ANÁLISIS Y APLICACIÓN DE TEXT MINING EN BASE DE DATOS SECTOR MANTENCIÓN Y REPARACIÓN OGC (07/07/18 – 31/08/19)

Alexander R. Ulloa Opazo
01/07/2020

RESUMEN

Al tener una gran gamma de tareas, la administración de Mantención y Reparación de OGC (Oil, Gas and Chemicals) ha tenido serios problemas con la gestión de información, pues para controlar las órdenes de trabajo (OT) se debe abrir un libro de Excel de 35 columnas y 23.000 registros, siendo *"Observación"* uno de los atributos más importantes, en el cual los técnicos registran en detalle el trabajo realizado en formato texto; por lo que si el administrador quiere ver el desempeño de una estación, tendría que filtrar la estación y leer cada una de sus observaciones, que pueden ser hasta 450.

A partir de este problema se decidió aplicar herramientas de aprendizaje no supervisado (clustering) pertenecientes a Text Mining, que fueron Topic Modeling y K-Means.

En primer lugar, y luego de limpiar y preparar la base de datos, se ejecutaron 3 modelos probabilísticos en relación con Topic Modeling (LDA Mallet, LDA Sklearn y LDA Gensim), en donde el primero fue el que entregó mejores resultados, revelando 24 tópicos con un nivel de coherencia sobre el 50%, que se interpretaron en conjunto con técnicos y administración. A partir de estos temas se pudo construir elementos visuales que permitieron demostrar cuales eran las fallas y/o actividades principales para cada atributo de una manera general y fácilmente interpretable.

Por otro lado, y debido a que la gerencia necesitaba conocer el detalle de las actividades, se implementó el algoritmo K-Means, el cual a partir de un elemento de un atributo (estación en estaciones , técnico en técnicos, etc) se selecciona el número óptimo de clusters (método del elbow) que luego se visualizan en formato texto, mostrando en primer lugar los clusters representativos y en su interior el conjunto de observaciones que componen cada cluster.

Se concluye que ambos modelos se complementan, puesto que con Topic Modeling (LDA) se pueden identificar patrones generales y fáciles de representar, mientras que con K-Means es posible entender el contexto y el porque de cada falla y/o actividad.

CONTENIDO

ANÁLISIS INICIAL DE BASE DE DATOS

ATRIBUTOS ACTUALES

Se procesó una base de datos con 35 atributos y 23.258 registros.

Cantidad de valores únicos

Recepción OT	350
Hora Recep.	768
OS	23258
OT	22208
Estacion	1013
Tipo Estación	3
Ubicación Técnica	1164
Origen del Servicio	8
Region	15
Prioridad SGS	15
Prioridad Cliente	14
Tecnico Asignado	79
Estado OT	4
Condición	12
Tipo de Falla	66
Observación	18516
Requerimiento	17070
Técnico Cierre	75
TDR	8991
TDS	10379
Fecha Arribo	462
Hora Arribo	1289
Fecha Cierre	455
Hora Cierre	1231
Cod. Repuesto	1
Repuesto	1
Cnt	1
Monto	1
Total	1
Ltr 93	434
Ltr 95	262
Ltr 97	346
Ltr D	394
Ltr K	104
Ind. Gestion	4

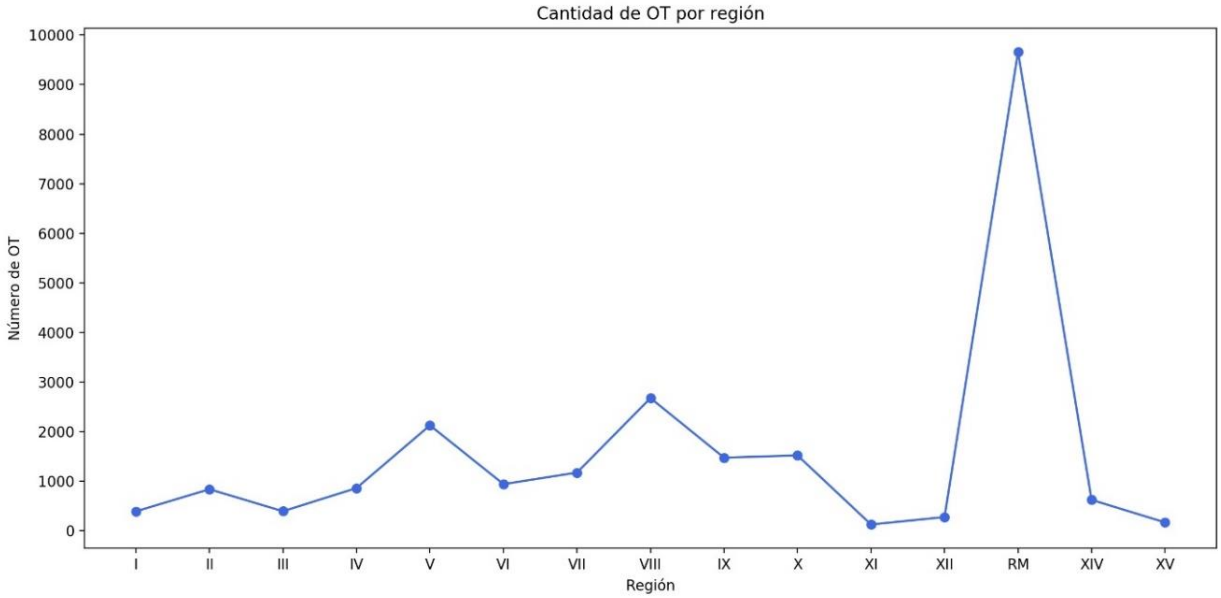
Cantidad de valores nulos

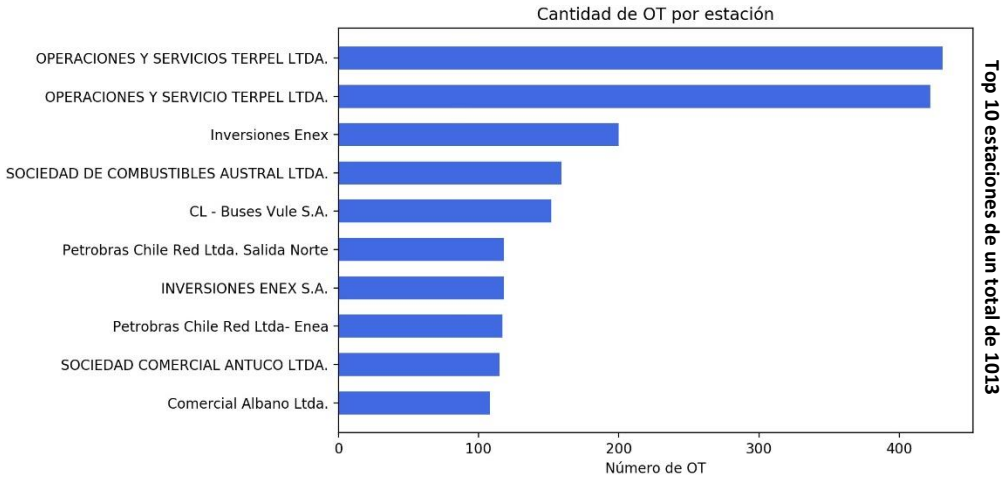
Recepción OT	0
Hora Recep.	0
OS	0
OT	1044
Estacion	0
Tipo Estación	2644
Ubicación Técnica	0
Origen del Servicio	0
Region	0
Prioridad SGS	7187
Prioridad Cliente	0
Tecnico Asignado	0
Estado OT	0
Condición	0
Tipo de Falla	0
Observación	1418
Requerimiento	0
Técnico Cierre	0
TDR	0
TDS	0
Fecha Arribo	1582
Hora Arribo	1582
Fecha Cierre	1953
Hora Cierre	1953
Cod. Repuesto	23258
Repuesto	23258
Cnt	23258
Monto	23258
Total	23258
Ltr 93	0
Ltr 95	0
Ltr 97	0
Ltr D	0
Ltr K	0
Ind. Gestion	13423

A partir del resultado de ambas tablas y de acuerdo con el criterio de gerencia fue posible reducir y transformar variables a : ***Region, Estacion, Tipo de Falla, Prioridad Cliente, Estado OT[1], Tecnico Asignado, Dia, Mes, Hora, Rango Recep/Arribo, Rango Arribo/Cierre, Observación*** y ***Requerimiento***; siendo estas dos últimas las columnas que contienen texto no estructurado (Observación hecha por técnicos y Requerimiento por parte de los clientes)

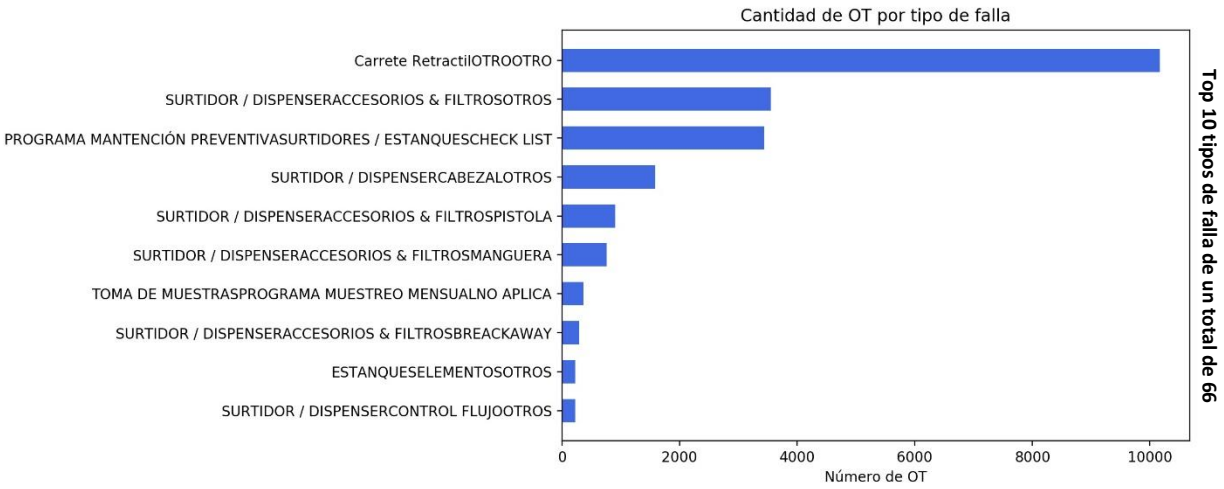
[1] OT: Orden de Trabajo

EXTRACCIÓN DE INFORMACIÓN

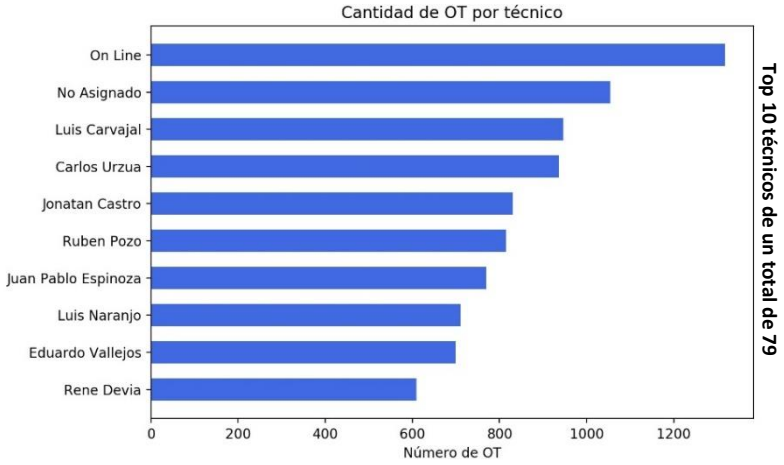


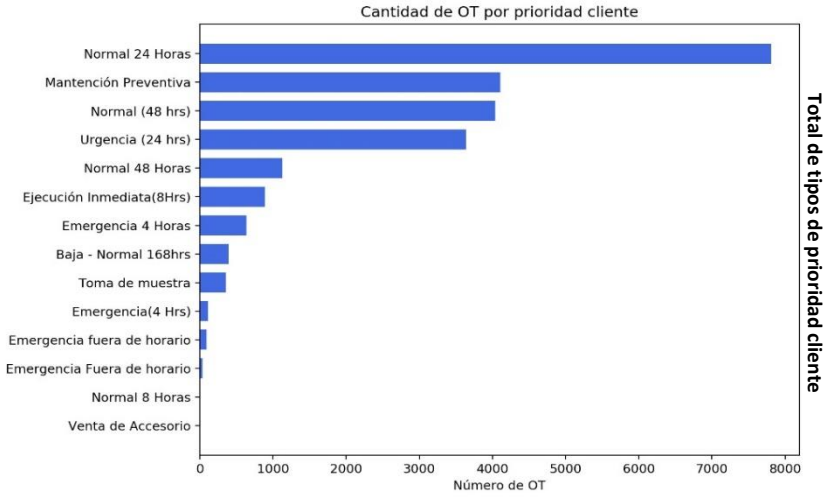


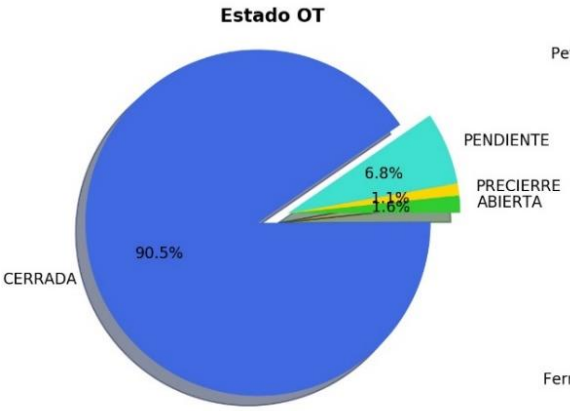
Top 10 estaciones de un total de 1013



Top 10 tipos de falla de un total de 66





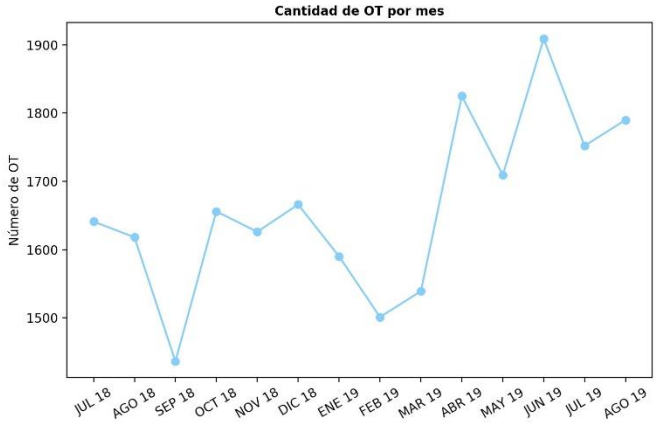


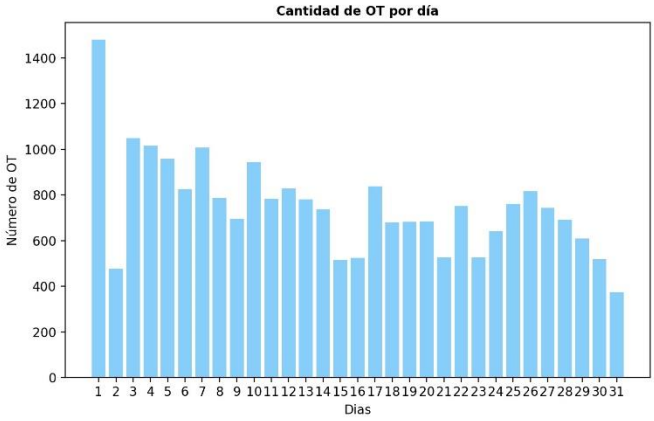
Se detectan estaciones con similar nombre, debido a que los técnicos relacionan un tipo particular de escritura con una región en específico. Por ejemplo:

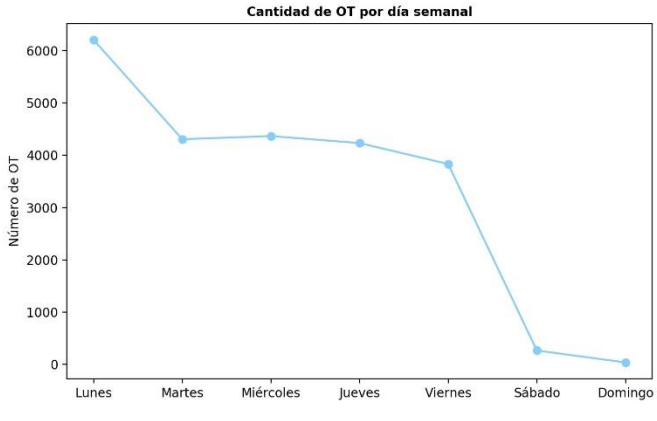
OPERACIONES Y SERVICIO**S** TERPEL LTDA (Santiago) =/= OPERACIONES Y SERVICIO TERPEL LTDA. (Regiones).

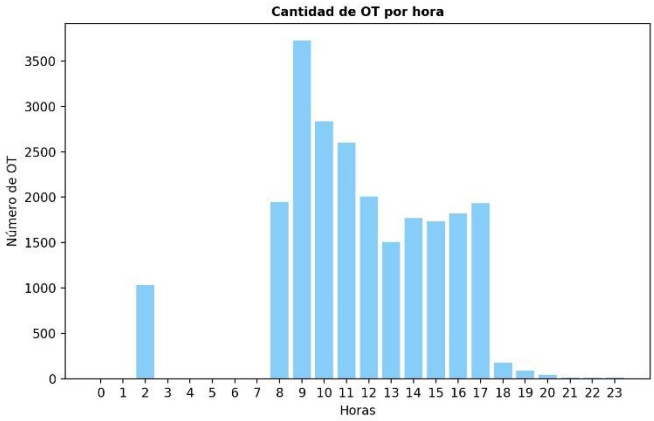
Diferencia de una 'S' en servicios y un punto al final.

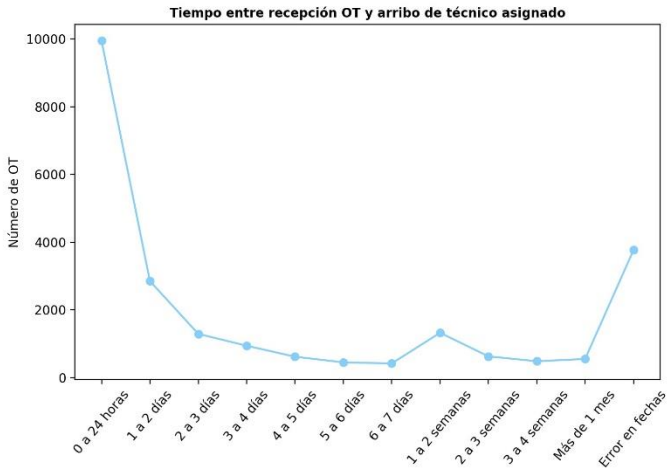
EVOLUCIÓN DE LAS ORDENES DE TRABAJO EN EL TIEMPO

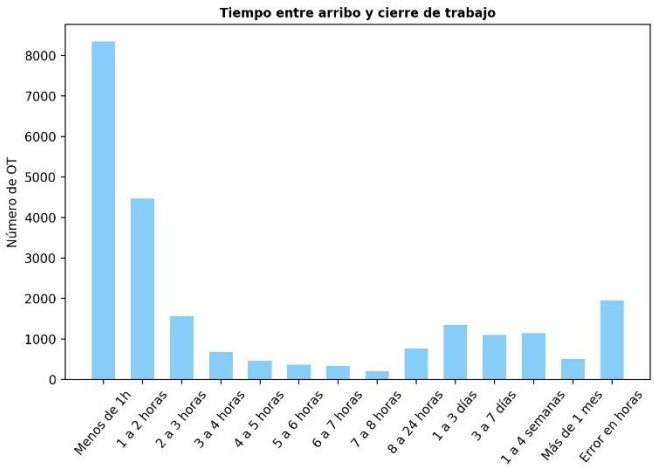












APLICACIÓN DE TEXT MINING

1. TOPIC MODELING – ALGORITMO LDA

Topic Modeling (o modelamiento de tópicos), perteneciente al aprendizaje no supervisado, es uno de los algoritmos de clustering más populares, permitiendo encontrar temas "ocultos" en largos volúmenes de texto, gracias a la aplicación de modelos probabilísticos.

Latent Dirichlet Allocation (LDA) corresponde a uno de esos modelos matemáticos antes mencionado , el cual ayuda a encontrar dichos tópicos por medio de 2 puntos: Localización de la mezcla de palabras que se asocia con cada tópico ($P[word \mid topics]$) , y determinación de la mezcla de tópicos que sirven para describir cada documento ($P[topic \mid documents]$)

Para el caso en estudio, solo se procesó la columna *"Observación"*, puesto que aquí es donde los técnicos describen en detalle la situación, a diferencia de *"Requerimiento"*, donde el registro es hecho por los clientes, siendo muchas veces genérico y poco útil al no explicar el problema real.

Se aplica eliminación de puntuación y patrones no deseados (por medio de Regex[2]) *Tokenization*[3], Eliminación de *Stopwords*[4] y *Lemmatization*[5] como parte de la limpieza y preprocesamiento de texto.

Original

' Se chequea falla, y se detecta Ortirak que cuando le da el sol directo se borra por com pleto, esto no permite el corrrecto uso del mismo.\nSe informa y solicita reparaciion a empresa encargada!

RegEx + 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']

Stopwords

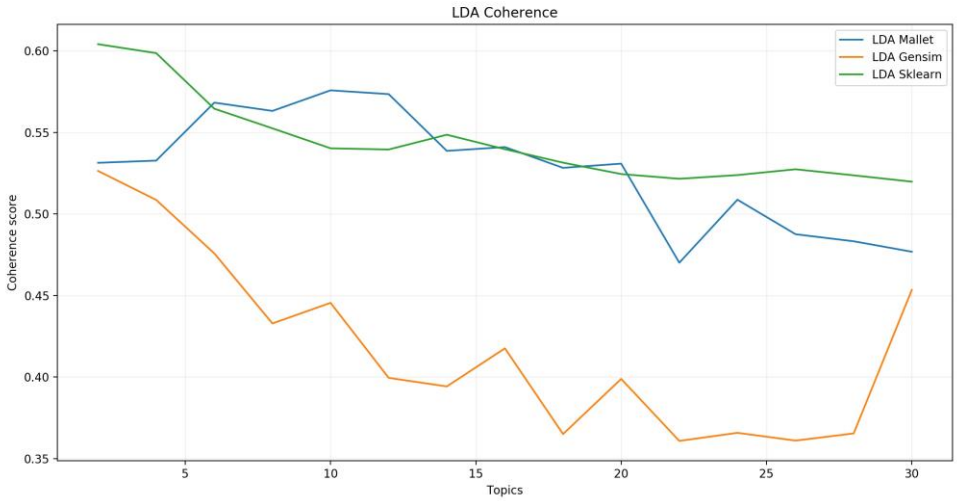
['falla', 'ortirak', 'sol', 'directo', 'borra', 'completo', 'permite', 'correcto', 'uso', 'solici ta', 'reparacion', 'empresa', 'encargada']

Lemmatization ['NOUN', 'ADJ', 'VERB', 'ADV','PROPN']

['fallo', 'ortirak', 'sol', 'directo', 'borrar', 'completar', 'permitir', 'correcto', 'usar', 'solicitar', 'repararacion', 'empresa', 'encargar']

[2] *Regular Expression*. Secuencia de caracteres que define patrones de búsqueda (ej. Fechas - dd/mm/yy)
[3] *Tokenization.* Dividir el documento (observación) en las palabras (o tokens) que lo componen
[4] Stopwords. Conjunto de palabras frecuentes que no agregan valor (ej. El, la los las, un, una,etc)
[5] *Lemmatization*. Reducción de las palabras a su forma más normal y representativa

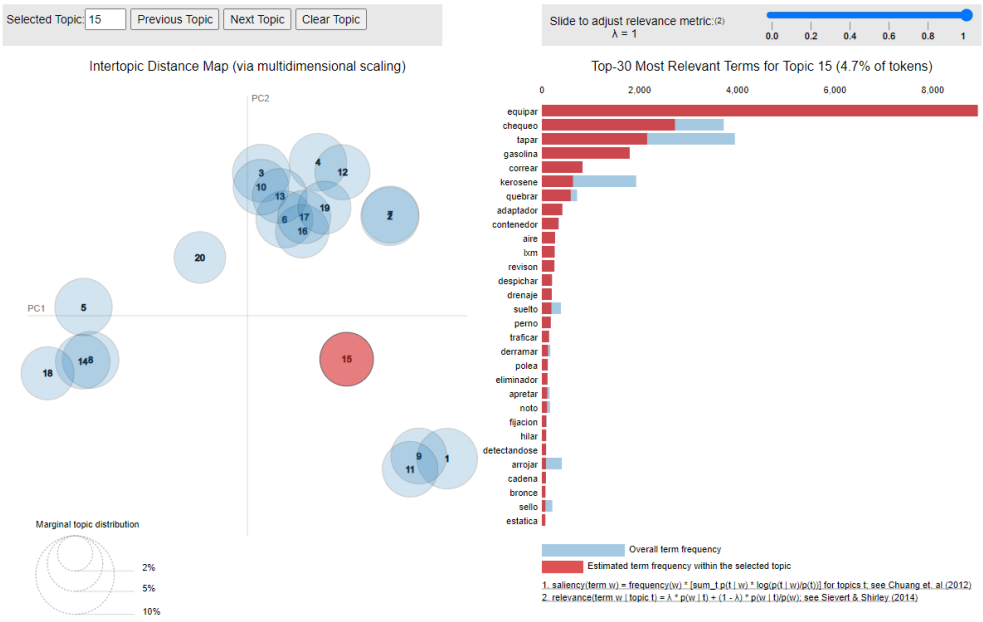
Para conseguir el mejor resultado, se compara 3 modelos LDA, de 3 paquetes diferentes, que son LDA Gensim, LDA Scikit-learn y LDA Mallet[6], donde cada uno de ellos sigue los mismos principios básicos de probabilidad, pero con diferencias en la calidad y coherencia en el producto final.



Se establecen 3 criterios para seleccionar el mejor modelo:

1. Nivel de coherencia (gráfico de arriba)
2. Calidad de las palabras
3. Calidad de los tópicos

A pesar de que el modelo de Scikit-Learn obtuvo mejores niveles de coherencia, a nivel global LDA Mallet generó resultados más adecuados a la empresa, por lo que a través de este modelo se logra identificar 24 tópicos, con un nivel de coherencia del 50,87 %



[6] *MAchine Learning for LanguagE Toolkit*

Texto representativo en 24 tópicos – Word Cloud

Texto representativo en 24 tópicos – Probabilidades

Tópicos del 0 a 11

```
[(0,
  '0.071*"bombo" + 0.065*"motor" + 0.042*"electrico" + 0.040*"tablero" + '
  '0.022*"normalizar" + 0.019*"cabezal" + 0.019*"sumergir" + 0.018*"electrica" '
  '+ 0.017*"cargar" + 0.016*"turbinar" + 0.014*"rele" + 0.013*"voltaje" + '
  '0.013*"medir" + 0.012*"cable" + 0.011*"voltio"'),
 (1,
  '0.080*"tk" + 0.070*"reglar" + 0.069*"root" + 0.067*"veeder" + 0.048*"sump" '
  '+ 0.048*"sensor" + 0.036*"medicion" + 0.034*"consola" + 0.032*"alarmar" + '
  '0.022*"tank" + 0.019*"dispenser" + 0.017*"tabla" + 0.014*"estanque" + '
  '0.013*"diferenciar" + 0.012*"telemedicion"'),
 (2,
  '0.091*"placa" + 0.050*"preset" + 0.042*"teclado" + 0.039*"sistema" + '
  '0.028*"bloquear" + 0.027*"control" + 0.026*"display" + 0.022*"orpak" + '
  '0.020*"encore" + 0.019*"gilbarco" + 0.018*"tarjeta" + 0.016*"personal" + '
  '0.016*"hidraulica" + 0.015*"igem" + 0.014*"comunicacion"'),
 (3,
  '0.172*"despachar" + 0.132*"venta" + 0.131*"normal" + 0.113*"isla" + '
  '0.074*"cliente" + 0.070*"cortar" + 0.067*"breakaway" + 0.041*"formar" + '
  '0.019*"publicar" + 0.017*"reconectable" + 0.015*"normalidad" + '
  '0.014*"condicion" + 0.013*"aire" + 0.009*"otcerrada" + 0.007*"corto"'),
 (4,
  '0.066*"incidencia" + 0.066*"cerrar" + 0.050*"ot" + 0.049*"trabajar" + '
  '0.047*"informar" + 0.045*"servicio" + 0.041*"encargar" + '
  '0.027*"administrador" + 0.020*"corresponder" + 0.020*"cotizacion" + '
  '0.017*"dia" + 0.015*"orden" + 0.012*"ltr" + 0.012*"autorizar" + '
  '0.012*"esperar"'),
 (5,
  '0.510*"dispensador" + 0.135*"medidor" + 0.103*"cambiar" + 0.043*"gasolina" '
  '+ 0.037*"chequeo" + 0.032*"advantage" + 0.021*"gilbarco" + 0.020*"encore" + '
  '0.020*"banco" + 0.011*"trabar" + 0.008*"trancar" + 0.005*"terminar" + '
  '0.004*"volumetricas" + 0.004*"manifold" + 0.002*"multiproducto"'),
```

```
 (6,
  '0.487*"producto" + 0.168*"prueba" + 0.073*"chequeo" + 0.048*"recircula" + '
  '0.043*"recirculacion" + 0.038*"despachar" + 0.032*"presentar" + '
  '0.015*"presenciar" + 0.010*"administracion" + 0.009*"pdiesel" + '
  '0.008*"proceder" + 0.007*"normalidad" + 0.007*"trabajo" + 0.005*"recovery" '
  '+ 0.004*"variar"'),
 (7,
  '0.473*"equipar" + 0.033*"ajustar" + 0.027*"palanca" + 0.022*"reparar" + '
  '0.021*"partir" + 0.020*"suelto" + 0.020*"soltar" + 0.016*"eliminar" + '
  '0.015*"accionamiento" + 0.014*"revison" + 0.013*"boto" + 0.013*"portar" + '
  '0.011*"breackaway" + 0.010*"apoyar" + 0.009*"perno"'),
 (8,
  '0.165*"tapar" + 0.064*"combustible" + 0.056*"descargar" + 0.043*"sellar" + '
  '0.036*"quebrar" + 0.030*"instalar" + 0.030*"nuevo" + 0.021*"adaptador" + '
  '0.020*"foto" + 0.018*"instalacion" + 0.017*"contenedor" + 0.016*"spill" + '
  '0.013*"adjuntar" + 0.011*"sello" + 0.010*"despichar"'),
 (9,
  '0.188*"fallo" + 0.157*"lado" + 0.154*"probar" + 0.153*"detectar" + '
  '0.044*"operar" + 0.031*"cambiar" + 0.028*"recirculando" + 0.022*"malo" + '
  '0.018*"pulser" + 0.016*"normalmente" + 0.014*"describir" + 0.013*"octuple" '
  '+ 0.011*"breakway" + 0.011*"doblar" + 0.008*"reconecta"'),
 (10,
  '0.120*"fugar" + 0.066*"detector" + 0.060*"linear" + 0.047*"presion" + '
  '0.045*"instalar" + 0.032*"adhesivo" + 0.029*"probar" + 0.027*"seriar" + '
  '0.025*"tk" + 0.024*"diesel" + 0.023*"psi" + 0.020*"instalacion" + '
  '0.018*"levantar" + 0.015*"check" + 0.014*"prueba"'),
 (11,
  '0.156*"aguar" + 0.144*"estancar" + 0.128*"litro" + 0.106*"retirar" + '
  '0.039*"tambor" + 0.034*"lts" + 0.028*"estanque" + 0.027*"camara" + '
  '0.017*"contaminar" + 0.015*"servicio" + 0.014*"diesel" + 0.014*"ltrs" + '
  '0.013*"almacenar" + 0.011*"sacar" + 0.011*"tk"'),
```

Tópicos del 11 a 23

```
(12,
 '0.260*"cambiar" + 0.147*"manguera" + 0.105*"reponer" + 0.045*"codigo" + '
 '0.032*"rotar" + 0.029*"bodega" + 0.024*"danada" + 0.024*"repuesto" + '
 '0.023*"utilizar" + 0.021*"danado" + 0.020*"desgastar" + 0.020*"vestir" + '
 '0.019*"pulgada" + 0.018*"posicion" + 0.014*"mts"'),
(13,
 '0.241*"disp" + 0.092*"remplazar" + 0.065*"surt" + 0.049*"opw" + '
 '0.032*"mostrar" + 0.025*"est" + 0.022*"conv" + 0.020*"dsl" + 0.015*"hose" + '
 '0.015*"lxm" + 0.014*"coordinar" + 0.014*"breckawey" + 0.014*"hacer" + '
 '0.014*"reinstalar" + 0.013*"proxima"'),
(14,
 '0.156*"valvula" + 0.085*"cambiar" + 0.052*"reparacion" + 0.052*"valvulas" + '
 '0.046*"proporcional" + 0.040*"oring" + 0.038*"kit" + 0.035*"wayne" + '
 '0.027*"filtracion" + 0.020*"centrifugar" + 0.019*"reparar" + 0.017*"check" '
 '+ 0.017*"union" + 0.015*"wip" + 0.014*"defectuoso"'),
(15,
 '0.128*"mantencion" + 0.089*"preventivo" + 0.082*"calibracion" + '
 '0.068*"limpieza" + 0.056*"verificacion" + 0.050*"spill" + '
 '0.050*"volumetrica" + 0.037*"mes" + 0.031*"litro" + 0.027*"matraz" + '
 '0.020*"programar" + 0.020*"pintar" + 0.020*"verificar" + 0.019*"accesorio" '
 '+ 0.019*"efectua"'),
(16,
 '0.186*"cambiar" + 0.119*"filtro" + 0.115*"filtrar" + 0.107*"flujo" + '
 '0.053*"lento" + 0.051*"alto" + 0.051*"ampolleta" + 0.035*"quemar" + '
 '0.023*"lpm" + 0.021*"entregar" + 0.019*"capacidad" + 0.018*"pls" + '
 '0.015*"saturar" + 0.015*"despues" + 0.014*"bajo"'),
(17,
 '0.062*"verificar" + 0.047*"venta" + 0.037*"presentar" + 0.031*"programar" + '
 '0.031*"fallo" + 0.029*"observar" + 0.028*"dejar" + 0.027*"error" + '
 '0.026*"hacer" + 0.018*"luego" + 0.018*"reset" + 0.017*"parametros" + '
 '0.016*"master" + 0.014*"anterior" + 0.014*"asistir"'),
```

```
(18,
 '0.049*"chequeo" + 0.044*"boca" + 0.043*"calibracion" + 0.037*"limpieza" + '
 '0.035*"chequean" + 0.034*"interior" + 0.032*"emergencia" + '
 '0.029*"descargar" + 0.026*"parar" + 0.025*"contar" + 0.024*"tapar" + '
 '0.024*"camaras" + 0.021*"pintar" + 0.021*"identificacion" + 0.020*"boton"'),
(19,
 '0.348*"gas" + 0.042*"filtración" + 0.032*"pd" + 0.031*"lts" + '
 '0.029*"servicio" + 0.024*"encontrar" + 0.022*"numeral" + 0.019*"cambiar" + '
 '0.018*"solo" + 0.016*"efectuar" + 0.012*"inspeccion" + 0.012*"posible" + '
 '0.011*"falto" + 0.010*"electronicos" + 0.010*"instalacion"'),
(20,
 '0.358*"diesel" + 0.356*"surtidor" + 0.072*"kerosene" + 0.041*"correar" + '
 '0.023*"limpio" + 0.013*"petrolear" + 0.013*"mp" + 0.013*"revolución" + '
 '0.006*"polea" + 0.005*"cebar" + 0.005*"jh" + 0.005*"cabezal" + '
 '0.005*"elemento" + 0.004*"caro" + 0.004*"comprar"'),
(21,
 '0.167*"mal" + 0.143*"reemplazar" + 0.081*"pos" + 0.076*"fundir" + '
 '0.065*"eds" + 0.053*"usar" + 0.038*"pendiente" + 0.035*"emitir" + '
 '0.034*"accesorio" + 0.031*"ods" + 0.022*"negro" + 0.017*"verde" + '
 '0.016*"power" + 0.016*"spin" + 0.014*"rojo"'),
(22,
 '0.269*"boca" + 0.232*"pistola" + 0.104*"cambiar" + 0.067*"filtracion" + '
 '0.043*"swivel" + 0.032*"gasolina" + 0.026*"destorcedor" + 0.015*"vida" + '
 '0.015*"util" + 0.014*"gatillo" + 0.013*"convencional" + 0.012*"pd" + '
 '0.011*"cano" + 0.011*"wipehose" + 0.011*"cortar"'),
(23,
 '0.134*"boca" + 0.128*"calibrar" + 0.120*"equipo" + 0.092*"preventivo" + '
 '0.059*"mantencion" + 0.050*"limpiar" + 0.042*"devolver" + '
 '0.038*"recirculado" + 0.028*"total" + 0.028*"producto" + 0.026*"tdf" + '
 '0.024*"eds" + 0.023*"mantenimiento" + 0.023*"tk" + 0.021*"post"')]
```

Interpretación inicial de tópicos

Para lograr interpretar los tópicos es necesario ver las probabilidades de ocurrencia de palabras (tablas anteriores), en donde al interior de cada tema se muestran en orden descendiente las 15 palabras que mejor representan al tópico (en el caso de Word Cloud , las que tienen mayor probabilidad aparecen más grande)

Por ejemplo en el tópico 3 se tiene *"despachar", "venta" , "normal", "isla"* y *"cliente* como sus primeras 5 palabras, por lo que es un buen indicio que el tema corresponde a despacho en ventas y atención a clientes

- **Tópico 0:** Bomba / Motor / Problemas eléctricos (tablero, conector, cortes, etc.)
- **Tópico 1:** Veeder root (consola, alarma, sensor, etc.)
- **Tópico 2:** Teclados preset/programación y cambios de placas.
- **Tópico 3:** Despacho venta cliente.
- **Tópico 4:** Detalles en incidencia (atendida normal bajo ot / no pertenece a sgs / atendida por otro contratista,etc)
- **Tópico 5:** Medidores (Wayne, advantage, etc.)
- **Tópico 6:** Revisión, chequeo y pruebas en equipos, surtidores y dispensadores.
- **Tópico 7:** Equipo operativo (ajustes y cambios menores)
- **Tópico 8:** Tapas de estanque / descarga.
- **Tópico 9:** Fallas generales (lados, breakaway,octuple,uniones, etc.)
- **Tópico 10:** Detector de fuga / Presión en líneas / Adhesivos de seguridad.
- **Tópico 11:** Retiro de agua.
- **Tópico 12:** Mangueras.
- **Tópico 13:** Toma/recepción/entrega de muestras.

- **Tópico 14:** Válvulas.
- **Tópico 15:** Mantención preventiva.
- **Tópico 16:** Filtros/ Flujo lento / Ampolletas.
- **Tópico 17:** Master reset / Reprogramación de parámetros.
- **Tópico 18:** Calibración y limpieza de surtidores (vinculado a mantención preventiva)
- **Tópico 19:** Equipos de gas 93/95/97 y problemas relacionados como verificación volumétrica, medición, cambio o reajuste en correas, etc.
- **Tópico 20:** Equipos de diesel/kerosene y problemas relacionados como verificación volumétrica, cambio o reajuste en correas, etc.
- **Tópico 21:** ODS/ OT mal emitidas, asignadas, duplicadas, etc.
- **Tópico 22:** Pistola / Swivel.
- **Tópico 23:** Calibración de bocas (vinculada en parte a mantenimiento preventivo)

Una vez identificados los tópicos, es mucho más fácil la obtención e interpretación de resultados.

¿Cómo se determina que una observación corresponde a cierto tópico?

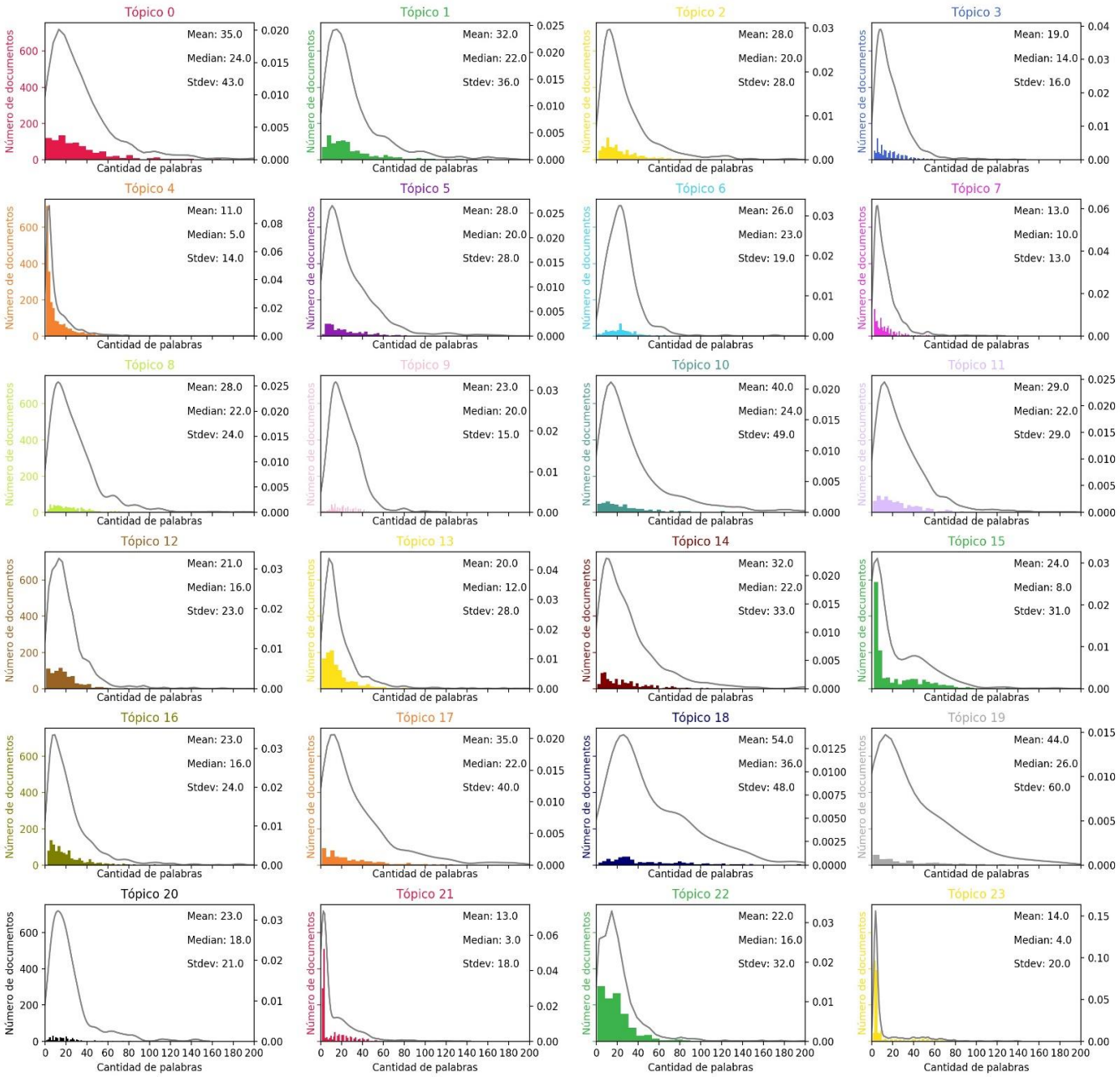
A modo de aclaración, el hecho de que a una observación se le asigne un tópico, quiere decir que dicha observación se encuentra compuesta por múltiples tópicos, pero sólo uno es capaz de explicar en mayor proporción el problema o actividad principal, por lo que se le denomina ***tópico dominante.***

Un ejemplo del concepto de tópico dominante se ve la observación:

"Se calibran 20 de 20 bocas en mantención preventiva, también se cambia manguera rota "

Donde se identifican a los tópicos 23 (calibración bocas), 15 (mantención preventiva) y 12 (mangueras), siendo finalmente el tópico 23 el **dominante.**

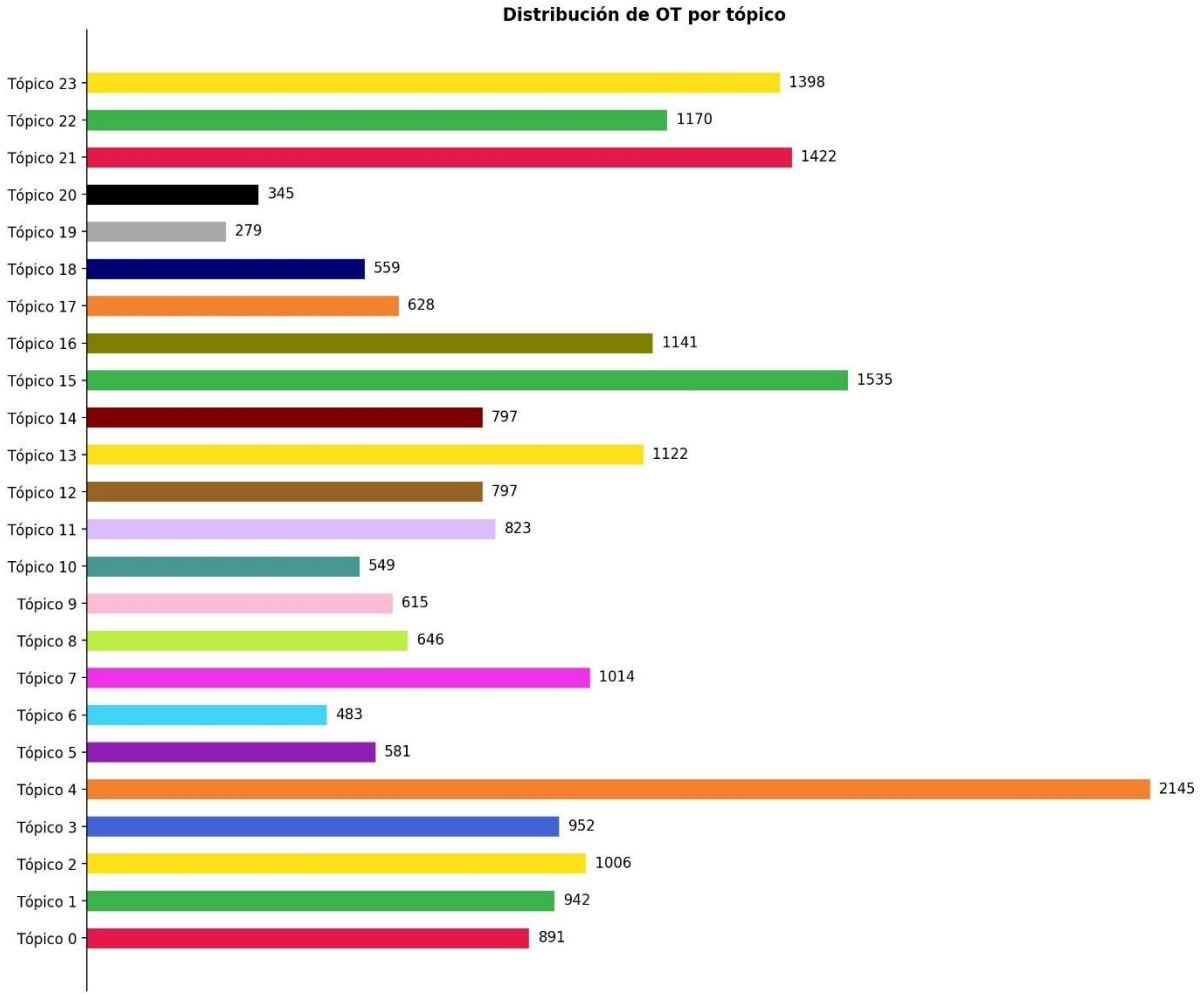
Distribución de la cantidad de palabras por tópico



Esto tiene mucho sentido, pues el **tópico 21 (ods mal emitida)** registra la menor cantidad de palabras por oración, con un promedio de 13 ya que no se requiere mucho detalle para describir ese problema, mientras que el **tópico 19 (actividades en estanques 93/95/97)** requiere describir en mayor grado la situación por lo que se tiene en promedio 44 palabras por oración.

Resultados del algoritmo LDA

Analizando la base de datos completa en los periodos de tiempo establecidos, se obtiene la siguiente distribución de tópicos:



El **tópico 4** (incidencia atendida/ detalle incidencia) es lo más registrado, seguido por el **tópico 15** (mantención preventiva) , **tópico 21** (ods[7] mal emitida/asignada) y el tópico 23 (calibración de bocas) por nombrar algunos.

Por otro lado se identifica que actividades específicas relacionadas con estanques de gas 93/95/97, diesel y kerosene, como verificación volumétrica y cambio en correas (**tópico 19** y **20**), son las que menos observaciones asociadas tienen.

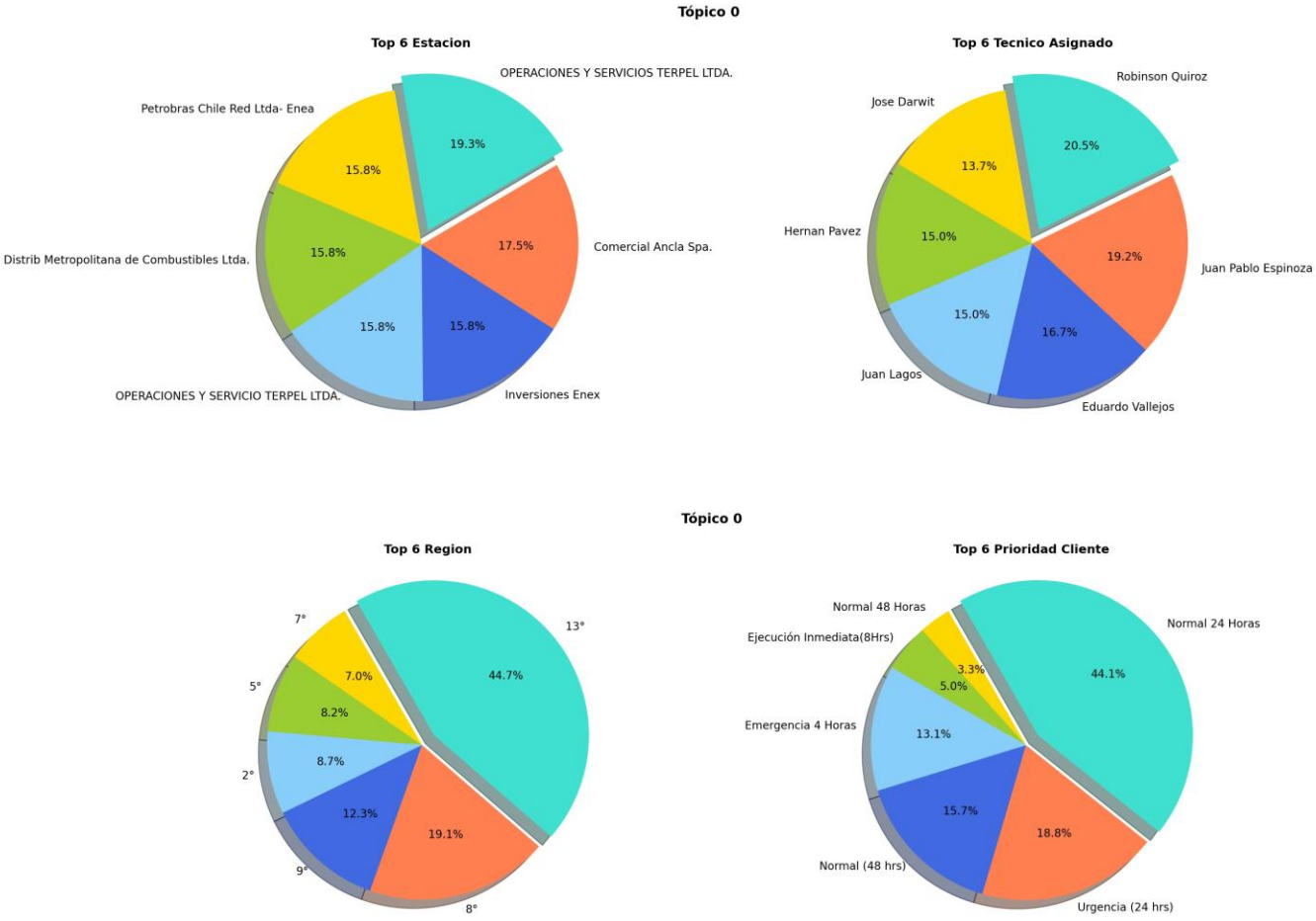
[7] ODS = OS = Orden de servicio (Sinónimo de OT)

Análisis del tópico 0 (bomba/motor/problemas eléctricos)

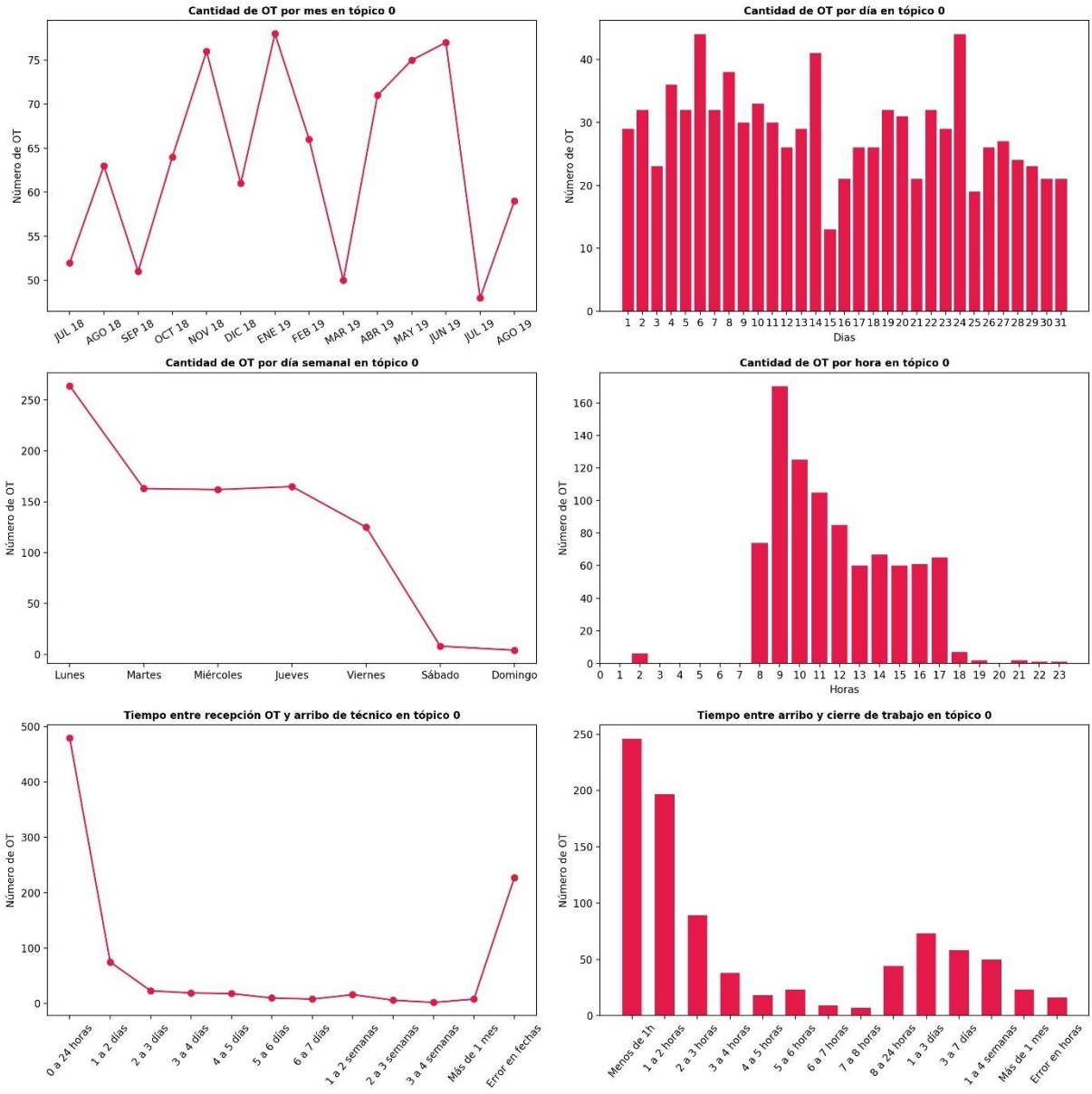
i. Texto representativo

```
Top 10 palabras en Tópico 0
['bombo, motor, electrico, tablero, normalizar, cabezal, sumergir, electrica, cargar, turbinar']

Texto representativo en Tópico 0
[['circuito', 'alumbrar', 'salar', 'arquear', 'curcuito', 'forzar', 'curcuito', 'forzar', 'minemarket', 'sevmide', 'consumir', 'mimen
'consumir', 'fase', 'automatico', 'general', 'circuito', 'subre', 'cargar', 'bajo', 'cortar', 'energia', 'electrica', 'ilumunacion',
'forzar', 'minemarket', 'envio', 'automatico', 'amp', 'amp', 'cambiar', 'automatico', 'ocaciones', 'consumir', 'superar', 'reponer',
r', 'salar', 'arquear', 'curcuito', 'forzar', 'curcuito', 'forzar', 'minemarket', 'sevmide', 'consumir', 'mimento', 'prender', 'herbi
'automatico', 'general', 'circuito', 'subre', 'cargar', 'bajo', 'cortar', 'energia', 'electrica', 'ilumunacion', 'forzar', 'salar', '
et', 'envio', 'automatico', 'amp', 'amp', 'cambiar', 'automatico', 'ocaciones', 'consumir', 'superar', 'reponer', 'comprar', 'zona']]
```

ii. Distribución del tópico 0 en los atributos de la base de datos



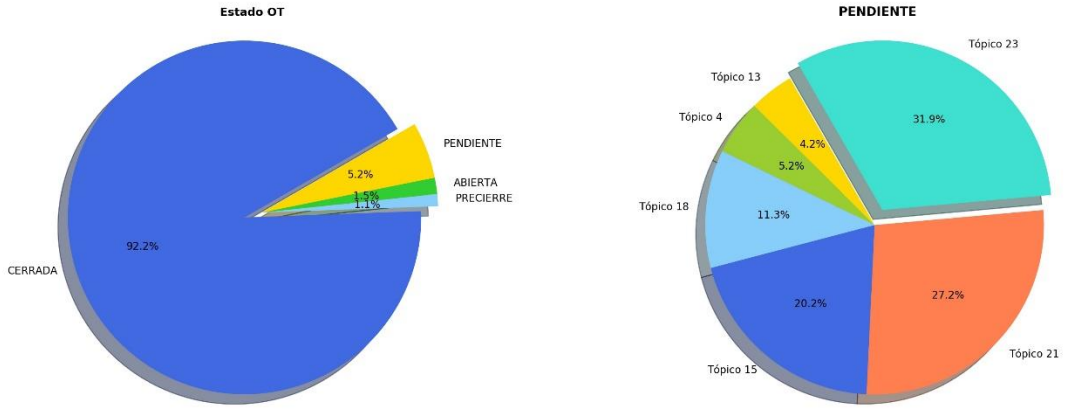
iii. Comportamiento del tópico 0 a través del tiempo



Es posible visualizar que en Noviembre del 2018, Enero del 2019 y Junio del 2019 se registra la mayor cantidad de OT/OS, las que tienden levemente a ocurrir antes del 15 de cada mes, por lo general los días lunes (aunque debido a que no se registra nada los fines de semana, se puede acumular para ese día), ingresándose al sistema entre a 9 AM y 11 AM.

Este tipo de problemas es atendido en cuestión de horas (tiempo entre recepción OT y arribo de técnico) y usualmente el técnico no tarda más de 1 o 2 horas en terminar el trabajo.

Distribución de tópicos en estado OT pendientes

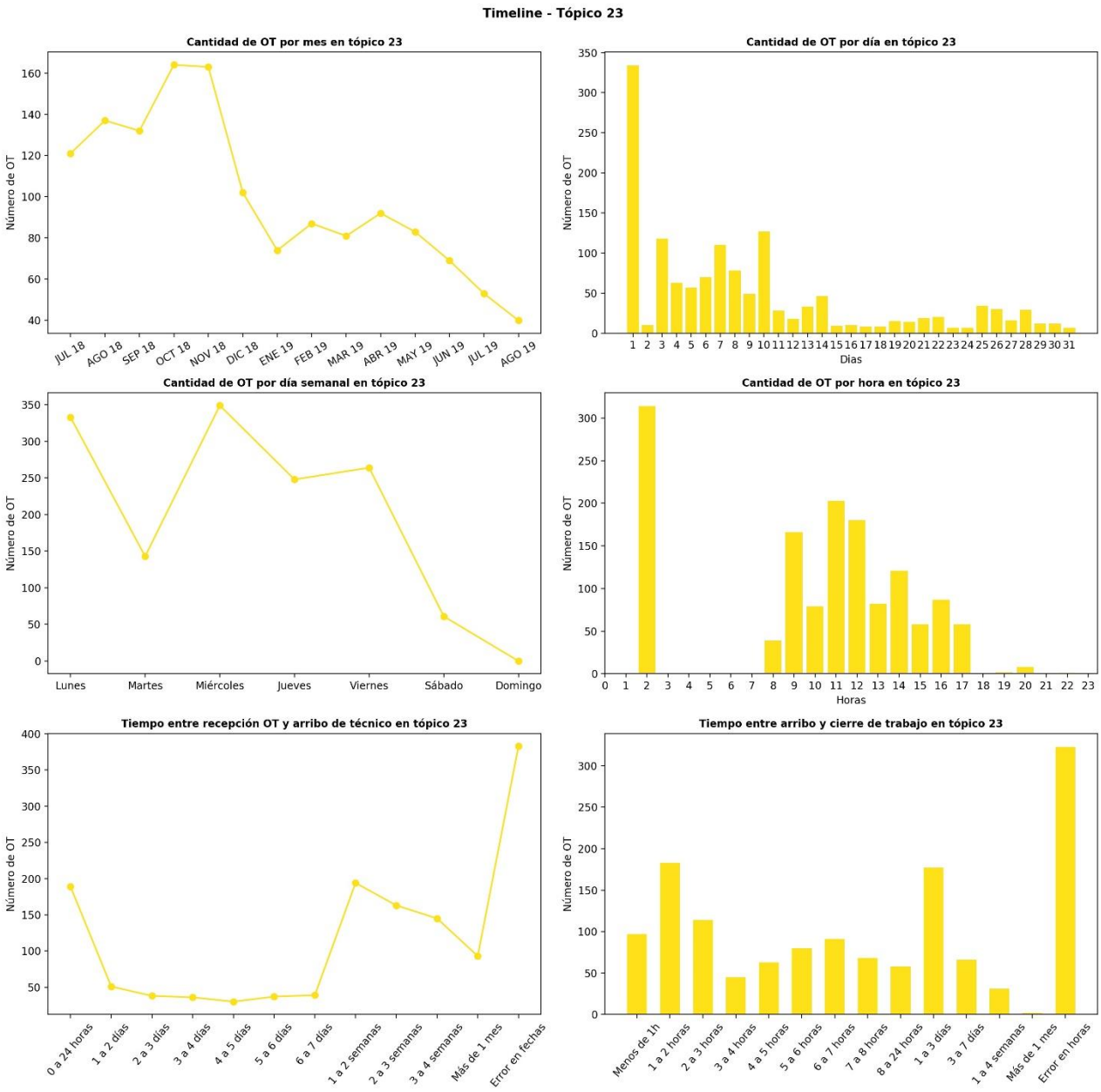


El **tópico 23** (calibración de bocas), **tópico 21** (ods mal emitidas) y **tópico 15** (mantención preventiva) son los registros con más ordenes pendientes. Para ser precisos, estos tres temas explican el 79.3% del total de top 6 registros pendientes.

```
Top 10 palabras en Tópico 23
['boca, calibrar, equipo, preventivo, mantencion, limpiar, devolver, recirculado, total, producto']

Texto representativo en Tópico 23
[['mantenimiento', 'preventivo', 'calibrar', 'boca', 'limpiar', 'equipo', 'producto', 'recirculado',
brar', 'boca', 'limpiar', 'equipo', 'producto', 'recirculado', 'devolver', 'tk', 'equipar', 'tdf', 'ma
eemplazar', 'pistola', 'filtracion', 'producto', 'recirculado', 'devolver', 'tk', 'equipar', 'tdf', 'm
'producto', 'recirculado', 'devolver', 'tk', 'equipar', 'tdf', 'mantenimiento', 'preventivo', 'calibra
'tk', 'equipar', 'tdf', 'mantenimiento', 'preventivo', 'calibrar', 'boca', 'limpio', 'equipar', 'reemp
'tk', 'equipar', 'tdf', 'tks']]
```

Para analizar el comportamiento histórico del mismo **tópico 23** anterior, simplemente basta ejecutar una línea de código en el módulo de visualización, consiguiendo la figura que se muestra en la página siguiente y que se puede interpretar como:

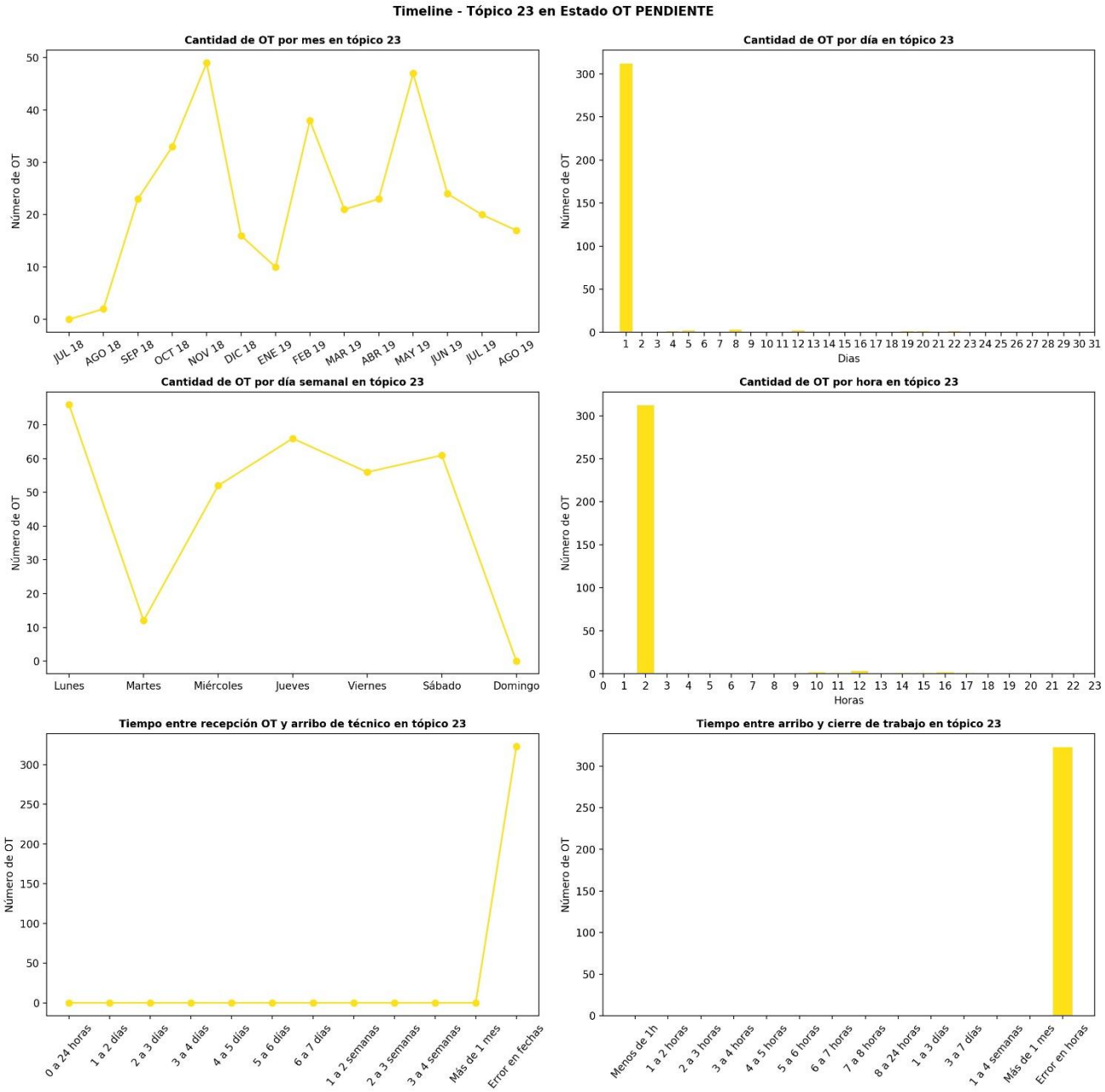
- En los meses de octubre y noviembre del 2018 fue el peak de OT ingresadas relacionadas con calibración de bocas y mantenimiento preventivo.
- OT recepcionadas en su mayoría los primeros 10 días de cada mes.
- Lunes y Miércoles como días en donde se suele registrar en base de datos.
- Entre 11 am y 12 pm se ingresan al sistema, aunque es las 2am cuando se ingresa la mayor cantidad.
- El tiempo entre recepción de OT y arribo de técnico suele ser entre 1 a 2 semanas para las que tienen fechas de recepción y arribo (no nulas), aunque existe una tendencia a que exista error en fechas (dado que nuca se llega a trabajar. En otras palabras tendencia a ordenes pendiente dado que las fechas son nulas)
- Usualmente tardan entre 1 a 2 horas o 1 a 3 días para las ot que no son pendientes



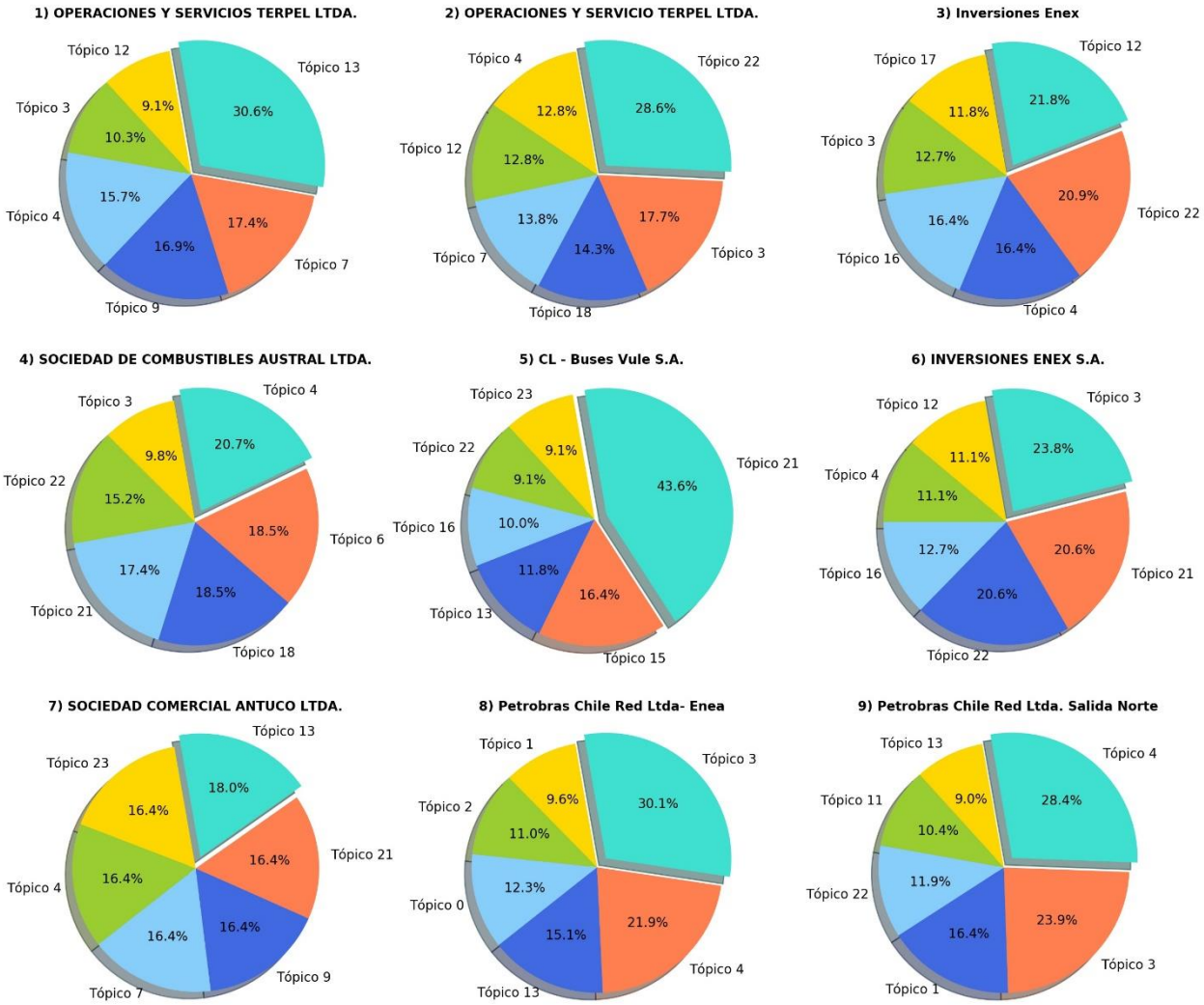
Lo anterior valida que la clasificación de las OT en tópicos de verdad funciona, ya que revela que existe una relación directa entre OT PENDIENTES y CALIBRACIÓN DE BOCAS / MANTENIMIENTO PREVENTIVO, por lo que se concluye que la **aplicación de este algoritmo de modelamiento de tópicos (LDA) es capaz identificar los principales problemas (de manera general) en cualquier atributo,** ya sea estación, técnico asignado, estado ot, etc.

La figura anterior demuestra la evolución del tópico de manera general, es decir se incluyen todas las OT de la base de datos relacionadas al tema en cuestión (tópico 23 - calibración de bocas). Sin embargo, a veces se quiere ver **cómo se desarrolla el tópico en un solo atributo seleccionado** (ya sea ver la evolución del tópico k en la estación K, el tópico y en el técnico Y, etc.)

A continuación se muestra el **comportamiento del tópico 23**, pero esta vez aplicado **exclusivamente a la OT pendientes**.

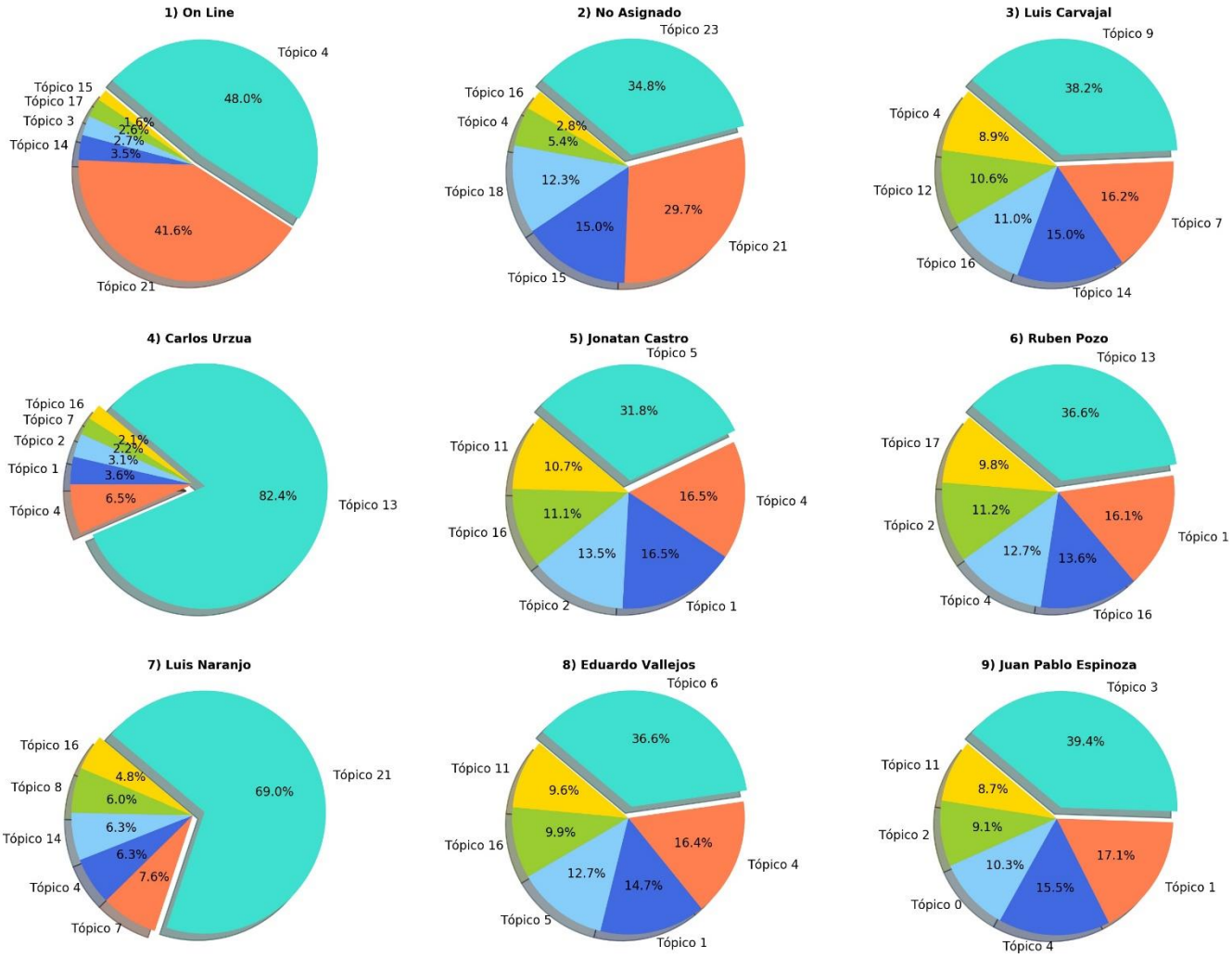


Distribución de tópicos en top 9 estaciones



Gracias al algoritmo se puede clasificar las principales fallas o tareas que tiene cada estación (categorizar fallas) donde al ver la evolución del tópico en el tiempo se pueden crear **estrategias preventivas eficaces que minimicen el impacto del problema en el futuro** (algo similar a un modelo predictivo)

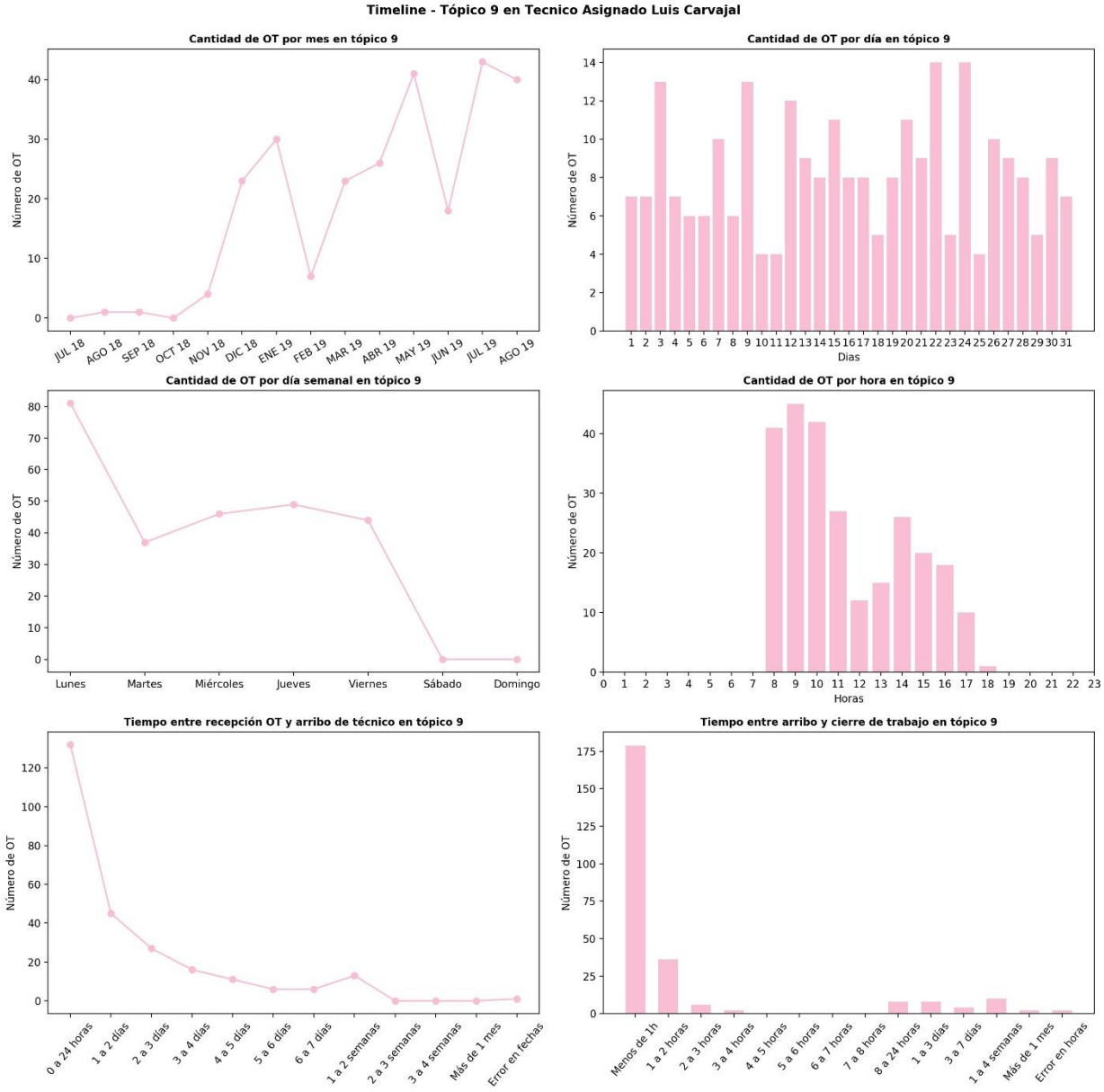
Distribución de tópicos en top 9 técnicos asignados



En base al registro histórico es posible identificar en que área un técnico tiene mayor experiencia y en base a eso asignarlo a un trabajo que mejor se adapte con sus capacidades. En otras palabras, es posible crear un **perfil para cada técnico asignado**.

A modo de demostración, el técnico Luis Carvajal (3° en la figura de arriba) es considerado un “maestro todoterreno” ya que trabaja en fallas de todo tipo (tópico 9 – 38.2%), remplazo de filtros/atención flujo lento (tópico 16 – 13.6%) y el cambio de válvulas (tópico 4 – 12-7%).

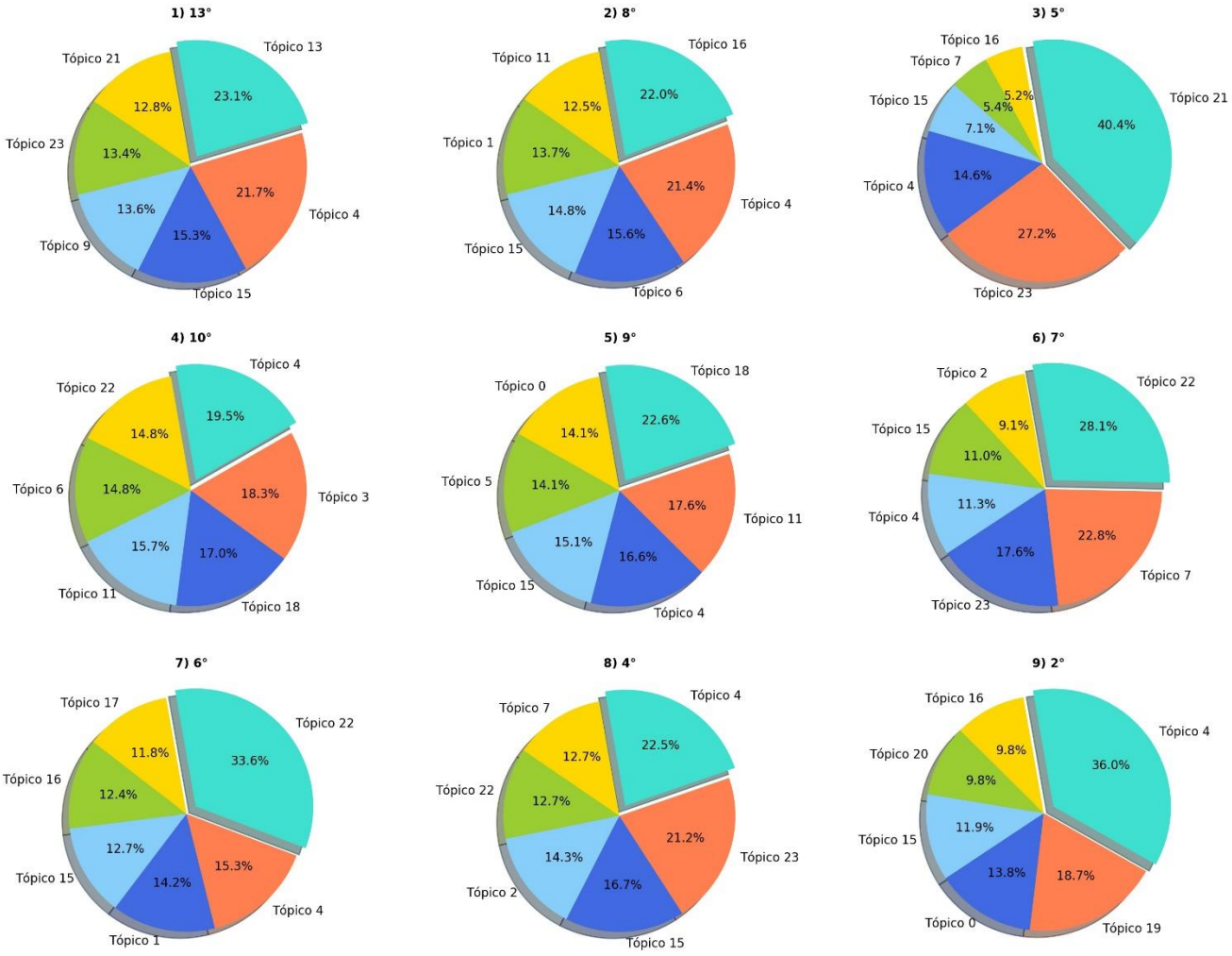
Analizando la línea de tiempo de la actividad de fallas generales (tópico 9) exclusivamente en Luis Carvajal.



Se puede ver su desempeño al decir que llega al lugar de trabajo en menos de 24 horas por lo general (tiempo recepción ot y arribo), y que concluye la tarea en menos de 2 horas (tiempo arribo y cierre), en horario de 8 am a 17 pm (aunque los trabajos los suele hacer en la mañana) .

Del mismo modo es posible analizar a cada atributo (estación, técnico, región, etc) de la base de datos.

Distribución de tópicos en top 9 regiones



Del mismo modo que para las estaciones, pero esta vez visto a un nivel macro, es posible crear **estrategias para las estaciones pensando en la zona o región** en que se encuentran. Por ejemplo, en la región de Valparaíso (5°) el mayor problema es que las ods están mal emitidas (tópico 21), por lo que si se investiga un poco más y se habla con administradores y encargados de estaciones de la zona, se puede llegar a la conclusión de que el sistema está fallando o que existe un evento atípico (solo por dar un ejemplo)

2. ALGORITMO K-MEANS

Si bien es cierto que la aplicación del algoritmo anterior ayuda a explicar perfectamente la composición general de los problemas presentes en el sector de Mantenimiento y Reparación de OGC, a veces es necesario un mayor grado de detalle en los clusters, **conservando el contexto y la semántica de la oración.**

Es por este motivo que se diseña un nuevo algoritmo, capaz de conservar todos los detalles y especificaciones de las OT/OS, al mantener la oración intacta (a diferencia del algoritmo LDA que extraer las palabras que más ocurren en cada oración)

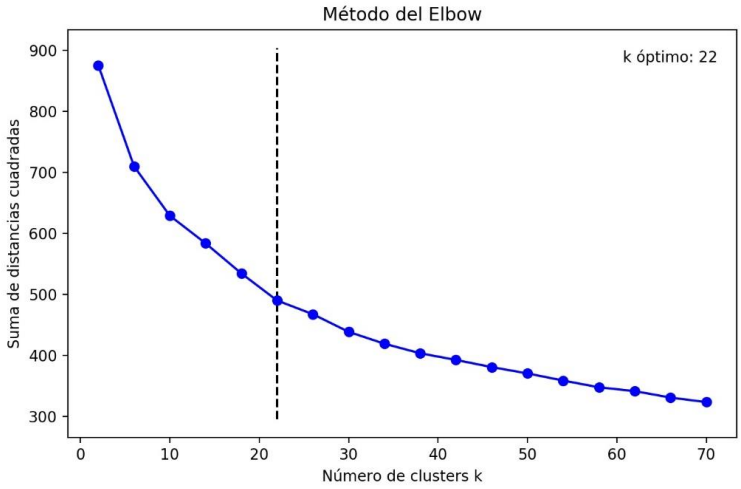
K-MEANS

Al igual que LDA se considera como un método de clusterización en el segmento de aprendizaje no supervisado, el cual corresponde a un algoritmo iterativo que intenta dividir el conjunto de datos en subgrupos (clusters) distintos y no superpuestos definidos por K, en los que cada punto de datos pertenece a un solo grupo. Busca que los elementos al interior del cada cluster sean lo más similar posible, manteniendo al mismo tiempo los clusters con elementos diferentes lo más alejado que se pueda. Asigna puntos de datos de tal manera que **la suma de la distancia cuadrada entre los puntos de datos y su centroide** (media aritmética de todos los puntos de datos que pertenecen a ese conjunto) **sea mínima**. Cuanta menos variación se tenga dentro de los clusters, sus elementos serán más homogéneos.

Aplicación de K-MEANS a técnico asignado ONLINE

Un ejemplo se puede ver al comparar los resultados de OT '**Online'** por ambos algoritmos (LDA y K-MEANS), en donde a través del primero se determinó que los tópicos 4 y 21 (detalle incidencia y ods mal emitida) son los que más destacaban.

Ejecutando K-Means se podrán identificar los mismos problemas, pero ahora con un mayor nivel de detalles, destacando el contexto, semántica y el sentido de la oración.



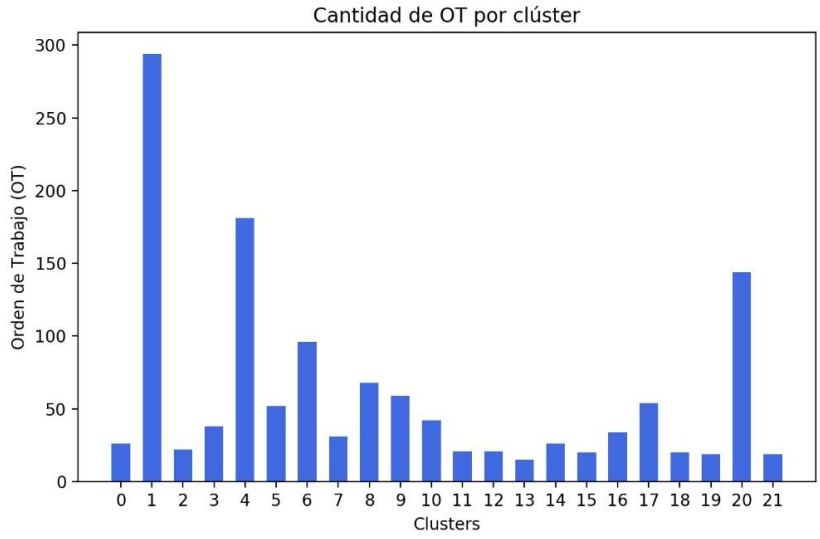
Antes de aplicar el algoritmo, es necesario conocer el número óptimo de clusters que necesitamos con el fin de maximizar la efectividad del modelo y optimizar los resultados, el cual es posible obtener mediante el **método de Elbow** (figura anterior).

El número óptimo de clusters para este caso es 22, por lo que se ingresa al modelo y se obtienen los resultados.

Los clusters para OT atendida de forma ONLINE son :

```
Cluster  0 : ot anulada por el cliente para ser atendida bajo ot 4
Cluster  1 : ods mal emitida
Cluster  2 : incidencia es atendida bajo ot 0
Cluster  3 : se cierra os ya que a partir de que se enlazaron los sistemas de sysqmr y csim se crearon os y ot automaticas dupl
icando las os de mp
Cluster  4 : incidencia debe ser atendida por estacion de servicio
Cluster  5 : ot atendida por otro contratista
Cluster  6 : incidencia mal emitida
Cluster  7 : se cierra por error de carga incidencia ya tenia ot
Cluster  8 : incidencia no corresponde a sgs por lo tanto se cierra ot
Cluster  9 : ot mal asignada
Cluster  10 : mantencion preventiva sera atendida bajo ot 1
Cluster  11 : ods mal ingresado
Cluster  12 : ot periodo vencido incidencia atendida bajo ot 20175
Cluster  13 : se envia informe a petrobras por variaciones en eds
Cluster  14 : incidencia es atendida bajo ot 8
Cluster  15 : incidencia duplicada
Cluster  16 : cliente reasigno incidencia ya que corresponde a otro contratista
Cluster  17 : incidencia sera atendida por estacion de servicio ya que corresponde a cambio de accesorio
Cluster  18 : incidencia atendida bajo la ot 2
Cluster  19 : incidencia fue atendida bajo la ot 6
Cluster  20 : incidencia es atendida bajo ot 1
Cluster  21 : incidencia atendida bajo la ot 9
```

Donde cada cluster contiene un conjunto de oraciones que se relacionan con él. El nombre del cluster corresponde a la oración que mejor representa el contenido de este.



Ahora si queremos ver el contenido de cada cluster, el algoritmo también lo permite, mostrando lo siguiente:

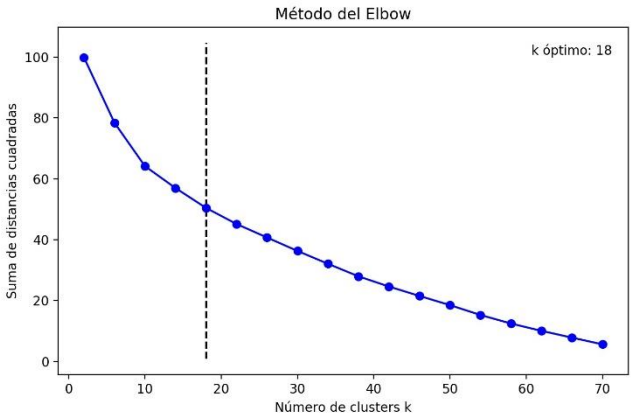
```
Cluster 0 : ot anulada por el cliente para ser atendida bajo ot 4
        sentence  0 :  incidencia anulada por enex
        sentence  1 :  orden anulada por cliente requerimiento se atender en mp de junio
        sentence  2 :  orden anulada por enex
        sentence  3 :  fuera de contrato se anula
        sentence  4 :  orden de trabajo fue anulada por cliente
        sentence  5 :  incidencia es anulada ya que anomalia corresponde a orpack
        sentence  6 :  incidencia es anulada
        sentence  7 :  ods anulada por cliente
        sentence  8 :  ods anulada por cliente
        sentence  9 :  ods anulada por cliente
        sentence  10 :  ods es anulada por encontrarse repetida
        sentence  11 :  ods anulada por cliente
        sentence  12 :  ods fue eliminada ya que eds atendio por su cuenta la ods se libero la ot 0 para anular esta
        sentence  13 :  ods fue eliminada ya que eds atendio por su cuenta la ods se libero la ot 0 para anular esta
        sentence  14 :  ods fue eliminada ya que eds atendio por su cuenta la ods se libero la ot 0 para anular esta
        sentence  15 :  anulada por enex
        sentence  16 :  ot anulada por cliente
        sentence  17 :  ot anulada por cliente
        sentence  18 :  ot anulada por cliente
        sentence  19 :  ot anulada por cliente bajo ot 8
        sentence  20 :  ods anulada por enex
        sentence  21 :  ot anulada por el cliente para ser atendida bajo ot 4
        sentence  22 :  anulada por cliente
        sentence  23 :  ot anulada por cliente
        sentence  24 :  orden anulada por enex
        sentence  25 :  ot anulada por el cliente
```

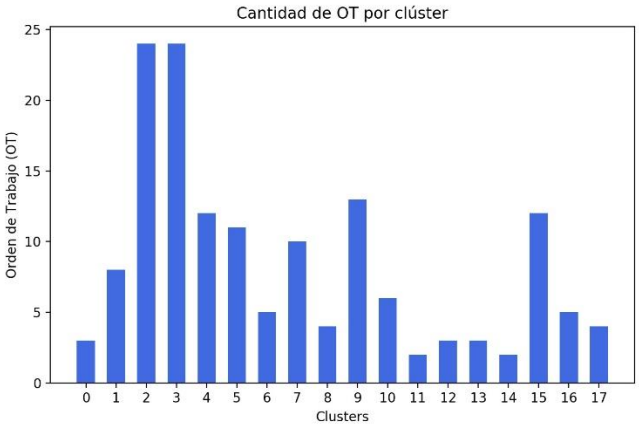
```
Cluster 15 : incidencia duplicada
        sentence  0 :  ods duplicada
        sentence  1 :  incidencia duplicada con ot 25540
        sentence  2 :  orden duplicada
        sentence  3 :  se cierra por duplicado de os
        sentence  4 :  orden duplicada
        sentence  5 :  cerrada por estar duplicada
        sentence  6 :  orden duplicada
        sentence  7 :  incidencia duplicada
        sentence  8 :  incidencia duplicada el requerimiento se atendera bajo la ot 18048
        sentence  9 :  os duplicada
        sentence  10 :  os duplicada
        sentence  11 :  incidencia duplicada
        sentence  12 :  incidencia duplicada
        sentence  13 :  incidencia duplicada
        sentence  14 :  incidencia duplicada
        sentence  15 :  incidencia duplicada
        sentence  16 :  incidencia duplicada
        sentence  17 :  duplicado
        sentence  18 :  duplicado
        sentence  19 :  incidencia duplicada
```

Este algoritmo es realmente últil cuando se requiere conocer detalles de la situación, ya que LDA o modelamiento de tópicos podría haber acertado en el tópico 21 de **ot anulada/mal asignada**, pero **con K-means se conoce el motivo**, que fue ***"Ya que eds atendio por su cuenta la ods se libero la ot 0 para anular esta",*** por dar un ejemplo.

Por otro lado, el algoritmo K-Means tiene por desventaja no ser tan visual como su contraparte LDA, el cual era mucho más facil explicar el comportamiento del sector. Con K-Means existe entropía, y se requiere tiempo y un buen dominio de los conceptos involucrados para interpretar cada cluster.

Aplicación de K-MEANS a estación “CL - Buses Vule S.A.”





Total clusters para *“CL – Buses Vule S.A.”*

Cluster 0 : se realiza tension de polea y correa de motor quedando equipo operativo
Cluster 1 : se realiza el cambio de 2 codos giratorio 1 2 swivel 1 y 2 filtros alta capacidad quedando equipo operativo
Cluster 2 : os mal emitida
Cluster 3 : ods mal emitida
Cluster 4 : mantencion preventiva se calibran 2 de 2
Cluster 5 : mp se realizara cuando 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 alt a 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 dent ro de la tolerancia establecida por la compania bocas verificadas 2 de 2 punto industrial queda operativo mp se realizara cuand o se repare surtidor n 2 se efectua proceso de mantencion preventiva de estanque y surtidores se realiza revision limpieza y ve rificacion volumetrica a los surtidores 1 y 2 con matraz de 20 lts se cambian 2 filtros de alta capacidad debido a flujo lento mejorando considerablemente 65 litros por minutos aprox se reaprietan porta pistolas boca 1 y 2 ademas se remarcan adhesivos bo caestanque
Cluster 6 : se realiza el cambio de 2 filtros spin on quedando equipo operativo
Cluster 7 : mantencion preventiva se realizara bajo ot 5
Cluster 8 : surt 1 d se realizo remplazo de un swivel de 1 opw surt 1 d se realizo remplazo de un swivel de 1 opw
Cluster 9 : se realiza el cambio de 1 pistola 1 quedando equipo operativo
Cluster 10 : 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 11 : se realiza visita a punto industrial en donde no se encuentra personal que autorice labores incidencia es cerrada y se queda a la espera de un nuevo reporte
Cluster 12 : en visita realizada se observa ducto de venteo por medio de canopy lo cual genera complicacion para efectuar labo res
Cluster 13 : se realiza revision de equipo encontrandose despacho operativo
Cluster 14 : se realiza reparacion y mantencion de valvula retencion quedando equipo operativo
Cluster 15 : se efectua el cambio de 2 filtros en surtidor 12 producto diesel se realiza cambio de pistola la cual presenta pr oblema en su cano
Cluster 16 : 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 17 : se realiza limpieza y mantencion de filtro malla quedando equipo operativo

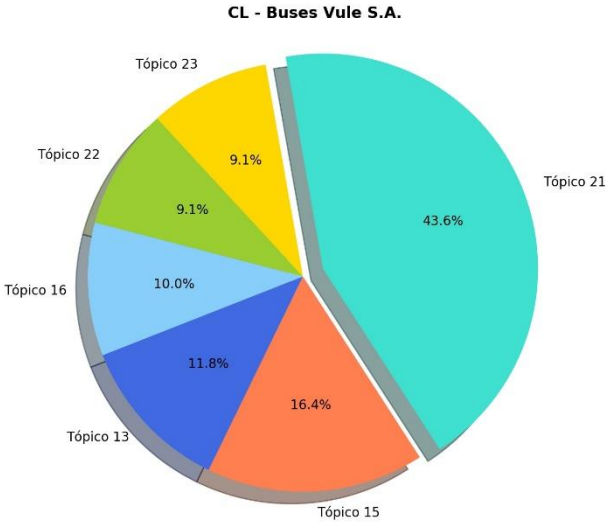
Analizando el contenido de alguno de los clusters, se tiene que:

```
Cluster 9 : se realiza el cambio de 1 pistola 1 quedando equipo operativo
        sentence  0 :  se rev surtidor 1 cara 1 la cual se encuentra con la correa cortada se cambia correa plus queda funcionando se rev s
        sentence  1 :  se revisa dispensador doble lado 1 con breakaway cortado por tiron de bus se cambia breakaway de 1 reconectabke y se
        sentence  2 :  se realiza el cambio de 1 correa wayne quedando equipo operativo
        sentence  3 :  se chquea falla y realiza el cambio de hn swiwel 1 con filtracion por uso se prueba quefando operativo en un 100 se
        sentence  4 :  se realiza el cambio de 1 kit reparacion bomba centrifuga blackmer 1 emisor de pulso y 1 filtro alta capacidad queda
        sentence  5 :  se realiza el cambio de 1 tapa adaptdor descarga 4 quedando equipo operativo
        sentence  6 :  se realiza el cambio de 1 manguera 1x5 mts 1 whipe hose 1 y 1 breackaway 1 quedando equipo operativo
        sentence  7 :  se realiza el cambio de 1 pistola 1 quedando equipo operativo
        sentence  8 :  se realiza el cambio de 1 pistola 1 quedando equipo operativo
        sentence  9 :  se realiza el cambio de 1 swivel 1 y 1 swivel 34 conv quedando equipo operativo
        sentence  10 :  se realiza el cambio de 1 swivel 1 y 1 swivel 34 conv quedando equipo operativo
        sentence  11 :  se realiza el cambio de 1 motor monofasico rele interno y 1 placa fuente de poder bp quedando equipo operativo
        sentence  12 :  se realiza el cambio de 1 swivel 1 quedando equipo operativo

Cluster 10 : 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 :  surt 12d se realizo remplazo de dos filtros de alta cap p14187 est 1d se realizo remplazo de una tapa 4 opw p16264 s
        sentence  1 :  surt 1 d se realizo remplazo de un filtro de alta cap p14187 surt 1 d se realizo remplazo de un filtro de alta cap p
        sentence  2 :  se acude por atencion de disp no se encuentra encargado la atencion es de 0800 a 1600 hrs se coordinara una proxima
        sentence  3 :  disp 12 34d se realizo remplazo de 4 filtros de alta cap p14187 disp 12 34d se realizo remplazo de 4 filtros de alta
        sentence  4 :  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  5 :  pendiente remplazo de 2 filtros disp 12d se realizo remplazo de 2 filtros de alta cap pendiente remplazo de 2 filtro

Cluster 12 : en visita realizada se observa ducto de venteo por medio de canopy lo cual genera complicacion para efectuar labores
        sentence  0 :  en visita realizada se observa ducto de venteo por medio de canopy lo cual genera complicacion para efectuar labores
        sentence  1 :  en visita realizada no se puede ejecutar labores ya que ducto de venteo pasa por el medio del canopy lo cual impide
        sentence  2 :  se acude por falla en venteo segun informacion de encargado ellos no realizaron insidencia por este tema o el disp c

Cluster 13 : se realiza revision de equipo encontrandose despacho operativo
        sentence  0 :  se realiza revision de equipo encontrandose despacho operativo
        sentence  1 :  se realiza revison de accesorios encontrandose despacho ok
        sentence  2 :  se realiza revison de equipo encontrandose despacho a buses ok
        sentence  0 :  se realiza reparacion y mantencion de valvula retencion quedando equipo operativo
        sentence  1 :  se realiza reparacion y fijacion de palanca de accionamiento quedando equipo operativo

Cluster 17 : se realiza limpieza y mantencion de filtro malla quedando equipo operativo
        sentence  0 :  se realiza limpieza y mantencion de valvulas check lomo tk y filtros malla quedando equipo operativo
        sentence  1 :  se coordinara una proxima visita para am surtidor con llave encargado se retiro surt 1 d se realizo remplazo de una
        sentence  2 :  se realiza limpieza y mantencion de filtro malla quedando equipo operativo
        sentence  3 :  surt 12 d se sacaron y reinstalaron filtro de malla en succion de las centrifugas para chequeo limpieza quedan surti
```

Como se puede ver, el hecho de que el cluster tenga por nombre la observación más representativa, no quiere decir que todas las observaciones al interior de él serán las mismas (cluster 9), sino que el algoritmo sabe que esa oración es la que mejor representa el conjunto ya que está más cerca del centroide, por lo cual tiene mayor valor.

Validando este algoritmo de K-Means se puede apreciar en la figura derecha lo obtenido por **LDA** (modelamiento de tópicos – algoritmo anterior), señalando que el **tópico 21** (ods mal emitida) y **15** (mantención preventiva) explican más el 60% de las actividades en esta estación.

Volviendo a K-Means, vemos en el histograma que la mayor cantidad de OT la registran los clusters 2, 3, 4 y 5 que efectivamente corresponden a ods mal emitidas y mantenimiento preventivo, pero con la ventaja de conocer el detalle.

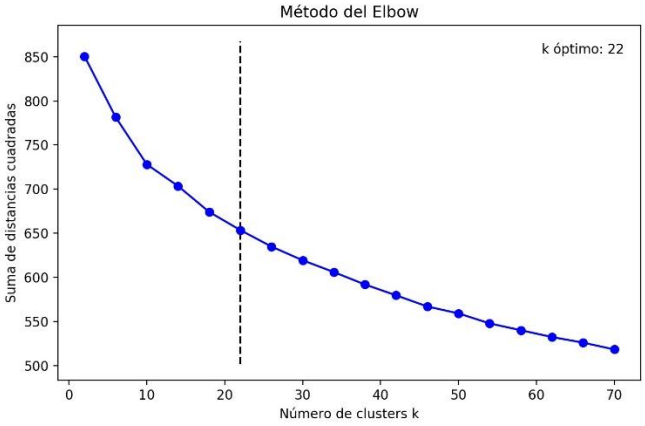


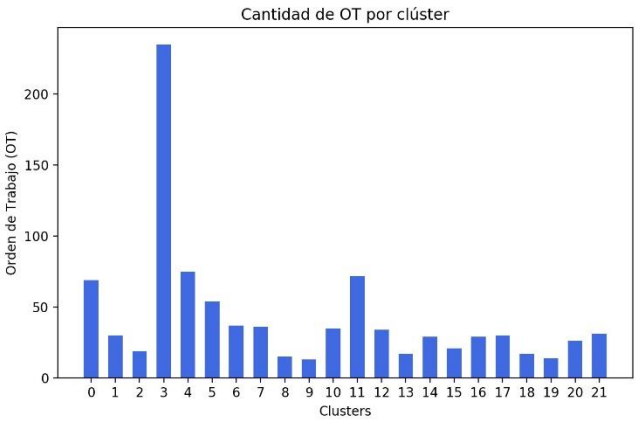
Algoritmo LDA – Topic modeling

Aplicación de K-MEANS a un técnico asignado

En las páginas 22 y 23 se analizó por medio del algoritmo LDA a Luis Carvajal, el cual se dijo que trabajaba mayormente en atención de *fallas generales, filtraciones / cambio de filtros* y *remplazo de válvulas.*

Ahora si volvemos a ver sus observaciones, pero esta vez a través de K-Means.





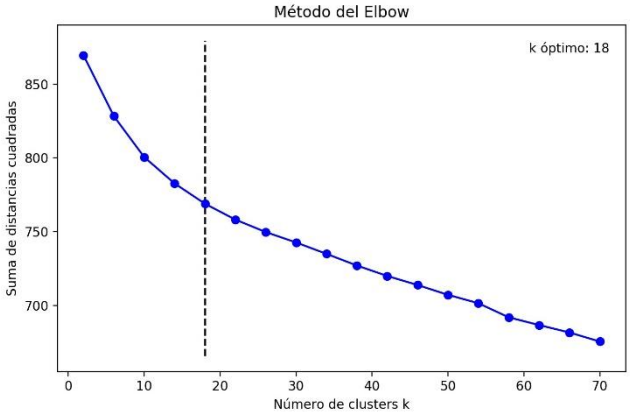
Los cluster representativos son:

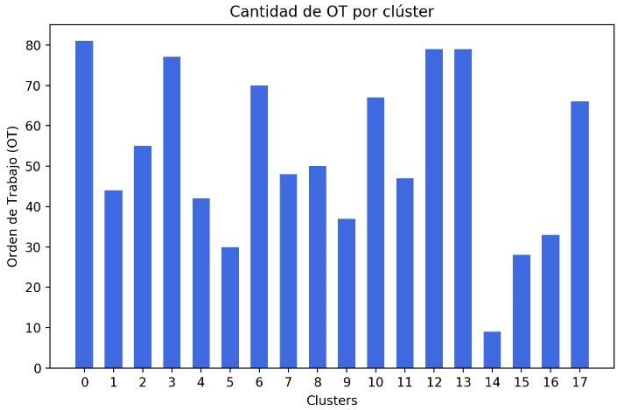
```
Cluster  0 : se chequea falla y se detecta problema de fuga en medidor se realiza el cambio de oring y mantencion completa a medidor 2 pm s
e encuentra manguera reseca con fuga en diesel y gas 95 se detecta fuga atravez de filtro velcon y pistola de 1 que no corta se nos solicit
Cluster  1 : se realiza el cambio de 1 manguera rv quedando equipo operativo
Cluster  2 : se realiza revision de equipo y accesorio colgante encontrandose despacho venta cliente
Cluster  3 : se realiza el cambio de 1 kit oring medidor eliminando filtracion quedando equipo operativo
Cluster  4 : se realiza el cambio de 1 valvula proporcional quedando equipo operativo
Cluster  5 : se chequea falla detectando surtidor bloqueado en gas 93 y gas 95 se realiza coldstar y se reprograma se prueba y queda operat
Cluster  6 : se chequea falla y se detecta breakway cortado y manguera con desgaste se cambia accesorios y se prueba queda operativo al 100
Cluster  7 : incidencia es atendida bajo ot 1
Cluster  8 : se realiza el cambio de 1 filtro baja capacidad quedando equipo operativo
Cluster  9 : se realiza el cambio de 1 ampolleta pls 9w quedando equipo operativo
Cluster  10 : se realiza el cambio de 1 pistola rv 2 mangueras rv y 1 breackaway rv quedando equipo operativo
Cluster  11 : se realiza el cambio de 1 pistola rv quedando equipo operativo
Cluster  12 : se chequea falla y se detecta falla se realiza el cambio de wipe hose rv se prueba y queda operativo se chequea falla y se de
Cluster  13 : se chequea falla y se detecta correa cortada se realiza cambio y se prueba quedando operativo se detecta ademas surtidor desc
Cluster  14 : se realiza el cambio de 1 valvulas ecologica quedando equipo despacho venta cliente operativo
Cluster  15 : se chequea falla detectando surtidores lentos y con fuga se realiza reapriete y sellado de todos los diesel se realiza cambio
Cluster  16 : se chequea falla y se detecta ampolletas quemadas se realiza el cambio de ampolletas y se prueba quedando operativa se cheque
Cluster  17 : se chequea falla y se detecta cable suelto se reaprieta y se prueba quedando 100 operativo se chequea falla y se detecta cabl
Cluster  18 : se chequea falla y se reconecta breakway cortado se prueba y queda operativo en un 100 se chequea falla y se reconecta breakw
Cluster  19 : se realiza el cambio de 1 motor rv quedando equipo operativo
Cluster  20 : se chequea falla y se detecta problema en filtro saturado de diesel se realiza cambio de filtro y se prueba queda operativo e
Cluster  21 : se chequea falla y se realiza el cambio de parametros se realiza prueba con ventas quedando operativo al 100 se chequea falla
```

Al observar el histograma, se aprecia que los clusters 3, 4, 0 y 11 ciertamente son las actividades en que mayor se desempeña , las que efectivamente coinciden con los tópicos encontrados en el algoritmo LDA.

Integración de modelos - K-MEANS en tópico 1 obtenido en LDA

La afinidad de ambos modelos es tanta que incluso es posible ejecutar el algoritmo de K-Means a cada uno de los tópicos encontrados en LDA, con el fin de conocer el detalle y composición de estos.





Extracto de 18 clusters encontrados para el ***tópico 1 (veeder root)***:

```
Cluster 11 : veeder root sin alarmas todo operativo veeder root sin alarmas todo operativo
        sentence  0 :  se visita eds para chequeo de alarma l2 la que fue retirada sensor y limpiado se instala quedando sin presencia de alarma
        sentence  1 :  alarma disipada alarmas corresponden a t1 y t2 aviso raz incr decf estanques 1 y 2 g93 sifoneados se verifican sondas de
        sentence  2 :  se procede a eliminar alarmas en veeder root las cuales no afectaban ventas quedando ok se procede a eliminar alarmas en
        sentence  3 :  se chequea alarma q4 se repara cable senal modulo de fuerza y se chequea funcionamiento con ventas a publico veeder root
        sentence  4 :  se realiza revision de sensor l7 el cual se encuentra en tank sump de est 2 93 se limpia y se elimina alarma en veeder ro
        sentence  5 :  veeder root presenta alarmas q1 q2 y q3 prueba anual se fuerza prueba alarmas no se disipan se requiere reseteo del equip
        ...

Cluster 12 : se revisan y limpian sondas medicion de estanques sin variacion falla corresponde a lectura de inventarios en veeder root se imprim
        sentence  0 :  se realiza calibracion con metodo calibex a 4 estanques
        sentence  1 :  se revisa regla de tanque n2 de p diesel enco tra do que esta parte marcando de 800 litros se solicitan tablas de calibra
        sentence  2 :  se requiere tabla de calibracion 93 lecturas no concuerdan con volumen de estanque configuracion en un punto se requiere
        sentence  3 :  se requierenntablas de calibracion chequeo de telemedicion se detectan diferencias entre alturas y volumen entre veeder r
        sentence  4 :  se realiza revision de tabla de calibracion en veeder root el cual se encuentra en 20 puntos se hace la comparacion de ta
        sentence  5 :  se verifica configuracion de estanque en veeder root equipo con tabla de 20 puntos regla en buen estado 20 puntos de veed
        ...

Cluster 13 : se realiza el cambio de una probeta telemedicion mag plus y un kit de flotadores 2 pulgadas en tk 1 gas 93 se chequea medicion con
        sentence  0 :  se retira regla y se mide ingresando nuevos valores a vedeer root en configuracion a 20 puntos equipo en gas 97 queda cal
        sentence  1 :  se realiza limpieza de sonda telemedicion se limpian flotadores y se chequea funcionamiento en consola tls 350 telemedici
        sentence  2 :  se limpian sondas de telemedicion se indica error en lectura de estanques 1 y 2 al momento de recepcionar producto se rev
        sentence  3 :  se recomienda no cargar tk a mas del 80 de su capacidad equipo operativo se retira sensor de telemedicion en gasolina 97
        sentence  4 :  se mantencion a sonda de telemedicion sonda se encuentra midiendo en modo normal se realiza intercambio de sondas estanqu
        ...
```

Antes de termianr, es necesario aclarar que el algoritmo LDA solo se ejecuta una vez para la base de datos completa,tardandose entre 15 a 30 minutos (dependiendo del modelo), y cuando finaliza, ya solo queda obtener elementos visuales de acuerdo a lo que se quiera analizar. Por el contrario, el algoritmo K-Means se debe ejecutar cada vez que se quiera analizar un item de un atributo nuevo (los analisis de los items procesados se pueden guardar), donde el tiempo de procesamiento oscila entre los 7 a 80 segundos.

CONCLUSIÓN

A partir de la columna de texto no estructurado y por medio de la aplicación de técnicas de clustering, se obtiene información crucial para la empresa con el objetivo de mejorar la toma de decisiones.

Se dice que la aplicación del algoritmo de Topic Modeling LDA, permite **obtener patrones generales** (por medio de la distribución de tópicos) en cualquier atributo del que se quiera obtener información, siendo una herramienta muy flexible y visual. Gracias a esto es posible identificar que actividades consumen mayor tiempo en cada estación, en que tarea se desempeña mejor un trabajador o como es el rendimiento en regiones, y así poder implementar estrategias y planes de acción focalizados que garanticen eficiencia y eficacia.

Por su parte el algoritmo **K-MEANS**, el cual nace de una necesidad específica, permite **conocer el detalle de los patrones anteriores**, descubriendo los motivos y el contexto de cada observación. Como se mencionó en un ejemplo anterior para las ordenes de trabajo (OT) Online, por medio de LDA se identifica que la ot anulada/mal asignada es la principal tendencia, pero gracias a K-Means es posible entender el motivo: *"Ya que eds atendio por su cuenta la ods se libero la ot 0 para anular esta"*. En consecuencia, con K-Means la gerencia puede estar mejor preparada para llevar a cabo planes estratégicos.

Finalmente, y como se mencionaba antes, la aplicación del modelo K-Means a cada uno de los tópicos es posible, por lo que si se tienen dudas respecto a los tópicos o se quiere asegurar respecto a una falla o problema, simplemente se debera ejecutar dicho algoritmo a las observaciones con el tópico en cuestión.