ElasticSearch

DOCUMENTACIÓN

Javier Martínez Álvarez | UO258092
Mª Isabel Fernández Pérez| UO257829
Lino Menéndez de Luarca Trabanco| UO216936
Gema Rico Pozas | UO238096

Ejercicio 1

Para la realización de este ejercicio utilizamos como keyword “alcoholism”. La consulta que hemos realizado es la siguiente:

```
{
        "size": 0,
        "query": {
            "query_string": {
                "default_field": "selftext",
                "query": query,
            }
        },
          "aggs": {
            "Title": {
              "significant_terms": {
                "field": "title",
                "size": number,
                 est: properties_est
              }
            },
            "Text": {
              "significant_terms": {
                "field": "selftext",
                "size": number,
                 est: properties_est
              }
            },
            "Subreddit": {
              "significant_terms": {
                "field": "subreddit",
                "size": number,
                 est: properties_est
              }
            }
          }
    }
```

Hacemos uso de agregaciones y también eliminamos palabras vacías. Con el estadístico de “mutual_information” obtenemos términos que son más frecuentes lo que implica que nos aparecen más palabras poco significativas. Con gnd es con el que obtenemos los términos más significativos. Con Jlh obtenemos un término medio, así como con chi cuadrado.

Ejercicio 2

A través de REST no se puede realizar una consulta “more like this” con agregaciones.

Para ello hemos hecho uso de elasticsearch-dsl. Se trata de una librería de alto nivel cuyo objetivo es ayudar a escribir y ejecutar consultas. Utilizamos la clase Search de la librería (que nos permite mediante el método query() definir cualquier tipo de query de elasticSearch) para definir una consulta “more like this” para que realice la misma tarea del ejercicio 1. A continuación añadimos las agregaciones mediante el método bucket() usando la propiedad .aggs que actúa como una agregación de nivel superior.

Procedemos a eliminar los resultados con palabras vacías como en el ejercicio anterior y ejecutamos.

Ejercicio 3

En un primer paso hemos probado con las palabras que creímos que más se podían repetir en un comentario sobre medicamentos como “med”, “pill”, “dosage”, etc. También, utilizando unos conocimientos básicos sobre medicamentos probamos a incluir en la consulta algunos de los sufijos más comunes en medicamentos como podían ser “*zepam”, “*ine”, “*tin”, etc. Pero solamente salían los medicamentos que acababan en los sufijos antes nombrados. También entonces probamos realizando una mezcla de ambas estrategias, pero tampoco obtuvimos los resultados esperados. Optamos por la validación automática utilizando Wikidata y la siguiente consulta "using OR prescribed OR dose OR mg -water". Con esta consulta y la validación obtuvimos el mejor resultado, aunque tuvimos que eliminar de los resultados “water” ya que aparece como un medicamento y consideramos que no lo es.

Ejercicio 4

Para este ejercicio al no haber una lista con la que poder validar los resultados hacemos uso del servicio de Google scholar, con el que sacamos artículos especializados en los temas para hacer dicha validación. Para el estudio del suicidio usamos el término “suicide” y para el caso de conductas autolesivas usamos el término “self-harm”. La estructura sigue la del ejercicio 3 pero en vez de validar con Wikidata utilizamos el servicio antes citado.