ESCUELA POLITÉCNICA NACIONAL

FACULTAD DE INGENIERÍA DE SISTEMAS

DESARROLLO DE UNA HERRAMIENTA DE EVALUACIÓN DE PROXIMIDAD DE INFORMACIÓN PARA VIGILANCIA ESTRATÉGICA

PROYECTO PREVIO A LA OBTENCIÓN DEL TÍTULO DE INGENIERO EN SISTEMAS INFORMÁTICOS Y DE COMPUTACIÓN

ALEXIS ADRIAN MIRANDA CARRILLO
alexis.miranda@epn.edu.ec

DIRECTOR: PhD EDISON LOZA AGUIRRE
edison.loza@epn.edu.ec

Quito, marzo 2021

DECLARACIÓN

Yo, Alexis Miranda, declaro bajo juramento que el trabajo aquí descrito es de mi autoría; que no ha sido previamente presentada para ningún grado o calificación profesional; y, que he consultado las referencias bibliográficas que se incluyen en este documento.

A través de la presente declaración cedo mis derechos de propiedad intelectual correspondientes a este trabajo, a la Escuela Politécnica Nacional, según lo establecido por la Ley de Propiedad Intelectual, por su Reglamento y por la normatividad institucional vigente.

Alexis Miranda

CERTIFICACIÓN

Certifico que el presente trabajo fue desarrollado por Alexis Miranda, bajo mi supervisión.

Edison Loza, PhD.

DIRECTOR DE PROYECTO

MSc. Carlos Montenegro,

CO DIRECTOR DE PROYECTO

ÍNDICE DE CONTENIDO

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

RESUMEN

Hoy en día, la sobrecarga de información lleva a los gerentes a no utilizar adecuadamente la información relevante que recopilan en procesos de Vigilancia Estratégica. Esta información se presenta usualmente en forma de pequeños fragmentos de texto que están dispersos en términos de tiempo, idioma y fuentes; y que deben ser enlazados para permitir un adecuado análisis. Dado que las características mencionadas generalmente impiden identificar conexiones en la información recopilada, en el presente proyecto se propone una herramienta para abordar este problema mediante el uso de técnicas de análisis de tópicos. La herramienta resultante proporciona una interfaz de lectura rápida, en la que se puede observar la relación de proximidad entre varios textos y se pueden visualizar fácilmente "constelaciones" de textos relacionados. Comparamos la herramienta desarrollado con dos mecanismos de agrupación y medición de proximidad existente, obteniendo resultados que muestran que nuestra herramienta destacó en términos de tiempo de ejecución y pertinencia de resultados.

PALABRAS CLAVE: Análisis de tópicos, teoría de grafos, proximidad de información, vigilancia estratégica.

ABSTRACT

Nowadays, information overload leads managers to not properly use the relevant information they collect in Strategic Scanning processes. This information is usually presented in the form of small pieces of text that are scattered in terms of time, language and sources; and that they must be linked to allow an adequate analysis. Since the characteristics mentioned generally prevent identifying connections in the information collected, in the present project we implemented a tool to address this problem through the use of topic analysis techniques. The resulting tool provides a quick-read interface, in which the proximity relationship between several texts can be observed and “constellations” of related texts can be easily visualized. We compare the developed tool with two existing proximity measurement and grouping mechanisms, obtaining results that show that our tool stood out in terms of execution time and relevance of results.

KEYWORDS: Topic analysis, graph theory, information proximity, strategic scanning.

1. INTRODUCCIÓN

1.1. Motivación del Estudio

Actualmente, la solidez y competitividad de las empresas se consolida por su capacidad de crear valor a través de la gestión de la información y el conocimiento [1]. Por ello, las organizaciones deben potenciar sus capacidades de adquirir, difundir y principalmente analizar información relacionada con su entorno socioeconómico [2]. Así, a través de las actividades de Vigilancia Estratégica (SScan), las organizaciones pueden recolectar y analizar información relevante que se encuentra en su entorno socioeconómico, y que, en general, puede esconder un alto valor estratégico [3]. La SScan se refiere entonces al proceso informacional que permite "la adquisición y el uso de información sobre eventos, tendencias y relaciones en el entorno externo de una organización, cuyo conocimiento ayudará a la dirección de una organización a planificar el curso de acción futuro de la misma" [4], [5].

Las actividades de SScan se vuelven menos efectivas cuando se enfrentan al problema de la sobrecarga de información, que surge del uso de tecnologías de la información cada vez más eficientes en las tareas de recolección de información [6]. Debido a la falta de herramientas apropiadas para lidiar con este problema de exceso, los gerentes no pueden usar la información relevante que recopilan para tomar decisiones estratégicas.

La sobrecarga de información conduce a una degradación de los procesos de toma de decisiones [7]. Varias investigaciones han demostrado que existe una cantidad óptima de información que se puede utilizar para tomar una decisión [8], [9]. Más allá de eso, la eficiencia del proceso de toma de decisiones disminuye tanto en términos de calidad (decisión racional en el contexto) como en términos del tiempo necesario para tomar la decisión (una decisión que llega demasiado tarde no es buena) [10].

En base a este problema de sobrecarga de información en procesos de SScan, en el presente trabajo se plantea el desarrollo de una herramienta que permita la evaluación de la proximidad de documentos mediante técnicas de minería de textos, específicamente, análisis de tópicos para relacionar la información recolectada mediante SScan. Para una fácil visualización de la relación de proximidad de información, se implementará además una representación gráfica mediante la aplicación de principios teoría de grafos, en la cual se podrá observar la relación de proximidad entre varios textos y construir "constelaciones" de textos relacionados.

Para el desarrollo de esta herramienta se seguirá una metodología de investigación basada en el enfoque de Ciencia del Diseño (Design Science, DS) [11], [12]. El principio fundamental de investigación de la DS es que tanto el conocimiento y la comprensión de un problema de diseño, así como la solución del problema en sí, se adquieren mediante la construcción de un artefacto [13], [14].

El desarrollo exitoso del artefacto en este proyecto permitirá abordar el desafío de la sobrecarga de información que afecta las actividades de SScan al reducir la cantidad de información que debe ser analizada por los gerentes. Lo que se traducirá en eficiencia a la hora de la toma de decisiones en base a la información recolectada por los dispositivos de SScan.

1.2. Objetivos

1.2.1. General

Desarrollar una herramienta de evaluación de proximidad de información para vigilancia estratégica utilizando técnicas de análisis de tópicos y teoría de grafos.

1.2.2. Específicos

- Diseñar una herramienta que permita visualizar la proximidad de información mediante análisis de tópicos y teoría de grafos.
- Implementar una herramienta que calcule la proximidad de información para SScan mediante análisis de tópicos.
- Implementar una herramienta que permita visualizar la proximidad de información de SScan mediante teoría de grafos.
- Evaluar la aceptabilidad de la herramienta propuesta en términos de utilidad percibida y facilidad de uso percibida mediante experimentaciones.

1.3. Alcance

En el presente proyecto de titulación se diseñará e implementará una herramienta que permitirá evaluar la proximidad de información orientada a soportar y mejorar el proceso de SScan. Además, permitirá visualizar la relación entre los documentos analizados.

La herramienta analizará: colecciones locales de documentos, los cuales estarán en un formato comprimido (.zip) o textos en un documento de Microsoft Excel (.xlsx), de los cuales se estimará la medida más próxima entre los documentos miembros de la colección. La herramienta no será un buscador de corpus.

1.4. Marco Teórico

En esta sección introduciremos los conceptos relacionados con cuatro temáticas. En primer lugar, la SScan y sus diferentes etapas serán presentadas. Luego, se presentará el modelo probabilístico Latent Dirichlet Assignment que constituye la base para el desarrollo de nuestra herramienta. A continuación, se presentarán la medida de proximidad de texto NM que ha sido desarrollada para solventar el problema de sobrecarga de información en SScan. Herramienta con la cual la presente solución será comparada. Enseguida, se presentarán criterios de Teoría de Grafos y, finalmente, el método de Louvain para clusterización de textos, el cual, si bien no ha sido diseñado para SScan, ofrece una alternativa de agrupamiento contra la cual nuestra herramienta también será comparada.

1.4.1. Vigilancia Estratégica

La SScan es el proceso informacional orientado a la adquisición y el uso de información acerca de eventos y tendencias del entorno externo de una organización, cuyo conocimiento puede ayudar a los administradores a planificar el proceder futuro de la organización [1].

Una organización no está exenta de los cambios y evoluciones que se producen en su entorno socioeconómico. Es por esto por lo que hoy en día, las organizaciones realizan, en mayor o menor medida, actividades de SScan con el fin de: mantenerse al día con las evoluciones y tendencias de su entorno [15], identificar amenazas y oportunidades [16], anticipar cambios y entender las fuerzas que los generan [17], y soportar los procesos de toma de decisión [18].

La efectividad de un dispositivo de SScan depende de la capacidad de las organizaciones para reconocer el valor anticipativo de una información, colectarla y analizarla pertinentemente de manera que permita la toma de una decisión estratégica oportuna [15].

Se pueden distinguir dos etapas principales en el proceso SScan [5]:

Adquisición de información

Esta etapa se divide en dos actividades:

- La primera implica la identificación adecuada de la extensión del entorno organizacional a monitorear en función de las necesidades de información de los miembros de la organización.

- La segunda actividad contempla la recolección de información pertinente del entorno empresarial, englobando todos los procesos que interactúan en él, como la transmisión, organización y filtrado de la información recogida.

Análisis de información

También se divide en dos actividades:

- La primera se encarga de realizar un pre-análisis con el propósito de reducir, optimizar y organizar la información recolectada para darle sentido, preparándola para la siguiente etapa.
- La segunda comprende es la construcción colectiva de significado, que consiste en reunir a los miembros de la organización que son capaces de comprender la información disponible desde un punto de vista organizacional y con un objetivo decisivo para abordar esta actividad, se utiliza el método Puzzle, que permite capturar el conocimiento de los usuarios de cada rama de la organización [19].

En este sentido, el presente trabajo pretende contribuir a la primera actividad de esta etapa, al aportar con una herramienta que facilite el pre-análisis de la información recolectado, permitiendo y potenciando la capacidad para identificar relaciones entre las piezas de información.

1.4.2. LDA

Latent Dirichlet Assignment (LDA) es un modelo probabilístico de aprendizaje no supervisado que permite modelar un corpus como una mezcla finita de tópicos [20] [21]. Esta técnica proporciona un procedimiento probabilístico mediante el cual se pueden generar documentos. Así, para generar un nuevo documento se elige una distribución de tópicos. Las palabras para generar el documento se eligen entonces de un tópico aleatorio siguiendo la distribución probabilística identificada [22].

LDA asume que cada documento contiene varios tópicos, y las palabras del documento se generan a partir de esos tópicos. Todos los documentos contienen un conjunto particular de tópicos, pero la proporción de cada tópico en cada documento es diferente [23]. La probabilidad de una secuencia de palabras no se ve afectada por el orden en que aparecen (concepto de Bolsa de palabras) [24].

En la representación gráfica de la Figura 1, las variables sombreadas y no sombreadas representan las variables observadas y latentes (es decir, no observadas) respectivamente. Los parámetros α y β son constantes del modelo. Las flechas indican las dependencias condicionales entre las variables, mientras que las "placas" (los

recuadros de la figura) se refieren a las repeticiones de los pasos de muestreo con las variables N, M representando al número de muestras. Por ejemplo, la placa interior de z y w ilustra el muestreo repetido de tópicos *z* y palabras *w* hasta que se hayan generado *N* palabras para un documento. La placa que rodea a *θ* ilustra el muestreo de una distribución de tópicos para cada documento d para un total de M documentos.

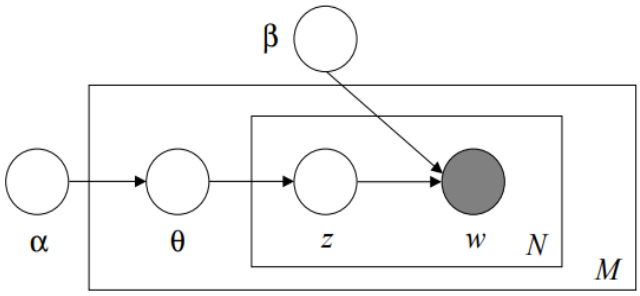


Figura. 1. Representación gráfica del modelo LDA [21]

1.4.3. NM Measure

Nearness Measure (NM) [25] es una medida que estima la proximidad entre textos en función del concepto de "información adyacente". Esta medida a sido desarrollada en trabajos previos con el objetivo de permitir la aproximación de información en VS [26].

En NM, la información adyacente se refiere “a dos elementos de información que se relacionan con el mismo tema, pero tienen pocas o ninguna palabra en común (a pesar de estar escritos en el mismo idioma)” [26]. En consecuencia, dos elementos de información en una base de datos relacionada con el mismo tema se denominan adyacentes si las listas de palabras lematizadas de cada texto manifiestan algunos o todos los criterios siguientes:

- Palabras en común: la misma palabra está presente en cada lista de palabras lematizadas para dos textos.
- Sinónimos (comparados mediante un thesaurus): una palabra de una lista lematizada tiene un sinónimo en la otra lista.
- Palabras concurrentes: la presencia de palabras de una lista lematizada que aparecen frecuentemente juntas en ambos textos.

Para dos textos, que denotamos como Ti y Tj, la NM entre ellos se establece de la siguiente manera:

- Para cada palabra en Ti, sumamos:
    - 0 si la palabra también está presente en Tj,
    - syn_value: si la palabra tiene un sinónimo en Tj,

- coo_value_min: buscamos la palabra en Tj con el coo_value más pequeño,

- Hacemos lo mismo con Tj,
- Finalmente, sumamos la primera suma (Ti comparado con Tj) a la segunda suma (Tj comparado con Ti) y dividimos por dos. Por lo tanto, sea: B una colección de texto, Ti y Tj dos textos lematizados sin stopwords, Mi una palabra de Ti, Mk una palabra de Tj y Mp una palabra en B.

$$NM(T_i, T_j) = \frac{\sum_{M_l \in T_i} m(M_l, T_j) \; + \; \sum_{M_k \in T_j} m(M_k, T_i)}{2}$$

$$m\big(M_l, T_j\big) = \begin{cases} 0 \; if \; M_l \in T_j \\ \frac{1}{2} \min_{M_p \in B} DCV\big(M_l, M_p\big) \; si \; T_j \; contiene \; un \; sinónimo \; de M_l \; (syn\_value) \\ \min_{M_k \in T_j} DCV(M_l, M_k) \; de \; lo \; contrario \; (coo\_value) \end{cases}$$

Para el cálculo de NM, es necesario recurrir a la Distancia de Cilibrasi y Vitanyi (DCV) [24]. DCV es una similitud semántica como “información mutua puntual” [27]. Se elige DCV porque su medida de co-ocurrencia evita los problemas de polisemia e implícitamente construye un universo semántico. DCV se define por:

$$DVC(M_l, M_k) = \frac{max(log(f(M_l)), log(f(M_k))) \; - \; log(f(M_l, M_k))}{log(n) \; - \; min(log(f(M_l), log(f(M_k)))}$$

Donde: f (x) es el número de textos que contienen x, f (x, y) es el número de textos que contienen x e y, n es el número de textos de la colección.

1.4.4. Teoría de grafos

Un grafo *G = (V, E)* es un conjunto *V* de vértices y un conjunto *E* de aristas, en el que una arista une a un par de vértices. Normalmente, los gráfos se representan con sus vértices como puntos en un plano y sus bordes como segmentos de líneas, o curvas, que conectan esos puntos [28]. Existen diferentes estilos de representación, adecuados para diferentes tipos de gráficos o diferentes propósitos de presentación. Para nuestra herramienta, nos concentramos en la clase más general de gráficos: gráficos no dirigidos, dibujados con bordes rectos.

Los grafos se componen de varios componentes relacionados llamados constelaciones. Cada constelación [25] consta de un círculo (que representa un texto) conectado a otros círculos por flechas (la Figura 2 presenta una constelación). Una serie de círculos y

flechas se convierte en una "rama" (o una ruta de avance, que se muestra en celeste en la Figura 2).

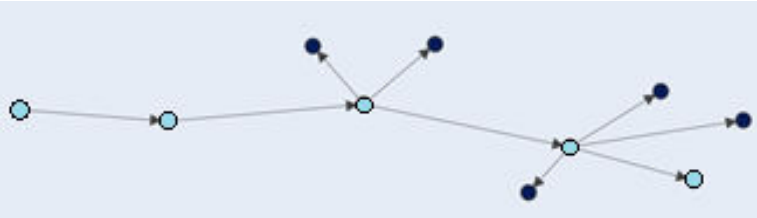


Figura 2. Ejemplo de una constelación.

1.4.5. Método de Louvain

La detección de comunidades es una operación ampliamente utilizada en el análisis de grafos. En este sentido el método de Louvain permite establecer comunidades o clústeres de documentos. Así, dado un grafo $G = (V, E)$, la de detección de comunidades en este método consiste en identificar una partición de vértices en comunidades de modo que los vértices relacionados se asignen a la misma comunidad y vértices dispares o no relacionados se dirijan a diferentes comunidades.

El problema de detección de comunidades consiste en identificar un conjunto de comunidades en un gráfico de entrada, donde las comunidades representan una partición de V. La bondad del clustering logrado por detección comunidades puede medirse mediante una métrica global denominada modularidad [29]. La modularidad de una partición es un valor escalar entre -1 y 1 que mide la densidad de enlaces dentro de comunidades en comparación con enlaces de las demás comunidades [30].

El método Louvain se divide en dos fases que se repiten iterativamente. Suponiendo que comenzamos con una red ponderada de N nodos. Primero, asignamos una comunidad diferente a cada nodo de la red. Entonces, en esta etapa inicial hay tantas comunidades como nodos. Luego, para cada nodo i consideramos los vecinos j de i y evaluamos la ganancia de modularidad que tendría lugar al eliminar i de su comunidad y colocarlo en la comunidad de j. Luego, el nodo i se coloca en la comunidad para la que esta ganancia sea máxima, pero solo si esta ganancia es positiva. Si no es posible una ganancia positiva, permanezco en su comunidad original.

Este proceso se aplica repetida y secuencialmente para todos los nodos hasta que no se puedan lograr más mejoras, completando así la primera fase. La segunda fase del algoritmo consiste en construir una nueva red cuyos nodos son ahora las comunidades encontradas durante la primera fase [31].

1.5. Herramientas de Desarrollo

En esta sección se presentan las diferentes herramientas y paquetes utilizadas en la presente investigación.

1.5.1. Pandas

Es un paquete de Python que proporciona estructuras de datos rápidas, flexibles y expresivas; diseñadas para que el trabajo con datos estructurados (tabulares, multidimensionales, potencialmente heterogéneos) y de series de tiempo sea fácil e intuitivo. Su objetivo es ser el bloque de construcción fundamental de alto nivel para realizar análisis de datos prácticos del mundo real en Python. Además, tiene el objetivo más amplio de convertirse en la herramienta de análisis y manipulación de datos de código abierto más potente y flexible disponible en cualquier idioma [32].

1.5.2. NLTK

Natural Language Toolkit (NLTK) es una biblioteca de Python de código abierto para el procesamiento del lenguaje natural. Proporciona interfaces fáciles de usar, recursos léxicos como WordNet, un conjunto de bibliotecas de procesamiento de texto para clasificación, tokenización, derivación, etiquetado, análisis y razonamiento semántico, wrappers para bibliotecas de procesamiento de lenguaje de nivel industrial, y un foro de discusión activo [33].

1.5.3. Spacy

Es una biblioteca para el procesamiento avanzado del lenguaje natural en Python y Cython. Se diseñó desde el primer día para su uso en productos reales. Spacy incluye modelos estadísticos previamente entrenados y vectores de palabras, y actualmente admite la tokenización para más de 60 idiomas. Cuenta con redes neuronales convolucionales de última generación para etiquetado, análisis y reconocimiento de entidades con nombre y una fácil integración de aprendizaje profundo [34].

1.5.4. Gensim

Su nombre proviene de la combinación de palabras Generate Similar. Gensim es una biblioteca de Python para modelado de tópicos, indexación de documentos y recuperación de similitudes con grandes corpus. El público objetivo de esta biblioteca es la comunidad de procesamiento del lenguaje natural y recuperación de información.

Algunas de las características más relevantes de esta librería son [35]:

- Todos los algoritmos son independientes de la memoria con respecto al tamaño del corpus (puede procesar una entrada más grande que la RAM, transmitida, fuera del núcleo).
- Provee interfaces intuitivas.
- Permite conectar fácilmente un propio corpus de entrada / flujo de datos (API de transmisión simple).
- Permite incorporar otros algoritmos de espacio vectorial (API de transformación simple).
- Permite implementaciones eficientes de múltiples núcleos de algoritmos populares, como el análisis semántico latente en línea (LSA/LSI/SVD), la asignación de Dirichlet latente (LDA), las proyecciones aleatorias (RP), el proceso de Dirichlet jerárquico (HDP) o el aprendizaje profundo de word2vec.
- Permite computación distribuida: puede ejecutar el análisis semántico latente y la asignación de Dirichlet latente en un grupo de computadoras.
- Ofrece una amplia documentación y tutoriales de Jupyter Notebook.

1.5.5. Networkx

NetworkX es un paquete de Python especializado en la creación, manipulación, estudio de la estructura, dinámica y funciones de redes complejas. Este paquete proporciona [36]:

- Herramientas para el estudio de la estructura y dinámica de redes sociales, biológicas y de infraestructura.
- Una interfaz de programación estándar e implementación de gráficos que es adecuada para muchas aplicaciones.
- Un entorno de desarrollo rápido para proyectos colaborativos y multidisciplinarios.
- La capacidad de trabajar sin problemas con grandes conjuntos de datos no estándar.

1.5.6. Dash

Dash es un marco productivo de Python para crear aplicaciones web. Escrito sobre Flask, Plotly.js y React.js, Dash es ideal para crear aplicaciones de visualización de datos con interfaces de usuario altamente personalizadas en Python puro. Es especialmente adecuado para cualquiera que trabaje con datos en Python.

A través de un par de patrones simples, Dash abstrae todas las tecnologías y protocolos necesarios para construir una aplicación interactiva basada en la web. Dash es lo suficientemente simple como para vincular una interfaz de usuario a su código Python.

Las aplicaciones de Dash se representan en el navegador web. Puede implementar sus aplicaciones en servidores y luego compartirlas a través de URL. Dado que las aplicaciones de Dash se ven en el navegador web, Dash es intrínsecamente multiplataforma y fácilmente utilizable en dispositivos móviles [37].

2. METODOLOGÍA

2.1. Ciencia del Diseño

Para el desarrollo de este trabajo se siguió la metodología de investigación basada en el enfoque de la Ciencia del Diseño (Design Science, DS) [11], [12]. El principio fundamental de DS es que tanto el conocimiento como la comprensión de un problema de diseño, así como la solución al problema en sí, se adquieren mediante la construcción de un artefacto [38]. El proceso para la implementación de la herramienta mediante DS consiste en:

- **Diseño y desarrollo:** La herramienta fue diseñada para permitir una fácil identificación de las "constelaciones" (gráficos) con sus documentos (nodos) y sus relaciones con los documentos más cercanos (bordes). Esto permitirá enfocar la atención de los gerentes a la información relevante lo que ayudará a confrontar el problema de la sobrecarga de información.

  La herramienta se desarrolló mediante un proceso iterativo de creación de prototipos funcionales. El artefacto se implementó mediante una arquitectura de dos niveles que utiliza un navegador web para visualizar los gráficos con sus respectivas relaciones y leer el documento seleccionado. El nivel lógico de la herramienta se desarrolló en Python con sus respectivos paquetes de soporte y el nivel de visualización utilizando Dash Open Source para facilitar la interacción con los resultados (gráficos) presentados a los usuarios finales.
- **Demostración:** La herramienta fue probada y mejorada gracias a las intervenciones con el director del proyecto. En cada una de las intervenciones se identificaron mejoras e implementación de características. Las intervenciones se llevaron a cabo hasta un punto de saturación en el que se validó el sistema como útil para abordar el problema de la sobrecarga de información.
- **Evaluación:** La herramienta fue evaluada por un grupo de expertos utilizando el Modelo de Aceptación de Tecnología (TAM) [39], en el cual se establecen criterios para comprender la intención de uso de los usuarios hacia una tecnología o herramienta. Esta intención está influenciada por una actitud individual que tiene dos determinantes: (1) la utilidad percibida, definida como la probabilidad subjetiva de una persona que, al utilizar un determinado sistema, mejore su desempeño laboral; y (2) la facilidad de uso percibida, que se refiere al grado en el que una persona cree que utilizar un determinado sistema no supondrá ningún esfuerzo [40].

También se formó un corpus de la base de datos de Scopus constituido sobre diferentes áreas de la Informática como se puede ver en la Tabla 2, con el cual se realizaron pruebas de desempeño de los algoritmos utilizados por la herramienta: LDA, Método Louvain y NM. Además, se realizó una comparativa entre dichos algoritmos como se puede observar en la Tabla 3.

2.2. Metodología de Evaluación

Como se mencionó en la sección anterior, para analizar el uso y comportamiento de los usuarios de la herramienta, se utilizó el Modelo TAM. El objetivo de TAM es proporcionar una explicación de los determinantes de la aceptación de la computadora que sea, en general, capaz de explicar el comportamiento del usuario en una amplia gama de tecnologías informáticas de usuario final y poblaciones de usuarios, al mismo tiempo que sea parsimoniosa y teóricamente justificada.

Idealmente, a un desarrollador le gustaría un modelo que sea útil no solo para la predicción, sino también para la explicación de la aceptación de una tecnología, de modo que los investigadores y los profesionales puedan identificar por qué un sistema en particular puede ser inaceptable, y seguir los pasos correctivos apropiados. Un propósito clave de TAM, por lo tanto, es proporcionar una base para rastrear el impacto de factores externos en las creencias, actitudes e intenciones internas.

La evaluación de una tecnología a través de TAM se enfoca en analizar las percepciones que un grupo de utilizadores tienen en torno a la utilidad y la facilidad de uso de esta. La utilidad percibida se define como la probabilidad subjetiva de los posibles usuarios de que el uso de un sistema de aplicación específico aumente el desempeño laboral dentro de un contexto organizacional. La facilidad de uso percibida se refiere al grado en que el posible usuario espera que el sistema de destino esté libre de esfuerzo [41].

Nuestra herramienta fue evaluada por un grupo de expertos utilizando TAM. El objetivo fue comprender la intención de uso de los usuarios hacia la herramienta. Con cada experto se procedió a entablar una reunión para conocer sus observaciones en cuanto a facilidad de uso y utilidad percibida por la herramienta, para conocer los comentarios de los expertos se desarrolló el siguiente cuestionario:

Tabla 1. Cuestionario para los evaluadores.

Evaluación

1. Analizando los clústeres de cada metodología, usted ha podido identificar clúster que correspondan con las siguientes temáticas (seleccione sólo aquellas que correspondan):

		Método		
Categoría	Sub categoría	LDA	Clustering	NM
Sistemas de Información	Sistemas de recomendación			
	Vigilancia Estratégica (SScan)			
Ingeniería de Software	Microservicios			
	Cloud Computing			
Seguridad de Información	Blockchain			
	Internet de las Cosas			
	Denegación de Servicios			
Sistemas Inteligentes	Reconocimiento Facial			
	Análisis de ondas cerebrales			
Interacción persona-computadora	Juegos Serios			
	Sobrecarga de información en pantalla			

2. ¿Considera que la herramienta utilizada es fácil de usar?

3. Tomando en cuenta el objetivo de permitir explorar con mayor eficiencia un corpus documental. ¿Considera usted que esta herramienta es útil?

4. ¿Tiene alguna recomendación de mejora?

3. RESULTADOS Y DISCUSIÓN

3.1. Desarrollo de la Herramienta

La herramienta se implementó siguiendo el método de creación de prototipos [43]. El método incluye un proceso iterativo de prototipado y evaluación que se repite hasta que la herramienta cumple con los objetivos para los que fue diseñada.

Las intervenciones se realizaron hasta un punto de saturación en el que se validó el sistema como útil para abordar la sobrecarga de información. En total, se requirieron cuatro iteraciones para lograr este estado, como se muestra en la Figura 3.

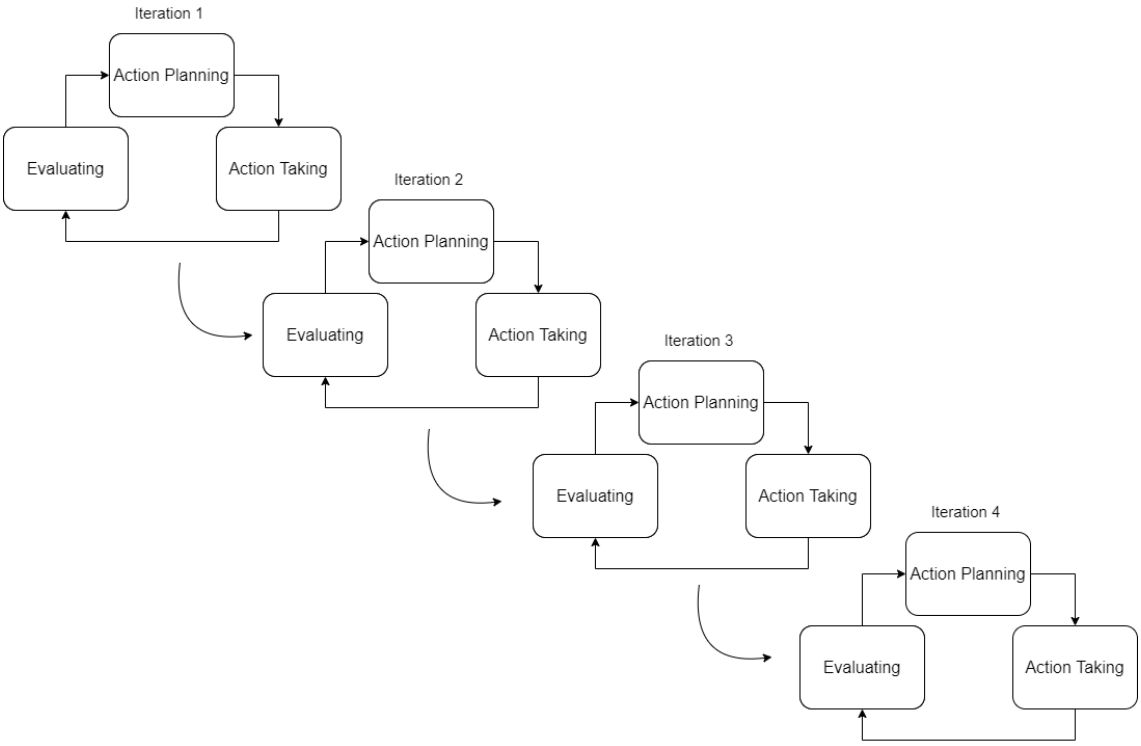


Figura 3. Desarrollo de la Herramienta.

Así, la herramienta se desarrolló siguiendo los pasos que se detallan a continuación:

1. Primera iteración

a) Diagnostico

El objetivo de esta etapa fue identificar el proceso para gestionar documentos miembros de un corpus. Como resultado de esta etapa, se concluyó que existía la necesidad de pre procesar los documentos para evitar trabajar con datos irrelevantes.

b) Planificación de acciones

Para solucionar las dificultades encontradas en la etapa de diagnóstico, se estableció que el corpus se almacene de manera local. Además, cada documento se dividió en tokens. Los tokens que sean de un único carácter son descartados, también aquellos tokens que contengan todos sus caracteres numéricos. A continuación, todos los tokens son lematizados. También, se descartaron las palabras comunes (Stopwords). Finalmente se agregan bigrams. Un bigram es un conjunto de dos palabras adyacentes que tienen significado.

c) Toma de acciones

En esta etapa se realizó la implementación del procesamiento de los documentos para lo cual se utilizó Python. La herramienta puede realizar la lectura de un corpus en formato MS Excel (.xlsx) y documentos de textos (.txt) agrupados en un archivo (.tar).

d) Evaluación

Se obtuvieron buenos resultados en cuanto al análisis de documentos. Para el próximo prototipo se plantea codificar el corpus.

2. Segunda iteración

a) Diagnostico

El objetivo en esta etapa fue identificar métodos para codificar el corpus. Para su posterior análisis y poder realizar operaciones de consulta de similitud entre documentos.

b) Planificación de acciones

Para solventar los requerimientos, se planteó codificar el corpus con los siguientes métodos: LDA, Método Louvain y NM Measure. A continuación, de la aplicación de los algoritmos se procedió a calcular el documento más próximo de todos los elementos del corpus y generar una matriz de similitud.

c) Toma de acciones

En esta etapa se implementó LDA, mediante el paquete Gensim, también el método Louvain y NM Measure con la ayuda del paquete Spacy. Estos métodos fueron implementados en Python. A continuación, cada método calcula su medida de proximidad entre todos los textos. Después, usamos estas medidas para poblar una matriz cuadrada del tamaño del número de textos que la denominamos matriz de similitud.

d) Evaluación

Se generó las diferentes matrices de similitud de todos los métodos sin ningún problema. Para el siguiente prototipo se planteó la representación gráfica de los métodos implementados.

3. Tercera iteración

a) Diagnostico

El objetivo en esta etapa fue implementar la representación gráfica de los algoritmos implementados en la etapa previa. También, poder visualizar el contenido del documento seleccionado por el usuario.

b) Planificación de acciones

Para abordar la representación gráfica de los algoritmos, se utilizó las matrices de similitud, en las cuales se representa el documento origen, su documento más próximo y su distancia. Además, se implementó la presentación al usuario de todos documentos analizados mediante técnicas de teoría de grafos, siendo así, cada documento un nodo y su representar su relación más próxima a otro documento, un borde dirigido. Cada nodo estará representado por su nombre o su ID en el corpus.

c) Toma de acciones

En esta etapa se implementó la representación gráfica de cada algoritmo, mediante el paquete Networkx, y su visualización mediante el paquete Matplotlib soportado por el lenguaje de programación Python.

d) Evaluación

Se obtuvo observaciones con los gráficos presentados por Matplotlib, respecto a la carencia de interacción con el usuario al momento de no poder observar el contenido de cada documento al seleccionarlo. Por lo cual, se optó por buscar alternativas para la representación gráfica y poder interactuar con los gráficos.

4. Cuarta iteración

a) Diagnostico

El objetivo en esta etapa fue mejorar la representación gráfica de los algoritmos y visualizar un documento seleccionado.

b) Planificación de acciones

Para la representación de nuevos gráficos se optó por cambiar a un paquete de visualización mediante tecnologías web.

c) Toma de acciones

En esta etapa se implementó la herramienta en dos niveles: nivel lógico basado en Python y el nivel de presentación mediante React con soporte del paquete Dash.

d) Evaluación

Con la actual arquitectura, el usuario puede interactuar con los grafos y ver su contenido. Los resultados se visualizan en un navegador web. Se tomaron en cuentas todas las observaciones presentadas a la herramienta. Como parte de la evaluación del sistema, se incluyó una discusión sobre su aceptación al final de cada intervención. Las preguntas se establecieron en base al modelo TAM respecto a facilidad de uso y utilidad brindada por la herramienta.

3.2. Descripción de la Herramienta

En esta sección presentaremos la operación de la herramienta tal como se muestra en la Figura 4. La etapa inicial de la herramienta de evaluación de proximidad recae en la lectura de los documentos que son miembros del corpus. Estos pueden cargarse desde un archivo de Microsoft Excel (.xlsx) o un archivo comprimido (.tar) con cada documento en formato de texto (.txt). La herramienta utiliza el paquete pandas para la lectura de los documentos [32]. Después de cargar el corpus, se obtienen dos listas: la primera lista contiene los documentos y la segunda contiene los nombres de los archivos de texto (.txt) o índices correspondientes del archivo MS Excel.

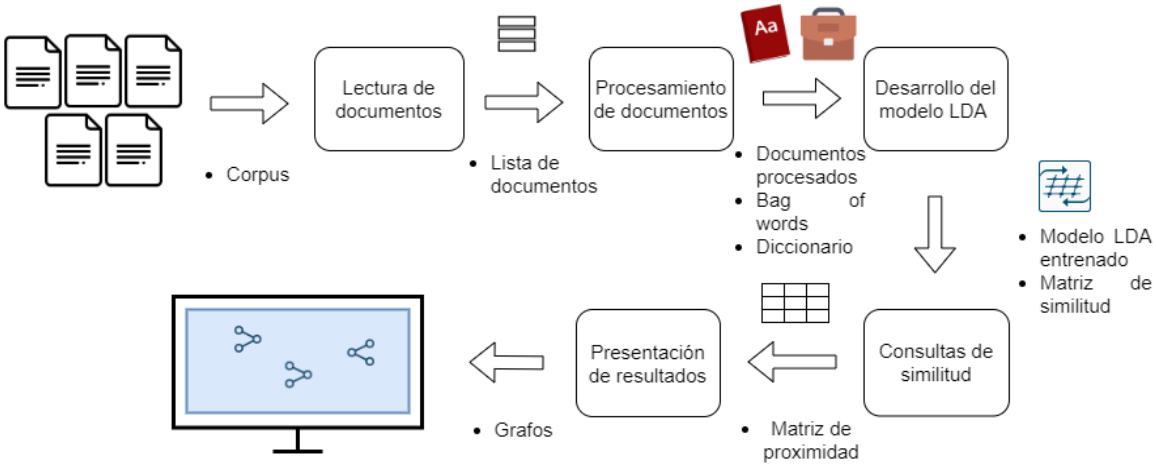


Figura 4. Proceso de implementación de la herramienta.

A continuación, se procesa la lista de documentos de la etapa anterior mediante las siguientes actividades:

- Dividir los documentos en tokens. Tokenizar consiste en dividir una secuencia de caracteres (strings) en frases, palabras, letras u otros elementos llamados tokens. Para analizar los documentos se procedió a separarlos en palabras.
- Eliminar números, pero no palabras con números.
- Eliminar palabras de un solo carácter.
- Lematizar palabras, este proceso consiste en extraer la raíz de una determinada palabra para reducir las variaciones de la palabra [42].
- Eliminar palabras útiles (Stopwords).
- Agregar bigrams, que son conjuntos de dos palabras adyacentes que tienen significado y deben estar presentes cinco o más veces en el corpus.

Esta etapa hace uso de los paquetes: NTLK [33] y Spacy [34]. A continuación, implementamos el diccionario de los documentos procesados del corpus, que contiene todas las palabras del corpus. Posterior, creamos la representación "Bolsa de palabras" de los documentos, como se puede observar en el Anexo III. Dicha representación contiene la frecuencia de palabras en el corpus. Al final de esta etapa, se obtiene una lista con los documentos procesados, la representación "Bolsa de palabras" y el diccionario del corpus.

En la siguiente etapa creamos el modelo LDA, para lo cual definimos los parámetros de entrenamiento:

- Corpus: Los documentos procesados.
- Id2word: El diccionario de corpus generado previamente.
- Chunksize: controla cuántos documentos se procesan al mismo tiempo en el algoritmo de entrenamiento. Aumentar el tamaño de la porción acelerará el entrenamiento, si el tamaño del trozo cabe fácilmente en la memoria.
- Num_topics: el número de tópicos en nuestro modelo. En nuestro caso seleccionamos 9 tópicos, ya que con este valor la coherencia del modelo es mayor al compararlo con otros valores, como la muestra la Figura 5.
- Pases: controla la frecuencia con la que entrenamos el modelo en todo el corpus.

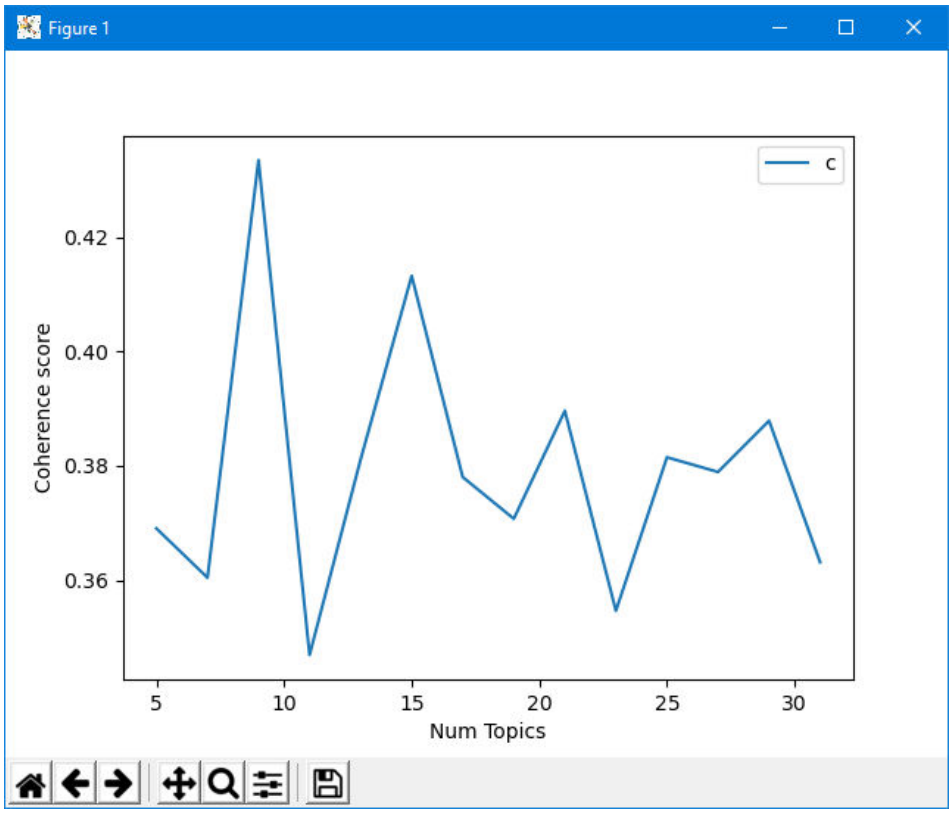


Figura. 5. Coherencia del modelo.

Esta etapa hace uso del paquete Gensim. Al final de esta etapa, se obtiene el Modelo LDA entrenado, que es una representación vectorial de n dimensiones (n depende del número de tópicos). Usando el modelo LDA entrenado, podemos realizar consultas de similitud entre un determinado documento y el corpus. Los documentos del modelo se ven como distribuciones de un conjunto de tópicos y cada palabra en un documento se genera en base a una distribución de palabras que es específica de cada tópico. Para comparar la similitud entre documentos trabajamos con sus correspondientes vectores que representan las contribuciones de cada palabra a un tópico.

Calculamos la similitud coseno entre todos los documentos. Usamos estas medidas para poblar una matriz cuadrada del tamaño del número de textos denominada matriz de similitud. Para cada fila de la matriz, buscamos la máxima similitud de coseno, es decir, para cada texto Ti buscamos el texto vecino más cercano que sea diferente de Ti (por definición, la similitud de coseno entre Ti y Ti es 1). La similitud coseno devuelve valores en el rango de [-1, 1] a un valor superior más similar (Figura. 6).

Realizamos la consulta entre el documento (representado en el espacio vectorial del modelo) y el corpus (matriz de similitud) utilizando la semejanza del coseno. Este proceso se realiza con todos los documentos del corpus. Los resultados se almacenan en una matriz denominada matriz de proximidad la cual contiene el identificador del

documento consultado, el identificador del documento más cercano al mismo y la distancia entre estos dos. En este proceso también se utiliza el paquete Gensim [35].

A continuación, la herramienta genera grafos para la representación visual de los documentos y sus relaciones. Cada grafo está formado por nodos, bordes y la posición de estos elementos. Los bordes se obtienen de la Matriz de Proximidad: [Identificador del documento de origen, Identificador del documento de destino, Distancia]. Los nodos están representados por los nombres o identificadores de los documentos. La posición de los elementos se calcula mediante el algoritmo Fruchterman-Reingold.

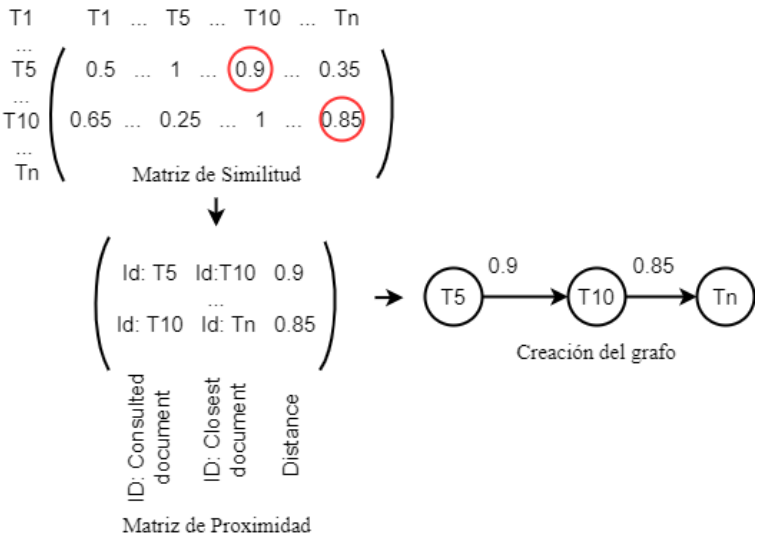


Figura 6. Creación de grafo.

Finalmente, para la visualización de resultados, la herramienta presenta cuatro interfaces:

1) Representación de proximidad entre documentos mediante LDA (Figura 7). En esta interfaz podemos visualizar las constelaciones con sus relaciones de proximidad y a su lado derecho una barra vertical que indica el número de conexiones de cada nodo (documento).

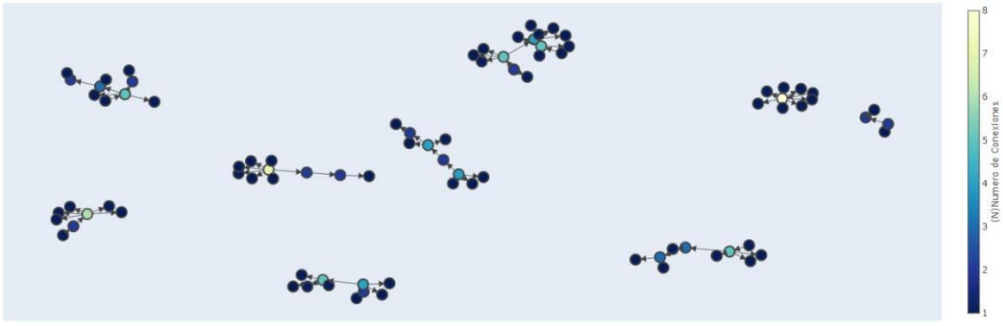


Figura 7. Visualización de la interfaz de proximidad de información mediante LDA.

En la parte derecha de la vista se puede observar una barra vertical, en el cual muestra el número conexiones de los nodos (textos).

2) Representación de proximidad entre documentos mediante método Louvain (Figura 8). Desarrollado para comparación, la interfaz permite visualizar las constelaciones con sus relaciones de proximidad y en su lado derecho una barra vertical que muestra el índice de un determinado Cluster.

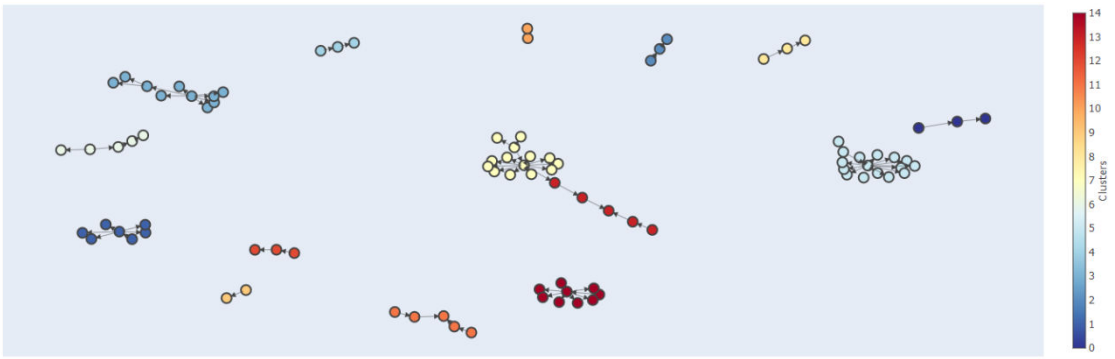


Figura 8. Visualización de la interfaz de proximidad de la información mediante método Louvain.

3) Representación de proximidad entre documentos mediante NM Nearness Measure (Figura 9). Desarrollado también con fines comparativos, la interfaz muestra las constelaciones con sus relaciones de proximidad y a su lado derecho una barra vertical que indica el número de conexiones de cada nodo (documento).

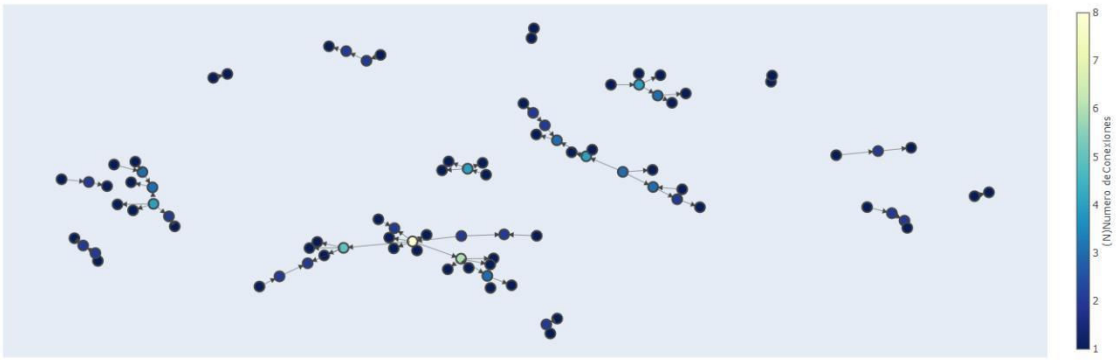


Figura 9. Visualización de la interfaz de proximidad de información a través de NM.

4) Visualización del documento. En las interfaces anteriores el usuario puede interactuar con los gráficos y ver el contenido de un documento que seleccione en esta interfaz.

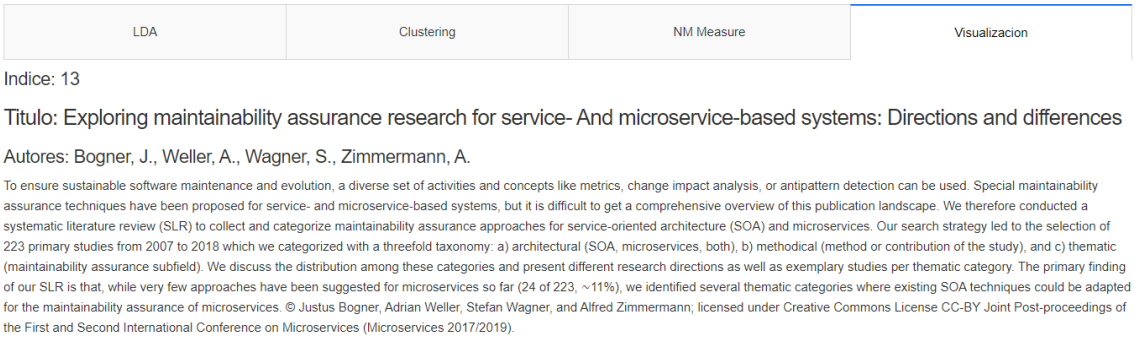


Figura 10. Visualización del documento seleccionado.

Para la visualización de gráficos la herramienta hace uso del paquete Dash. Para ver los resultados, la herramienta utiliza un navegador (Chrome o Firefox).

3.3. Pruebas funcionales

Para evaluar la herramienta, la etapa inicial fue la creación de un corpus de prueba. Los documentos se obtuvieron de la base de datos Scopus de diferentes áreas:

Tabla 2. Elementos del corpus.

Categoría	Subcategoría
Sistemas de información	- Sistemas de recomendación - SScan
Ingeniería de software	- Microservicios - Computación en la nube
Seguridad de la información	- Blockchain - Internet de las Cosas
Sistemas inteligentes	- Reconocimiento Facial - Análisis de ondas cerebrales
Interacción humano - computador	- Juegos Serios

De cada subcategoría se obtuvieron diez documentos con un total de noventa documentos. El corpus construido corresponde al área de conocimientos de los expertos que evaluarán la herramienta posteriormente. A continuación, ejecutamos la herramienta con el corpus generado anteriormente obteniendo los siguientes resultados:

Tabla 3. Resultados obtenidos de la herramienta.

Método	Tiempo de Ejecución	Número de Constelaciones
LDA	0: 00: 05.743872	10
Método Louvain	0: 00: 06.095323	14
NM	7: 05: 51.688682	15

Las pruebas se realizaron en la computadora con las siguientes características:

Tabla 4. Características de la computadora para las pruebas de desempeño.

Dispositivo	Descripción
CPU	Intel Core i7-7700 3.60 GHz
RAM	16 GB
Arquitectura	x64
Disco duro	SSD 120 GB
GPU	GeForce GTX 1060 6GB

Posterior, la herramienta fue probada mediante reuniones con diez profesores investigadores miembros de la Escuela Politécnica Nacional de la Facultad de Sistemas pertenecientes al Departamento de Informática y Ciencias de la Computación.

Durante cada prueba, los profesores llenaron una matriz en la cual marcaron una subcategoría de la Tabla 2 cuando identificaron que una determinada constelación correspondía a uno de los temas propuestos. Para fines de control, se agregaron dos subcategorías adicionales: Denegación de servicios en Seguridad de la información y sobrecarga de información en pantalla en Interacción persona-computadora. Eso significa que se añadieron estas subcategorías como se muestra en la Tabla 5, pero no se adicionaron artículos relacionados con estos temas al corpus. Al final de la prueba, se solicitó también a cada evaluador que respondieran preguntas sobre su percepción de utilidad, facilidad de uso y recomendaciones de mejora. Los resultados de las constelaciones identificadas durante las pruebas se resumen en la Tabla 5.

Al finalizar todas las reuniones, se procedió a contabilizar los resultados expuestos por todos los profesores.

Tabla 5. Resultados de la detención de constelaciones.

		Número de veces identificado		
Categoría	**Subcategoría**	**LDA**	**Lovian**	**NM**
Sistemas de información	Sistemas de recomendación	10	9	6
	SScan	5	4	1
Ingeniería de software	Microservicios	10	8	6
	Computación en la nube	10	10	6
Seguridad de la información	Blockchain	10	6	6
	Internet de las Cosas	8	5	6
	Denegación de servicios*	1	1	1
Sistemas inteligentes	Reconocimiento Facial	10	7	8
	Análisis de ondas cerebrales	10	6	6
Interacción humano – computador	Juegos serios	8	7	6
	Sobrecarga de información en pantalla*	0	0	0
	Total	9	6,72	5,90

*Denota una subcategoría de control

En la Tabla 5, se muestra el promedio del número de veces que los docentes identificaron constelaciones pertenecientes a las diferentes subcategorías, omitiendo las subcategorías de control: ‘Denegación de servicios’ y ‘Sobrecarga de información en pantalla’. Los resultados obtenidos señalan que mediante LDA los profesores identificaron constelaciones relacionadas con las subcategorías con mayor facilidad en comparación con los otros algoritmos (Método Louvain y NM Measure).

Respecto a las subcategorías de control se obtuvo que un evaluador seleccionó una vez la subcategoría: Denegación de servicios. Sin embargo, casi en su totalidad de evaluaciones las subcategorías de control no fueron seleccionadas.

Para evaluar la aceptación de la herramienta, se procedió a clasificar las observaciones sobre la facilidad de uso en base a dos categorías: críticas positivas y críticas negativas. A su vez, para determinar la utilidad de la herramienta las observaciones de los docentes se clasificaron también en las mismas categorías (críticas positivas y negativas).

A continuación, se presentan las observaciones brindadas por los docentes referentes a la facilidad de uso de la herramienta:

Tabla 6. Observaciones respecto a la facilidad de uso.

Facilidad de uso	
Positivas	Negativas
- Una vez se conoce la lógica o razón de ser de la herramienta, es fácil de usar. - Con una breve explicación en el primer uso, fue fácil entender la funcionalidad de esta. - Sí, debido a que cuenta con el manual de usuario, cada pestaña posee la etiqueta y la información que se muestra del resumen del paper identifica la categoría a la que pertenece. - Es fácil de usar, previo a la demostración. - Con relación a su facilidad de uso, en una escala de 1 al 10, en donde 1 es muy fácil y 10 muy complicada, la herramienta se la puede determinar en un 3. - Si, presenta una interfaz clara para interactuar con esta. - Si, fácil. - La herramienta es lo suficientemente intuitiva de usar para un usuario no entrenado, aunque pudiera presentar ciertas dificultades hasta que el usuario se acople. - Considero la herramienta es fácil de utilizar, las interfaces son simples y la navegación no presenta complejidad en su acceso.	- Un poco confusa en el aspecto que las constelaciones están desordenadas lo que en ocasiones hace que se visite la misma más de una vez. - Parcialmente, requiere supervisión inicial y puede mejorar aumentando el tamaño del nodo que contenga más relaciones. - Requiere desarrollar la interfaz visual para entender de mejor manera los datos presentados.

También, la Tabla 7 muestra las observaciones de los docentes respeto al uso percibido o brindado por la herramienta para explorar un corpus:

Tabla 7. Observaciones respecto a la utilidad.

Uso percibido	
Positivos	Negativos
- La herramienta es útil dado que permite conocer los artículos relacionados entre sí y sobre todo coloca en el centro el paper que podría ser más útil. - Si me pareció útil, tomando en cuenta que tiene tres algoritmos distintos, los cuales producen distintos resultados, ampliando las posibilidades de encontrar trabajos relevantes. -Sí, es muy útil por ejemplo cuando se trabaja con varios documentos para elaborar una tesis, o una investigación se requiere de la inversión de mucho tiempo en la lectura individual de cada peper, sin embargo, con el uso de la herramienta se puede simplificar la lectura de papers y leer uno que contenga en resumen a otros papers. -Es una herramienta de una utilidad más que aceptable, debido a que se puede ahorrar tiempo en la revisión de artículos, debido a que los relaciona entre sí, asignando grados de relación. -La herramienta es útil, pero no podría decir si permite explorar con mayor eficiencia un corpus. Permite explorar un corpus, pero se tendría que especificar a qué se hace referencia con el término “eficiencia”. - Sí, para propósitos de investigación puede ser muy útil una vez se explore en qué casos debo escoger NM, Clustering o LDA conforme mi estrategia de investigación o necesidad. -Si, se puede explorar fácilmente un corpus. - Es útil, ya que optimizaría el tiempo de búsqueda de temas relacionados, permitiendo encontrar los documentos claves. - La herramienta es bastante útil ya que permite encontrar coincidencias en un corpus de documentos basados en sus temas y su contenido. - Considero que la herramienta es útil.	

Posterior a su clasificación en observaciones positivas y negativas se procedió a contabilizar los resultados (Figura 11). Estos resultados sugieren que los docentes consideran que la herramienta es útil para explorar con mayor eficiencia un corpus documental

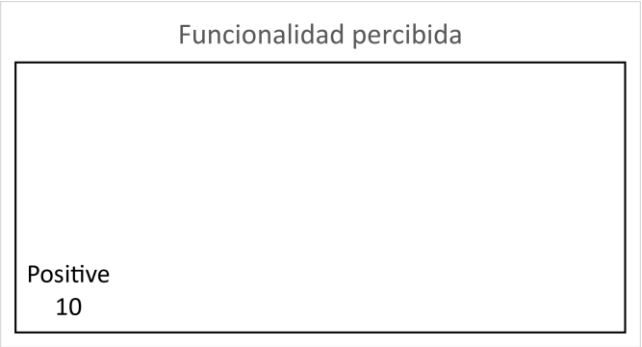


Figura 11. Funcionalidad percibida de la herramienta.

Respecto a la facilidad de uso (Figura 12), se obtuvo una gran mayoría de críticas positivas en comparación a pocas críticas negativas. Lo cual sugiere que los docentes pudieron manejar la herramienta sin ninguna complicación y sin requerir mucho esfuerzo.

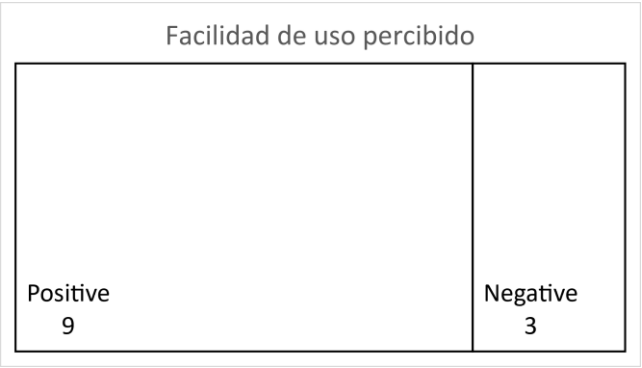


Figura 12. Facilidad de uso percibida de la herramienta.

También, los profesores agregaron algunas recomendaciones a la herramienta (Tabla 8), las cuales se detallan a continuación. Estas recomendaciones serán implementadas en versiones futuras de la herramienta.

Tabla 8. Recomendaciones respecto a la herramienta

Recomendaciones
- Para una mayor diferenciación de los documentos los colores que indican la cantidad de relaciones deberían contrastarse más, o usar otro elemento diferenciador. - Los nodos deberían ser de mayor tamaño para que los usuarios los visualicen mejor y tomando en cuenta que la gerencia suele usar laptops pequeñas o tablets. - Los clusters visitados deberían cambiar de tonalidad para indicar al usuario que ya fueron visitados y revisados.

4. CONCLUSIONES Y RECOMENDACIONES

4.1. Conclusiones

- La sobrecarga de información es un problema que afecta la efectividad de SScan. En este proyecto de titulación, proponemos una herramienta para abordar este problema. La solución propone un mecanismo de medición para estimar la proximidad entre varias piezas de información textual, utilizando técnicas de análisis de tópicos (LDA), y permite la creación de grupos de documentos relacionados representados mediante grafos.
- Comparamos nuestra herramienta basada en LDA con el método Louvain y NM Measure, que son opciones alternativas disponibles para relacionar textos. En términos de tiempo de ejecución, los métodos basados en LDA y Louvain mostraron, con mucho, un tiempo de convergencia bajo (alrededor de cinco a seis segundos). El número de constelaciones creadas por LDA fue más cercano (10) a los grupos de documentos incluidos en el corpus (9), a diferencia de los otros métodos que propusieron más de un 50% de constelaciones.
- Durante las pruebas con evaluadores, nuestro algoritmo basado en LDA obtuvo un mejor desempeño al momento de permitir identificar constelaciones de documentos similares.
- Tanto NM como LDA aboradar el problema de sobrecarga de información en SScan. Sin embargo, la interpretación de sus resultados es diferente. NM tiene un mecanismo para identificar el núcleo de una constelación basado en mostrar un par de documentos que resumen el contenido de la constelación. Por el lado nuestro algoritmo basado en LDA, el núcleo identificable al centro de la constelación está vinculado a la cantidad de conexiones que tiene un documento. En cualquiera de los dos escenarios parece ser que los administradores pueden ahorrar tiempo leyendo el núcleo para obtener una imagen rápida de todos los textos de la constelación. Esta aseveración debe aún ser probada para nuestro algoritmo.

4.2. Recomendaciones

- Respecto a la formación del corpus, se recomiendan realizar pruebas con documentos de diferente naturaleza y contexto. Se podría incluso elegir varios documentos que hablen de diferentes tópicos para verificar la generación de las constelaciones y observar cómo éstas están compuestas y proceder a las comparaciones con los diferentes métodos de la herramienta.

- Referente a la visualización de los grafos se podría trabajar por representaciones en 3D y verificar como influye en la facilidad de uso para los usuarios de la herramienta. Este desarrollo puede ser útil en corpus grandes.
- Se podría trabajar en añadir funcionalidad a la herramienta respecto a obtener los documentos, mediante API de bases de datos científicas o almacenes de datos respecto a vigilancia estratégica para evitar tener documentos de manera local y trabajar con enlaces.
- Por último, es necesario realizar más pruebas con profesionales de SScan en un escenario real para validar la relevancia de nuestra solución.

5. REFERENCIAS BIBLIOGRÁFICAS

[1] H. A. Simon, «Bounded Rationality and Organizational Learning*,* » *Organization Science*, vol. II, nº. 1, pp. 125-134, Feb. 1991. [En línea]. Disponible: http://pubsonline.informs.org/doi/abs/10.1287/orsc.2.1.125

[2] H. Lesca, « Veille stratégique pour le management stratégique : Etat de la question et axes de recherche, » *Economies et Sociétés*, vol. XX, nº. 5. pp. 31–50, 1994.

[3] N. Lesca y M.-L. Caron-Fasan, «Strategic scanning project failure and abandonment factors: lessons learned, » *European Journal of Information Systems*, vol. XVII, nº 4, pp. 371-386, 2008.

[4] F. J. Aguilar, Scanning the business environment, ser. Studies of the modern corporation. New York, Etats-Unis: Macmillan, 1967.

[5] C. W. Choo, «The art of scanning the environment, » *Bulletin of the American Society for information Science and Technology*, vol. XXV, nº. 3, pp. 21–24, 1999.

[6] G.-W. Bock, M. Mahmood, S. Sharma, and Y. J. Kang, «The Impact of Information Overload and Contribution Overload on Continued Usage of Electronic Knowledge Repositories, » *Journal of Organizational Computing and Electronic Commerce*, vol. XX, no. 3, pp. 257–278, 2010.

[7] P. Hemp, «Death by information overload, » *Harvard Business Review*, vol. LXXXVII, nº. 9, pp. 82–89, 121, Sep. 2009. [En línea]. Disponible: http://www.ncbi.nlm.nih.gov/pubmed/19736853

[8] G. A. Miller, «The magical number seven, plus or minus two: some limits on our capacity for processing information, » *Psychological Review*, vol. LXIII, nº. 2, pp. 81–97, 1956.

[9] A. S. Kelton and R. R. Pennington, «Internet financial reporting: The effects of information presentation format and content differences on investor decision making, » Computers *in Human Behavior*, vol. XXVIII, nº. 4, pp. 1178–1185, Jul. 2012. [En línea]. Disponible: http://www.sciencedirect.com/science/article/pii/S0747563212000301

[10] S. Paul and D. L. Nazareth, «Input information complexity, perceived time pressure, and information processing in GSS-based work groups: An experimental investigation using a decision schema to alleviate information overload conditions, » *Decision Support Systems*, vol. XLIX, nº. 1, pp. 31–40, 2010. [En línea]. Disponible: http://www.sciencedirect.com/science/article/pii/S0167923609002668

[11] S. March y G. Smith, «Design and Natural Science Research on Information Technology, » *Decision Support Systems,* vol. XV, pp. 251-266, 1995.

[12] H. A. Simon, The Sciences of the Artificial, The MIT Press, 1996.

[13] A. R. Hevner, S. Ram, S. March y J. Park, «Design Science in Information Systems Research, » Management Information Systems Quarterly, vol. XXVIII, nº 1, pp. 75-105, 2004.

[14] S. Gregor y A. Hevner, «Positioning and Presenting Design Science Research for Maximum Impact, » MIS Quarterly, vol. XXXVII, nº 2, pp. 337-356, 2013

[15] N. Lesca, M.-L. Caron-Fasan y S. Falcy, «How Managers Interpret Scanning Information, » *Information & Management*., vol. XLIX, nº 2, pp. 126-134, 2012.

[16] M. Xu, G. Kaye y Y. Duan, «UK executives' vision on business environment for information scanning: a cross industry study, » *Information & Management*, vol. XL, nº 5, pp. 381-389, 2003

[17] C. W. Choo, «Environmental scanning as information seeking and organizational learning, » *Inf. Res.*, vol. VII, nº 6, pp. 1-37, 2001.

[18] B. A. Walters, J. J. Jiang y G. Klein, «Strategic Information and Strategic Decision Making: The EIS/CEO Interface in Smaller Manufacturing Companies, » *Inf. Manage*., vol. XL, nº 6, pp. 487-495, 2003.

[19] K. Rouibah y S. Ould-ali, «PUZZLE: a concept and prototype for linking business intelligence to business strategy, » *Journal of Strategic Information Systems*, vol. XI, nº 2, pp. 133-152, 2002.

[20] M. Steyvers y T. Griffiths, Probabilistic topic models. Handbook of latent semantic analysis., Mahwah: Lawrence Erlbaum Associates Publishers, 2007.

[21] D. Blei, A. Ng y M. I. Jordan, «Latent Dirichlet Allocation, » *Journal of Machine Learning Research*, vol. III, pp. 993-1022, 2003.

[22] M. Steyvers and T. Griffiths, Rational Analysis as a Link between Human Memory and Information Retrieval., The Probabilistic Mind: Prospects for Bayesian cognitive science., 2002.

[23] T. Griffiths and M. Steyvers, «Finding Scientific Topics. Proceedings of the National Academy of Sciences of the United States of America, » vol. CI, p. 5228-5235, 2004.

[24] H. Wallach, «Topic modeling: Beyond bag-of-words, » de Proceedings of the 23rd International Conference on Machine Learning, 2006.

[25] A. Casagrande, E. Loza y L. Vuillon, «Improving Strategic Scanning information analysis: an alternative measure for information proximity evaluation, » *Third International Conference on Enterprise Systems,* 2015.

[26] H. Lesca, N. Lesca, y H. Lesca, *Weak signals for strategic intelligence: anticipation tool for managers*. London: ISTE; Hoboken, N.J.: Wiley, 2011.

[27] R. Cilibrasi and P. Vitanyi, «The Google Similarity Distance, » *IEEE Transactions on Knowledge and Data Engineering*, vol. XIX, no. 3, pp. 370–383, 2007.

[28] T. M. J. Fruchterman y E. M. Reingold, «Graph Drawing by Force–Directed Placement, » SOFTWARE - PRACTICE AND EXPERIENCE, vol. XXI, pp. 1129-1164, 1991.

[29] M. E. J. Newman y M. Girvan, «Finding and evaluating community structure in networks, » *Phys. Rev. E*, vol. LXIX, nº 2, p. 026113, 2004.

[30] Girvan M and Newman M E J, 2002 *Proc. Natl. Acad. Sci*. USA 99 7821.

[31] V. D. Blondel, J.-L. Guillaume, R. Lambiotte y E. Lefebvre, «Fast unfolding of communities in large networks, » *Journal of Statistical Mechanics: Theory and Experiment,* vol. X, 2008.

[32] Python Software Foundation, «pypi.org, » 2020. [En línea]. Disponible: https://pypi.org/project/pandas/. [Último acceso: 10 agosto 2020].

[33] S. Bird, E. Klein y E. Loper, «nltk.org, » 2019. [En línea]. Disponible: https://www.nltk.org/api/nltk.html. [Último acceso: 10 agosto 2020].

[34] Python Software Foundation, «pypi.org, » 2020. [En línea]. Disponible: https://pypi.org/project/spacy/. [Último acceso: 10 agosto 2020].

[35] R. Rehurek y P. Sojka, «Software framework for topic modelling with large corpora, » de THE LREC 2010 WORKSHOP ON NEW CHALLENGES FOR NLP FRAMEWORKS, 2010.

[36] NetworkX developers, «networkx.github.io, » 2014. [En línea]. Disponible: https://networkx.github.io/documentation/latest/. [Último acceso: 10 agosto 2020].

[37] Plotly, «plotly.com,» 2020. [En línea]. Disponible: https://dash.plotly.com/introduction. [Último acceso: 26 agosto 2020].

[38] Hevner R, Salvator T, Jinsoo Park, y Sudha Ram. Design Science in Information Science. 2004

[39] F.D. Davis, R. Bagozzi and PP. Warshaw, «User Acceptance of Computer Technology: A Comparison of Two Theoretical Models», *Manage. Sci.,* vol. XXXV, nº 8, pp. 982-1003, Aug.1989.

[40] V. Venkatesh, M.G. Morris, G.B. Davis and F.D. Davis, «User Acceptance of Information Technology: Toward a Unified View», *MIS Q.*, vol. XXVII, nº 3, pp. 425–478, Sept. 2003.

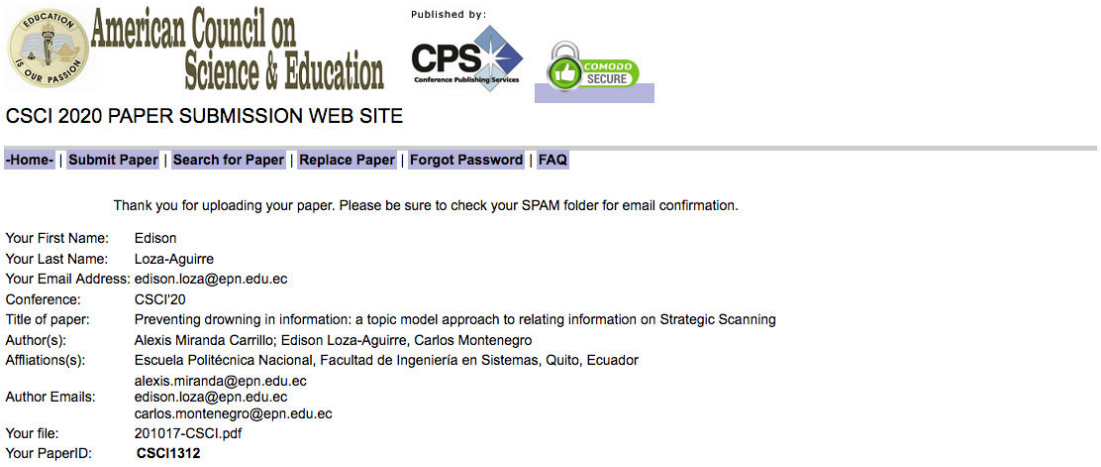
[41] S.A. Brown, A. PP. Massey, M. M. Montoya-weiss y J. R. Burkman, «Do I really have to? User acceptance of mandated technology», *Eur. J. Inf. Syst.,* vol. 11, nº 4, pp. 283-295, 2002.

[42] I. Purnajiwa, « Lemmatization in Balinese Language., » *Jurnal Elektronik Ilmu Komputer Udayana*, vol. IIX, nº 3, 2020.

[43] I. Sommerville, SOFTWARE ENGINEERING, Seventh ed., United Kingdom: Pearson Education Limited, 2005, pp. 375-376.

6. ANEXOS

6.1. Anexo I. Articulo Científico



American Council on Science & Education

Published by: CPS Conference Publishing Services

COMODO SECURE

CSCI 2020 PAPER SUBMISSION WEB SITE

-Home- | Submit Paper | Search for Paper | Replace Paper | Forgot Password | FAQ

Thank you for uploading your paper. Please be sure to check your SPAM folder for email confirmation.

Your First Name:	Edison
Your Last Name:	Loza-Aguirre
Your Email Address:	edison.loza@epn.edu.ec
Conference:	CSCI'20
Title of paper:	Preventing drowning in information: a topic model approach to relating information on Strategic Scanning
Author(s):	Alexis Miranda Carrillo; Edison Loza-Aguirre, Carlos Montenegro
Affliations(s):	Escuela Politécnica Nacional, Facultad de Ingeniería en Sistemas, Quito, Ecuador
Author Emails:	alexis.miranda@epn.edu.ec edison.loza@epn.edu.ec carlos.montenegro@epn.edu.ec
Your file:	201017-CSCI.pdf
Your PaperID:	**CSCI1312**

Figura 13. Notificación de recepción del articulo a Computational Science & Computational Intelligence (CSCI'20)

De: Council Secretariat
Enviado: domingo, 8 de noviembre de 2020 12:53
Para: CARLOS ESTALESMIT MONTENEGRO ARMAS; EDISON FERNANDO LOZA AGUIRRE; ALEXIS ADRIAN MIRANDA CARRILLO
Asunto: Your paper is accepted (CSCI1312); IEEE CPS, Xplore, Scopus, Ei - CSCI 2020

Dear Drs. Alexis Miranda Carrillo, Edison Loza-Aguirre*, and
Carlos Montenegro:

(This is being sent from two different servers to make sure that
you receive it.)

We are pleased to inform you that the following paper which you
submitted to:

O. The 2020 International Conference on Computational Science and
Computational Intelligence
(CSCI'20: December 16-18, 2020, Las Vegas, USA)
https://www.american-cse.org/csci2020/
Publisher: IEEE CPS - https://www.computer.org/conferences/cps
Science Indexations: IEEE Xplore, Ei Compendex, Scopus, and others.

Research Track/Symposium on

CSCI-ISCS: Computational Science
https://www.american-cse.org/csci2020/symposiums-ISCS

has been accepted in the category shown below; i.e., accepted for both,
publication in the IEEE CPS proceedings and presentation (the referees'
report appears between the two rows of stars "*" in this email):

Paper ID #: CSCI1312
Title/Authors: Preventing Drowning in Information: A Topic Model Approach
to Relating Information on Strategic Scanning
Alexis Miranda Carrillo, Edison Loza-Aguirre*, Carlos
Montenegro
Escuela Politecnica Nacional, Facultad de Ingenieria en
Sistemas,
Quito, Ecuador;
Escuela Politecnica Nacional, Departamento de Informatica y
Ciencias de la Computacion Quito, Ecuador

Paper Category: REGULAR RESEARCH PAPER
(maximum of 6 pages - if the authors need extra pages,
then the editors would permit one extra page above and

Figura 14. Email de notificación de Aceptación a CSCI'20

6.2. Anexo II. Manual de Usuario

HERRAMIENTA DE EVALUACIÓN DE PROXIMIDAD DE INFORMACIÓN PARA VIGILANCIA ESTRATÉGICA

Versión 1.1

CONTENIDO

OBJETIVO

El objetivo del documento es presentar una guía a usuarios finales para el uso de la herramienta de evaluación de proximidad.

JUSTIFICACIÓN

Con la finalidad de adquirir información respecto a la utilidad y facilidad de uso percibida de la herramienta se presenta esta guía, para que los usuarios finales puedan interactuar con la herramienta.

ALCANCE

El presente documento aplicará para los usuarios finales que harán uso de la herramienta de evaluación de proximidad.

DEFINICIONES

LDA Latent Dirichlet Allocation o Asignación Latente de Dirichlet (ALD).

NM Measure: Nearness Measure (Medida de proximidad).

Cluster: Conjunto de nodos.

Zoom: Efecto de acercamiento de determinados elementos de la vista.

DESCRIPCIÓN

Con acceso a la herramienta de evaluación de proximidad se podrá ver una interfaz inicial como se muestra en la Figura 1.

Figura 1. Vista inicial de la herramienta de evaluación de proximidad.

La herramienta posee cuatro pestañas para su interacción: LDA, Clustering, NM Measure y Visualización.

Figura 2. Opciones de interacción de la herramienta.

En las pestañas: LDA, Clustering y NM Measure se podrá visualizar costelaciones conformadas por nodos y sus relaciones. Los nodos son representados por círculos de diferentes colores los cuales son documentos del corpus. Las relaciones son graficadas mediante flechas las cuales indican su relación con el nodo(documento) más próximo del corpus.

En dichas pestañas (LDA, Clustering y NM Measure) podemos realizar zoom de áreas rectangulares específicas, mediante un clic sostenido el usuario podrá dibujar un rectángulo, como se muestra en la Figura 3.

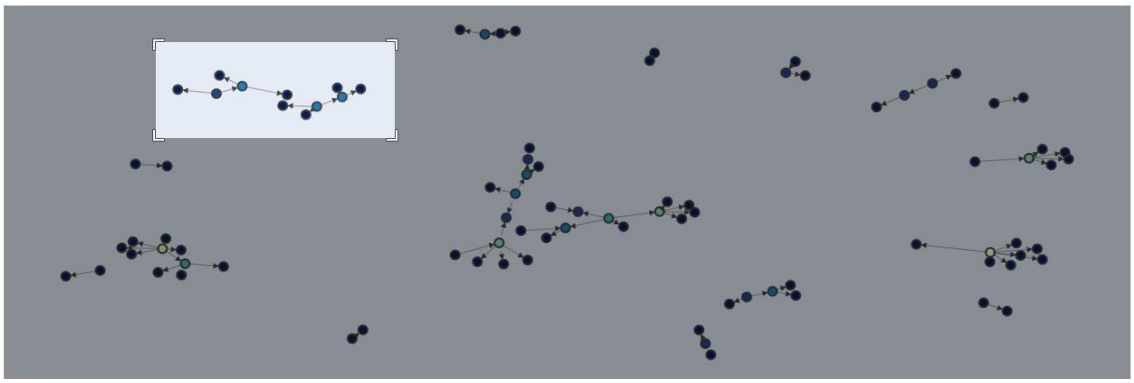


Figura 3. Aplicar zoom a áreas específicas de la vista.

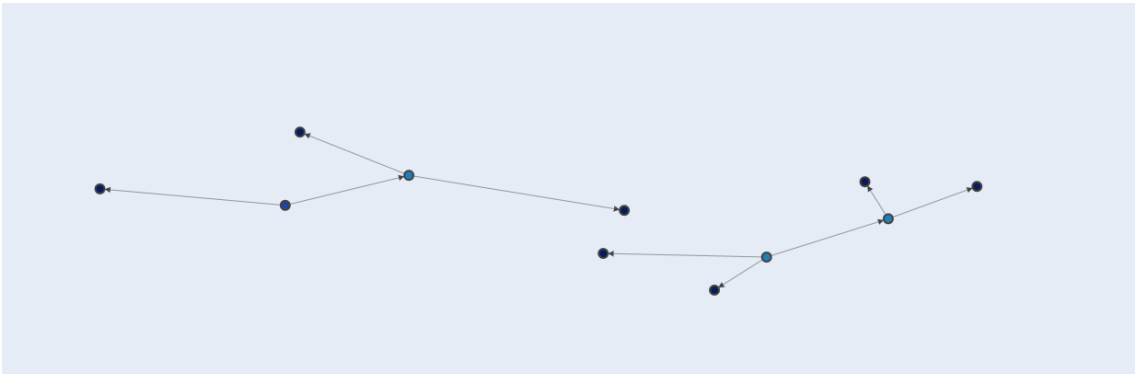


Figura 4. Visualización del zoom aplicado.

Para regresar a la vista principal (Antes del zoom aplicado) el usuario realizará doble clic en cualquier parte de la vista excepto en los nodos o relaciones.

Para poder visualizar el nodo (documento) realice un clic en dicho nodo y la herramienta automáticamente lo dirigirá a la pesta*ña* de Visualización, en la cual podrá observar su contenido.

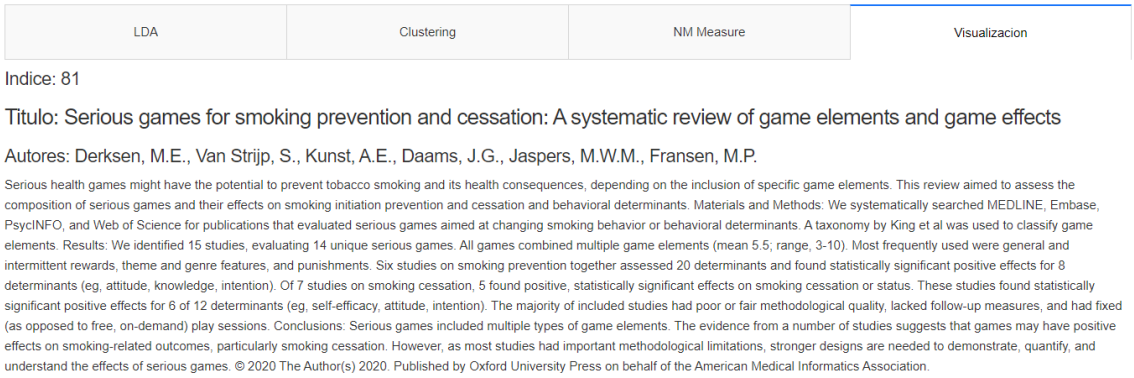


Figura 5. Contenido de la pesta*ña* Visualización.

LDA

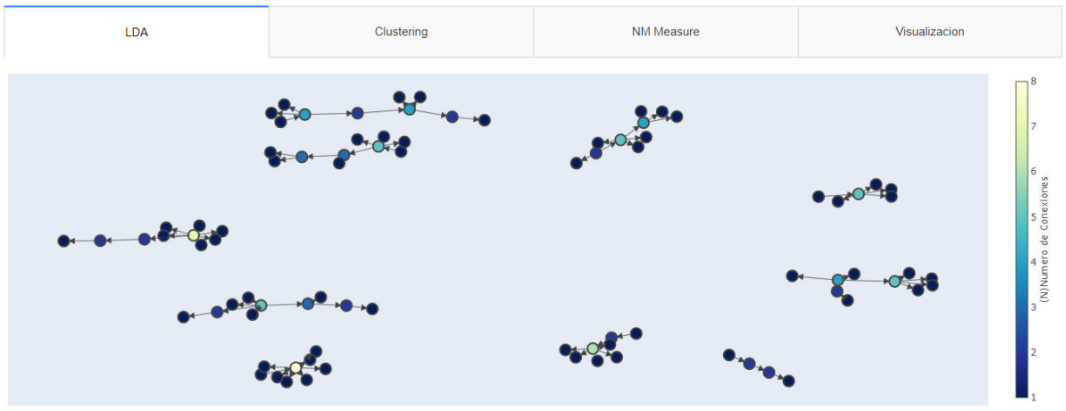


Figura 7. Vista principal de la pestaña LDA.

En la pestaña LDA se podrá visualizar constelaciones, con sus respectivos nodos y relaciones. El usuario podrá mover el cursor sobre los nodos y observará etiquetas en las cuáles se muestra el índice o nombre del documento y el número de conexiones del nodo correspondiente. (Índice: Nombre del documento o índice en el archivo Excel, N: Número de conexiones del nodo).

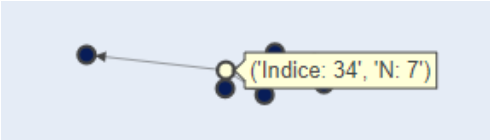


Figura 8. Etiqueta del nodo de la pestaña LDA.

En la parte derecha de la vista se puede observar una barra vertical, en el cual muestra el número conexiones de los nodos.

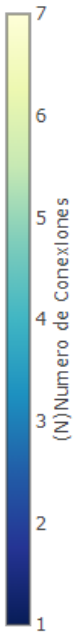


Figura 9. Barra de la pesta*ña* LDA con Número de conexiones.

Los colores de los nodos corresponden al número de conexiones en dicho gráfico, respecto a la barra vertical.

CLUSTERING

En la pestaña Clustering se podrá visualizar constelaciones, con sus respectivos nodos y relaciones. El usuario podrá mover el cursor sobre los nodos y observará etiquetas en las cuáles se muestra el índice o nombre del documento.

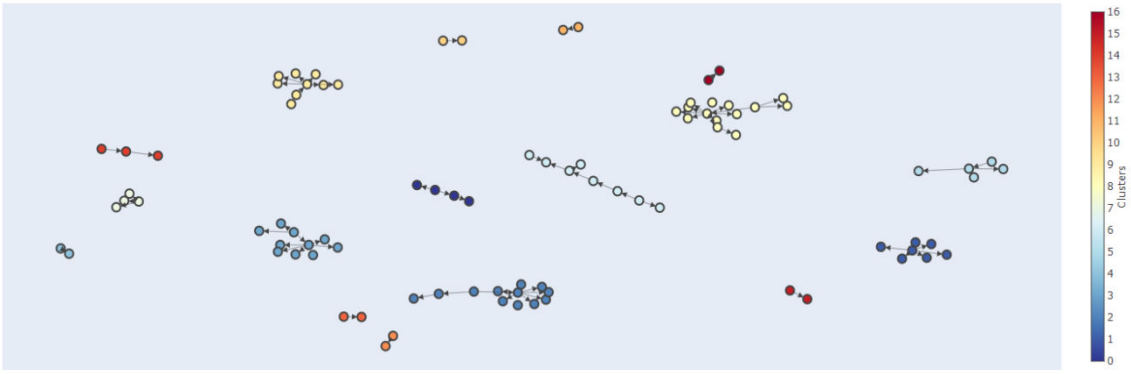


Figura 10. Vista principal de la pestaña Clustering.

Figura 11. Etiqueta del nodo de la pestaña Clustering.

En la parte derecha de la vista se puede observar una barra vertical, en el cual muestra el número del Cluster.

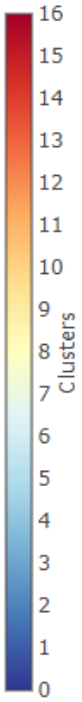


Figura 12. Barra de la pesta*ña* Clustering con Número de Clusters.

Los colores de los nodos corresponden al Cluster que pertenecen.

NM MEASURE

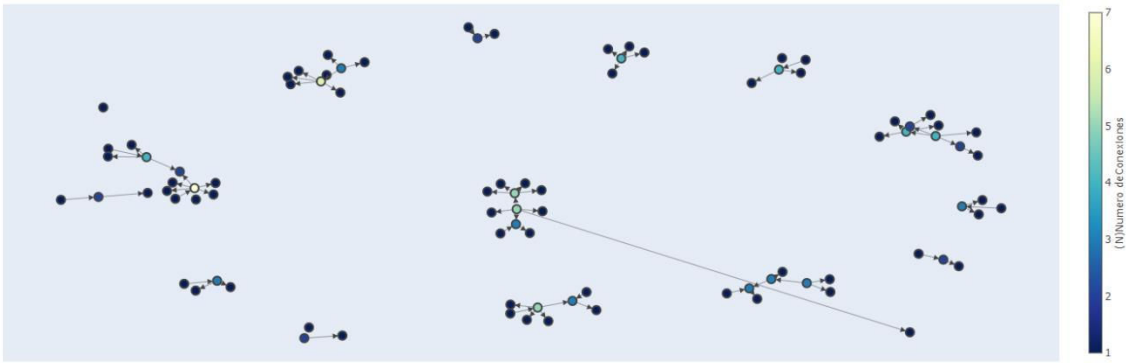


Figura 13. Vista principal de la pestaña NM Measure.

En la pestaña NM Measure se podrá visualizar constelaciones, con sus respectivos nodos y relaciones. El usuario podrá mover el cursor sobre los nodos y observará etiquetas en las cuáles se muestra el índice o nombre del documento.

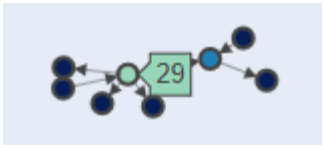


Figura 14. Etiqueta del nodo de la pestaña NM Measure.

En la parte derecha de la vista se puede observar una barra vertical, en el cual muestra el número conexiones de los nodos.

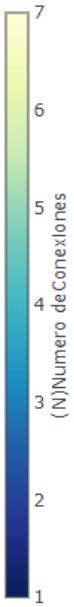
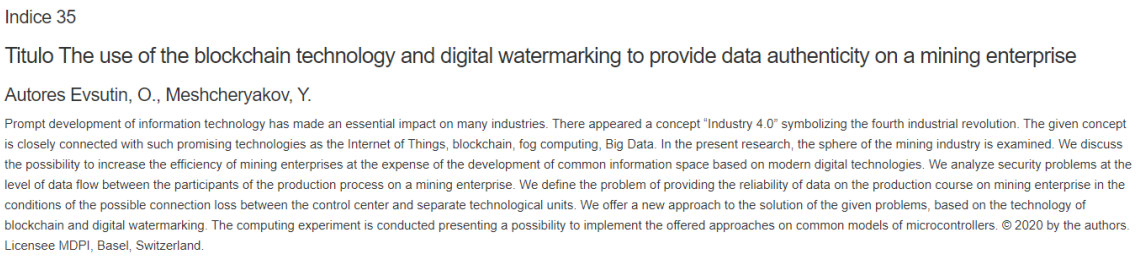


Figura 15. Barra de la pesta*ña* NM Measure con Número de conexiones.

Los colores de los nodos corresponden al número de conexiones en dicho gráfico, respecto a la barra vertical.

VISUALIZACIÓN



Indice 35

Titulo The use of the blockchain technology and digital watermarking to provide data authenticity on a mining enterprise

Autores Evsutin, O., Meshcheryakov, Y.

Prompt development of information technology has made an essential impact on many industries. There appeared a concept "Industry 4.0" symbolizing the fourth industrial revolution. The given concept is closely connected with such promising technologies as the Internet of Things, blockchain, fog computing, Big Data. In the present research, the sphere of the mining industry is examined. We discuss the possibility to increase the efficiency of mining enterprises at the expense of the development of common information space based on modern digital technologies. We analyze security problems at the level of data flow between the participants of the production process on a mining enterprise. We define the problem of providing the reliability of data on the production course on mining enterprise in the conditions of the possible connection loss between the control center and separate technological units. We offer a new approach to the solution of the given problems, based on the technology of blockchain and digital watermarking. The computing experiment is conducted presenting a possibility to implement the offered approaches on common models of microcontrollers. © 2020 by the authors. Licensee MDPI, Basel, Switzerland.

Figura 16. Vista principal de la pestaña Visualización.

Para poder visualizar el contenido de los nodos (documentos) el usuario realizará un clic en los mismos, la herramienta automáticamente desplegará la pestaña Visualización. Después de la visualización de un determinado nodo, al regresar a la misma pestaña donde fue seleccionado la vista mostrará únicamente el nodo seleccionado de color, como se muestra en la Figura 17.

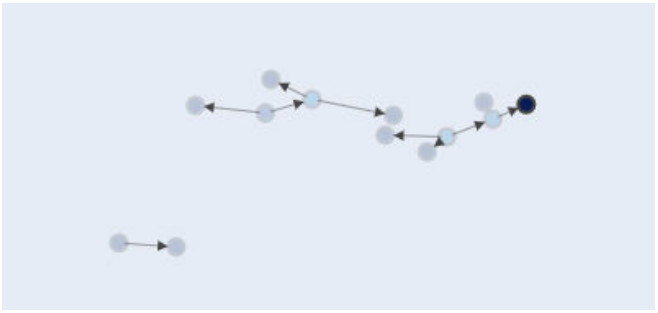


Figura 17. Nodo seleccionado.

Para visualizar nuevamente los nodos coloreados, se deberá seleccionar (Un clic) en el único nodo de color.

Ejemplo: El usuario seleccionó un determinado nodo en la pestaña LDA, la herramienta mostrará su contenido en la pestaña Visualización. Ahora, desea conocer el contenido de un nodo de la pestaña NM Measure, para lo cual se necesita ir a la pestaña LDA y seleccionar el nodo previamente seleccionado en dicha pestaña. Finalmente, ir a la pestaña del nodo que desea conocer su contenido (NM Measure).

Nota: En el caso de no volverse a colorear los nodos, el usuario podrá actualizar la página (Parte superior izquierda del navegador), con lo cual la herramienta regresará al estado inicial.

6.3. Anexo III. Pseudocódigo

Input: texts of corpus
Function processDocuments(texts)
for all texts in our corpus **do**
    docs ← split texts into words
    docs ← change words to lowercase
end for
for all words w of docs **do**
    if word is not numeric **then**
    docs ← docs + word
    end if
    if word has at least char **then**
        docs ← docs + word
    end if
    docs ← docs + lemmatized word
    if word is not in stopwords **then**
        docs ← docs + word
    end if
    if word W_i and word W_{i+1} are bigram **then**
        docs ← docs + (W_i + W_{i+1})
    end if
end for
dictionary ← Dictionary(docs)
end Function
Output: processed docs

Function CreateGraph(list of docs)
n: Number of docs
SimilarityMatrix: Each algorithm (LDA, Louvain, NM) has its matrix
for all texts t_i in similarityMatrix **do**
    for all texts t_j in *similarityMatrix* **do**
        if t_i is older higher than all **then**
            proximityMatrix[n, n] ← *proximityMatrix*[n, n] (id: tj, id: ti, ti)
        end if
    end for
end for
for all texts in *proximityMatrix* **do**
    graph (node: tj, node: ti, edge: ti)
end for
end function