

# Elasticsearch assignment

UO258425, Carlos Manrique Enguita

UO236405, Daniel Rückert García

UO258454, Violeta Ruiz Martínez

December 13, 2019

## Abstract

This is the abstract of the project, here we will resume our work briefly

## Contents

<b>1</b>	<b>Exercise 1</b>	<b>1</b>
1.1	Percentage . . . . .	2
1.2	Google normalized distance . . . . .	2
1.3	Chi square . . . . .	3
<b>2</b>	<b>Exercise 2</b>	<b>3</b>
<b>3</b>	<b>Exercise 3</b>	<b>4</b>
<b>4</b>	<b>Exercise 4</b>	<b>4</b>

## 1 Exercise 1

We choose "Rehab" as topic to develop this first exercise.

To recover all the posts related to the topic we do a first query in order to find the most significant words related to it. To do that we look for the most significant words in posts that contain the words **Rehab** and **Rehabilitation**. We perform an aggregation in the selftext field. We tried this first query using three different similarity metrics: Chi square, Google normalized distance and percentage. We have created an index without stopwords so we perform this first query in this index to get the most optimum results.

## 1.1 Percentage

When using percentage we set the maximum number of significant terms to 100 because it is the one that gets the fewer results. With this metric we get 10 terms where we consider 5 as relevant:

- Rehabilitation
- Rehab
- Rehabs
- Criminals
- Librium (a medical drug)

## 1.2 Google normalized distance

When we use the Google normalized distance we set the maximum number of significant terms to 20 as this metric and the Chi square get the most number of results. With this metric we got 20 terms where we consider 9 as relevant:

- Rehab
- Rehabilitation
- Jail
- Facility
- Criminals
- Outpatient
- Heroin
- Homicide
- Detox

### 1.3 Chi square

When we used the Chi square metric we also set the maximum number of significant terms to 20 as we did for the Google one. Using this metric we also get 20 terms and we consider 9 as relevant:

- Rehab
- Rehabilitation
- Detox
- Alcoholism
- Sober
- Drinking
- Criminals
- Outpatient
- Hospital

After getting the significant terms we perform a second query searching for these words in the selftext.

## 2 Exercise 2

MLT queries do not directly produce a list of significant terms, they produce however a list of documents including the relevant terms used to perform the query. This is similar to how Google produces its search results. As we know GND uses the documents returned by google to generate the list of significant terms.

Therefore, to perform a similar operation to that of GND we just have to do a query for significant terms over the MLT query in order to get the most significant terms of those documents.

After performing such operation, we compared the results to those obtained by a GND query, obtaining a list of 20 significant terms for each query. Comparing the results, we obtained that 75% of the significant terms were identical. However, increasing the number of significant terms to more than 100 reduced this accuracy down to 40%.

Results with 20 significant terms:

```

Windows PowerShell
PS E:\repos\Elasticsearch\Exercise2> python .\Exercise2.py
MLT:
['aa', 'alcoholic', 'bac', 'cocked', 'criminals', 'detox', 'drinking',
'hospital', 'injunction', 'inpatient', 'meetings', 'nex', 'outpatient', 'reh
ab', 'rehabilitation', 'skyrocketing', 'sober', 'staples', 'voc', 'vocational']
]
-----
GND:
['aa', 'bac', 'cocked', 'criminals', 'detox', 'dui', 'facility', 'her
oin', 'homicide', 'injunction', 'inpatient', 'jail', 'nex', 'outpatient', 'reh
ab', 'rehabilitation', 'skyrocketing', 'staples', 'voc', 'vocational']
]
Total matches: 15, similarity: 75.0%

```

Results with 150 significant terms:

```

Windows PowerShell
PS E:\repos\Elasticsearch\Exercise2> python .\Exercise2.py
MLT:
['aa', 'abuse', 'addict', 'addiction', 'addicts', 'ago', 'agreeing',
'alcohol', 'alcoholic', 'alcoholics', 'alcoholism', 'also', 'ambien', 'another',
', april', 'arrested', 'bac', 'back', 'bender', 'bottle', 'called', 'came', 'care',
', center', 'charts', 'clarified', 'clean', 'cocked', 'codependent', 'cok
e', 'concluded', 'criminals', 'dad', 'day', 'days', 'decided', 'detox', 'detoxe
ed', 'done', 'drank', 'drink', 'drinking', 'drug', 'drugs', 'drunk', 'dui', 'd
xm', 'enabling', 'escort', 'even', 'facility', 'family', 'feel', 'finally', 'f
ired', 'first', 'found', 'get', 'go', 'going', 'got', 'hack', 'harass', 'hawai
i', 'help', 'heroin', 'home', 'homicide', 'hospital', 'hospitalized', 'house',
', injunction', 'inpatient', 'jail', 'job', 'just', 'know', 'last', 'librium',
', life', 'like', 'liquor', 'living', 'loaded', 'long', 'lost', 'made', 'meeting
s', 'meth', 'methadone', 'month', 'months', 'much', 'narc', 'needed', 'never',
', nex', 'night', 'now', 'one', 'outpatient', 'police', 'problem', 'program',
', recovering', 'recovery', 'rehab', 'rehabilitation', 'rehab', 'reprimanded',
', roots', 'sent', 'since', 'skyrocketing', 'slps', 'sober', 'sobriety', 'sponsor
', 'stalker', 'staples', 'started', 'still', 'suboxone', 'take', 'thought', 't
ime', 'times', 'told', 'took', 'towns', 'treatment', 'tried', 'two', 'ultimatu
m', 'vivitrol', 'voc', 'vocational', 'vodka', 'want', 'week', 'weeks', 'well',
', went', 'wife', 'will', 'withdrawals', 'work', 'working', 'year', 'years']
]
-----
GND:
['aa', 'addict', 'addicts', 'agreeing', 'alcoholic', 'alcoholics', 'a
lcoholism', 'ambien', 'apt', 'arrested', 'bac', 'bender', 'charts', 'clarified
', 'cocked', 'codependent', 'coke', 'concluded', 'criminals', 'detox', 'detoxe
d', 'dui', 'dxi', 'enabling', 'escort', 'facility', 'hack', 'harass', 'hawaii
', 'heroin', 'homicide', 'hospitalization', 'hospitalized', 'injunction', 'inpa
tient', 'jail', 'librium', 'liquor', 'loaded', 'meetings', 'meth', 'methadone',
', narc', 'nex', 'nil', 'opted', 'outpatient', 'overdosed', 'oxycodone', 'prog
ram', 'recovering', 'rehab', 'rehabilitation', 'rehab', 'reprimanded', 'roots
', 'skyrocketing', 'slps', 'snf', 'sponsor', 'stalker', 'staples', 'suboxone',
', towns', 'ultimatum', 'vivitrol', 'voc', 'vocational', 'vodka', 'withdrawals'
]
]
Total matches: 63, similarity: 42.0%

```

### 3 Exercise 3

### 4 Exercise 4