KVR Search Engine

Hidde Hensel, Mike Trieu, Freddy de Greef

Elasticsearch

- Opensource search engine
- Build on Apache Lucene
- Easy to use API

Building the Elasticsearch index

- Inhoud / Bibliografische omschrijving
- Trefwoorden
- Doc ID
- Vraag
- Antwoord
- Indiener
- Partij indiener
- Jaar

Database

```
"antwoord" : {
  "type" : "string",
  "analyzer" : "dutch"
},
"doc_id" : {
  "type" : "string",
  "index" : "not_analyzed'
},
"indiener" : {
  "type" : "string"
},
"inhoud" : {
  "type" : "string",
  "analyzer" : "dutch"
},
"jaar" : {
  "type" : "string"
},
"partij" : {
  "type" : "string"
},
"rubriek" : {
  "type" : "string",
  "analyzer" : "dutch"
},
"trefwoorden" : {
  "type" : "string",
  "analyzer" : "dutch"
},
"vraag" : {
  "type" : "string",
  "analyzer" : "dutch"
}
```

Queries in elasticsearch

- Full text search
- Build-in dutch stemmer
- Build-in tokenizer
- Ranked retrieval based on combination of boolean search and vector space model

```
$getParams['body'] = [
    'from' => $currentpagevalue*10,
    'size' => 10,
    'query' => [
        'bool' => [
            'should' => [
                ['match' => [
                    'vraag' => $query
                ]],
                ['match' => [
                    'antwoord' => $query
                ]],
                ['match' => [
                    'indiener' => $query
                ]]
            ]
        ]
    ],
```

Facets

- Elasticsearch Aggregations
- Creates facets specified by field
- Can easily filter in aggregations

```
'aggs' => [
    'aggspartij' => [
        'terms' => [
            'field' => 'partij',
            'size' => 0
        ]
    ],
    'aggsjaar' => [
        'terms' => [
            'field' => 'jaar',
            'size' => 0,
            'order' => [
                '_term' => 'asc'
            ]
        ]
    ]
]
```

```
'filter' => [
    'term' => [
        'jaar' => $fjaar
    ]
]
```

Wordclouds

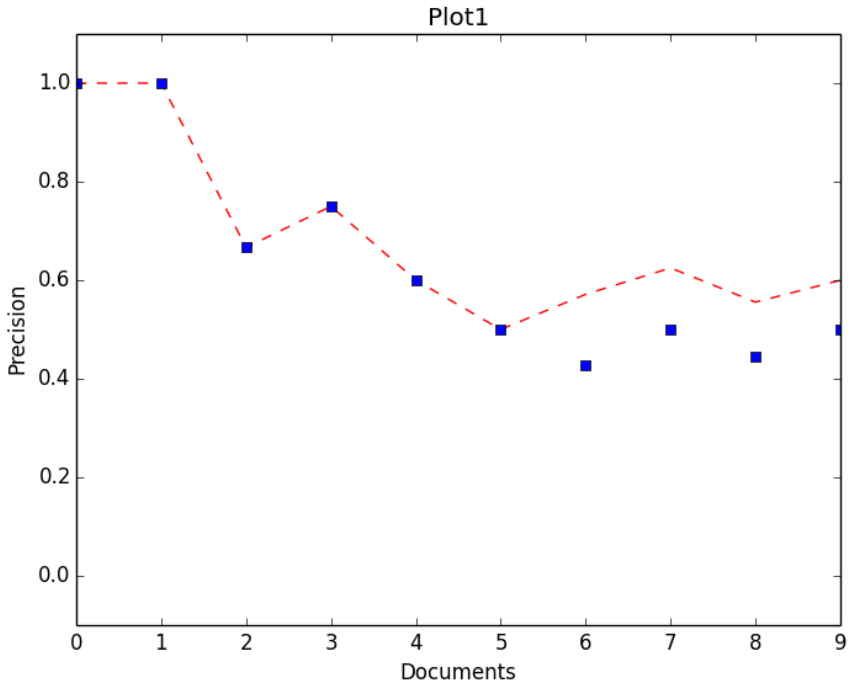
- Elasticsearch did not work

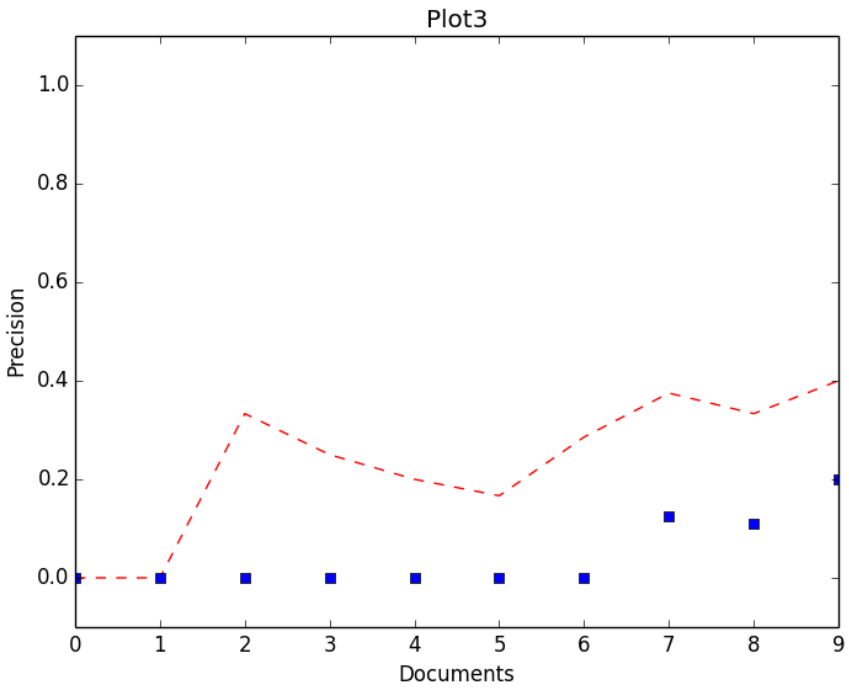
- pytagcloud package
- Easy to use
- But, only filters english stopwords

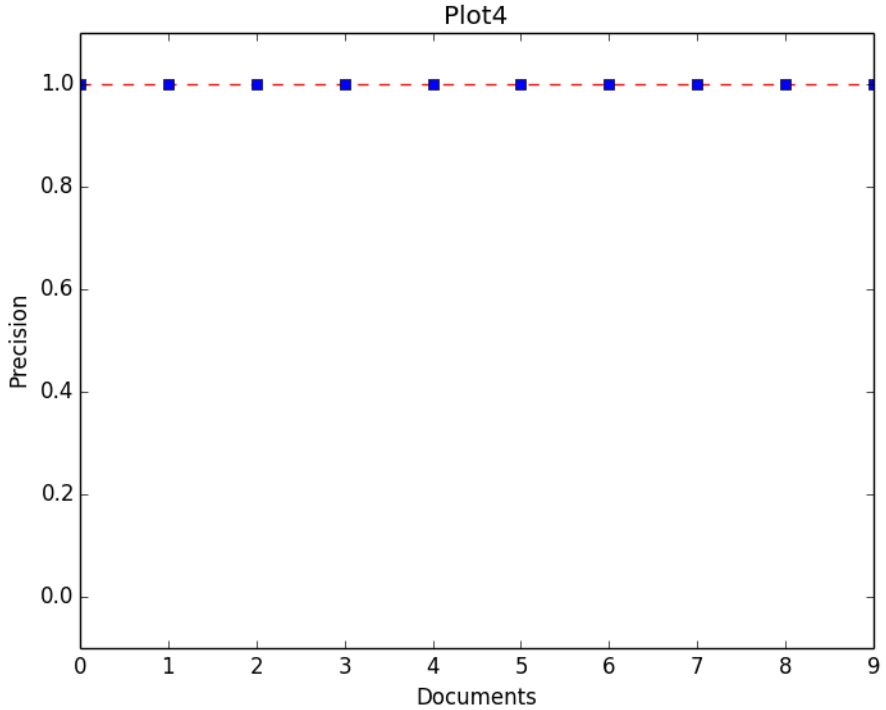
Queries

- problemen van chinezen in de samenleving
- file problemen in het verkeer
- export visserij nederland
- in stand houden van de wilde zwijnen populatie
- veiligheid van geert wilders

Problemen van chinezen in de samenleving




Plot3
Precision
Documents
0.0
0.2
0.4
0.6
0.8
1.0
0
1
2
3
4
5
6
7
8
9


Plot4
Precision
Documents
0.0
0.2
0.4
0.6
0.8
1.0
0
1
2
3
4
5
6
7
8
9

Evaluation

	Relevant	Not relevant
Relevant	26	2
Not relevant	5	17

Cohens kappa: 0.81

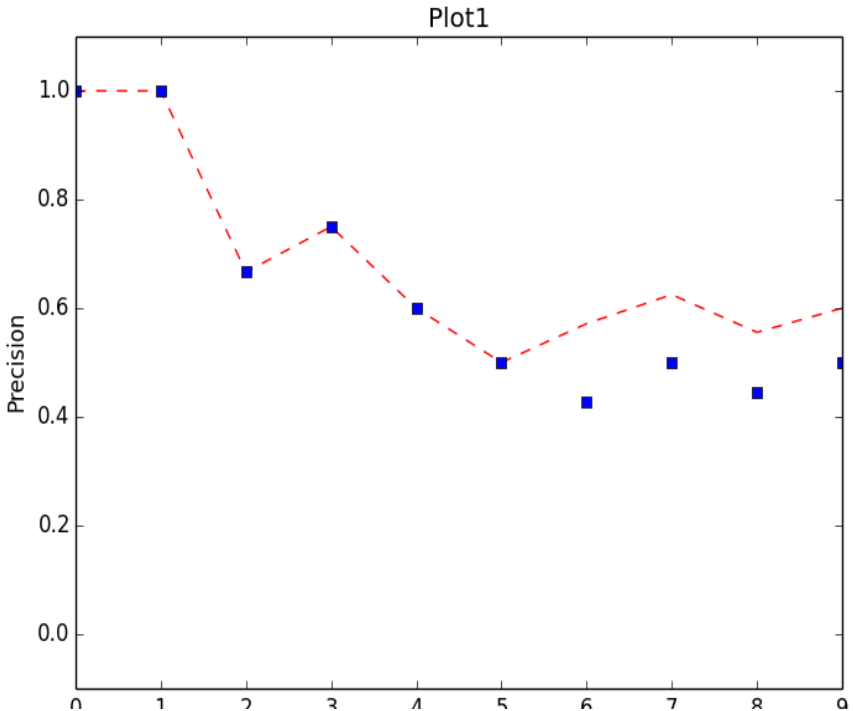
Precision considered one of the judges agree: 66.0

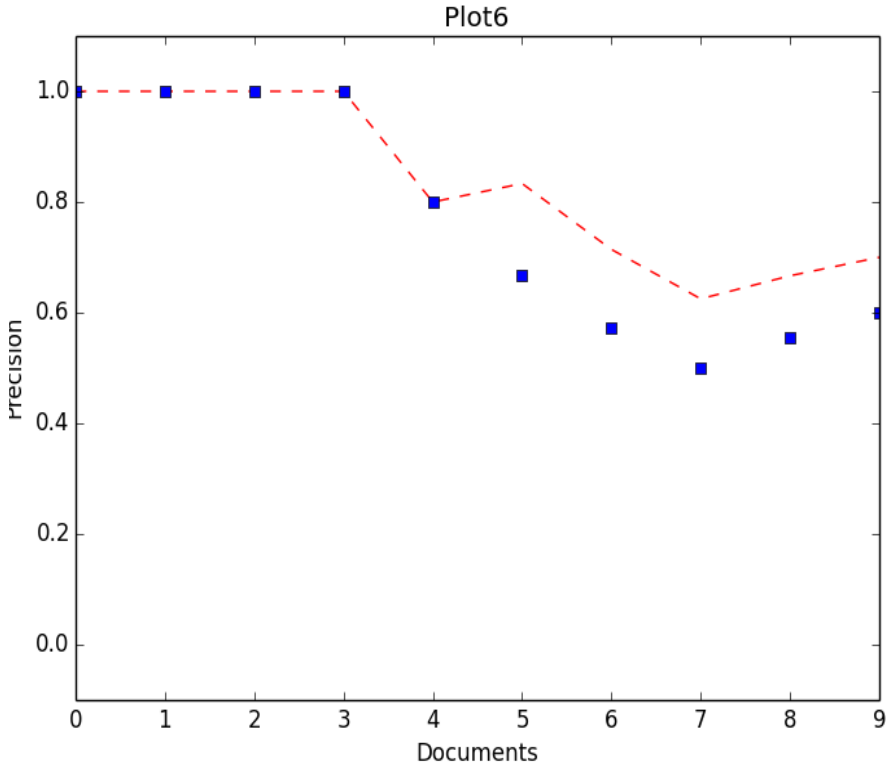
Precision considered two of the judges agree: 52.0

Different ranking

Only look at the answers

Problemen van chinezen in de samenleving





Evaluation

	Relevant	Not relevant
Relevant	21	3
Not relevant	3	23

Cohens kappa: 0.84

Precision considered one of the judges agree: 54.0

Precision considered two of the judges agree: 42.0