UNIVERSIDAD DEL BÍO-BÍO
FACULTAD DE INGENIERÍA
DEPARTAMENTO INGENIERÍA INDUSTRIAL

Profesor Guía
Fredy Troncoso Espinosa

“Informe Ejecutivo”

“Diseño, desarrollo e implementación de un sistema de clasificación de fallas usando Text Mining en servicios de mantenimiento y reparación”

“Executive Report”

“Design, development and implementation of a fault classification system by using Text Mining in maintenance and repair services”

Alexander Robinson Ulloa Opazo

Concepción, 02 de Septiembre de 2020

Resumen

El estudio se enfoca en dar solución al problema de identificación y clasificación de las fallas y actividades en los trabajos de mantenimiento y reparación realizados por la empresa SGS Chile, las cuales se encuentran contenidas en campos de texto no estructurados. Para ello se utilizaron dos técnicas de clusterización de minería de textos, Topic Modeling y K-Means, basándose en la metodología CRISP-DM. Para empezar se obtuvo una caracterización del negocio y de los datos, identificando el proceso de una orden de trabajo, sus atributos más importantes y realizando un pre-procesamiento de los campos de texto no estructurados. Posteriormente, se procedió a modelar ambos algoritmos. En primer lugar, para Topic Modeling, después de generar múltiples modelos, se determinó que el mejor modelo fue LDA Mallet con 24 tópicos, donde cada tópico representó una falla o actividad. En base a esto se crearon módulos de visualización para representar las principales fallas y tareas por atributo. Paralelamente, se diseñó un algoritmo basado en K-Means que permitió conocer el detalle técnico y el contexto de a la información obtenida. A diferencia de LDA que se aplicó una sola vez y a la totalidad de los registros, K-Means está pensado para ejecutarse tantas veces como atributos se quiera analizar, por lo tanto el mejor modelo se determina en función del atributo seleccionado. Los clusters encontrados por K-Means se visualizan en formato texto. Se concluye que ambos algoritmos contribuyen de manera conjunta en el proceso de planificación estratégica y toma de decisiones del sector, ya que por medio de LDA se pueden identificar patrones generales y fáciles de representar, mientras que con K-Means es posible entender el contexto y el motivo detrás de cada falla o actividad.

Palabras clave: Minería de textos, clusterización, modelamiento de tópicos, k-medias, clasificación de fallas y actividades, mantención y reparación.

Abstract

The study seeks for a solution to the problem of identification and classification of faults and activities in the maintenance and repair work done by SGS Chile, which are contained in unstructured text fields. For this purpose, two text mining clustering techniques were used, Topic Modeling and K-Means, based on the CRISP-DM methodology. Initially a characterization of the business and data was obtained, identifying the general process of a work order, its most important attributes and performing a pre-processing of the unstructured text fields. Subsequently, both algorithms were modeled. First, for Topic Modeling, after generating multiple models, it was determined that the best model was LDA Mallet with 24 topics, where each topic represented a fault or activity. Based on this, visualization modules were created that allowed to identify the main failures and tasks by attribute, in a general and easily interpretable way. At the same time, an algorithm based on K-Means was designed that allowed to know the technical detail of the information obtained by the previous algorithm, by preserving the context and semantics of each observation. Unlike LDA that was applied only once and to all the database, K-Means is designed to be executed as many times as attributes to be analyzed, so the best model is determined according to the selected attribute. The clusters found by K-Means are displayed in text format. It is concluded that both algorithms contribute together in the strategic planning and decision making process, since through LDA it is possible to identify general patterns and easy to represent, while with K-Means it is possible to know the context and the reason behind each failure or activity.

Keywords: Text mining, clustering, topic modeling, k-means, classification of faults and activities, maintenance and repair.

1. Introducción

El origen del tema de investigación nace de la necesidad de mejorar la gestión de información en los trabajos de mantenimiento y reparación realizados por la línea de negocios OGC (Oil, Gas & Chemicals) de la empresa SGS Chile. En este contexto, se identifica que una de las principales problemáticas es la falta de clasificación de las fallas y actividades más recurrentes en los trabajos de este tipo, las cuales se encuentran contenidas en campos de texto no estructurados en la base de datos del sector.

A pesar de que el sistema cuenta con la opción de ingresar el tipo de falla, existe la tendencia de que las fallas se clasifiquen incorrectamente, ya que por lo general se atienden múltiples desperfectos a la vez y porque no existen etiquetas claras. Por este motivo, para conocer la verdaderas fallas y el detalle técnico de las actividades, los gerentes y jefes de área deben revisar manualmente una a una las observaciones emitidas por los técnicos al finalizar cada trabajo, observaciones que corresponden a campos de texto no estructurados.

De acuerdo con Hotho, Nurnberger, & Paaß (2005), la información almacenada en los campos de texto no estructurado no puede ser procesada por computadoras de manera tan simple, puesto que generalmente consideran el texto como secuencias de caracteres. Bajo esta premisa, la disciplina de Text Mining se ha convertido la mejor opción para extraer información valiosa a partir de este tipo de formato, al ser capaz de saltar la barrera semántica relacionada a la lingüística y al significado de las palabras (Berry & Kogan, 2010).

Campbell & Borthwick (2012) definen a Text Mining como a una gama de métodos y tecnologías interdisciplinarias que se utilizan con el fin de obtener información a partir de una colección de documentos no estructurados, tales como emails, informes, materiales académicos, o cualquier otro tipo de texto; permitiendo identificar patrones que conducen al descubrimiento de nuevos conocimientos. Por otro lado, Feldman & Dagan (1995) nombran a esta disciplina como Knowledge Discovery from Text, refiriéndose al proceso de extraer información de calidad a partir de texto por medio de diferentes técnicas y algoritmos.

Dentro de las múltiples técnicas en que las que se basa Text Mining destaca una llamada Clustering. Han, Kamber & Pei (2012) definen a Clustering como un conjunto de algoritmos de aprendizaje no supervisado que tienen por objetivo encontrar grupos (clusters) de similares características en una colección de documentos, de tal forma que los elementos de un grupo sean parecidos entre sí, pero distintos a los elementos de los otros grupos. En otras palabras, es un proceso de agrupación automática. Los algoritmos de Clustering se dividen distintos tipos. No obstante, los más populares son los algoritmos de agrupación probabilística, como Topic Modeling, y los algoritmos de partición, como K-Means.

Por un lado Topic Modeling, o modelamiento de tópicos en español, permite encontrar estructuras escondidas en largos volúmenes de texto. Existen diversas técnicas para el modelado de tópicos, siendo LDA una de las más empleadas. Latent Dirichlet Allocation (LDA) es un modelo bayesiano jerárquico de tres niveles que busca extraer información temática a partir de una colección de documentos (corpus). De acuerdo con Blei, Ng, & Jordan (2003), la idea básica de LDA es que los documentos se representan como una mezcla aleatoria de k grupos o temas latentes (tópicos), donde cada tema corresponde a una distribución de probabilidad sobre las palabras.

Por otra parte y al igual que el algoritmo anterior, K-Means facilita la obtención de patrones de comportamiento a partir de grandes cantidades de información. MacQueen (1967) define K-Means como un algoritmo iterativo que intenta dividir el conjunto de datos en k grupos (clusters) distintos y no superpuestos, en los que cada elemento pertenece a un solo grupo. El algoritmo busca que los elementos al interior del cada cluster sean lo más similar posible, y a su vez, que los clusters con elementos diferentes estén lo más alejado que se pueda. Para ello, asigna puntos de datos de tal manera que la suma de la distancia cuadrada (SSE) entre los puntos de datos y el centroide del conjunto sea mínima. Cuanta menos variación se tenga dentro de los clusters, los elementos serán más homogéneos.

La finalidad del estudio es la de demostrar que la aplicación de Text Mining puede contribuir al proceso de planificación estratégica y la toma de decisiones del sector OGC, al extraer información valiosa a partir de texto no estructurado. Para ello, y con el fin de dar solución a la situación planteada en un principio de falta de identificación y clasificación de las fallas y actividades en los trabajos de mantenimiento y reparación, se propone la creación de un sistema de algoritmos que permita reconocer y agrupar automáticamente las principales tareas del sector, basándose en los algoritmos de Topic Modeling y K-Means.

Existe una gran variedad de estudios vinculados a este tema. Por ejemplo, Fattori, Pedrazzi & Turra (2003) analizan el caso de PackMOLE, una herramienta de Text Mining diseñada para extraer información sobre patentes en el campo del embalaje, demostrando que tiene ventajas sobre las técnicas de análisis manual. Sin embargo, la calibración de los procesos de agrupación interna es difícil y consume tiempo. Esto lleva al uso de un sistema híbrido entre técnicas de Text Mining y clasificaciones manuales.

Por otra parte, Constanza Contreras (2014) logra extraer conocimiento relevante a partir de los reclamos de clientes en el Servicio Nacional del Consumidor (SERNAC), por medio de ejecutar distintos algoritmos de Topic Modeling, tales como LDA, PYTM, LSA y NMF. Se identifican problemas comunes entre los consumidores, temas de contingencia nacional y problemas específicos de productos o servicios, los cuales permiten caracterizar el comportamiento de empresas y consumidores frente a ciertas problemáticas.

María Sepúlveda (2019) presenta una herramienta visual que permite analizar y descubrir patrones de robos en la ciudad de Santiago de Chile, por medio de aplicar Topic Modeling a los campos de texto no estructurados de la base de datos de la Asociación de Aseguradoras de Chile (AACh). Se crean y comparan dos interfaces, los cuales obtienen un buen desempeño en cuanto a la resolución de las tareas propuestas, siendo capaces de caracterizar los modus operandi de los delincuentes.

Como objetivo general se tiene “Generar un sistema de algoritmos basados en Text Mining que permitan extraer información valiosa de las actividades de mantenimiento y reparación del sector OGC, clasificando los tipos de fallas más recurrentes y definiendo estadísticos que ayuden al proceso de toma de decisiones”. Los objetivos específicos son a) Diseñar las estructuras lógicas involucradas en el proceso completo de recolección, limpieza, extracción y manejo de la información, b) Identificar términos, palabras claves y contenido semántico por medio de la aplicación de los algoritmos LDA y K-Means, agrupando fallas y actividades c) Analizar y evaluar la calidad de los modelos y sus resultados, y d) Visualizar la información obtenida en un formato fácil de interpretar.

2. Material y Métodos

El estudio se centró exclusivamente en las actividades de mantenimiento y reparación de la línea de negocio OGC (Oil, Gas & Chemicals) de la empresa SGS Chile, ubicada en Puerto Madero 130, en la comuna de Pudahuel, Santiago de Chile; la cual es una empresa multinacional que entrega servicios de inspección, verificación, análisis, certificación y capacitación, contando con más de 97.000 trabajadores y una red de más de 2.600 oficinas y laboratorios en todo el mundo (SGS SA, s.f.). Actualmente, la organización está presente en 15 regiones de Chile con 9 líneas de negocios activas.

La investigación fue de tipo aplicada-tecnológica, debido a que se buscó generar conocimientos que se pudieran poner en práctica en el sector productivo de una empresa. Al mismo tiempo fue de tipo descriptiva, ya que se representó la situación actual de los trabajos de mantenimientos y reparación del sector OGC. En cuanto a los datos empleados, estos fueron de tipo cualitativo, puesto que se procesaron los campos de texto no estructurados correspondientes a las observaciones de los técnicos al finalizar cada trabajo.

Los datos fueron recolectados directamente desde la base de datos proporcionada por el área que propuso el tema (OGC), en donde se analizaron las órdenes de trabajo (OT) comprendidas entre los meses de julio del 2018 y agosto del año 2019, abarcando un total de 23.258 registros. También se realizaron diversas reuniones con ejecutivos del área, con el fin de comprender detalles técnicos de los conceptos descritos.

La metodología en la que se basó el estudio se denomina CRISP-DM (Cross Industry Standard Process for Data Mining), la cual es un proceso cíclico y flexible de seis pasos que implica el descubrimiento general de información a partir de una fuente de datos (Chapman, y otros, 2000). Las seis etapas que componen esta metodología se definen de la siguiente forma: 1) Comprensión del negocio, 2) Comprensión de los datos, 3) Preparación de los datos, 4) Modelado, 5) Evaluación de resultados y 6) Despliegue o Implementación.

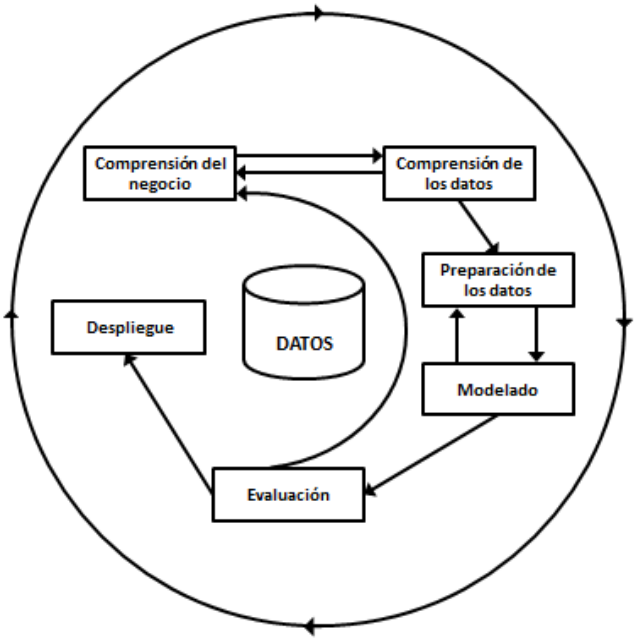

Figura 2.1. Metodología CRISP-DM. Fuente: CRISP-DM 1.0: Step-by-step data mining guide

Para la **etapa 1** se realizó una caracterización del proceso general de una orden de trabajo (OT), identificando seis fases: 1) Detección de falla, 2) Emisión de OT, 3) Recepción de OT, 4) Asignación de técnico, 5) Ejecución del trabajo y 6) Reporte de cierre.

En la **etapa 2** se estudió la base de datos de manera general. Primero, se seleccionaron las variables más relevantes para el estudio, reduciendo de 35 a 13 atributos. Luego, se analizó la distribución de las OT en los atributos escogidos, reconociendo los elementos con mayor cantidad de trabajos requeridos, como por ejemplo las estaciones más atendidas y los técnicos que más trabajan. En esta etapa se comprobó que las fallas son clasificadas de forma ineficiente, ya que como muestra la Figura 2.2., aproximadamente el 50% de las actividades son clasificadas como “otros”.

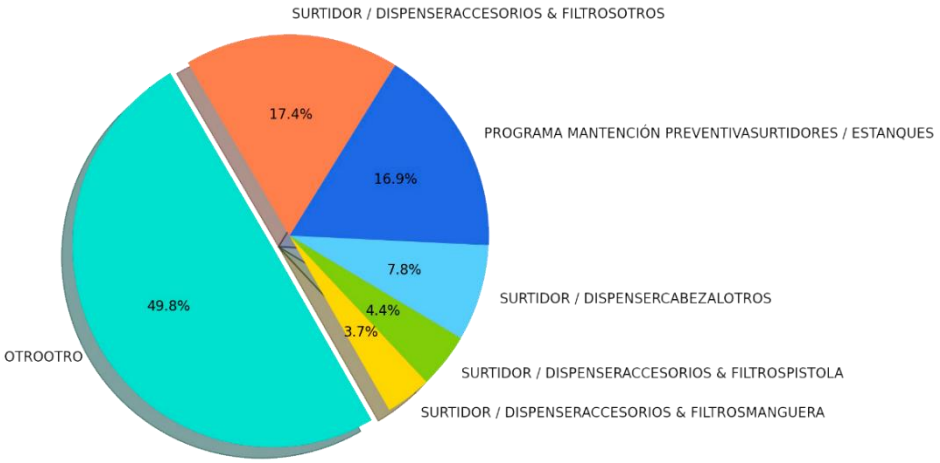

Figura 2.2. Clasificación de fallas en sistema actual. Fuente: Elaboración propia

Posteriormente, en la **etapa 3** se diseñaron las estructuras lógicas involucradas en el proceso de limpieza y transformación del texto no estructurado, contenido en los registros del atributo “Observación”. Para la limpieza, se creó una función basada en técnicas de expresiones regulares (RegEx) que permitió remover la información que no aportaba valor (stopwords, emails, fechas, etc.). Para la transformación, se decidió emplear el método de Lemmatization, el cual fue capaz de reducir la complejidad de la información al llevar las palabras a su forma de diccionario base (lema), mediante el análisis morfológico.

Tabla 2.1. Ejemplo de limpieza y transformación de texto

Texto Original	*[' Se chequea faalla, y se detecta Ortirak que cuando le da el Sol directo se borra por completo, esto no permite el corrrecto uso del mismo.\nSe informa y solicita reparaciion a empresa encargada el día 24/03/18 ']*
Limpieza y Tokenization	*['se', 'chequea', 'falla', 'se', 'detecta', 'ortirak', 'que', 'cuando', 'le', 'da', 'el', 'sol', 'directo', 'se', 'borra', 'por', 'completo', 'esto', 'no', 'permite', 'el', 'correcto', 'uso', 'del', 'mismo', 'se', 'informa', 'solicita', 'reparacion', 'empresa', 'encargada', 'el', 'dia']*
Eliminación de Stop Words	*['falla', 'ortirak', 'sol', 'directo', 'borra', 'completo', 'permite', 'correcto', 'uso', 'solicita', 'reparacion', 'empresa']*
Lemmatization	*['fallo', 'ortirak', 'sol', 'directo', 'borrar', 'completar', 'permitir', 'correcto', 'usar', 'solicitar', 'repararacion', 'empresa']*

Fuente: Elaboración propia

La **etapa 4** correspondió al modelamiento. Por un lado, para el algoritmo LDA de Topic Modeling se probaron tres librerías en Python (Gensim, Mallet y Scikit-Learn), en donde en cada una de ellas se ejecutaron 15 modelos LDA, aplicados al total de registros, iterando en dos el parámetro de número de tópicos K, desde K=2 hasta K=30. Lo anterior se hizo para encontrar los K temas que representaran de mejor forma los trabajos a nivel general.

Para encontrar el K óptimo en LDA se utilizaron las métricas Topic Coherence (Syed & Spruit, 2017), que calcula un valor de coherencia C_V (entre 0 y 100 por ciento) por cada nivel de K basándose en que las palabras con significados similares tienden a aparecer en contextos similares, y Cosine Similarity (Perone, 2013), la cual es una medida de orientación que permite identificar qué tan similares son los tópicos al calcular el coseno del ángulo que forman (valores entre -1 y 1). Además se diseñó una métrica denominada Juicio de Expertos, que corresponde a una escala de valoración que evalúa la coherencia de tópicos en base al criterio humano. Para ello se encuestaron a diez ejecutivos de OGC, los cuales asignaron un valor (entre de 1 a 10) a cado tópico encontrado en los mejores modelos LDA.

Por otra parte, y a petición particular de OGC de conservar el estado original de las observaciones para obtener detalles técnicos, se diseñó un algoritmo basado en K-Means, el cual a diferencia de LDA está pensado para ejecutarse tantas veces como atributos se quiera analizar; por lo tanto no existe un K óptimo general como en LDA, sino que se determina función de los datos ingresados. En otras palabras, con LDA se modeló para obtener patrones generales, mientras que con K-Means se modeló para obtener una clasificación específica y detallada de los registros del atributo seleccionado.

La Figura 2.3. muestra la estructura lógica del algoritmo K-Means, en donde para encontrar el K óptimo se hace uso de dos métricas. Por un lado el Método del Codo, o Elbow Method, es una herramienta visual que utiliza las distancias medias de las observaciones con respecto a su centroide (distancia intra-cluster o SSE), teniendo por objetivo encontrar el punto de inflexión donde un incremento de *k* no mejore sustancialmente esta distancia (Kodinariya & Makwana, 2013). Lo anterior se complementa con el algoritmo “Kneedle”, creado por Satopaa et al. (2011), que permite identificar de manera automática el “codo” de una curva. Por otro lado el Análisis de Silueta es un método que mide la distancia de separación entre los clusters (distancia inter-cluster). Esta medida se encuentra en el rango $[-1, +1]$, donde un valor alto significa que se hizo un buen clustering (Dabbura, 2018).

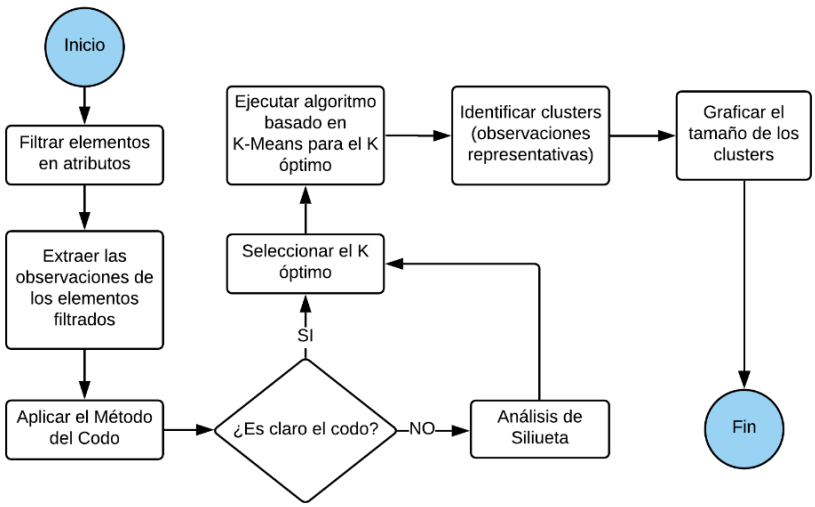

Figura 2.3. Diagrama de flujo para la aplicación del algoritmo K-Means. Fuente: Elaboración propia

En esta **etapa 5** se procedió a evaluar los modelos diseñados. Primeramente, para el algoritmo LDA se seleccionó el mejor modelo en base a las métricas de la etapa 4. Luego, se analizaron los tópicos encontrados y su distribución en los atributos más importantes. Paralelamente se presentan algunos ejemplos del algoritmo K-Means, mostrando la calidad de los clusters creados y las utilidad de emplear esta técnica en conjunto con la anterior.

Finalmente, en la **etapa 6** se busca implementar este sistema de algoritmos en la administración de OGC. Para ello se crean dos prototipos, uno para cada algoritmo, los cuales contienen módulos de visualización interactivos y de fácil comprensión, facilitando la identificación de fallas y actividades en los trabajos de mantenimiento y reparación.

3. Resultados

3.1. Topic Modeling

3.1.1. Selección de modelo óptimo LDA

Se utilizó las métricas de coherencia de tópicos c_v, similitud del coseno y juicio de expertos.

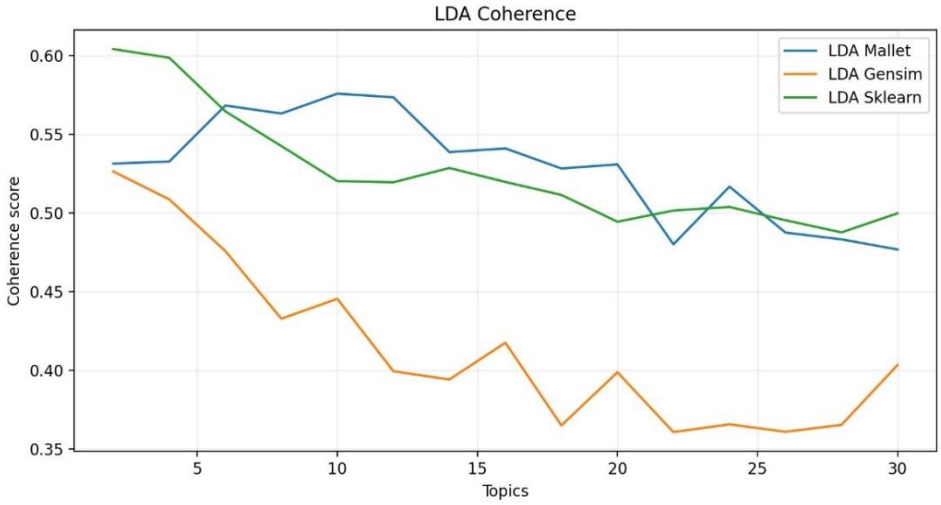

Figura 3.1. Coherencia de tópicos C_V en modelos LDA. Fuente: Elaboración propia

Por medio de los coeficientes de coherencia c_v se seleccionaron los seis mejores modelos: LDA Mallet para K=10, K=12, K=16, K=20, K=24: y LDA Sklearn para K=30. Se decidió ignorar los modelos para K=2 y K=4, debido a su contenido general y abstracto.

Luego, se crearon matrices de similitudes del coseno para cada modelo seleccionado, procesando las 1000 palabras de mayor probabilidad por tópico. A continuación se muestra un ejemplo para las matrices con K=10 y K=30.

	0	1	2	3	4	5	6	7	8	9
0	1.00	0.11	0.18	0.15	0.16	0.13	0.24	0.15	0.19	0.16
1	0.11	1.00	0.12	0.09	0.10	0.11	0.15	0.10	0.10	0.10
2	0.18	0.12	1.00	0.15	0.14	0.13	0.27	0.14	0.19	0.17
3	0.15	0.09	0.15	1.00	0.12	0.11	0.22	0.13	0.17	0.14
4	0.16	0.10	0.14	0.12	1.00	0.14	0.21	0.15	0.16	0.14
5	0.13	0.11	0.13	0.11	0.14	1.00	0.17	0.11	0.12	0.12
6	0.24	0.15	0.27	0.22	0.21	0.17	1.00	0.23	0.31	0.26
7	0.15	0.10	0.14	0.13	0.15	0.11	0.23	1.00	0.17	0.15
8	0.19	0.10	0.19	0.17	0.16	0.12	0.31	0.17	1.00	0.18
9	0.16	0.10	0.17	0.14	0.14	0.12	0.26	0.15	0.18	1.00

Figura 3.2. Matriz de similitud del coseno para LDA Mallet con K=10. Fuente: Elaboración propia

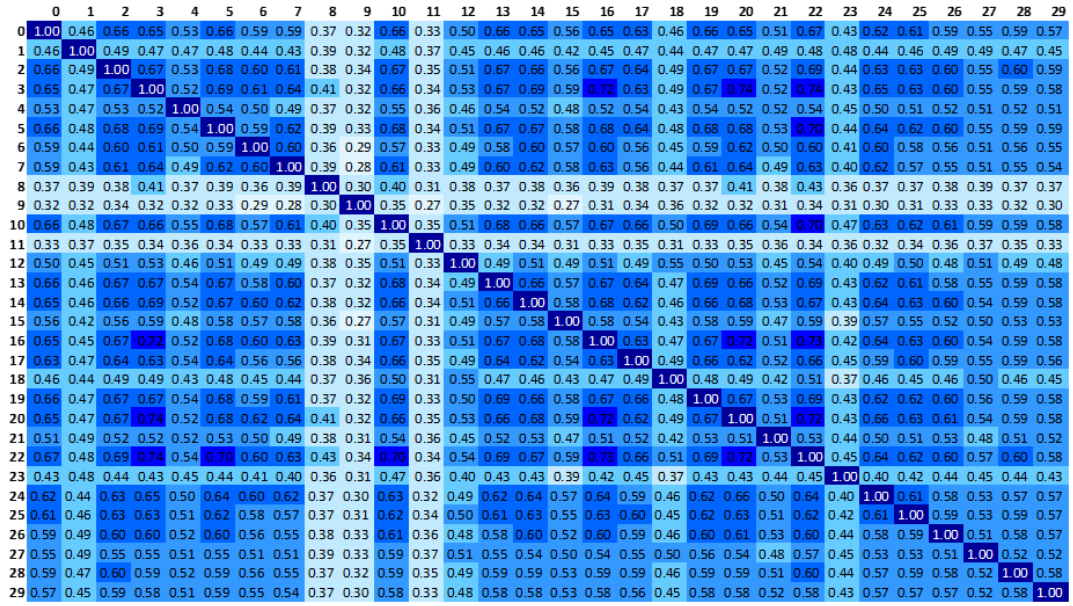

Figura 3.3. Matriz de similitud del coseno para LDA Sklearn con K=30. Fuente: Elaboración Propia

Por último, se evaluó nuevamente la coherencia y calidad de los tópicos para los seis modelos seleccionados, pero esta vez basándose en criterios humanos. A partir de esto se obtuvo un calificación general para cada modelo, las cuales se representan en la Tabla 3.1.

Tabla 3.1. Juicio de expertos para modelos LDA seleccionados

Modelo LDA	Mallet K=10	Mallet K=12	Mallet K=16	Mallet K=20	Mallet K=24	Sklearn K=30
Calificación	6,49	6,742	7,294	7,395	7,521	6,957

Fuente: Elaboración Propia

En base a lo anterior, se determinó que el modelo que mejor representó la situación de las mantenciones y reparaciones fue LDA Mallet, para un número de tópicos K igual a 24.

3.1.2. Interpretación de tópicos

Posteriormente, se le asignó una etiqueta a cada uno de los 24 tópicos generados a partir del modelo LDA seleccionado, en función de sus 15 palabras más representativas. Estas etiquetas son descritas en la Tabla 3.2.

La Figura 3.4. muestra el formato en que las palabras más representativas fueron ilustradas, dando un ejemplo para los primeros cuatro tópicos.

Figura 3.4. Nube de palabras en tópicos identificados. Fuente: Elaboración propia

Tabla 3.2. Asignación de etiquetas para tópicos encontrados en LDA Mallet con K=24

Tópico	Etiqueta
0	Bomba, motor y problemas eléctricos
1	Mediciones en estanques y ajustes en consola veeder root
2	Teclado preset y placas de control
3	Despacho del producto en ventas a clientes
4	Confirmación de atención de incidencia
5	Medidores en dispensadores
6	Revisión y pruebas del producto
7	Ajustes y reparaciones menores en equipos
8	Tapas de descarga
9	Chequeo de fallas generales y detección de problemas asociados
10	Detectores de fuga, presión en líneas y adhesivos de seguridad
11	Retiro de agua en estanques
12	Cambio de mangueras
13	Remplazo de piezas generales (como breakaway y gráficas), y toma de muestras
14	Cambio y reparación de válvulas
15	Programa mantención preventiva
16	Cambio de filtros, ampolletas y problemas de flujo lento
17	Master reset y reprogramación de parámetros
18	Calibración y limpieza general de equipos
19	Verificación volumétrica y ajustes generales en estanques de gas 93,95 y 97
20	Verificación volumétrica y ajustes generales en estanques de diesel y kerosene
21	OT mal emitidas, ingresadas, asignadas o duplicadas
22	Cambio y reparación de pistolas y swivel
23	Calibración de bocas

Fuente: Elaboración propia

3.1.3. Distribución de tópicos

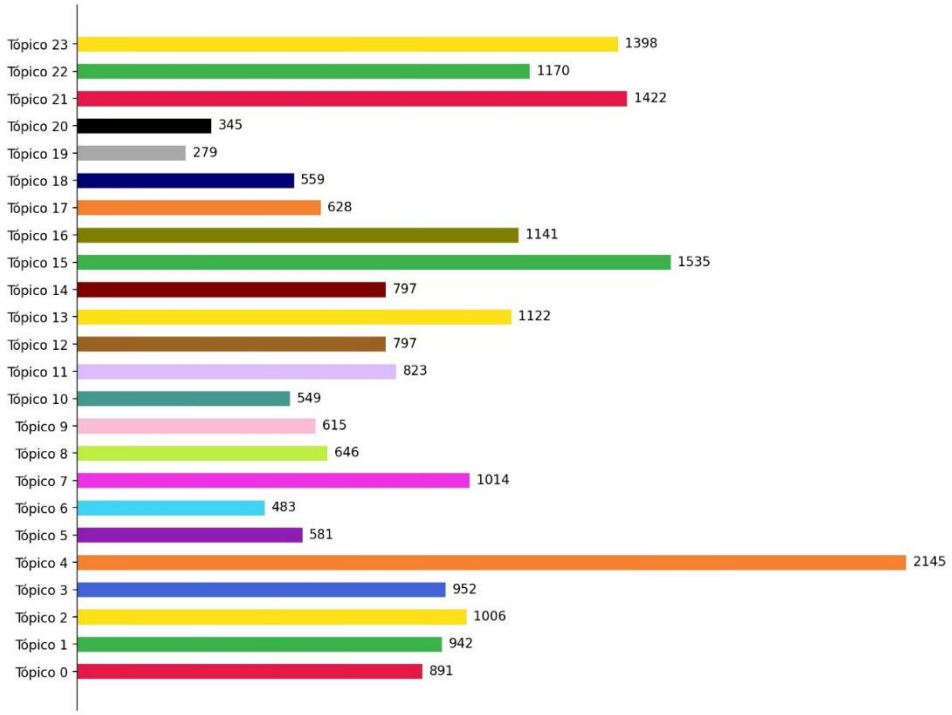

Figura 3.5. Distribución de las órdenes de trabajo por tópico. Fuente: Elaboración propia

La Figura 3.5. muestra cómo se distribuyen las principales fallas y actividades en la totalidad de registros de la base de datos. Para llevar a cabo esta tarea, a cada observación (texto no estructurado) se le asignó de manera automática el tópico que mejor representara el trabajo descrito (tópico dominante). A partir de la asignación anterior fue posible además identificar los tópicos más activos por atributo, como se representa en la Figura 3.6, donde se ven los tópicos principales para la estación *"CL – Buses Vule S.A."*, el técnico asignado *"Luis Carvajal"* y la prioridad de cliente *"Emergencia 4 Horas"*. Este proceso se puede replicar para cada uno de los atributos presentes en el sistema de información.

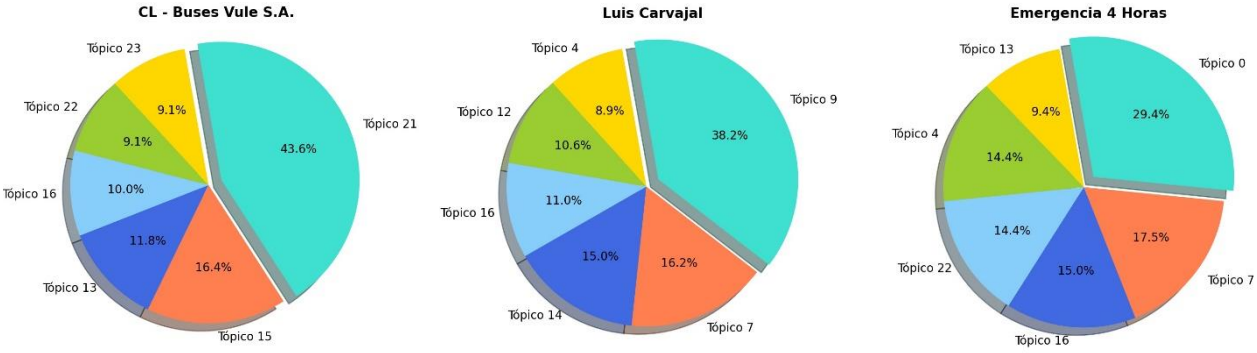

Figura 3.6. Distribución de tópicos por atributo. Fuente: Elaboración propia

3.1.4. Extracción de información desde un tópico

Con el fin de analizar el comportamiento de un tópico en particular, se diseñaron distintos módulos de visualización, los cuales pueden ser ejecutados de la misma forma en cada uno de los tópicos extraídos. A modo de ejemplo, se aplicó este análisis en el tópico 0 (Bomba, motor y problemas eléctricos), como muestra la Figura 3.7., en donde se concluye lo siguiente:

- En los meses de noviembre del 2018, enero, mayo y junio del 2019, se ingresó una mayor cantidad de OT vinculadas con el tópico 0, mientras que en julio del 2019 se alcanzó su punto más bajo.
- La cantidad de OT recepcionadas tiende a ser igual durante todo el mes.
- Los lunes son los días con mayor demanda, caso contrario a los fines de semana donde casi no existen OT ingresadas.
- Por lo general las OT tienden a emitirse antes del mediodía.
- El tiempo entre *Recepción de OT* y *Arribo de Técnico* suele ser menor a 24 horas.
- El tiempo en concluir un trabajo, es decir, el tiempo entre *Arribo de Técnico* y *Cierre de Trabajo*, usualmente es menor a una hora o se encuentra en un rango de una a dos horas.

Lo anterior hace referencia al comportamiento histórico de la totalidad de las OT vinculadas al tópico 0. Sin embargo, este análisis puede ser aún más específico, al evaluar el comportamiento del tópico exclusivamente para las OT de un atributo en particular, como por ejemplo el comportamiento del tópico 0 exclusivamente en la estación *"CL – Buses Vule S.A."*

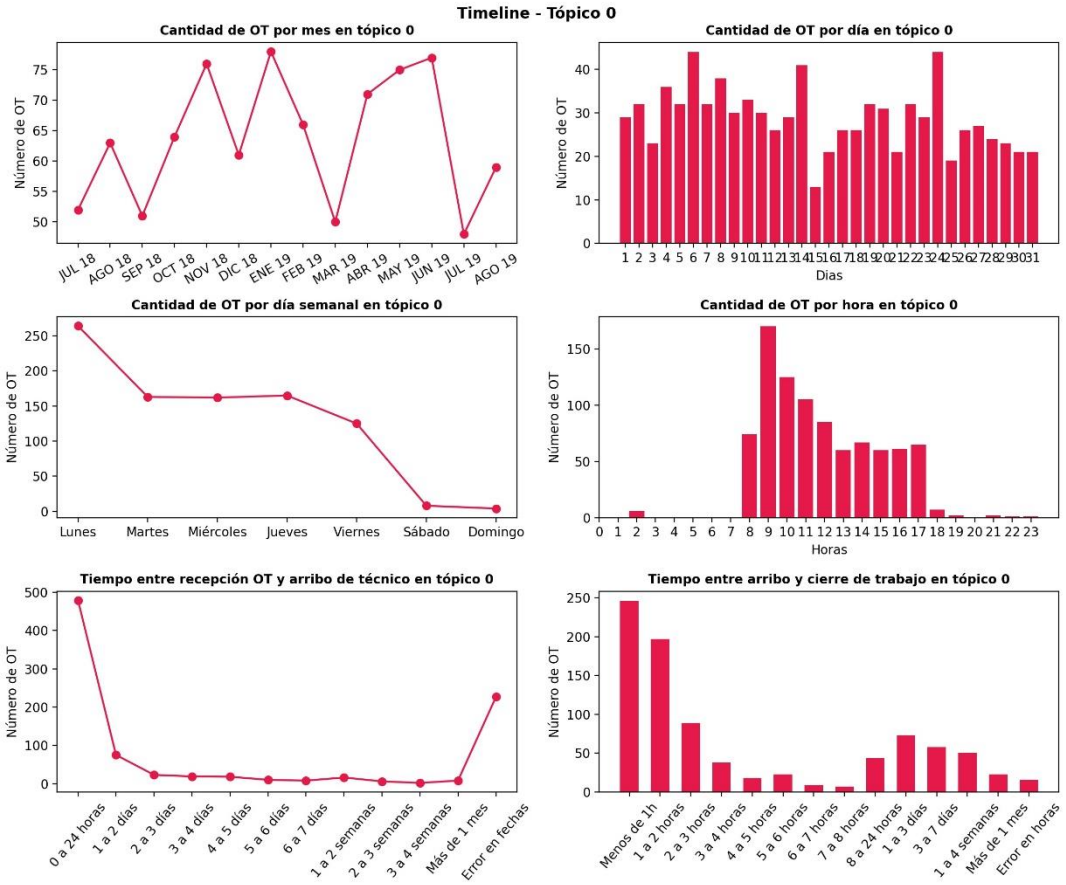

Figura 3.7. Comportamiento del tópico 0 en el tiempo. Fuente: Elaboración propia

3.2. K-Means

Para llevar a cabo la demostración de este algoritmo, se seleccionó la estación "*CL - Buses Vule S.A.*" como objeto de estudio. Una vez filtrados los registros para este atributo, se continuó con el diagrama de flujo mostrado en la Figura 2.3.

3.2.1. Selección de modelo óptimo K-Means

Se emplearon las métricas del método del codo y el análisis de silueta, obteniendo lo siguiente:

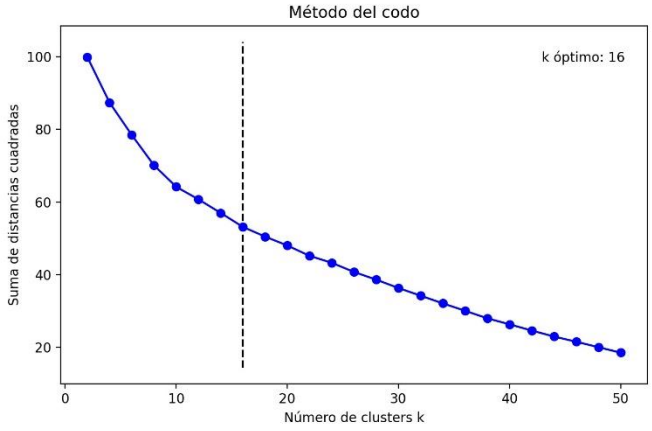

Figura 3.8. Método del codo para OT en estación CL Buses Vule S.A. Fuente: Elaboración propia

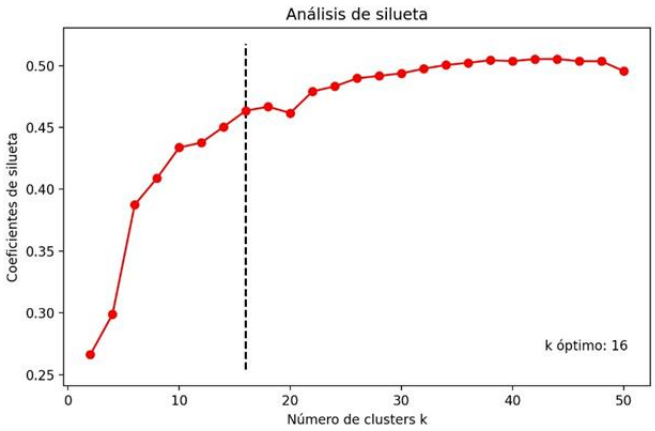

Figura 3.9. Análisis de silueta para OT en estación CL Buses Vule S.A. Fuente: Elaboración propia

A través del método del codo de la Figura 3.8 se determina de forma preliminar que el K óptimo es 16. Sin embargo, el punto de inflexión no es claro, por lo que se decide aplicar el análisis de silueta que se muestra en la Figura 5.9, donde en base a los coeficientes de siluetas se confirma que a partir del K=16 la curva comienza a aplanarse. Por ende, basándose en ambos métodos se considera que 16 es el número de clusters óptimo.

A continuación, la Figura 3.10 grafica el análisis de silueta detallado para el algoritmo K-Means con K=16. A partir de esta figura se identifica que los clusters 0, 2, 3, 4 y 9; son los de mejor calidad, dado que sus coeficientes son cercanos a uno, mientras que el cluster 6 es el que obtiene la peor evaluación, ya que el valor de su coeficiente de silueta es negativo, haciendo referencia a que las observaciones al interior de este cluster están mal asignadas.

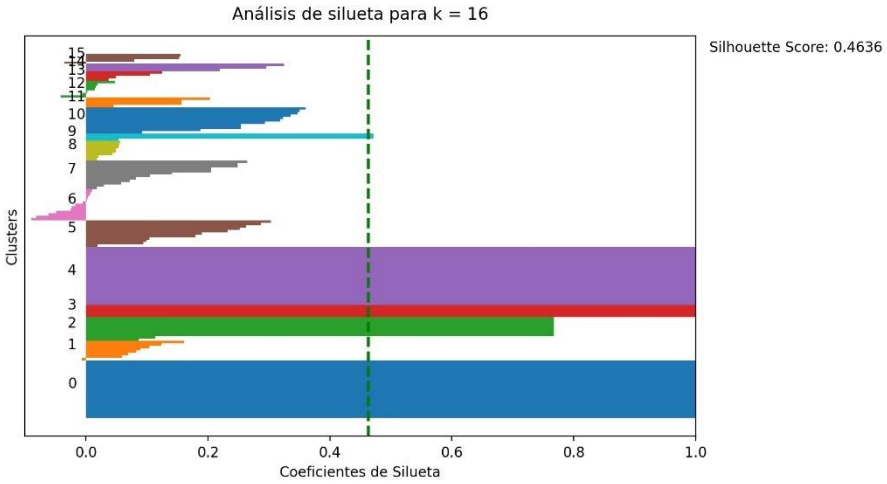

Figura 3.10. Análisis de silueta detallado para K=10 en CL Buses Vule S.A. Fuente: Elaboración Propia

3.2.2. Representación de clusters

Una vez encontrado el número de clusters K óptimo para los registros de la estación "*CL - Buses Vule S.A.*", se ejecutó el algoritmo K-Means para el valor K seleccionado. A partir de esto se extrajeron las observaciones más representativas para cada uno de los clusters generados, como se representa en la Figura 3.11.

```
Cluster  0 : ods mal emitida
Cluster  1 : surt 1 d se realizo remplazo de un filtro de alta cap surt 1 d se realizo remplazo de un filtro de alta cap
Cluster  2 : mantencion preventiva se calibran 2 de 2
Cluster  3 : se realiza el cambio de 2 filtros spin on quedando equipo operativo
Cluster  4 : os mal emitida
Cluster  5 : mantencion preventiva se realizara bajo ot 5
Cluster  6 : en surtidores 1 y 2 diesel se cambia 02 filtros p14187 quedando 100 operativos en surtidores 1 y 2 diesel se cambi
a 02 filtros p14187 quedando 100 operativos
Cluster  7 : se realiza el cambio de 2 codos giratorio 1 2 swivel 1 y 2 filtros alta capacidad quedando equipo operativo
Cluster  8 : en surtidor numero 1 se detecto y se cambiaron 02 pistolas convencionales de 1 con sensor malo se revisan y se lim
pian los dos filtros de malla de entrada de bomba se ajust a presion general realizando pruebas con preset y con ventas dispens
adores queda con despacho normal repuestos utilizados 02 pistolas convencionales de 1 p14781 en surtidor numero 1 se detecto y
se cambiaron 02 pistolas convencionales de 1 con sensor malo se revisan y se limpian los dos filtros de malla de entrada de bom
ba se ajust a presion general realizando pruebas con preset y con ventas dispensadores queda con despacho normal repuestos util
izados 02 pistolas convencionales de 1 p14781
Cluster  9 : se realiza el cambio de 1 tapa adaptdor descarga 4 quedando equipo operativo
Cluster  10 : mp se realizara cuando se repare surtidor n 2 se efectua proceso de mantencion preventiva de estanque y surtidore
s se realiza revision limpieza y verificacion volumetrica a los surtidores 1 y 2 con matraz de 20 lts se cambian 2 filtros de a
lta capacidad debido a flujo lento mejorando considerablemente 65 litros por minutos aprox se reaprietan porta pistolas boca 1
y 2 ademas se remarcan adhesivos bocaestanque regla incompleta comienza desde los 200 litros aprox y estanque sin agua bocas de
ntro de la tolerancia establecida por la compania bocas verificadas 2 de 2 punto industrial queda operativo mp se realizara cua
ndo se repare surtidor n 2 se efectua proceso de mantencion preventiva de estanque y surtidores se realiza revision limpieza y
verificacion volumetrica a los surtidores 1 y 2 con matraz de 20 lts se cambian 2 filtros de alta capacidad debido a flujo lent
o mejorando considerablemente 65 litros por minutos aprox se reaprietan porta pistolas boca 1 y 2 ademas se remarcan adhesivos
bocaestanque
Cluster  11 : pendiente remplazo de 2 filtros disp 12d se realizo remplazo de 2 filtros de alta cap pendiente remplazo de 2 fil
tros disp 12d se realizo remplazo de 2 filtros de alta cap
Cluster  12 : pendiente el cambio de un motor trifasico a surtidor numero 2 se instala 1 motor trifasico surtidor n 2 al moment
o de operar equipo se detecta filtracion en bomba centrifuga se instala empaquetadura quedando en tiempo de secado para un buen
sellado se repara equipo cambiando motor trifasico ya que se encontraba defectuoso produciendo corte energia electrica ademas s
e repara filtracion en bomba centrifuga quedando 100 operativo el surtidor pendiente el cambio de un motor trifasico a surtidor
numero 2 se instala 1 motor trifasico surtidor n 2 al momento de operar equipo se detecta filtracion en bomba centrifuga se ins
tala empaquetadura quedando en tiempo de secado para un buen sellado se repara equipo cambiando motor trifasico ya que se encon
traba defectuoso produciendo corte energia electrica ademas se repara filtracion en bomba centrifuga quedando 100 operativo el
surtidor
Cluster  13 : se realiza limpieza y mantencion de filtro malla quedando equipo operativo
Cluster  14 : se realiza revision de equipo encontrandose despacho operativo
Cluster  15 : se acude por falla en venteo segun informacion de encargado ellos no realizaron insidencia por este tema o el dis
p con falla venteo se encuentra en pilae canopy se acude por falla en venteo segun informacion de encargado ellos no realizaron
insidencia por este tema o el disp con falla venteo se encuentra en pilae canopy
```

Figura 3.11. Clusters para OT en CL - Buses Vule S.A. usando K-Means. Fuente: Elaboración propia

La Figura 3.12. muestra el tamaño de los clusters encontrados, donde se puede identificar que los clusters 0 y 4 son los registros más ingresados para la estación en estudio, mientras que el cluster 9 es el que menos observaciones asociadas tiene.

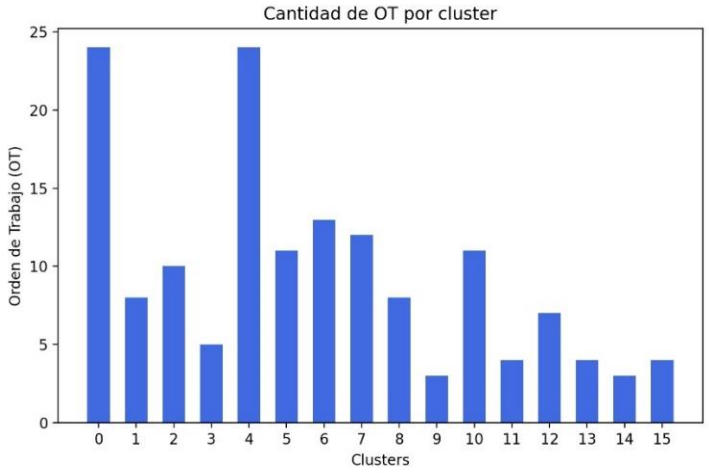

Figura 3.12. Cantidad de OT por cluster en CL - Buses Vule S.A. Fuente: Elaboración propia

Además, por medio del algoritmo es posible ver el contenido al interior de cada cluster en el formato que se muestra en la Figura 3.13.

```
Cluster 0 : ods mal emitida
        sentence  0 :  ods mal emitida
        sentence  1 :  ods mal emitida
        sentence  2 :  ods mal emitida
        ...

Cluster 1 : surt 1 d se realizo remplazo de un filtro de alta cap surt 1 d se realizo remplazo de un filtro de alta cap
        sentence  0 :  se coordinara una proxima visita para am surtidor con llave encargado se retiro surt 1 d se reali
        sentence  1 :  surt 1 d se realizo remplazo de una pistola de 1 opw p14781 surt 1 d se realizo remplazo de una p
        sentence  2 :  surt 1 d se realizo remplazo de un filtro de alta cap p14187 surt 1 d se realizo remplazo de un f
        ...

Cluster 2 : mantencion preventiva se calibran 2 de 2
        sentence  0 :  mantencion preventiva se calibran 2 de 2
        sentence  1 :  mantencion preventiva se calibran 2 de 2
        sentence  2 :  mantencion preventiva no se realiza por no encontrarse personal encargado del punto
        ...

Cluster 3 : se realiza el cambio de 2 filtros spin on quedando equipo operativo
        sentence  0 :  se realiza el cambio de 2 filtros spin on quedando equipo operativo
        sentence  1 :  se realiza el cambio de 2 filtros spin on quedando equipo operativo
        sentence  2 :  se realiza el cambio de 2 filtros spin on quedando equipo operativo
        ...
```

Figura 3.13. Extracto del contenido de los clusters usando K-Means. Fuente: Elaboración propia

4. Discusión

El resultado esperado era generar un sistema capaz de identificar y clasificar fallas a partir de las observaciones ingresadas por los técnicos en tareas de mantenimiento y reparación, en la línea de negocio OGC de la empresa SGS Chile, ubicada en Santiago de Chile.

Se propuso un sistema de algoritmos basados en Text Mining que permitió encontrar patrones de comportamiento en campos de texto no estructurado. En primer lugar, se empleó el algoritmo LDA, basado en la técnica de agrupación probabilística de modelamiento de tópicos, el cual fue capaz de identificar las actividades más recurrentes del sector de manera general. Posteriormente, se diseñó un algoritmo basado en la técnica de agrupación de partición K-Means, que permitió dar a conocer detalles específicos de las actividades por medio de analizar las observaciones para un atributo en particular.

Por medio del algoritmo LDA se identificaron 24 actividades recurrentes. Para encontrar este número se emplearon las métricas de coherencia de tópicos, similitud del coseno y juicio de expertos. Para la coherencia de tópicos se obtuvo un valor c_v de 51,7% con K=24, donde luego de 24 comenzaba a decrecer drásticamente. En cuanto a la similitud del coseno se identificó que para K=10 los tópicos eran muy independientes y generales, mientras que con K=30 los tópicos comenzaban a ser redundantes y parecidos, por lo que se determinó que el K óptimo debía estar en un rango intermedio. Luego, en juicio de expertos la mejor calificación fue para el modelo LDA Mallet K=24, con una nota de 7,52.

Posteriormente, se le asignó uno de los 24 tópicos a cada observación en función de la actividad descrita. A partir de esto se pudo construir elementos visuales identificando las principales fallas y tareas de manera general y fácilmente interpretable. Algunos ejemplos de lo que se extrajo en LDA se muestra en las Figuras 3.5, 3.6 y 3.7., donde se puede ver la distribución de los tópicos en la base de datos total, la distribución de tópicos por atributo y el comportamiento de los tópicos en el tiempo respectivamente. Por ejemplo, se identificó que las cuatro tareas más recurrentes en la estación "*CL – Buses Vule S.A.*" fueron las OT mal emitidas (tópico 21 – 43.6%), mantenimiento preventivo (tópico 15 – 16,4%), remplazo de piezas / toma de muestras (tópico 13 – 11.8%) y cambio de filtros (tópico 16 – 10%).

Por otra parte por medio del algoritmo K-Means fue posible conocer detalles técnicos y específicos para cada atributo seleccionado, siguiendo el diagrama de flujo de la Figura 2.3. A diferencia de LDA que se aplicó una sola vez y a toda la base de datos encontrando un número de clusters K óptimo general, en K-Means el K óptimo (mejor modelo) se debe determinar en función de los datos del atributo seleccionado. Por ende el número de clusters encontrados varía dependiendo de que elemento a analizar. A modo de ejemplo se aplicó este algoritmo en la estación "*CL – Buses Vule S.A.*", descubriendo que el K óptimo fue de 16. A partir de estos 16 clusters se identificó que los problemas más recurrentes en esta estación fueron las OT mal emitidas (cluster 0 y 4), mantenimiento preventivo (cluster 2, 5 y 10) y remplazo filtros y piezas generales, como se representa en la Figura 3.11.

Al ver los resultados del K-Means se puede comprobar que efectivamente corresponden a los tópicos identificados por medio de LDA para la misma estación. Por una parte, la ventaja de K-Means sobre LDA está en el nivel de detalles que se pueden obtener a partir de los clusters. Por ejemplo, gracias a LDA se sabía que el remplazo de filtros era una actividad frecuente en la estación analizada. No obstante, aplicando K-Means fue posible conocer el tipo de filtro cambiado (malla , alta capacidad, spin on, etc.) y que al cambiar filtros por lo general se realizan otros remplazos asociados. Por otro lado, la desventaja de K-Means está en la representación visual de los resultados. Como se observó antes, con LDA es fácil dar a conocer los clusters identificados en los atributos, gracias a los gráficos y herramientas visuales. No es el caso de K-Means, el cual presenta los resultados en forma de texto, requiriendo de tiempo y un buen dominio de los conceptos para interpretar cada cluster.

La validez interna de la solución planteada se sustenta en primera instancia en la compatibilidad de ambos modelos, ya que como se demuestra con anterioridad, a partir de las dos técnicas utilizadas se llega a resultados similares. Por otra parte, el proceso para llegar a estos resultados estuvo bajo supervisión directa de los encargados de la línea de negocios OGC, validando cada una de las etapas de la metodología aplicada.

El sistema de algoritmos propuesto cumple con el objetivo elemental de identificar y clasificar las fallas en los trabajos de mantenimiento y reparación del sector, las cuales son representadas de mejor forma en el prototipo creado que contiene módulos interactivos de visualización para ambos modelos. Dicho sistema puede replicarse en cualquier sector que necesite procesar campos de texto no estructurados, tales como áreas de Marketing, Finanzas o Atención al cliente, por ende no presenta limitaciones externas. Sin embargo, actualmente para ejecutar el prototipo se requiere de conocimiento básico en Python.

Dentro de los resultados encontrados en la investigación bibliográfica destaca el trabajo de Constanza Contreras (2014), donde se aplicó distintas técnicas de modelamiento de tópicos en la base de datos del Servicio Nacional del Consumidor (SERNAC), tales como LDA, PYTM, LSA y NMF; evaluando los modelos para distintos número de tópicos K. Se seleccionó el modelo PYTM encontrando 25 temas relacionados a problemas comunes entre los consumidores, temas de contingencia nacional y problemas específicos de productos o servicios. La semejanza de este trabajo con la investigación propuesta está en la forma de determinar el mejor modelo, ya que se emplearon métricas similares. Sin embargo, el contexto y la presentación de resultados es distinta.

En la publicación realizada por María Sepúlveda (2019) se aplicó nuevamente el algoritmo LDA de modelamiento de tópicos con el fin de encontrar patrones de robo de vehículos. No obstante, el objetivo principal de este trabajo estaba en crear una herramienta visual

interactiva, por lo que el número de tópicos óptimo K y las métricas implicadas pasaron a segundo plano. Se diseñaron dos interfaces gráficas de usuarios (GUI) las cuales fueron sometidas a una evaluación con usuarios, donde se midió el desempeño en cuanto a tiempo, interacción y rendimiento de cada una al resolver múltiples tareas sobre tendencias, agregación y sobre información puntual. A pesar de que la investigación presentada sobre las actividades de mantenimiento y reparacion en OGC no contemplaba la creación de un interfaz, el trabajo de María Sepúlveda (2019) contribuyó a la creación de los módulos de visualización para el prototipo, los cuales a pesar de ser básicos y no tan sofisticados como los de María, ayudan a explicar de mejor forma la situación del sector OGC.

Shanie, Suprijadi, & Zulhanif (2017) presentan un estudio donde se utiliza K-Means para agrupar texto en el análisis de patentes respecto al té verde. Se analizaron datos en formato texto mediante dos fases: una fase de preparación de los datos y una fase de análisis. La fase de preparación de datos utilizó métodos de minería de textos y la etapa de análisis de datos utilizó técnicas estadísticas, donde incorporaron el algoritmo de clusterización de K-Means. Los resultados de este estudio demostraron que, sobre la base del valor máximo de Silhouette, se generaron 87 conglomerados asociados a quince términos. Contrastando este trabajo con la investigación actual, se identifica que en ambos está presente el análisis de silueta para determinar el número óptimo de clusters. Sin embargo, la diferencia está en la metodología empleada, ya que el trabajo de patentes incorpora la curva Zipf que clasifica la información en trivial, interesante y no importante, presentando los resultados en gráficos similares a los desarrollados por el algoritmo LDA, mientras que en el trabajo actual los clusters son visualizados como texto, mostrando la observación más representativa.

A partir de este sistema de algoritmos se pueden derivar planes estratégicos a largo plazo relacionados con el análisis predictivo de fallas por estación, el perfil del trabajador, el análisis de mantenciones a nivel regional y el estudio de las actividades más importantes.

Se plantea la opción de mejorar aspectos técnicos a modo de trabajo futuro, tales como mejorar la calidad del preprocesamiento de las palabras, agregar nuevas técnicas para el modelamiento de tópicos como los algoritmos PYTM y NMF e incorporar nuevas métricas que permitan realizar una mejor comparación del valor y la calidad de los modelos. Además se propone crear un interfaz gráfico de usuario (GUI) que permita representar los clusters de manera automática y fácil de interpretar, sin necesidad de ejecutar códigos en Python.

5. Conclusiones

El trabajo presenta el diseño, desarrollo e implementación de un sistema de algoritmos basados en Text Mining, específicamente en las técnicas de clusterización de Topic Modeling LDA y K-Means, que a partir de campos de texto no estructurado en trabajos de mantenimiento y reparación de la empresa SGS Chile, permiten la identificación y clasificación de las fallas y actividades más recurrentes, contribuyendo con la planificación estratégica y el proceso de toma de decisiones.

Para llevar a cabo el estudio se aplicó la metodología CRISP-DM (Cross Industry Standard Process for Data Mining), la cual destaca por ser un proceso cíclico y flexible de seis pasos que implica el descubrimiento general de información a partir de una fuente de datos. Se considera que esta es la metodología apropiada para el caso dado que contempla una etapa de entendimiento del negocio y además porque al ser un proceso cíclico permite mejorar constantemente los aspectos de cada una de las etapas.

Con el desarrollo de esta propuesta se da cumplimiento al objetivo principal del proyecto que era el de clasificar las principales fallas y tareas relacionadas con trabajos de mantenimiento y reparación del sector OGC. Por un lado gracias al algoritmo LDA de Topic Modeling se pueden identificar patrones generales y fáciles de representar, mientras que por otro lado con K-Means es posible entender el contexto y el motivo detrás de cada falla o actividad, permitiendo además conocer el detalle de los patrones encontrados en LDA.

La aplicación conjunta de estos algoritmos facilita la obtención de información valiosa en tiempo real, la cual contribuye con la elaboración de planes estratégicos a corto y largo plazo, y ayuda al proceso de toma de decisiones. Por medio de estos algoritmos se pueden crear planes preventivos para las estaciones al conocer la frecuencia y tendencia de las fallas, se puede evaluar el trabajo de un técnico al asociarlo con una falla y el tiempo en que se demora en realizar una tarea, o se pueden identificar que actividades son las críticas del sector al analizar su importancia en la columna de prioridad, por dar algunos ejemplos.

En cuanto a los objetivos específicos, el primero hacía referencia al diseño de las estructuras lógicas involucradas en el proceso completo de recolección, limpieza, extracción y manejo de la información. Estas estructuras fueron creadas e implementadas en virtud de ejecutar exitosamente los algoritmos. Este objetivo está implícitamente contenido en las primeras tres etapas de la metodología CRISP-DM que corresponden a la comprensión del negocio, comprensión de los datos y preparación de los datos respectivamente.

El segundo objetivo específico correspondió a la identificación de términos, palabras claves y contenido semántico por medio de la aplicación de los algoritmos LDA y K-Means, agrupando fallas y actividades creación de los modelos que permitan clasificar las fallas y actividades correspondientes. Para ello se diseñaron las estructuras lógicas en ambos algoritmos, estableciendo parámetros, incorporando librerías de código abierto e iterando los modelos testeados. Como resultado se lograron agrupar las respectivas fallas y tareas, de manera general en LDA, y de forma más detallada en K-Means. Este objetivo se encuentra en la etapa cuatro de la metodología, denominada como modelamiento.

Para el análisis y evaluación de la calidad de los modelos y sus resultados, que corresponde al tercer objetivo específico, se emplearon distintas métricas en función del algoritmo aplicado. En el algoritmo LDA se identificó que el mejor modelo fue el de la librería Mallet para 24 tópicos, los cuales luego fueron distribuidos en la totalidad de los registros. Para el algoritmo K-Means se identificó que no existe un modelo general, sino que el mejor modelo varía en función del atributo seleccionado, en base a esto se prueba la estación “*CL – Buses Vule S.A.”*, reconociendo 16 clusters. Se comparan ambos algoritmos para dicha estación y se concluye que los resultados son compatibles. La etapa cinco de CRISP-DM hace referencia a este objetivo, ya que corresponde a la fase de evaluar resultados.

Por último y en relación con el cuarto objetivo específico y con la última etapa de la metodología CRISP-DM, se crea e implementa un prototipo que incluye módulos de visualización interactivos capaces de representar de manera simple los resultados obtenidos, orientado a ejecutivos del sector y personal que trabaje con las ordenes de trabajo de mantenciones y reparaciones. Sin embargo, se dice que el prototipo no es apto para todo el personal, dado que se requiere de un conocimiento básico de programación en Python, por lo que se considera que actualmente no puede ser replicado en gran volumen. Este prototipo fue probado por un grupo reducido de ejecutivos los cuales lograron extraer información de interés y a su vez permitieron retroalimentar el proyecto.

En base a lo anterior se plantea el desafío de crear una interfaz gráfica de usuario (GUI) la cual pueda ser utilizada sin la necesidad de conocer como programar en Python u otro tipo de lenguaje de programación, entregando resultados aún más simples e interactivos a través de herramientas visuales mejoradas e intuitivas. Paralelamente se propone mejorar la calidad de los resultados al analizar un número mayor de ordenes de trabajo, aplicar nuevas métricas más específicas y mejorar la fase de pre-procesamiento, incorporando técnicas focalizadas como la identificación de entidades y otros métodos asociados.

Finalmente y debido a que el sistema de algoritmos es capaz de procesar todo tipo de campos de texto no estructurado, se deja abierta la posibilidad de replicar este sistema en otras áreas de la empresa, tales como Marketing, Finanzas, Recursos Humanos y Atención al Cliente, donde no necesariamente los campos de texto deben estar contenidos en una base de datos como lo fue para este proyecto, sino que pueden estar presentes en correos electrónicos, redes sociales, documentos en formato Word, o cualquier otro tipo de formato asociado con texto, donde la única diferencia en comparación con lo realizado en este informe estaría en la forma de recolectar y procesar la información.

6. Referencias

Berry, M. W., & Kogan, J. (2010). Survey of Text Visualization Techniques. En *Text Mining: Applications and Theory* (págs. 107-129). Wiley.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research, 3*, 993-1022.

Campbell, R., Pentz, E., & Borthwick, I. (2012). Metadata and Text Mining. En *Academic and Professional Publishing* (págs. 278-279). Chandos Publishing.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *CRISP-DM 1.0: Step-by-step data mining guide.* SPSS.

Contreras Piña, C. (2014). *Extracción de conocimiento nuevo desde los reclamos recibidos en el Servicio Nacional del Consumidor mediante técnicas de Text Mining.* Universidad de Chile, Santiago.

Dabbura, I. (17 de Septiembre de 2018). *K-means Clustering: Algorithm, Applications, Evaluation Methods, and Drawbacks*. Obtenido de Towards Data Science: https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a

Fattori, M., Pedrazzi, G., & Turra, R. (2003). Text mining applied to patent mapping: A practical business case. En *World Patent Information* (4 ed., Vol. 25, págs. 335-342).

Feldman, R., & Dagan, I. (1995). Knowledge discovery in texts. En *KDD'95: Proceeding of the First International Conference on Knowledge Discovery and Data Mining* (págs. 112-117). Montreal: AAAI Press.

Han, J., Kamber, M., & Pei, J. (2012). Cluster Analysis: Basic Concepts and Methods. En *Data Mining: Concepts and Techniques* (3rd ed., págs. 443-494). Morgan Kaufmann.

Hotho, A., Nurnberger, A., & Paaß, G. (Mayo de 2005). A Brief Survey of Text Mining. *LDV Forum, 20*, 19-62.

Kodinariya, T. M., & Makwana, P. R. (Noviembre de 2013). Review on determining number of Cluster in K-Means Clustering. *International Journal of Advance Research in Computer Science and Management Studies, 1*(6), 90-95.

MacQueen, J. (1967). Some Methods for Classification and Analysis of Multivariate Observations. En L. M. Le Cam, & J. Neyman, *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (págs. 281-299). University of California Press.

Perone, C. S. (12 de Septiembre de 2013). *Machine Learning : Cosine Similarity for Vector Space Models*. Obtenido de Terra Incognita: http://pyevolve.sourceforge.net/wordpress/?p=2497

Satopaa, V., Albrecht, J., Irwin, D., & Raghavan, B. (2011). Finding a “Kneedle” in a Haystack: Detecting Knee Points in System Behavior. En *Proceedings of the 2011 31st International Conference on Distributed Computing Systems Workshops* (págs. 166–171). IEEE Computer Society. Obtenido de https://doi.org/10.1109/ICDCSW.2011.20

Sepúlveda Ramirez, M. F. (2019). *Herramientas de Analítica Visual para Modelos de Tópicos sobre Colecciones de Documentos.* Pontificia Universidad Católica de Chile, Santiago.

SGS SA. (s.f.). *SGS en resumen*. Recuperado el 27 de Septiembre de 2019, de SGS: https://www.sgs.cl/es-es/our-company/about-sgs/sgs-in-brief

Shanie, T., Suprijadi, J., & Zulhanif. (2017). Text grouping in patent analysis using adaptive K-means clustering algorithm. *American Institute of Physics Conference Series. 1827.* AIP Publishing. doi:https://doi.org/10.1063/1.4979457

Syed, S., & Spruit, M. (2017). Full-Text or Abstract? Examining Topic Coherence Scores Using Latent Dirichlet Allocation. *2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA)* (págs. 165-174). Tokyo: IEEE.