Elastic Search

Robin Hansma, Lars Lokhoff, Daan van Ingen

Resources

- Dataset: Telegraaf (1918-1994)
- Elastic Search
- Website: Flask

Indexeren

```
namespaces = {'pm': 'http://www.politicalmashup.nl', 'dc': 'http://purl.org/dc/elements/1.1/'}
files_indexed = []
print glob.glob( os.path.join(folder, '*.gz'))
# Loop over all files in folder
for filename in glob.glob( os.path.join(folder, '*.gz') ):
    files_indexed.append(filename)
    # Open them and parse XML
    with gzip.open(filename) as xml:
        tree = ET.parse(xml)

        # Get all articles and extract data from it
        root = tree.getroot()
        for article in root.findall('pm:root', namespaces):
            date = (article.find('pm:meta/dc:date', namespaces)).text
            subject = (article.find('pm:meta/dc:subject', namespaces)).text
            title = (article.find('pm:content/title', namespaces)).text
            if title == None:
                title = ''

            if (article.find('pm:content/text/p', namespaces)):
                text = (article.find('pm:content/text/p', namespaces)).text
                # Construct JSON data structure and store in ES
                data = {}
                data['date'] = date
                data['subject'] = subject
                data['title'] = unicodedata.normalize('NFKD', unicode(title)).encode('ascii','ignore')
                data['text'] = unicodedata.normalize('NFKD', unicode(text)).encode('ascii','ignore')
                json_data = json.dumps(data)
                res = es.index(index=INDEX, doc_type='article', body=json_data)
return True, files_indexed
```

Zoeken

```
55
56  def search(es, text, fields=[], filter_query={}):
57      """
58      Search
59
60      text -- Search query
61      fields -- fields to search in
62      filter_query -- filter
63
64      return [documents]
65      """
66      body = {
67        'query': {
68          'query_string': {
69              'query': text
70          }
71        },
72        'filter': filter_query
73      }
74
75      return es.search(index=INDEX, body=body)
```

Demo

Verbeterpunten

- Te doen:
  - Word cloud
  - Timeline
  - Geavanceerd zoeken
  - Opmaak
- Toekomst:
  - Verbeteren samenvatting
  - Gerelateerde zoekopdrachten