Práctica Entorno ELK
(Máster Data Science)

Beatriz Visitación, Natalia Alonso
y Susana Albarrán

Contenidos

1 Introducción

GESTIÓN DEL DOCUMENTO			
Autor	Beatriz Visitación, Natalia Alonso y Susana Albarrán	**Fecha**	21/06/2020
Nombre fichero	Memoria_Práctica_ELK.pdf	**Versión**	1.0
Descripción: En esta memoria se incluyen todos los pasos que se han seguido para realizar la práctica de la asignatura Recuperación de la Información. La realización de la práctica se ha llevado a cabo en un entorno **Windows**. Y se han usado las siguientes versiones: **elasticsearch-7.6.2**, **kibana-7.6.0-windows-x86_64**, **logstash-7.6.2** y **cerebro-0.9.1.**			

En esta práctica se partirá de una fuente de datos real extraída de Kaggle: https://www.kaggle.com/mchirico/montcoalert. Este dataset contiene las llamadas realizadas en un período de tiempo al 911, número de Emergencias en Montgomery County (Pennsylvania).

La práctica consta de tres partes fundamentales: recogida y tratamiento de los datos, almacenamiento y análisis de los datos, y visualización.

El objetivo de este trabajo es utilizar esta información para cargar datos con Logstash, realizar consultas en Elasticsearch y visualizaciones con Kibana. Paralelamente se ha usado Cerebro (otra aplicación para ELK) para monitorizar los índices de Elasticsearch.

2 Recogida y procesamiento de los datos

La recogida de los datos consiste en descargar el fichero **911.csv** de Kaggle: https://www.kaggle.com/mchirico/montcoalert. Este fichero tiene un tamaño de 114Mb y consta de los siguientes campos:

lat	Latitud
long	Longitud
desc	Descripción de la emergencia
zip	Código Postal
title	Título de la Emergencia
timeStamp	Fecha y Hora de la llamada
twp	Distrito
add	Dirección
e	Siempre toma valor 1

En nuestro caso se ha procesado el fichero 911.csv desde Logstash. Estos son los pasos que se han seguido:

1. Desde **Kibana**, se crea el índice 911 y se realiza el mapeo para el campo que vamos a considerar como "geo_point". Se ha decido transformar este campo a tipo **geo_point** para poder hacer una visualización en un mapa. Este campo, como se detallará a continuación, se ha conseguido mediante la unión de la latitud y la longitud.

```
PUT /911
{
  "mappings": {
    "properties": {
      "location" : {"type":"geo_point"}
    }
  }
}
```

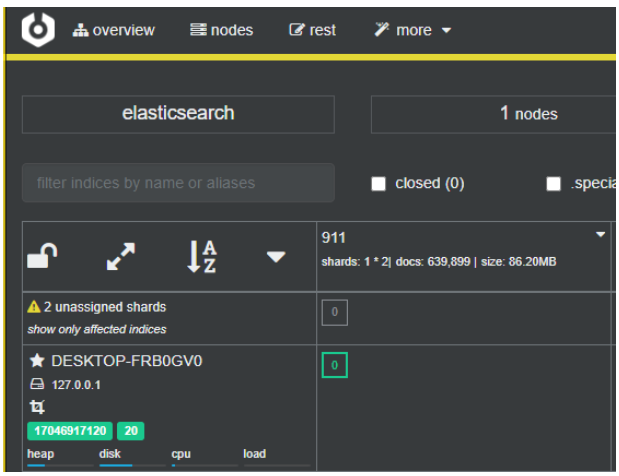
2. Con **Logstash** se realiza la carga de datos a través del fichero **logstash911.conf.**

   Cabe destacar que se ha realizado una disección del campo **"timeStamp"** para dividirlo en fecha **("Date")** y Hora **("Hour"),** y para el campo **"title"** que se divide en tipo de emergencia **("TypeEmerg")** y motivo de la llamada **("Motive").** También se ha añadido un campo nuevo **"location"** mediante la unión de los campos **"lat"** y **"lng".** Estas decisiones se han tomado pensando en las consultas y visualizaciones que se pretenden realizar (por ejemplo: filtro por días o meses o filtro por tipo de emergencia).

   A continuación se muestra una captura del fichero **logstash911.conf:**

```
input {
    file {
        path => "C:/Users/susi_/Desktop/Rec_Informacion/Practica/Datos_911/911.csv"
       start_position => beginning
       sincedb_path => "NUL"
    }
}

filter {
    csv {
        separator=> ","
        columns =>["lat","lng","desc","zip","title","timeStamp","twp","addr","e"]
        }
    dissect {
          mapping => {"timeStamp"=> "%{Date} %{Hour}"}
          mapping => {"title" => "%{TypeEmerg} %{Motive}"}
          }

    mutate {
            convert => {
               "lat" => "float"
               "lng" => "float"
               "zip" => "integer"
               }
             add_field =>{
               "location" => "%{lat},%{lng}"
             }
       remove_field => ["message", "@version","@timestamp", "host", "path", "lat", "lng",
"e","desc","timeStamp","Hour"]
     }
}
output {
    elasticsearch {
       hosts => ["localhost:9200"]
        index => "911"
     }
    stdout{}
}
```

Se comprueba la correcta ingesta de datos desde la aplicación “Cerebro”, en el puerto 9000. **"Cerebro**" es una herramienta de administración web de Elasticsearch creada con Scala, Play Framework, AngularJS y Bootstrap. Esta aplicación nos ha ayudado a comprobar de una manera rápida y sencilla los índices creados y su estructura (mappings). Esta gestión también se puede hacer también desde kibana, sin embargo, Cerebro tiene una interfaz muy manejable. Se comprueba que se han cargado un total de **639.899** documentos en el índice **911**:



3 Análisis de los datos

En esta etapa se realizan consultas sobre el índice que se ha creado en Elasticsearch, 911. Las consultas se realizan desde la API de Kibana.

3.1 Consulta 1.

Se realiza un "**match**" para el motivo de emergencia "ABDOMINAL PAINS" sobre el campo **"Motive.keyword"** (campo no analizado, es decir, encontrará documentos con la misma cadena que la buscada, en mayúscula y plural). Por último, se ordena la búsqueda por fecha en modo ascendente y se muestran los campos **"TypeEmerg", "Motive"** y **"Date".** Por si la salida fuera muy grande, se mostrarán solo los 100 primeros documentos que cumplan los criterios.

```
POST 911/_search
{
  "from":0,"size":100,
  "query": {
    "match": {"Motive.keyword": "ABDOMINAL PAINS"}
  },
  "sort":[
    {"Date":"asc"}
  ],
  "_source": {
    "includes": ["Motive","Date", "TypeEmerg"]

  }
}
```

El primer documento obtenido como resultado de esta consulta siguiendo los criterios:


```
{
  "took" : 1,
  "timed_out" : false,
  "_shards" : {
    "total" : 1,
    "successful" : 1,
    "skipped" : 0,
    "failed" : 0
  },
  "hits" : {
    "total" : {
      "value" : 8716,
      "relation" : "eq"
    },
    "max_score" : null,
    "hits" : [
      {
        "_index" : "911",
        "_type" : "_doc",
        "_id" : "lXWZ6nIBTusOmQIlMOcL",
        "_score" : null,
        "_source" : {
          "Motive" : "ABDOMINAL PAINS",
          "TypeEmerg" : "EMS:",
          "Date" : "2015-12-10"
        },
        "sort" : [
          1449705600000
        ]
      }
    ]
  }
}
```

3.2 Consulta 2.

Se realiza un "**multi-mach**" para buscar simultáneamente la cadena "TRAFFIC" sobre el campo **"TypeEmerg"** y "DISABLED" en el campo **"Motive",** se obtendrán sólo aquellos documentos que cumplan ambas condiciones, ya que se ha usado la cláusula "**must**". En este caso, la búsqueda se ha realizado sobre los campos analizados (sin .keyword). Finalmente se mostrarán los campos **"TypeEmerg", "zip", "Date","twp", "Motive".**

```
POST 911/_search
{
  "query": {
    "bool": {
      "must": [
        {
          "multi_match": {
            "query": "TRAFFIC",
            "fields": ["TypeEmerg"]
          }
        },
        {
          "multi_match": {
            "query": "DISABLED",
            "fields": ["Motive"]
          }
        }
      ]
    }
  },
  "_source": {
    "includes": ["TypeEmerg", "zip", "Date","twp", "Motive"]
  }
}
```

El primer documento obtenido como resultado de esta consulta siguiendo los criterios:


```
{
  "took" : 4,
  "timed_out" : false,
  "_shards" : {
    "total" : 1,
    "successful" : 1,
    "skipped" : 0,
    "failed" : 0
  },
  "hits" : {
    "total" : {
      "value" : 10000,
      "relation" : "gte"
    },
    "max_score" : 3.7110016,
    "hits" : [
      {
        "_index" : "911",
        "_type" : "_doc",
        "_id" : "4HWZ6nIBTusOmQIlL-P0",
        "_score" : 3.7110016,
        "_source" : {
          "zip" : null,
          "Motive" : "DISABLED VEHICLE -",
          "twp" : "WEST CONSHOHOCKEN",
          "TypeEmerg" : "Traffic:",
          "Date" : "2015-12-12"
        }
      }
    ]
  }
}
```

3.3 Consulta 3.

En esta consulta se pretende realizar un recuento del número de llamadas por código postal, usando la cláusula “aggregations”. A continuación, se muestran los motivos asociados al distrito con mayor número de llamadas (código postal 19401).

```
POST 911/_search
{
  "size" : 0,
  "aggregations" : {
    "zip" : {"terms" : {"field": "zip"}}
  }
}

POST 911/_search
{
  "query":{
    "match":{"zip":"19401"}
  },
  "_source": {
    "includes": ["twp","zip","Motive"]
  }
}
```

La primera parte de la consulta muestra que el código postal 19401 es el distrito con mayor número de llamadas con un total de 43.814. También se muestra el resultado de la segunda parte de la consulta, después de aplicar un filtro “zip” : “19401”.


```
{
  "took" : 1,
  "timed_out" : false,
  "_shards" : {
    "total" : 1,
    "successful" : 1,
    "skipped" : 0,
    "failed" : 0
  },
  "hits" : {
    "total" : {
      "value" : 10000,
      "relation" : "gte"
    },
    "max_score" : null,
    "hits" : [ ]
  },
  "aggregations" : {
    "zip" : {
      "doc_count_error_upper_bound" : 0,
      "sum_other_doc_count" : 300046,
      "buckets" : [
        {
          "key" : 19401,
          "doc_count" : 43814
        },
        {
          "key" : 19464,
          "doc_count" : 42202
        },
```



```
{
  "took" : 3,
  "timed_out" : false,
  "_shards" : {
    "total" : 1,
    "successful" : 1,
    "skipped" : 0,
    "failed" : 0
  },
  "hits" : {
    "total" : {
      "value" : 10000,
      "relation" : "gte"
    },
    "max_score" : 1.0,
    "hits" : [
      {
        "_index" : "911",
        "_type" : "_doc",
        "_id" : "ZA5v4HIBCud1ZNNO-2KH",
        "_score" : 1.0,
        "_source" : {
          "zip" : 19401,
          "Motive" : "CARDIAC EMERGENCY",
          "twp" : "NORRISTOWN"
        }
      }
    ]
  }
}
```

3.4 Consulta 4.

En esta consulta se realiza un filtro mediante la cláusula “filter” para localizar las llamadas donde en el campo **“Motive”** se encuentre la palabra “abdominal” y devuelva las llamadas realizadas con este motivo durante el año 2018.

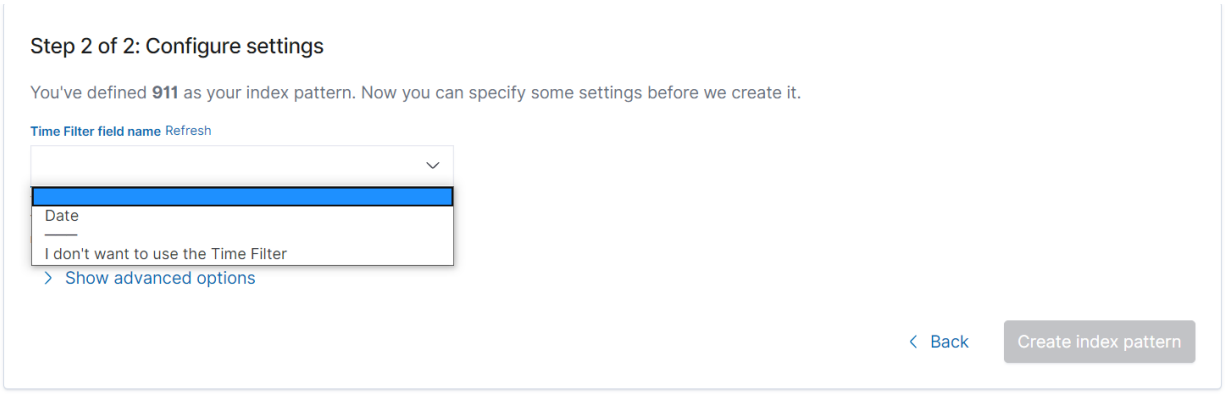
```
POST 911/_search
{
  "from":0,"size":1000,
  "query": {
    "bool": {
      "filter" :[
        {"match": {"Motive" : "abdominal"}},
        {"range" : {"Date" : {"gte" : "2018-01-01", "lte" : "2018-12-31"}}}
      ]
    }
  },
  "_source": {
    "includes": ["Motive", "Date"]

  }
}
```

El primer documento obtenido como resultado de esta consulta siguiendo los criterios:


```
{
  "took" : 1,
  "timed_out" : false,
  "_shards" : {
    "total" : 1,
    "successful" : 1,
    "skipped" : 0,
    "failed" : 0
  },
  "hits" : {
    "total" : {
      "value" : 1870,
      "relation" : "eq"
    },
    "max_score" : 0.0,
    "hits" : [
      {
        "_index" : "911",
        "_type" : "_doc",
        "_id" : "7nqa6nIBTusOmQIlt1Fo",
        "_score" : 0.0,
        "_source" : {
          "Motive" : "ABDOMINAL PAINS",
          "Date" : "2018-01-01"
        }
      }
    ]
  }
}
```

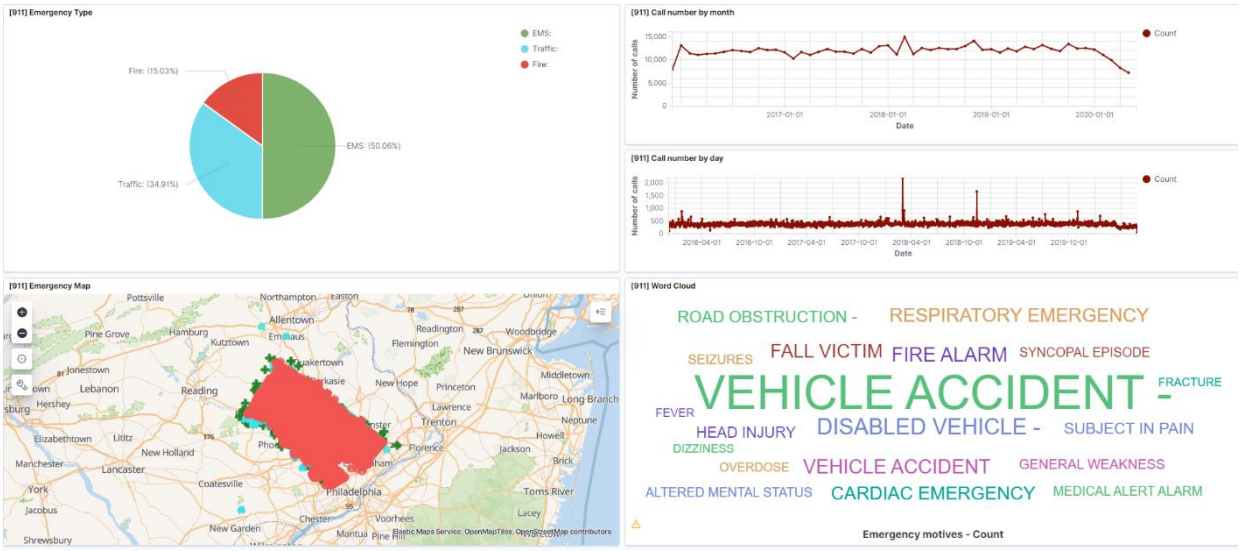
4 Visualización

Para cargar los datos en **Kibana**, se crea el índice ya cargado en Elasticsearch “911”. A la hora de crear dicho índice, se ha seleccionado la opción “I don’t want to use the Time Filter”:

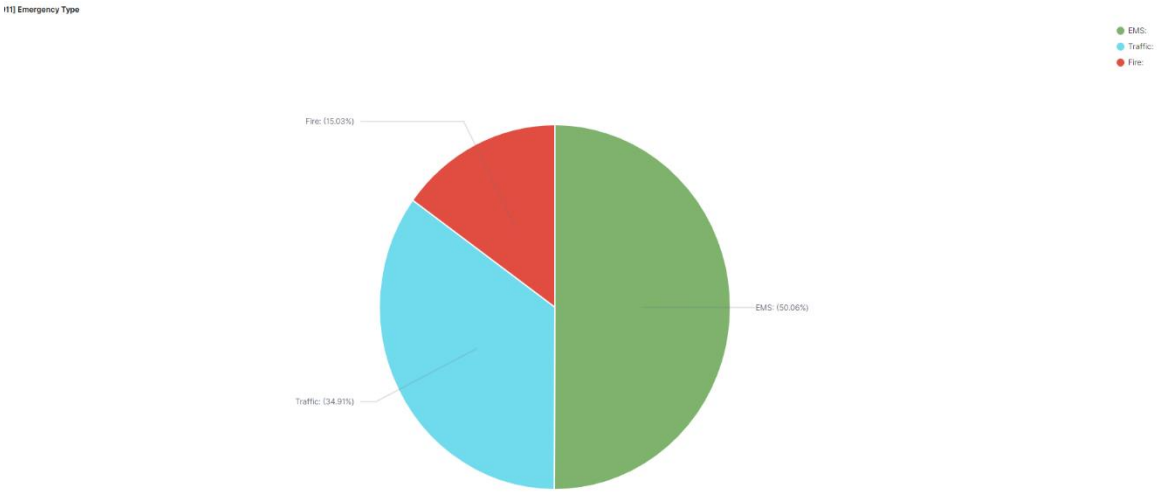


Además, de manera adicional, se ha cambiado el formato del campo fecha de la siguiente manera:

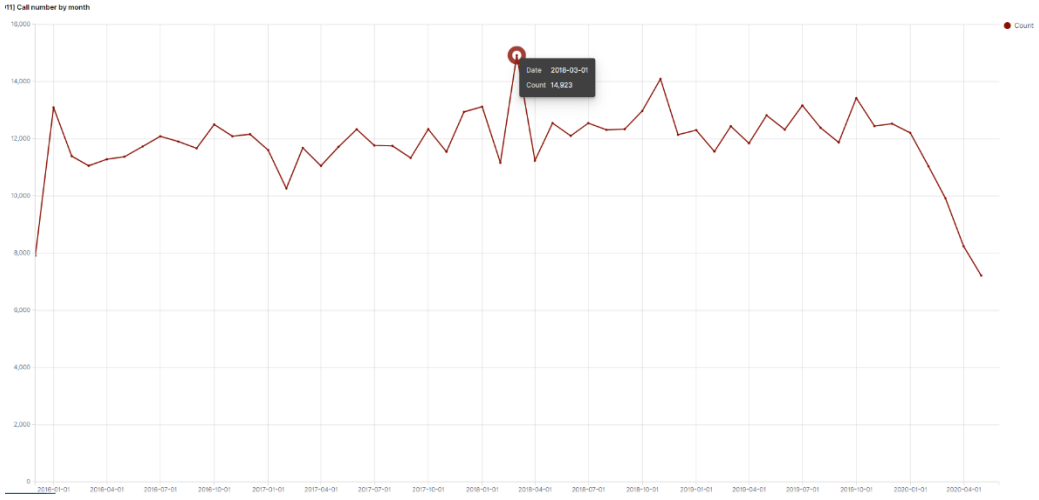
En esta práctica se realizan 5 gráficos que se presentarán en un Dashboard. El resultado del Dashboard es el que se muestra a continuación. Se ha realizado una visualización de tipo pie, donde se muestra el porcentaje de cada tipo de emergencia: EMS (Emergency Medical Services), Fire, o Traffic Emergency). En segundo lugar, se ha creado un Date Histogram donde se puede observar el número de llamadas por día y por mes. También se ha visualizado un mapa con cada tipo de emergencia donde se puede ser la dirección (add), el distrito (twp) y el motivo de la emergencia. Por último, se ha realizado una Word Cloud donde se pueden ver los motivos asociados a las llamadas más frecuentes.



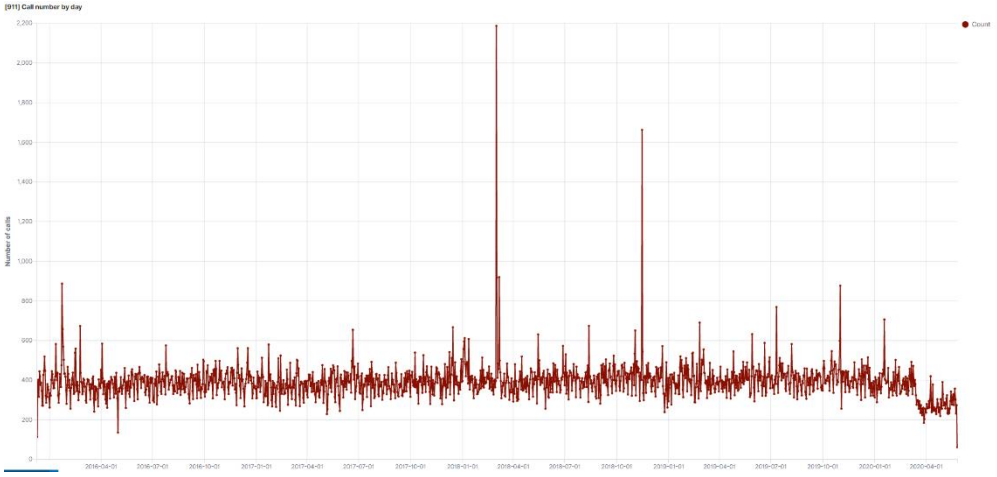
- [911] Emergency Type.



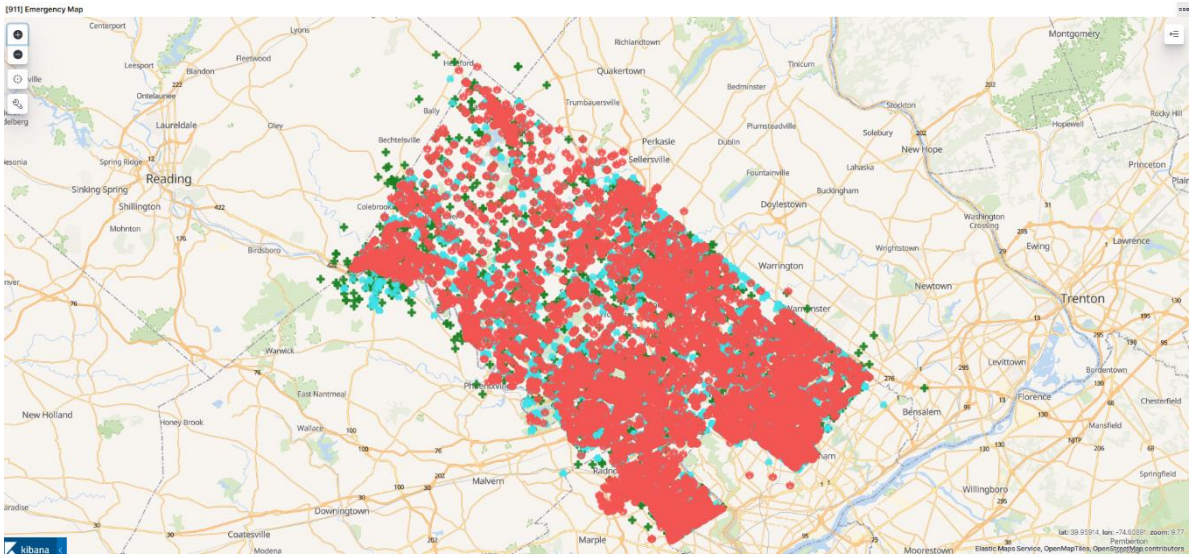
- [911] Call number by Month.



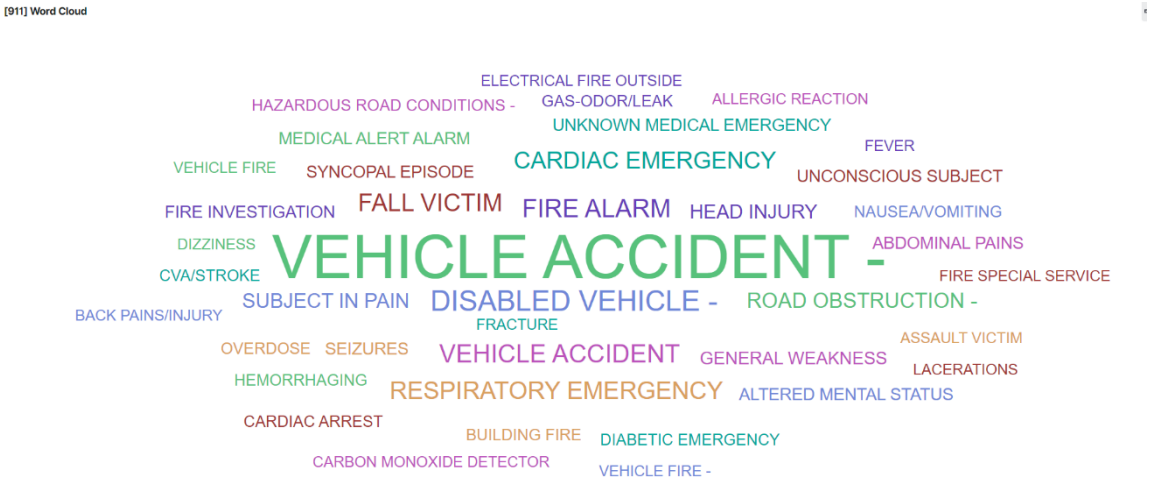
- [911] Call number by Day.



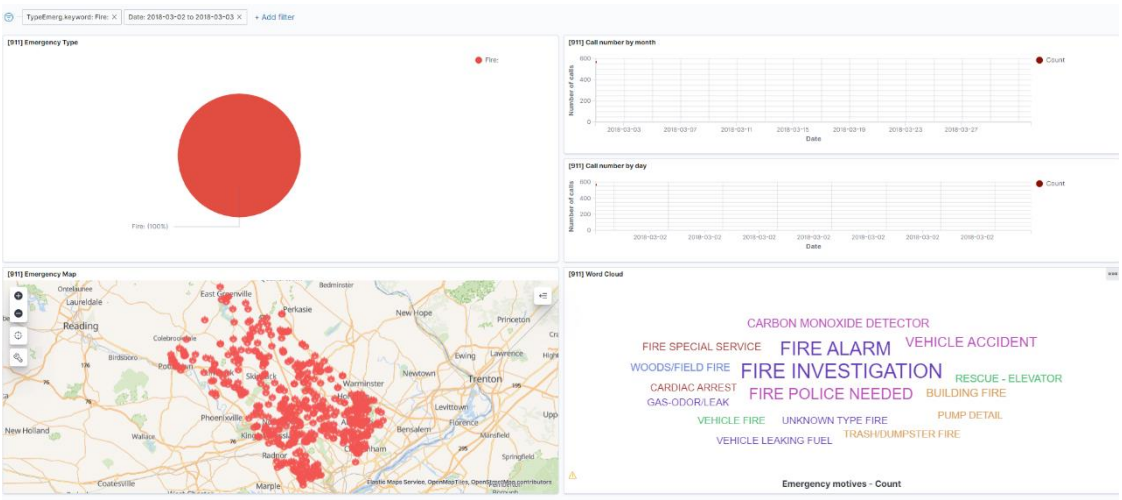
- [911] Emergency Map.



- [911] Word Cloud.



- [Filtro]: en la siguiente figura se puede observar el resultado de la visualización tras aplicar un filtro por tipo de emergencia (Fire) y para un día en concreto (02/03/2018).



El enfoque que se ha seguido a la hora de elaborar las consultas y las visualizaciones ha sido poder monitorizar en tiempo real (si se contaran con datos reales en streaming) los lugares desde donde se realizan las llamadas con sus datos asociados (dirección, tipo de emergencia…).