REPOSITORIOS DE INFORMACIÓN.

ELASTICSEARCH.

Integrantes del grupo:

Adán Fernández Sánchez – UO251162

Pelayo García Torre - UO251143

Álvaro Manuel Camporro Ayala - UO251562

Contenido

1. INTRODUCCIÓN

El objetivo de nuestro trabajo es devolver el 'id' de los usuarios que más han twiteado acerca de un tema que será introducido previamente por parámetro.

Para ello, hemos utilizado ElasticSearch, habiendo probado previamente nuestras consultas mediante 'Cerebro' para que, una vez seguros de que estaban bien formuladas, pasarlas a lenguaje PHP.

Hemos utilizado la colección de tweets del campus virtual.

2. DESARROLLO DEL TRABAJO

En primer lugar, hicimos la siguiente consulta para buscar los tweets en inglés que llevaran la palabra 'sport'. La cláusula 'bool' nos permite añadir más condiciones.

```
{
    "query":
    {
        "bool": {
            "must": [
                {
                    "match": {
                        "lang":"en"
                    }
                },
                {
                    "match": {
                        "text":"sport"
                    }
                }
            ]
        }
    }
}
```

Posteriormente, formamos una agregación mediante el uso de 'aggs' utilizando la función 'significant_terms'. Mediante esta función analizamos los datos obtenidos y encontramos los términos que aparecen con una frecuencia que es estadísticamente anómala en comparación con los datos de fondo.

```
{
    "size":0,
    "query":
    {
        "bool": {
            "must": [
                {
                    "match": {
                        "lang":"en"
                    }
                },
                {
                    "match": {
                        "text":"sport"
                    }
                }
            ]
        }
    },
    "aggs":
    {
        "Palabras":{
            "significant_terms":
            {
                "field":"text"
            }
        }
    }
}
```

Tras comprobar su correcto funcionamiento en Cerebro, procedimos a la transformación del código a PHP. Para esto solamente tuvimos que cambiar los caracteres '=' por '=>' y las llaves por corchetes. De tal manera:

```
$params = [
    "index" => "2008-feb-02-04",
    "type" => "tweet",
    "body" => [
        "size" => 0,
        "query" => [
            "bool" => [
                "must" =>
                [
                    [
                        "match"=>
                        [
                            "lang"=>"en"
                        ]
                    ]
                    ,
                    [
                        "match"=>
                        [
                            "text"=>"sport"
                        ]
                    ]
                ]
            ]
        ],
        "aggs"=>[

            "Palabras"=>[
                "significant_terms"=>
                [
                    "field"=>"text",
                    "size"=> 100
                ]
            ]

        ]
    ]
];

$results = $client->search($params);

print_r($results);
```

También podríamos imprimir los datos mediante:

foreach($results['aggregations']['Palabras']['buckets'] as $value)

echo $value['key'], ':', $value['score'], "\n";

en lugar del *print_r($results);*

Ahora nos toca, una vez sacados los datos más relevantes, sacar los usuarios. Para ello, metemos en la cláusula 'match' cada término sacado en particular. Además, también añadimos un 'boost'. El 'boost' se refiere al valor que tiene cada término devuelto anteriormente, que variará en función de la importancia. Por ejemplo, a 'sport' le metemos un 'boost' de 400, aunque lo ideal sería que fuera la puntuación que nos devolviera la consulta anterior ('score'). Posteriormente lo haremos de esa manera.

```
{
    "size":100,
    "query":
    {
        "bool": {
            "should": [
                {
                    "match": {
                        "text" : {
                            "query": "sport",
                            "boost": 400
                        }
                    }
                },
                {
                    "match": {
                        "text" : {
                            "query": "21dfh7",
                            "boost": 3
                        }
                    }
                },
                {
                    "  ": {
                    "text" : {
                            "query": "entrance",
                            "boost": 2
                        }
                    }
                }
            ]
        }
    },
    "aggs": {
        "numero_usuarios": {
            "terms": {
                "field": "user_id_str"
            }
        }
    }
}
```

En este caso, lo hemos hecho con el término ‘sport’ y metiendo los ‘boost’ sin ningún criterio. Lo correcto sería hacerlo como anteriormente hemos comentado.

Meterlo de manera manual sería demasiado laborioso. Lo más óptimo sería hacerlo de manera dinámica (en tiempo de ejecución).

Para hacerlo en tiempo de ejecución tendríamos que recorrer el resultado de la consulta (por ejemplo mediante un bucle ‘for’). Nosotros optaremos por un bucle ‘foreach’. En este bucle, recorreremos la variable ‘results’ hasta llegar a la parte donde está el valor de cada expresión (‘score’).

```
$terminos = [];
foreach($results['aggregations']['Palabras']['buckets'] as $value) {
    echo $value['key'], ': ', $value['score'];
    $terminos[] = [
    'match'=>[
        'text' => [

            'query'=> $value['key'],
            'boost'=> $value['score']

            ]
        ]
    ];
}
```

De esta manera estaría en tiempo de ejecución.

Para finalizar, solamente tendremos que utilizar la variable ‘términos’ de la misma forma en que utilizamos los términos más repetidos con su ‘boost’ correspondiente. De esta forma, solo tendremos que sustituir todos los ‘match’ por el array.

```
$params = [
    "index" => "2008-feb-02-04",
    "type" => "tweet",
    "body" => [
        "size" => 0,
        "query" => [
            "bool" => [
                "should" => $terminos
                ]
            ],
        "aggs"=>[
            "numero_usuarios"=>
            [
                "terms"=>
                [
                    "field"=> "user_id_str"
                ]
            ]
        ]
    ]
];
```

Así, ya tendríamos los usuarios que más twitearon sobre el término pasado por parámetro en la colección de tweets de la que disponemos.

Si quisiéramos imprimirlos, solamente tendríamos que añadir

```
$results = $client->search($params);

print_r($results);
```

3. OBSERVACIONES

En algunas ocasiones nos encontramos con algún error en lo que creemos que era la configuración de Cerebro al realizar alguna consulta. Investigando por la web (https://www.elastic.co/guide/en/elasticsearch/reference/current/fielddata.html), descubrimos que no tenía algunos apartados activados.

Para activarlos de nuevo, tuvimos que poner las siguientes líneas de código (PUT).

```
{
  "properties":
  {
    "user_id_str":
    {
      "type":   "text",
      "fielddata": true
    }
  }
}
```

Especificando que era de tipo ‘tweet’.

A la de hora de sacar los términos más relevantes, nos encontramos con otros términos que creemos que eran de URL’s u otras expresiones puestas por robots (como pueda ser el término ‘2ldfh7’ que utilizamos anteriormente).

4. ANÁLISIS ACERCA DE LAS QUERYS DEL CAMPUS VIRTUAL

Al realizar las consultas que se encuentran el fichero 'Consultas.txt' con un tamaño 'size=10' casi todas nos van a devolver los mismos resultados, ya que un tweet tendrá más 'score' si aparecen únicamente los términos buscados en el tweet y no otras palabras. Luego para la relevancia y precisión, hablaremos de un tamaño muchísimo más elevado, no para los 10 resultados que nos salen.

4.1 American football conference

Ejecutando la **consulta 1** del fichero, no nos proporciona ningún resultado, ya que en la colección de tweets no hay ninguno que contenga las tres palabras.

Ejecutando la **consulta 2**, los resultados que aparecen, son tweets que contienen los términos especificados, pero no van a ser relevantes, ya que va a haber tweets que contengan el término 'football' pero que no te hable de la Conferencia de Fútbol Américano, por ejemplo que hablen del fútbol escocés, que no nos interesa para nada.

Ejecutando la **consulta 3** nos van a salir resultados que nos interesan más, ya que al menos, van a hablar de fútbol américano, obteniendo tweets más relevantes que con la anterior consulta y con más precisión.

4.2 David Tyree, Laurence Maroney, Eli Manning, Miami Dolphings, Michael Straham, Plaxico Burress, Randy Moss.

Ejecutando la **consulta 1** con los términos David Tyree, Laurence Maroney, Eli Manning… nos devuelve los tweets en los que aparece el nombre especificado.

Ejecutando la **consulta 2**, nos salen muchos más resultados, siendo muchos de estos pocos relevantes ya que habrá tweets acerca de un 'David' que no es el que estamos buscando. Lo mismo pasa con Tyree.

La consulta que nos daría resultados más relevantes sería la primera, ya que nos hablan de David Tyree.

4.3 Defensive end, Running back, Wide receiver

Ocurre exactamente lo mismo que con la querie anterior, para la **consulta 1** nos da un nº de tweets relevantes que hablan de 'defensive end', 'Running back' y 'Wide receiver' pero en cambio ejecutando la **consulta 2** nos da resultados poco relevantes, ya que al igual que antes, nos salen tweets en los que aparece 'end', pero que no tienen nada que ver con lo que estamos buscando.

4.5 Football, quarterback, NFL, XLII.

Al ejecutar las **consulta 1 y 2**, nos aparecen tweets que contienen la palabra football, pero como tienen más 'max_score' los tweets en los que aparece la palabra 'football' más veces, o tiene más peso en proporción a las demás, pues nos salen tweets que únicamente contienen 'football', y si quieres buscar información acerca del fútbol, pues no la vas a encontrar en tweets que únicamente muestren 'football'.

Las dos consultas tienen la misma precisión, ya que es un único término.

4.6 Gleedale

Mismo resultados y conclusión que Football. Se le podría añadir algún parámetro más, como el de la **consulta 3**, y obtener, por ejemplo, información de la ciudad de Glendale y no de alguien que se llame así.

4.7 National Football Conference.

Ejecutando la **consulta 1**, nos sale un único tweet que contiene los 3 términos.

Hemos puesto un 'boost' en el término 'conference' para que nos muestre los tweets de National Football, pero dándole más importancia al término 'conference', para poder distinguirlo, por ejemplo, de la National Football League.

Aplicando la **consulta 2**, nos aparecen resultados más relacionados con el término 'conference', ya que le estamos dando más importancia a ese término. Muestra resultados, que hablan de conference, pero no de la National Football Conference, que es lo que estamos buscando. Luego con ésta consulta, no obtenemos información relevante.

Ejecutando la **consulta 3**, aparecen resultados como en la consulta 2. Es información poco relevante, ya que únicamente hablan de 'conference'.

4.8 National Football League.

Mismos resultados y conclusiones que con la anterior querie.

Con la **consulta 1**, nos aparecen resultados realmente relevantes, que nos hablan acerca de lo que estamos buscando.

Para las demás consultas, saca resultados con el término league, que no se refieren en absoluto a lo que estamos buscando.

4.9 New England Patriots.

Ejecutando la **consulta 1**, nos aparecen tweets relevantes a cerca de los New England Patriots, en los que se habla acerca de ellos, pero hay tweets que únicamente contiene 'New England Patriots' y no aportan nada de información acerca de lo que se comenta sobre la querie especificada.

Si ejecutamos la **consulta 2**, nos salen resultados que pueden tener o no relación con el tema que estaos buscando. Salen tweets que nos hablan de England, que no tienen nada que ver con lo que estamos buscando.

Con la **consulta 3**, obtenemos resultados parecidos a la consulta 2, pero más específicos del tema que estamos tratando, ya que, al menos, nos van a salir resultados de los England Patriots, y no de lo que pasa en England.

4.10 New York, Super Bowl.

Con la **consulta 1**, nos saldrán resultados que hablen de Nueva York, pero al igual que sucedía con New England Patriots, pueden salir tweets con únicamente 'New York' que no aportan nada.

Ejecutando la **consulta 2**, nos salen muchos más resultados, que como comentamos anteriormente, tienen menor precisión, ya que pueden salir resultados que contengan 'new' o 'york' pero no se refieran a la ciudad. Poco relevante

4.11 New York Giants.

Aplicando la **consulta 1**, aparecen tweets que hablan acerca de los Giants, es información relevante, aunque los tweets de más relevancia, los primeros que se muestran, no aporten información acerca de lo que estamos buscando.

Si ponemos un boost en las otras dos consultas en el término 'Giants', obtendremos resultados que nos hablen acerca de los 'Giants', que es de lo que queremos hablar, y obtendremos más información en menos tweets acerca de ellos. Como en las demás consultas, puede volver a pasar que saquen resultados con el término 'new' que no nos sirven para nada.

4.12 Univerity of Phoenix.

Aplicando la **consulta 1** podemos ver que aparecen una serie de resultados que nos aportan información acerca de lo que estamos buscando.

Aplicando la **consulta 2** vemos que nos salen un montón de resultados. Muchos de ellos son debidos a que la palabra 'of', es una palabra muerta que no aporta nada a la consulta, lo que lleva a que salgan resultados que no se asemejan en nada a la búsqueda que estamos haciendo.

Podemos quitar la palabra 'of', ya que pensamos que al ser una palabra 'muerta', no va a aportar nada a los resultados.

Ejecutando la **consulta 3** vemos que nos salen menos resultados pero mucho más relevantes y con más precisión que en la consulta 2. Aun así, puede haber tweets que nos hablen de 'University' pero que no se refieran a la universidad de 'Phoenix'.

5. REALIZACIÓN DEL TRABAJO

El trabajo ha sido realizado por todos los miembros del grupo, siendo repartido en partes equitativas, habiéndonos reunido tanto para formular la idea sobre la que queríamos basarnos para hacer el trabajo como en las posteriores consultas y trabajo de realización y búsqueda de documentación.