elk_base_enron_template

November 1, 2023

[ ]:
```
!pip install elasticsearch
```

```
Requirement already satisfied: elasticsearch in /opt/conda/lib/python3.11/site-
packages (8.10.1)
Requirement already satisfied: elastic-transport<9,>=8 in
/opt/conda/lib/python3.11/site-packages (from elasticsearch) (8.10.0)
Requirement already satisfied: urllib3<3,>=1.26.2 in
/opt/conda/lib/python3.11/site-packages (from elastic-
transport<9,>=8->elasticsearch) (2.0.7)
Requirement already satisfied: certifi in /opt/conda/lib/python3.11/site-
packages (from elastic-transport<9,>=8->elasticsearch) (2023.7.22)
```

[ ]:
```
!pip install gdown
```

```
Requirement already satisfied: gdown in /opt/conda/lib/python3.11/site-packages
(4.7.1)
Requirement already satisfied: filelock in /opt/conda/lib/python3.11/site-
packages (from gdown) (3.13.1)
Requirement already satisfied: requests[socks] in
/opt/conda/lib/python3.11/site-packages (from gdown) (2.31.0)
Requirement already satisfied: six in /opt/conda/lib/python3.11/site-packages
(from gdown) (1.16.0)
Requirement already satisfied: tqdm in /opt/conda/lib/python3.11/site-packages
(from gdown) (4.66.1)
Requirement already satisfied: beautifulsoup4 in /opt/conda/lib/python3.11/site-
packages (from gdown) (4.12.2)
Requirement already satisfied: soupsieve>1.2 in /opt/conda/lib/python3.11/site-
packages (from beautifulsoup4->gdown) (2.5)
Requirement already satisfied: charset-normalizer<4,>=2 in
/opt/conda/lib/python3.11/site-packages (from requests[socks]->gdown) (3.3.0)
Requirement already satisfied: idna<4,>=2.5 in /opt/conda/lib/python3.11/site-
packages (from requests[socks]->gdown) (3.4)
Requirement already satisfied: urllib3<3,>=1.21.1 in
/opt/conda/lib/python3.11/site-packages (from requests[socks]->gdown) (2.0.7)
Requirement already satisfied: certifi>=2017.4.17 in
/opt/conda/lib/python3.11/site-packages (from requests[socks]->gdown)
(2023.7.22)
Requirement already satisfied: PySocks!=1.5.7,>=1.5.6 in
/opt/conda/lib/python3.11/site-packages (from requests[socks]->gdown) (1.7.1)
```

0.1 Setting up the Elasticsearch

[ ]:
```
from datetime import datetime
from elasticsearch import Elasticsearch

ELASTIC_PASSWORD = "p2iFCHUbC7ze1QoIMVw"

es = Elasticsearch("http://elasticsearch:9200",
                   basic_auth=("elastic", ELASTIC_PASSWORD))

es.info()
```

[ ]:
```
ObjectApiResponse({'name': 'es-node', 'cluster_name': 'tdt4117-ir-data-cluster',
'cluster_uuid': 'rAkxpaFNSfypbZhfl9rXuA', 'version': {'number': '8.4.2',
'build_flavor': 'default', 'build_type': 'docker', 'build_hash':
'89f8c6d8429db93b816403ee75e5c270b43a940a', 'build_date':
'2022-09-14T16:26:04.382547801Z', 'build_snapshot': False, 'lucene_version':
'9.3.0', 'minimum_wire_compatibility_version': '7.17.0',
'minimum_index_compatibility_version': '7.0.0'}, 'tagline': 'You Know, for
Search'})
```

0.2 Setting up the documents

[ ]:
```
# Downloading the data
"""
The data is uploaded at a google drive directory, by running this code box,
the data will be downloaded and you will have access to it at
(enron_short/maildir)
"""
# import gdown

# googleDriveURL = "https://drive.google.com/file/d/
  ↪10OwgK91eOlNRsrAV31sSJu7KJUujBTnE/view?usp=sharing"
# output = 'enron_short.tar.gz'
# gdown.download(googleDriveURL, output, quiet=False, fuzzy=True)
```

[ ]:
```
'\nThe data is uploaded at a google drive directory, by running this code box,
\nthe data will be downloaded and you will have access to it
at\n(enron_short/maildir)\n'
```

[ ]:
```
# check if the directory exists
from os import path

zipfile_path = "enron_short.tar.gz"
if not path.isfile(zipfile_path):
    print("the zipfile is not here, please ensure you download it first.")
```

[ ]:
```
!tar xf enron_short.tar.gz
```

[ ]:
```
# check if the directory exists
from os import path

documents_path = r"enron_short/maildir"
if not path.exists(documents_path):
    print("the directory is not here, please ensure you have the documnets␣
 ↪first.")
```

0.3 Your Code:

[ ]:
```
# Index all emails folder enron_short/maildir
es.indices.delete(index="index", ignore_unavailable=True) # Delete the index if␣
 ↪it exists
# Create the index with mapping for email body
es.indices.create(index='index', body={
    "mappings": {
        "properties": {
            "body": {
                "type": "text"
            }
        }
    }
})

import os
from tqdm import tqdm

i = 1
for root, dirs, files in tqdm(os.walk(documents_path)):
    for file in files:
        with open(os.path.join(root, file), "r", encoding="utf-8",␣
 ↪errors="ignore") as f:
            data = f.read()
            es.index(index="index", id=i, body={"content": data})
            i += 1

es.get(index="index", id=1)
```

```
192821it [25:36, 125.48it/s]
```

[ ]:
```
ObjectApiResponse({'_index': 'index', '_id': '1', '_version': 1, '_seq_no': 0,
'_primary_term': 1, '_ignored': ['content.keyword'], 'found': True, '_source':
{'content': 'Message-ID: <1706910.1075842665022.JavaMail.evans@thyme>\nDate:
Wed, 16 Jun 1999 08:32:00 -0700 (PDT)\nFrom: bieraugel@efortress.com\nTo:
gerald.nemec@enron.com\nSubject: Catching up\nMime-Version: 1.0\nContent-Type:
text/plain; charset=us-ascii\nContent-Transfer-Encoding: 7bit\nX-From: "Paul
Bieraugel" <bieraugel@efortress.com>\nX-To: Gerald Nemec\nX-cc: \nX-bcc: \nX-
```

```
Folder: \\Gerald_Nemec_Dec2000_June2001_1\\Notes Folders\\Personal\nX-Origin:
NEMEC-G\nX-FileName: gnemec.nsf\n\nGerald,  good to hear from you and very
belated congrats on your law\ndegree/bar exam/ascent to layerhood deal.  That\'s
way cool.  Sounds like you\nare really doing well out there.  Tool belts, house
restoration, sounds like\nfun.  Too bad the Navy won\'t allow me the time to do
it.   Regardless,\nthey\'ve been moving me around so much lately that it\'s all
pointless.  I got\nback from Italy in Feb.  I started school here in Newport,
RI.  Next I\'l\nspend 3 months in Northern VA going to school.  Then, I\'ll meet
my ship in\nthe Persian Gulf for a fun filled 4 months in the Desert.  I\'m
getting\npretty down on the Navy and may be considering a job change in the next
two\nyears.  Well, to answer your question, Michele and I have three kids
now.\nKelly is four, Carson is three, and Katherine (Kate) is two months
(born\n4/15/99).  I figure we\'re done for now but will have to check back when
Kate\nis too big to cuddle.  Good to hear George and Ernie are still doing
well.\nCrazy days we had together.  Funny how we all tend to grow up
eventually.\nMy parents are fine.  They are almost finished with the house on
the ranch.\nWhen that is done, they plan to move out there permanently.  John is
working\nsemi-part time as a governemnt employee counting fish on alaskan
fishing\nboats.  Dangerous job/low pay.  But he likes it and he has his summers
for\nsurfing expeditions (i.e.  he is still bumming around).   Mark is
the\nbiggest suprise, he went back to school and got a Masters degree in
library\nscience and information technology.  He did really well and now has a
job\nwith Microsoft working as an archiver (apparently librarians and
computer\ngeeks too now).  Hey good to hear from you.   don\'t know when/how we
can\nwork in a visit.  Possibly after March next year unless you\'re out in
San\nDiego while Im in the Gulf and you can look Michele up.  Keep in touch
and\nwell work something out.  Take care Gerald.  Paul.'}})
```

[ ]:
```python
def pretty_print(query, res):
    print("Query: ", query)
    print("Number of results: ", res["hits"]["total"]["value"])
    print("Results: ")
    for hit in res["hits"]["hits"]:
        print("Score: ", hit["_score"])
        print("Content: ", hit["_source"]["content"][:100] + "...")
```

[ ]:
```python
# Query for "Norwegian and University and Science and Technology" and␣
↪"Norwegian University Science Technology"

query = "Norwegian and University and Science and Technology"
res = es.search(index="index", body={"query": {"match": {"content": query}}})
pretty_print(query, res)
```

```
Query:  Norwegian and University and Science and Technology
Number of results:  10000
Results:
Score:  22.428768
Content:  Message-ID: <21996921.1075855468318.JavaMail.evans@thyme>
```

```
Date: Fri, 21 Dec 2001 04:43:38 -0800 (PST…
Score:  22.428768
Content:  Message-ID: <24164547.1075840774929.JavaMail.evans@thyme>
Date: Fri, 21 Dec 2001 04:43:38 -0800 (PST…
Score:  21.451519
Content:  Message-ID: <33145748.1075855468410.JavaMail.evans@thyme>
Date: Sun, 23 Dec 2001 14:45:10 -0800 (PST…
Score:  21.451519
Content:  Message-ID: <21179930.1075840774807.JavaMail.evans@thyme>
Date: Sun, 23 Dec 2001 14:45:10 -0800 (PST…
Score:  21.329004
Content:  Message-ID: <24069195.1075846942557.JavaMail.evans@thyme>
Date: Mon, 24 Apr 2000 11:26:00 -0700 (PDT…
Score:  21.329004
Content:  Message-ID: <9729632.1075847069332.JavaMail.evans@thyme>
Date: Mon, 24 Apr 2000 11:26:00 -0700 (PDT)…
Score:  21.329004
Content:  Message-ID: <4007612.1075855889171.JavaMail.evans@thyme>
Date: Mon, 24 Apr 2000 11:26:00 -0700 (PDT)…
Score:  20.510086
Content:  Message-ID: <17207457.1075840768406.JavaMail.evans@thyme>
Date: Fri, 11 Jan 2002 02:15:10 -0800 (PST…
Score:  20.136003
Content:  Message-ID: <2414413.1075840790149.JavaMail.evans@thyme>
Date: Fri, 18 Jan 2002 13:06:28 -0800 (PST)…
Score:  18.675121
Content:  Message-ID: <15334861.1075856302114.JavaMail.evans@thyme>
Date: Mon, 7 Aug 2000 03:20:00 -0700 (PDT)…
```

[ ]:
```
query = "Norwegian University Science Technology"
res = es.search(index="index", body={"query": {"match": {"content": query}}})
pretty_print(query, res)
```

```
Query:  Norwegian University Science Technology
Number of results:  9744
Results:
Score:  20.404533
Content:  Message-ID: <21996921.1075855468318.JavaMail.evans@thyme>
Date: Fri, 21 Dec 2001 04:43:38 -0800 (PST…
Score:  20.404533
Content:  Message-ID: <24164547.1075840774929.JavaMail.evans@thyme>
Date: Fri, 21 Dec 2001 04:43:38 -0800 (PST…
Score:  19.514381
Content:  Message-ID: <33145748.1075855468410.JavaMail.evans@thyme>
Date: Sun, 23 Dec 2001 14:45:10 -0800 (PST…
Score:  19.514381
Content:  Message-ID: <21179930.1075840774807.JavaMail.evans@thyme>
Date: Sun, 23 Dec 2001 14:45:10 -0800 (PST…
```

```
Score:  18.755571
Content:  Message-ID: <24069195.1075846942557.JavaMail.evans@thyme>
Date: Mon, 24 Apr 2000 11:26:00 -0700 (PDT…
Score:  18.755571
Content:  Message-ID: <9729632.1075847069332.JavaMail.evans@thyme>
Date: Mon, 24 Apr 2000 11:26:00 -0700 (PDT)…
Score:  18.755571
Content:  Message-ID: <4007612.1075855889171.JavaMail.evans@thyme>
Date: Mon, 24 Apr 2000 11:26:00 -0700 (PDT)…
Score:  18.455051
Content:  Message-ID: <17207457.1075840768406.JavaMail.evans@thyme>
Date: Fri, 11 Jan 2002 02:15:10 -0800 (PST…
Score:  18.111769
Content:  Message-ID: <2414413.1075840790149.JavaMail.evans@thyme>
Date: Fri, 18 Jan 2002 13:06:28 -0800 (PST)…
Score:  16.287764
Content:  Message-ID: <15334861.1075856302114.JavaMail.evans@thyme>
Date: Mon, 7 Aug 2000 03:20:00 -0700 (PDT)…
```