizadas en su redacción como conectores; las cuales son identificadas por un experto en el tema. Para trabajar con ellas se almacenaron en tres listas (filtros.txt, filtro_señalamiento y fitros_recomendaciones) dado que estas cambian dependiendo del elemento que se esté analizando.
- La tokenización se ha llevado a cabo con el uso de la función *split*, que transforma los corpus a una lista cuyo contenido serán los tokens de las palabras incluidas en los títulos, palabras clave y resúmenes de los documentos.
- Se aplica radicalización para eliminar los sufijos y prefijos, y queda la raíz de la palabra.

Para trabajar con los datos en la etapa del preprocesamiento se utiliza la biblioteca NLTK 3.2.5 que es un conjunto de técnicas para el análisis y manipulación del lenguaje natural. Permite realizar tareas de transformación y limpieza de documentos tales como: eliminar caracteres especiales, signos de puntuación, convertir todo el texto en minúscula, eliminar palabras comunes o sin significado (conocidas como palabras de paradas) de la lengua en la que está escrito.

Los datos preprocesados se guardan en los diccionarios de textos limpios, quedando listos para la etapa de modelado. A continuación, se muestra el pseudocódigo del algoritmo del preprocesamiento.

Pseudocódigo

Entrada: Logros, Senalamientos, Recomenda

Salida: Diccionarios Preprocesado en Lpp, Spp y Rpp

Crear *diccionarios Lpp, Spp y Rpp para guardar los datos procesados*

Crear *una variable Lt, St y Rt para guardar listas de tokens*

Transformar *texto a minúscula*

Eliminar *signo de puntuación*

Eliminar *palabras de parada*

Transformar *en lista de tokens y* ***guardar*** *en Lt, St y Rt*

Radicalizar *Lt, St y Rt*

Guardar *Lt, St y Rt* ***en*** *Lpp, Spp y Rpp*

Fin Para

Retornar *Lpp, Spp y Rpp*

2.1.4-Construción del modelo

A continuación, se ilustran los pasos a seguir para la construcción del modelo de temas en ICC.

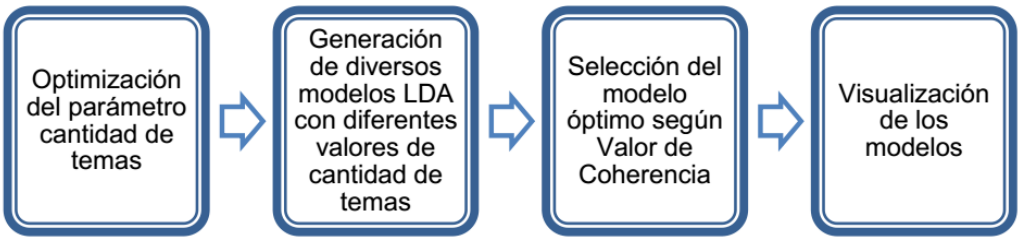


Figura 7: Construcción del modelo.

El modelo LDA requiere que se le defina de antemano la cantidad de temas a buscar por lo cual es importante hacer un análisis de este parámetro, previo a ejecutar el modelo. Por lo tanto, centraremos la optimización del parámetro *num_topics* "cantidad de temas" que es un parámetro clave en el modelo en cuestión.

La cantidad de temas para un modelo puede ser estimado usando la función HDP de Gensim 3.7.2. El Proceso Jerárquico de Dirichlet (HDP por sus siglas en inglés) fue propuesto por Y.Teh, que consideró que la estructura de HDP es similar a la estructura de LDA e introdujo el uso de HDP para encontrar la cantidad óptima de temas T en el algoritmo LDA (36).

Una de las formas más comunes para obtener la cantidad de temas es construir varios modelos LDA con diferentes cantidades de temas (num_topics) y tomar aquel que devuelva el valor de coherencia más grande considerando el contexto. Dicho valor de coherencia es una forma simple de ver qué tan bueno es el modelo.

En el (epígrafe 3.1 Comparación del modelo mediante valor de coherencia) se detallan los análisis.

Posteriormente, utilizando el parámetro adecuado, se genera el modelo en cuestión observando los temas obtenidos y finalmente analizaremos dichos temas mediante visualizaciones.

Para construir el modelo LDA utilizando la biblioteca de Python Gensim 3.7.2 (ver epígrafe 1.6.1 Lenguajes de programación) se necesita conocer:

- el diccionario de palabras preprocesadas que es la salida de la etapa anterior, (*id2word_logros*, *id2word_senalam, id2word_recomenda*).
- el corpus es un mapeo del documento representado a través un vector que almacena el conjunto de pares (id_palabra, frecuencia_palabra), donde el id_palabra es un id único asignado a cada palabra del documento y frecuencia_palabra es la cantidad de veces que aparece la palabra en el documento (*corpus_logros*, *corpus_senalam*, *corpus_recomenda*).
- la cantidad de temas o temas que pudiera contener el documento *(topic_logros*, *topic_senalam*, *topic_recomenda)*.
- Alfa y Beta: son hiperparámetros que afectan la densidad de temas. El valor por defecto de ambos es 1/num_topics.
- Chuncksize: es el número de documentos a ser utilizados en cada pasada de entrenamiento.
- Passes: la cantidad de pasadas por el corpus durante el entrenamiento.

Código para la construcción del modelo LDA de los señalamientos

```
id2word_senalam=corpora.Dictionary([lista_senalamiento_limpia])

corpus_senalam =[id2word_senalam.doc2bow(text) for text in
[lista_senalamiento_limpia]]

lda_model=gensim.models.ldamodel.LdaModel(corpus=corpus_senalam,
id2word=id2word_senalam, num_topics=20, random_state=100,
update_every=1, chunksize=100, passes=10, alpha='auto',
per_word_topics=True)

print (lda_model.print_topics())
```

El modelo obtenido es representado mediante una lista de n_temas, cada tema es una combinación de palabras claves y cada palabra clave contribuye con un peso al tema. La lista de palabras dentro del tema es ordenada descendentemente por su peso, indicando la importancia de la palabra dentro del tema.

En el método propuesto se construyeron tres modelos, uno para cada tópico de los informes de controles a clases. Los modelos seleccionados para cada tema fueron: (En el epígrafe 3.1 se detallan los análisis realizados).

- Tema Señalamiento: Modelo generado con 20 temas y valor de coherencia 0.5065.
- Tema Logros: Modelo generado con 25 temas y valor de coherencia 0.5104.
- Tema Recomendaciones: Modelo generado con 15 temas y valor de coherencia 0.4885.

A continuación, se describirá detalladamente el modelo señalamientos obtenidos a través de una representación visual de la distancia intertópicos. Los modelos obtenidos en los tópicos logros y recomendaciones siguen una representación visual similar a los de señalamientos.

Visualización de los temas del Modelo Señalamiento con 20 temas

Para tener una visualización interesante de los temas y sus palabras claves, se utilizó la librería pyLDAvis, que nos provee de una forma simple de analizar los resultados obtenidos. El gráfico muestra un mapa de la distancia intertópicos y las palabras clave de los temas. En la *Figura 8* se muestra el modelo señalamiento con 20 temas y las palabras claves del tema 1.

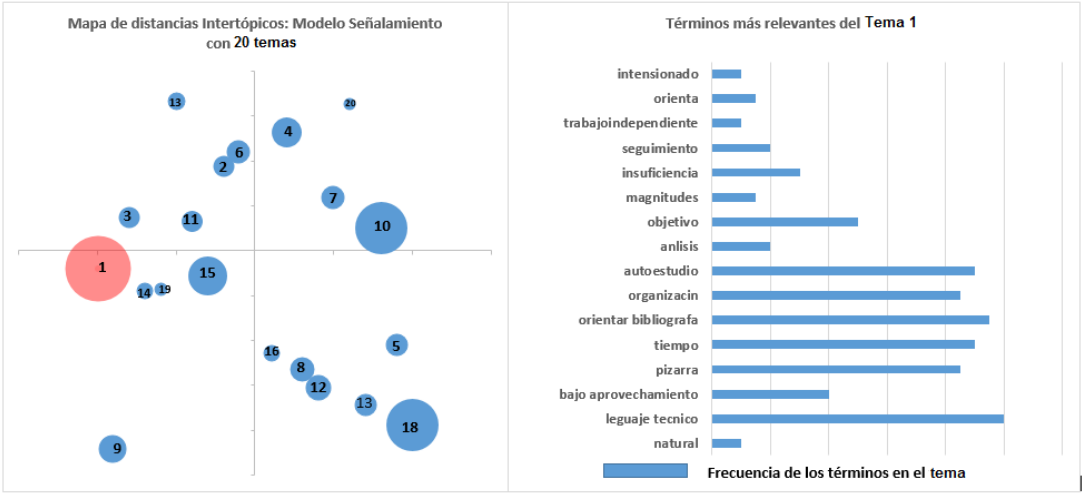


Figura 8: Visualización de la salida del modelo LDA (Señalamientos).

Cada burbuja en la parte izquierda de la visualización representa un tema. Cuanto más grande es la burbuja, más predominante es ese tema en el modelo. Un buen modelo de temas es aquel que tiene burbujas bastante grandes y dispersas en todo gráfico, es decir, que no estén todas juntas, solapadas y agrupadas en un único cuadrante(37). Un modelo con muchos temas seguramente tendrá burbujas pequeñas, ubicadas en una misma región del gráfico y con muchos casos de solapamiento.

En el modelo se observa un equilibrio entre las burbujas grandes y las pequeñas. Las burbujas se encuentran dispersas en el gráfico y existen pocos casos de solapamiento; lo cual indica que estamos en presencia de un modelo adecuado. Las burbujas más predominantes en el modelo son la 1, 10 y la 18. Las palabras claves están asociadas claramente al tema de los ICC “Señalamientos”, siendo más específicos, el 1 está asociado a “deficiente uso del lenguaje técnico, insuficiencias en la orientación de la bibliografía a utilizar para el autoestudio”. El 10 está asociado a “deficiencias en el control del tiempo asignando a las actividades, deficiente organización de la pizarra” y el 18 a “Insuficiente seguimiento a estudiantes con bajo aprovechamiento”. Las burbujas solapadas indican que hay palabras repetidas en los temas. Sin embargo, un análisis interno del tema puede mostrar las diferencias entre ellos. Por ejemplo, el solapamiento de los temas 8 y 12 está asociado a un tema dentro del tópico señalamientos que podríamos llamar “Insuficiente análisis de los resultados”. El tema 8 se refiere al “tiempo dedicado al cálculo” y el 12 a la “interpretación de los resultados”.

Las barras en el gráfico de la derecha indican la frecuencia del término en el tema, es decir cuan relevante es la palabra. Como se puede apreciar el más predomínate (tema 1), está caracterizado por las palabras “lenguaje técnico, orientar bibliografía y autoestudio” que son las de mayor frecuencia.

Otra forma de visualizar el modelo, es la nube de palabras. Esta representa las palabras de mayor relevancia en general para todos los tópicos. Las palabras de mayor tamaño indican la importancia de la palabra en el modelo. Esta representación es más fácil de interpretar para los usuarios finales de la herramienta. La nube de palabras, de conjunto con la lista de términos ordenadas por frecuencia, permitirá extraer los temas más relevantes tratados en los datos (Figura 11). Por tal motivo se utilizará la nube de palabras para visualizar el modelo en la herramienta. Para generar la nube de palabras se utilizará la biblioteca WordCloud 1.5.0 de Python.

2.2-Herramienta Info_Controles a Clases

La herramienta para la detección de información útil ICC (Info_Controles a Clases) soporta el método diseñado para el agrupamiento temático. Su objetivo es proporcionar a los directivos docentes información implícita de un conjunto grande de ICC disminuyendo el tiempo de análisis.

2.2.1-Descripción de la herramienta

Es una aplicación de escritorio, diseñada con el objetivo de facilitar el trabajo de los directivos docente en la extracción de información útil desde un conjunto grande de ICC. Esta permite detectar cuales son los principales temas descritos en los informes. Puede ser utilizada como herramienta de apoyo en la evaluación de la calidad de la ejecución del proceso docente educativo.

La solución consiste en una herramienta realizada en lenguaje Python que permite la lectura de archivos en formato .xlsx con información referente de los controles a clases, para finalmente obtener los temas principales de dichos controles.

Para visualizar el modelo se utiliza la nube de términos la cual permite observar las palabras de mayor importancia en el modelo. La importancia de palabra en la nube está dada por el tamaño de la misma (entre más grande más importante es).

2.2.2-Guía para el uso la herramienta

La herramienta permite extraer información de un conjunto de ICC siempre que se cumpla el siguiente requisito:

Los datos a procesar deben estar en el DATASET.xlsx donde el texto de cada tópico (logros, señalamientos y recomendaciones) esté almacenado en una hoja diferente (ver Figura 6).

Para su utilización el directivo docente deberá escoger la dirección donde se encuentra el DATASET y cargarlos, como se muestra en la Figura 9.

Figura 9: Cargar datos en la herramienta.

Cuando los datos estén cargados correctamente el directivo selecciona el botón referente al elemento que se quiera mostrar la nube de términos para su posterior análisis (Figura 10).

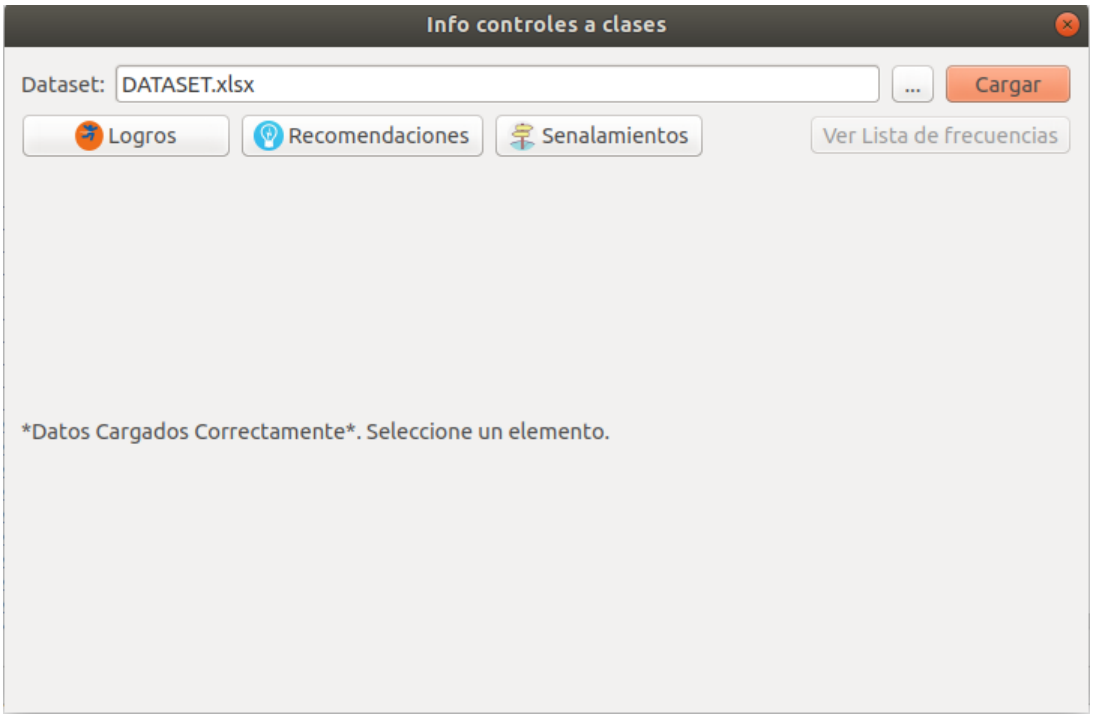


Figura 10: Botones para mostrar Nube de términos.

Cuando se selecciona el botón que se desee (Logros, Recomendaciones, Señalamientos) se mostrará la nube de términos resultante. Como se observa en la figura siguiente (Ver Figura 11), las palabras de mayor tamaño son las de mayor frecuencia en el texto y por tanto tienen mayor relevancia dentro del tema al que

pertenece. Estando en la nube de términos se puede ver la lista de frecuencia (botón ¨Ver Lista de frecuencias¨).

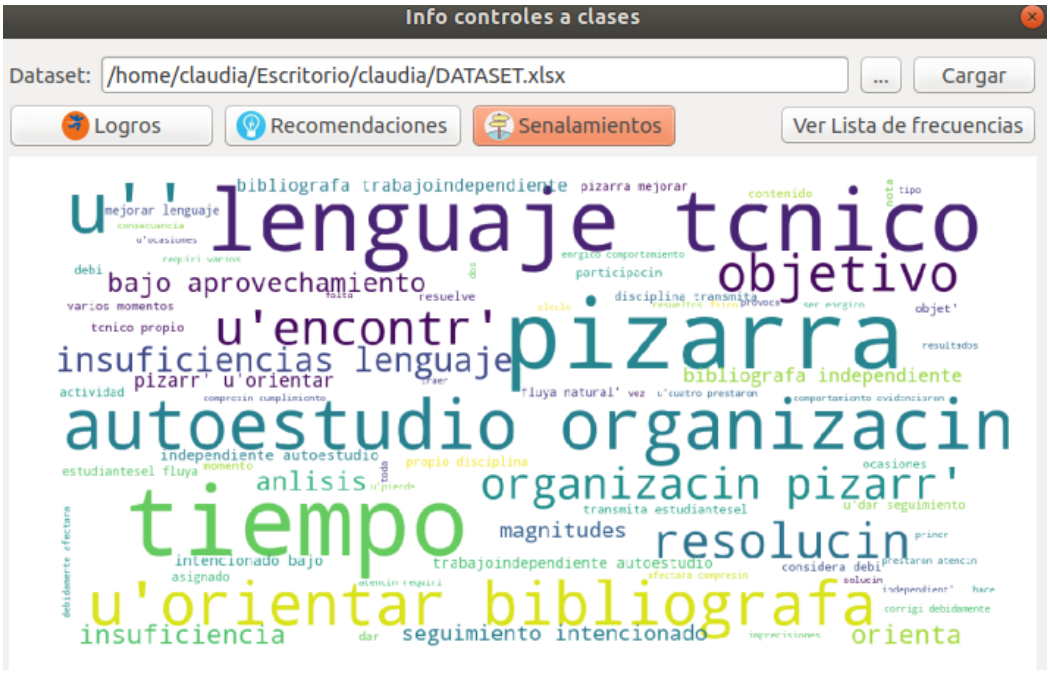


Figura 11: Nube de términos (Señalamientos).

Para interpretar la nube de términos teniendo la lista de frecuencia, se escoge un término en la nube y para saber el contexto en que se enmarca se busca en la lista de términos, ejemplo:

- insuficiencias lenguaje técnico
- organización pizarra
- tiempo análisis resultados
- orientar bibliografía a utilizar en el autoestudio

Conclusiones del capítulo

Como resultado de la elaboración de la propuesta de solución se obtuvo:

- El método para el agrupamiento de temas en ICC, que genera tres modelos LDA que representan los principales temas tratados en un conjunto de ICC.
- La herramienta Info_Controles a Clases que soporta el método y facilita el trabajo de los directivos docente para la extracción de información útil.
- Una guía para facilitar el correcto uso por los directivos docentes.

De esta forma la herramienta Info_Controles a Clases puede ser utilizada por los directivos docentes en la evaluación de la calidad de la ejecución del proceso docente educativo; con el objetivo de utilizar la información extraída en los futuros

análisis correctivos del trabajo metodológico. De esta forma se da cumplimiento a los objetivos específicos 3 y 4 de la investigación.

CAPÍTULO 3: VERIFICACIÓN DE LA VIABILIDAD DE LA PROPUESTA DE SOLUCIÓN

Introducción

En el presente capítulo se muestra el análisis correspondiente a la evaluación de los modelos obtenidos mediante el valor de coherencia y la valoración de la viabilidad del uso de la herramienta mediante casos de estudios. Se comienza escogiendo el valor de coherencia que proporcione mejor resultado para realizar el modelo de temas. Luego se estima el tiempo y se calcula la precisión para valorar la viabilidad del uso de la herramienta.

En la Figura 12 se propone la estrategia a seguir para valorar la viabilidad del uso de la herramienta.

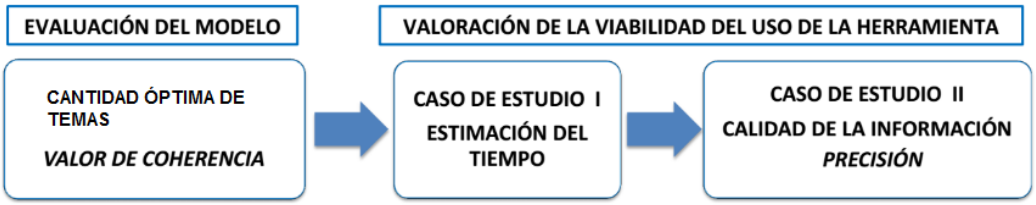


Figura 12: Estrategia de valoración de viabilidad del uso de la herramienta.

3.1 Comparación del modelo mediante valor de coherencia

A continuación, se explicarán los pasos realizados para validar la solución planteada.

1. Selección de la cantidad de temas.

Para formar el conjunto de valores posibles de cantidad de temas para obtener un buen modelo con pocos casos de solapamiento y burbujas dispersas en el gráfico, se comenzó creando el valor de tema que generó el hdp.

```
hdp = HdpModel(corpus_senalam,id2word_senalam)

topics_inf = hdp.print_topics()

hdp_topic_señalamiento = len(topics_inf)
```

Se obtuvo como resultado del hdp los posibles valores óptimos de cantidad de temas:

- hdp_topic_Señalamiento= 20
- hdp_Logros= 20
- hdp_Recomendaciones= 10

A partir de ellos se generaron siete valores de temas comenzando en cinco, de forma tal que los tres valores generados por el hdp estuviesen en el conjunto.

2. Evaluación de los modelos mediante la métrica valor de coherencia.

Los modelos de temas no garantizan que los temas extraídos del corpus se hayan interpretados correctamente, por lo tanto, se proponen medidas de coherencia para distinguir entre los temas buenos y malos. LDA es una técnica no supervisada, lo que significa que se desconoce antes de ejecutar el modelo cuántos temas existen en nuestro corpus. La coherencia de los temas es una de las principales técnicas utilizadas para estimar la cantidad de temas(38).

La medida de valor de coherencia del tema es una buena manera de comparar modelos de temas en función de su interpretación humana. Las coherencias de los temas capturan el número óptimo de temas (cantidad de temas) definido como puntuación de coherencia. A mayor valor de coherencia mayor interpretación de los temas del modelo.

En esta investigación se utilizó la biblioteca Gensim para calcular el valor de coherencia de cada modelo como se muestra en el código a continuación.

```
coherence_model_lda = CoherenceModel(model=lda_model, texts=docs, dictionary=dictionary, coherence='c_v')

coherence_lda = coherence_model_lda.get_coherence()

print ('\nCoherence Score: ', coherence_lda)
```

3. Análisis comparativo.

Los valores de coherencia obtenidos para cada modelo usando el conjunto V_temas se muestran en la Tabla 3.

Tabla 3: Valores de coherencia para los cantidad de temas.

Número de temas	Valor de Coherencia Señalamientos	Valor de Coherencia Logros	Valor de Coherencia Recomendaciones
5	0,4378	0,4356	0,4084
10	0,4695	0,4584	0,4205
15	0,4794	0,4994	0,4885
20	0,5065	0,5065	0,4185
25	0,4769	0,5104	0,4058
30	0,4853	0,4505	0,4265

La Figura 13 representa los diferentes valores de coherencia obtenidos en la tabla anterior.

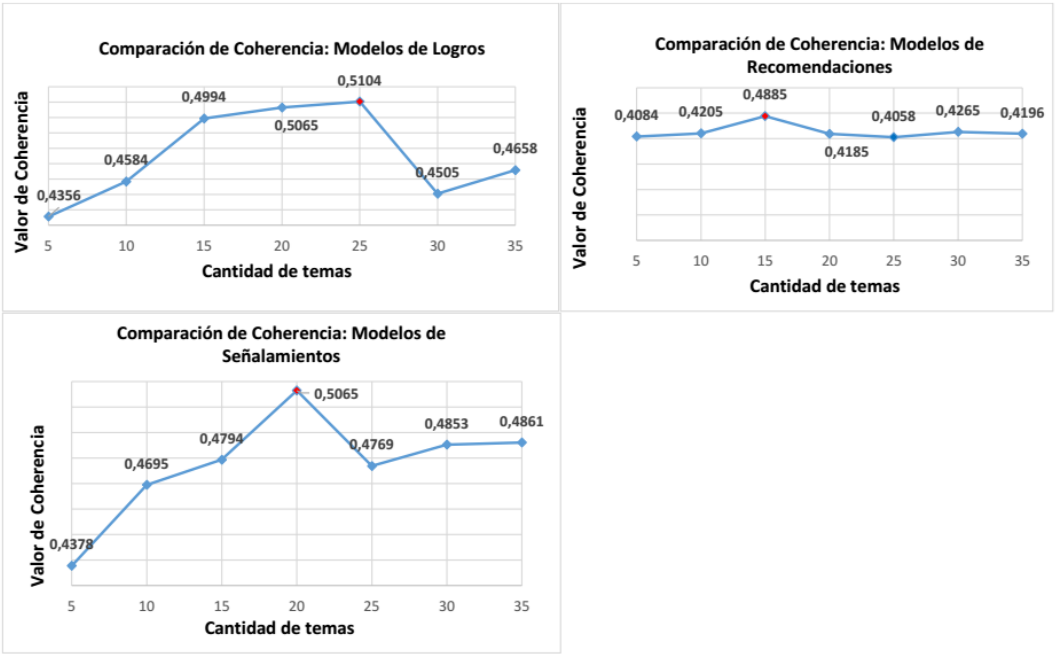


Figura 13: Valor de coherencia para diferentes temas

Como se observa en el gráfico el valor óptimo de la cantidad de temas obtenido fue 20 para señalamiento, 25 para logros y 15 para recomendaciones. Cuando el valor de coherencia sigue aumentando, lo mejor es elegir el número de temas que representa el máximo valor de coherencia antes de que la curva se empiece a hacer más llana(38).

Al generar los modelos con los distintos valores de temas y hallarles el valor de coherencia, solo en el caso de señalamientos el valor de coherencia más alto lo

tiene el valor de temas generado por la función hdp ya que esta función se utiliza para encontrar una posible cantidad óptima de temas T en el algoritmo LDA(36), en el resto de los casos no coincide.

3.2- Valoración de la viabilidad del uso de la herramienta

En el marco de estudio se entiende por Información útil:

- Diminución del tiempo para extraer información de un alto volumen de datos.
- Calidad de la información.

Caso de estudio

Teniendo en cuenta los elementos que componen la obtención de información útil de un conjunto de ICC, la valoración de la viabilidad del uso de la herramienta en un entorno real se evaluará mediante la realización de dos casos de estudio con el propósito de probar si:

- El uso de la herramienta propuesta disminuye el tiempo que emplean los directivos para extraer información de un conjunto amplio de ICC.
- Determinar la calidad de la información extraída a través de las medidas de exhaustividad y precisión de la herramienta Info_Controles a Clases.

3.2.1- Caso de estudio para estimar el tiempo

Con el objetivo de probar si el uso de la herramienta propuesta disminuye el tiempo que emplean los directivos para extraer información de un conjunto amplio de ICC, se diseñó un caso de estudio que compone dos escenarios: un primer escenario donde se estima el tiempo usando el método actual (escenario más simple) y un segundo escenario donde se estima el tiempo usando de la herramienta con una (escenario complejo). Finalmente se hará una comparación gráfica de los tiempos estimados.

Estimación del tiempo con el método actual

Para estimar el tiempo mínimo que se demora un directivo para extraer la información del conjunto de ICC se toma como caso de estudio:

Un directivo docente que deberá obtener información de un conjunto de 11 ICC que representa el 10% de la población usando el método actual (revisión manual documento a documento) el cual se compone de las siguientes tareas:

- Revisar todos los ICC y extraer los datos.
- Analizar el conjunto de datos para extraer los temas tratados.
- Determinar la frecuencia de ocurrencia de cada tema tratado.
- Determinar los temas de mayor importancia a partir de la frecuencia de ocurrencia.

Se desarrollará una implementación sencilla del método actual: Solo un directivo docente, con el 10% de los ICC dedicado a analizar uno de los Temas: “Señalamientos”. En la Tabla 4 se detalla el tiempo mínimo que requirió el directivo para realizar las tareas y el tiempo total requerido para emitir un juicio sobre los temas más importantes extraídos en el conjunto de datos.

Tabla 4: Tiempo estimado para obtener la información usando el método actual.

Tareas	**Tiempo (horas)**
Revisar 9 ICC y extraer los datos.	0,16horas
Analizar el conjunto de datos para extraer los temas tratados.	0,50horas
Determinar la frecuencia de ocurrencia de cada tema tratado	0,33horas
Determinar los temas de mayor importancia a partir de la frecuencia de ocurrencia.	0,23horas
Tiempo total	1,19horas

Estimación del tiempo utilizando la herramienta Info_Controles a Clases

En este caso siguiendo un escenario complejo se estimará el tiempo máximo que demora un directivo para extraer la información del conjunto de ICC en la herramienta.

Dicho directivo docente que deberá obtener información de un conjunto de 90 ICC que representa el total de la población usando la herramienta. Se analizarán los tres temas analizados en los ICC. En la Tabla 5 se detalla el tiempo máximo que requirió un directivo para analizar cada tema y el tiempo total requerido para emitir un juicio sobre los temas más importantes tratados en el conjunto de datos.

Tabla 5: Tiempo estimado para obtener la información usando la herramienta Info_Controles a Clases.

Temas	Tiempo
Logros	0,08horas
Señalamientos	0,06horas
Recomendaciones	0,05horas
Tiempo Total	**0,19horas**

Evaluación del impacto del tiempo requerido para evaluar

A partir de las estimaciones realizadas anteriormente se puede evaluar el impacto del tiempo requerido para realizar la evaluación del 10% de los ICC analizando solamente los señalamientos y 100% de los ICC con la herramienta.

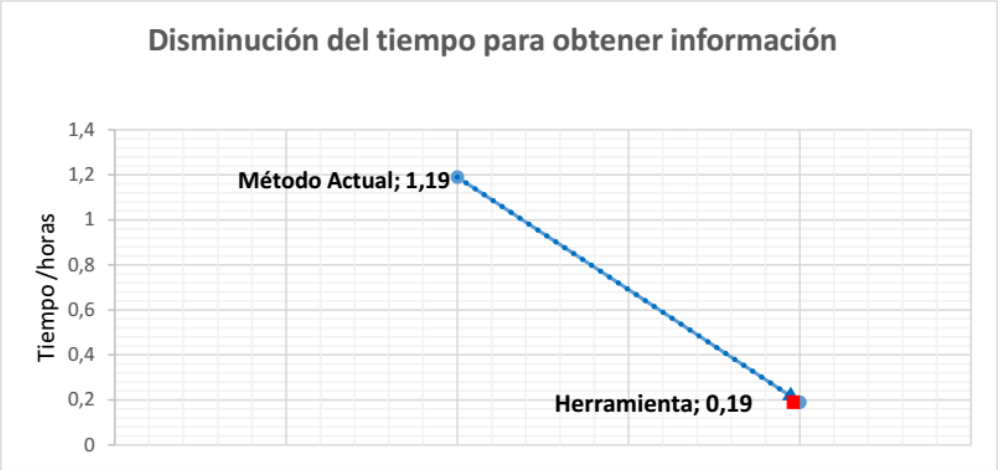


Figura 14: Análisis comparativo del tiempo.

Como se puede apreciar en la figura (Figura 14) anterior usando el método actual un directivo docente demora aproximadamente 1,19 horas para obtener la información dedicados a esta tarea. Sin embargo, usando la herramienta propuesta solo necesitan 0,19 horas para obtener esta información. Lo que indica que utilizando la herramienta Info_Controles a clases hay una reducción significativa del tiempo para obtener la información.

3.2.2- Caso de estudio para evaluar la calidad de la información

Con el objetivo de tener una medida de la calidad de la información que se puede obtener con el uso de la herramienta propuesta se diseñó un caso de estudio para medir la precisión de la herramienta al extraer la información de los datos.

En el caso de estudio intervendrán dos directivos docentes. Uno de ellos deberá extraer los temas más importantes de una muestra de 27 ICC seleccionados al azar de la población (90 ICC) que representa el 30%, a los cuales se les denomina “Temas relevantes. El otro directivo deberá extraer los temas más importantes usando la herramienta con toda la población de ICC, se les denomina “Temas recuperados”. Este análisis se realizará solo para el tema “Señalamientos”.

Precisión

Este concepto fue definido por Kent (39),como factor de pertinencia. Hay otros autores que se refieren a él, como ratio3 de aceptación. Para Salton (40)la precisión es la proporción de material recuperado realmente relevante, del total de los temas recuperados. A esta definición añade que el resultado de esta operación está entre 0 y 1. Así, la recuperación perfecta es en la que únicamente se recuperan los documentos relevantes y por lo tanto tiene un valor de 1.

Esta medida está relacionada con dos conceptos, el de ruido y el de silencio informativo. De este modo, cuanto más se acerque el valor de la precisión a 0, mayor será el número de temas recuperados que no le sirvan al usuario y por lo tanto el ruido que encontrará será mayor.

Acorde con la definición de Salton se tiene la siguiente expresión:

$$Precisión = \frac{Número\ de\ temas\ relevantes recuperados}{Número total\ de temas recuperados}$$

Aplicando la fórmula anterior a los datos, se obtienen los siguientes resultados:

$$Precisión = \frac{6}{7} = 0{,}85$$

Como se puede apreciar se obtuvo un 85% de precisión, valor considerado adecuado para investigaciones en ciencias sociales(42). Por lo tanto, podemos concluir que la calidad de la información extraída por la herramienta Info_Controles a Clases es adecuada.

[3] La palabra **ratio**, significa 'razón o cociente entre dos números'.

Conclusiones del capítulo

Como resultados de la evaluación del modelo y validación de la viabilidad del uso de la herramienta se obtuvo que:

- La calidad del modelo fue evaluada de aceptable al escoger un número de temas óptimo mediante el valor de coherencia.
- Es viable el uso de la herramienta Info_Controles a Clases por los directivos docente teniendo en cuenta el tiempo empleado en obtener la información necesaria.
- El modelo propuesto realizó un 85% de precisión lo que indica que es considerado como resultado aceptable.
- El modelo propuesto disminuye significativamente el tiempo requerido de los directivos a la hora de extraer temas del conjunto de ICC.

De esta forma se da cumplimiento al cuarto objetivo específico de la investigación relacionada con la valoración de la viabilidad de la herramienta con respecto al caso de estudio planteado.

CONCLUSIONES

Como resultado de la elaboración de la fundamentación teórica se concluye que:

- Existe la necesidad de desarrollar una herramienta para la extracción de información útil de informes de controles a clases, basado en el algoritmo de agrupamientos por temas: Modelo de Asignación Latente de Dirichlet. Utilizando las bibliotecas de Python para el trabajo con minería de texto y guiado por la metodología CRISP-DM.

Sobre la propuesta de solución:

- Se presenta un Método para el agrupamiento por temas en informes de controles a clases, mediante el cual se obtienen tres modelos de temas que representan los principales temas tratados en un conjunto de ICC.
- Se presenta una herramienta para la extracción de información útil en ICC que facilita el trabajo de los directivos docentes. La cual puede ser utilizada como herramienta de apoyo en la evaluación de la calidad del proceso docente educativo.

Sobre los resultados obtenidos:

- La calidad de los modelos fue evaluada de adecuado teniendo en cuenta la elección de la cantidad óptima de temas a partir del valor de coherencia.
- La herramienta propuesta mostró un 85% de precisión al ser utilizada para extraer información de un conjunto amplio de ICC. Además, disminuye significativamente el tiempo empleado en la extracción de información. Por lo que se considera viable su uso como herramienta de apoyo en el trabajo de los directivos docentes.

RECOMENDACIONES

Proponer una plantilla Látex para la redacción de los Informes de controles a clases.

Integrar la herramienta actual al Sistema de Gestión Universitaria (Akademos).

BIBLIOGRAFÍA

1. JUAN CARLOS GONZÁLEZ, LUISA MARÍA BAUTE ÁLVAREZ y RAÚL ALPIZAR FERNÁNDEZ. Una mirada a la gestión del trabajo metodológico de los jefes de departamento docente universitario. In: *Revista Científica de la Universidad de Cienfuegos*. 2014, Vol. 6, no. 4.

2. FIDEL CASTRO RUZ. Discurso pronunciado por Fidel Castro Ruz. In: *1981*. La Habana, 7 julio 1981. Discurso pronunciado por Fidel Castro Ruz, presidente de la República de Cuba, en el acto de graduacion de 10 658 egresados del destacamento pedagogico universitario «manuel ascunce domenech», en el poligono de ciudad libertad, el 7 de julio de 1981, «ano del xx aniversario de Giron».

3. ELENA SOBRINO PONTIGO y JOSÉ FRANCISCO ECHEMENDÍA GALLEGO. *Visión de la labor y algunos resultados concretos de la aplicación del trabajo metodológico*. 2017. S.l.: Cap. Silverio Blanco Núñez.

4. Gaceta oficial de la república de cuba ministerio de justicia. In: [online]. 21 junio 2018, Vol. No. 25. Available from: http://www.gacetaoficial.cu/.

5. RAÚL ALPIZAR FERNÁNDEZ. *Modelos de gestión para la formación y desarrollo de los directivos académicos de la Universidad de Cienfuegos*. Tesis Doctoral. La Habana: Universidad de La Habana, 2004.

6. Módulo de Gestión Académica. In: [online]. 2016. [Accessed 29 noviembre 2018]. Available from: https://www.siu.edu.ar/siu-guarani/.

7. *SIGA Principal Sistema Integrado de Gestión Académica. [En línea]. https://www.unisabana.edu.co/menu-superior-2/enlaces-rapidos/siga/informacion-siga* [online]. 2018. S.l.: s.n. Available from: https://www.unisabana.edu.co/menu-superior-2/enlaces-rapidos/siga/informacion-siga.

8. YANOSKI CALDERÍN DELGADO. *GESTACAD. Sistema para la gestión académica*. Facultad De Informática, CUBA: Universidad Matanzas “Camilo Cienfuegos”, 2015.

9. VANESSA DANAE MUÑOZ CASTILLO, HILDA GARCÍA BARRIOS, ORLANDO RUBIERA HERNÁNDEZ, CARLOS RAMÓN LÓPEZ PAZ y INGRID WILFORD RIVERA. SIGENU-DSS-LITE: Nuevas capacidades de integración de información docente en Instituciones de Educación Superior en Cuba. In: *revistaci@idict.cu* [online]. agosto 2015, Vol. 46, no. núm. 2. Available from: Disponible en: http://www.redalyc.org/articulo.oa?id=181441052003.

10. ANA MARIA SÁNCHEZ GONZÁLEZ. «Sistema de Gestión Universitaria,» [En línea]. Available: https://www.monografias.com/trabajos92/gestion-universitaria/gestion-universitaria.shtml. In: . 2012,

11. RONEN FELDMAN, MOSHE FRESKO, YAKKOV KINAR, YEHUDA LINDELL, ORLY LIPHSTAT, MARTIN RAJMAN, YONATAN SCHLER y OREN ZAMIR. *Text Mining at the Term Level*. S.l.: s.n., 1998. European Symposium on Principles of Data Mining and Knowledge DiscoveryPKDD 1998: Principles of Data Mining and Knowledge Discovery pp 65-73Online: 19 October 2006

12. AH-HWEE TAN. Text mining: Promises and chanllenges. In: [online]. 1999, Available from: http://textmining.krdl.org.sg. In South East Asia Regional Computer Confederation (SEARCC'99). Westin Stamford Hote

13. RONEN FELDMAN y JAMES SANGER. *The text mining, HANDBOOK* [online]. S.l.: s.n., 2007. ISBN 978-0-521-83657-9. Available from: www.cambridge.org/9780521836579. Book Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data Cambridge University Press New York, NY, USA

14. PAULA ANDREA BENAVIDES CAÑÓN y SANDRA RODRÍGUEZ CORREO. *Procesamiento del lenguaje natural en la recuperación de información*. mayo 2007. S.l.: s.n.

15. ANIL K JAIN, NARASIMBA MURTY y PTRICK J. FLYNN. Data clustering: a review. In: *Journal ACM Computing Surveys*. 1999, Volumen 31, Issue 3, Sept 1999, New YorkACM New York, NY, USA

16. DAVID M. BLEI. *Probabilistic Topic Models*. Communications of the ACM.: s.n., 2012.

17. FABRIZIO SEBASTIANI. Machine learning in automated text categorization. In: . 2002, pp. 1-47

18. FRANCA DEBOLE y FABRIZIO SEBASTIANI. Supervised term weighting for automated text categorization. In: . S.l.: s.n., 2003. SAC '03 Proceedings of the 2003 ACM symposium on Applied computing, pp. 784-788.

19. XUE DEJUN y SUN MAOSONG. Chinese text categorization based on the binary weighting model with non-binary smoothing. In: *European Conference on Information Retrieval ECIR 2003: Advances in Information Retrieval pp 408-419*. S.l.: s.n., 2003. ECIR'03 Proceedings of the 25th European conference on IR research, pp. 408-419.

20. SCOTT DEERWESTER, SUSAN T. DUMAIS, GEORGE W. FURNAS, THOMAS K LANDAUER y RICHARD HARSHMAN. Indexing by Latent semantic Analysis. In: *Journal of the American Society for Information Science, 41 (6), pp. 391-407*. 1990,

21. PETER .W FOLTZ. Using latent semantic indexing for information filtering. In: *COCS '90 Proceedings of the ACM SIGOIS and IEEE CS TC-OA conference on Office information systems,*. 1990, pp. 40- 47.

22. FRANK MCCAREY, MEL Ó CINNÉIDE y NICHOLAS KUSHMERICK. Recommending library methods: an evaluation of the vector space model (VSM) and latent semantic indexing (LSI). In: *ICSR'06 Proceedings of the 9th international conference on Reuse of Off-the-Shelf Components, pp. 217-230.* S.l.: s.n., 2006.

23. KARI TORKKOLA. Discriminative features for document classification. In: *16th International Conference on Pattern Recognition, 1, pp. 472-475.* 2002,

24. THOMAS HOFMANN. Probabilistic latent semantic indexing. In: . 2005, California, United States

25. THOMAS HOFMANN. Probabilistic latent semantic analysis. In: *UAI'99 Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence, pp. 289-296.* 1999,

26. BRANTS, T. Test Data Likelihood for PLSA Models. In: *Information Retrieva*. 2005, pp. 181- 196.

27. BUNTINE, W. Estimating Likelihoods for Topic Models. In: *ACML '09 Proceedings of the 1st Asian Conference on Machine Learning: Advances in Machine Learning,*. 2009, pp. 51- 64.

28. MARK GIROLAMI y ATA KABÁN. On an equivalence between PLSI and LDA. In: *SIGIR '03 Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval*. S.l.: s.n., 2003. pp. 433-434.

29. MASADA, T, MIYAHARA, S y KIYASU, S. Comparing LDA with pLSI as a dimensionality reduction method in document clustering. In: *LKR'08 Proceedings of the 3rd international conference on Large-scale knowledge resources: construction and application,*. 2008, pp. 13-26.

30. DAVID M. BLEI y MICHAEL I JORDAN. Variational inference for Dirichlet process mixtures. Bayesian Analysis,. In: . 2006, pp. 121- 144.

31. YUE LU, QIAOZHU MEI y CHENGXIANG ZHAI. Investigating task performance of probabilistic topic models: an empirical study of PLSA and LDA. Information Retrieval. In: . 2011, Vol. Vol 14, pp. 178-203.

32. JUAN MIGUEL MOINE. *Metodologías para el descubrimiento de conocimiento en bases de datos: un estudio comparativo.* S.l.: s.n., 2013.

33. DAVID L OLSON y DURSUN DELEN. Advanced Data Mining Techniques. Berlin, Germany : Springer publishing, 2008. In: . 2008,

34. *Librerías de Machine Learning con Python - Ligdi González.htm*. S.l.: s.n.

35. *Download PyCharm". JetBrains. 7 May 2019.* S.l.: s.n.

36. YEE WHYE TEH, DAVID M. BLEI, MICHAEL I JORDAN y MATTHEW J BEAL. Hierarchical Dirichlet Processes. In: *online: 1 enero 2012*. 2007, Vol. 101. Journal of the American Statistical Association, 101 (476), pp. 1566-1581.

37. LUCIANO HAMMOE. Detección de tópicos utilizando el Modelo LDA. In: . 2018, Buenos Aires

38. KAMAL KUMAR. Evaluation of topic modeling: Topic Coherence. In: *3 mayo de 2018*. 2018, Program Manager -BI Data Analytics / Process excellence

39. E.M. KEEN. Measures and Averaging Methods Used in Performance Testing Indexing System. In: . 1996, Crandfield, Eng.,Aslib Crandfield Project

40. GERARD SALTON. Evaluation parameters. In: . 1993,

41. GERARD SALTON y MICHAEL J. MCGILL. *Introduction to Modern Information Retrieval.* S.l.: s.n., 1986. ISBN 0-07-054484-0.

42. GRAU. La eficiencia en la graduación universitaria analizada con descubrimiento de conocimientos en la base de estudiantes de la universidad central de las villas. In: . 2012, Santa Clara. Cuba

ANEXOS

Anexo 1- Temas relevantes: Los temas extraídos de la muestra por el directivo 1 seleccionados al azar de la población (Señalamientos).

1. Orientar correctamente la bibliografía
2. Organización en la pizarra
3. Disciplina
4. Bajo aprovechamiento
5. Insuficiente lenguaje técnico.
6. Se pierde tiempo

Anexo 2- Temas recuperados: los temas extraídos por el directivo 2 usando la herramienta.

1. Orientar bibliografía para autoestudio
2. Lenguaje técnico
3. Organización pizarra
4. Insuficiencias en el tiempo
5. Disciplina
6. Bajo aprovechamiento
7. Análisis estudiantes