



|  |
| --- |
| ElasticSearch  2019 |
|  |
| 15th of december  Information repositories – Uniovi 19-20  Written by: Óscar Sánchez Campo (UO265078)  Daniel Finca Martínez (UO264469) |

# Prologue

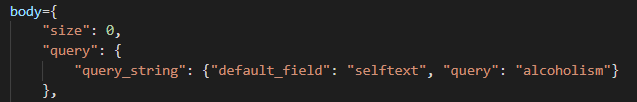
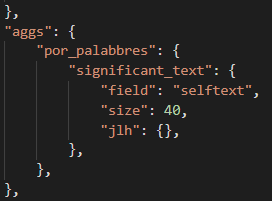
For this project, the provided script “bulkIndexer4.py” has been used to generate the index. Process such as stemming, removing stopwords and n-grams generation are used in such script.

Note: Lowercase operation has been commented out and the reason why this has been done will be explained later on (Exercise 3).

# Exercises

## Exercise 1

In this exercise “alcoholism” topic has been chosen. As an initial query to retrieve significant terms, a simple structure has been used as shown here:



To expand the results provided with such request, Elasticsearch query language aggregations feature is added to above body as stated on the image to the right. Using this, most common words used in posts retrieved from above results are grouped by frequency. These new outcomes then serve as the expanded set of terms about the topic chosen as the theme of the exercise.

In addition to above aggregation snippet, different similarity metrics have been used. These are gnd, chi\_square and jlh. Different aggregation sizes have been tested to assess their performance. Jlh metric was used with above query. This will be explained later on.