CS 5604: Information Storage and Retrieval
Elasticsearch

•••

Soumya Arvind Kumar

Yuan Li

Nicholas Gill

Satvik Chekuri

Tianrui Hu

Instructor: Dr. Edward A. Fox

TA: Ziqian Song

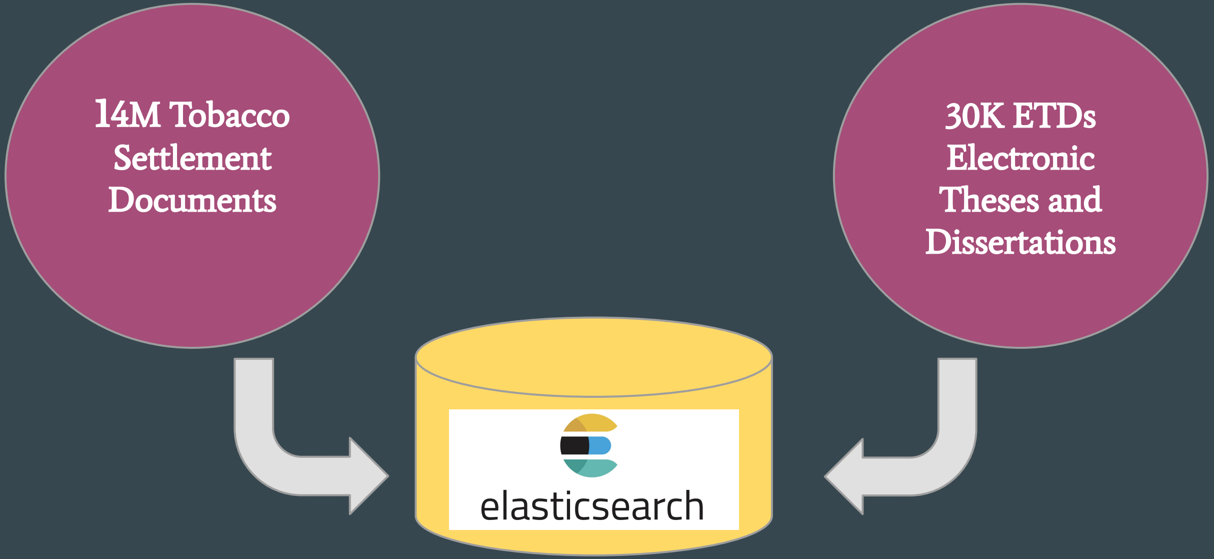
12/10/19

Virginia Tech, Blacksburg, VA, 24060

PROJECT OVERVIEW

Problem Statement

Build an Information and Retrieval System that will act as a search engine to support ranking, searching, browsing and recommendations for two large collections of data:



14M Tobacco Settlement Documents

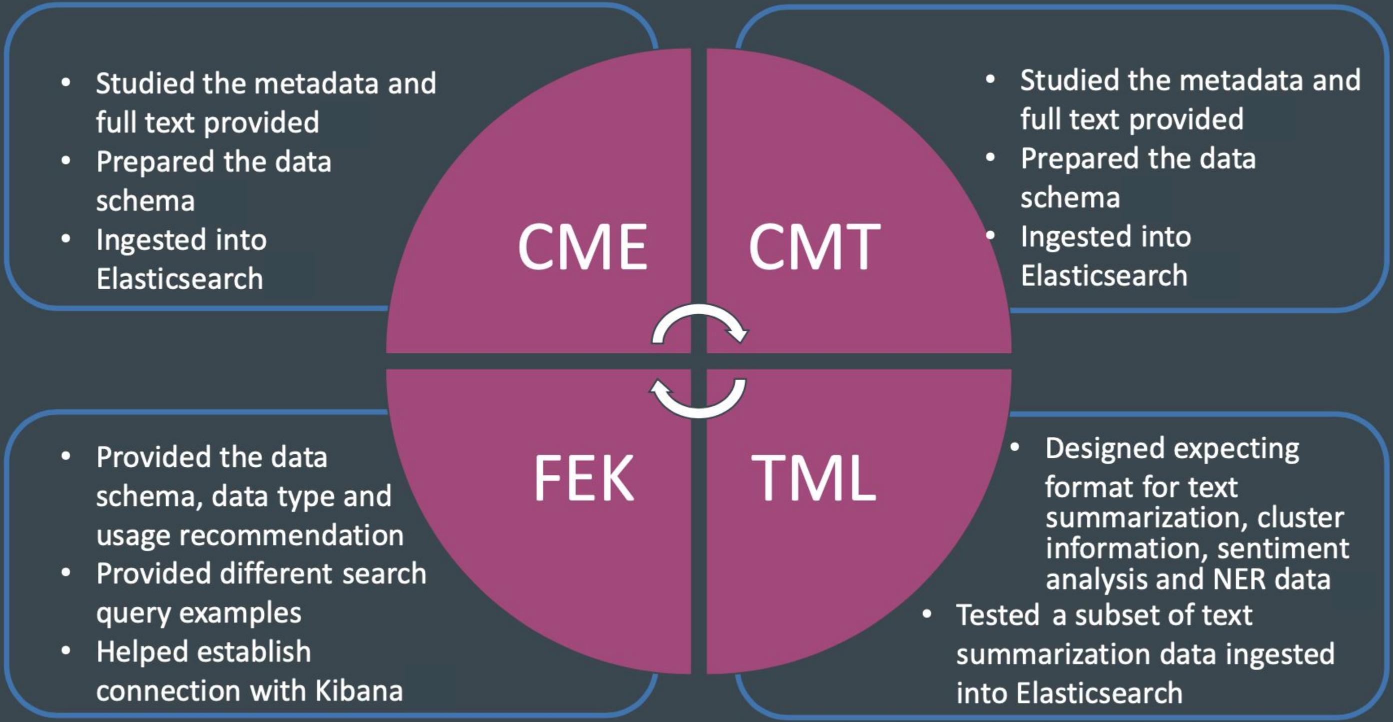
30K ETDs Electronic Theses and Dissertations

elasticsearch

Requirements for Elasticsearch

- Ingest data provided by the CME and CMT teams into Elasticsearch in the correct format.
- Decide the relevancy and importance of fields related to the ETD and tobacco dataset and provide feedback on the same.
- Incorporate additional data from TML team related to text summarisation, name entity recognition, sentiment analysis, and clustering information.
- For enhanced search accuracy, perform boosting to assign higher weights to important fields.
- Implement nested queries for in-depth search inside each document.
- Establish connection with Kibana to support searching, browsing and information visualisation.
- Implement automatic ingesting and updating scripts to monitor a designated directory on ceph for new incoming files.

Contribution to Other Teams



CME

- Studied the metadata and full text provided
- Prepared the data schema
- Ingested into Elasticsearch

CMT

- Studied the metadata and full text provided
- Prepared the data schema
- Ingested into Elasticsearch

FEK

- Provided the data schema, data type and usage recommendation
- Provided different search query examples
- Helped establish connection with Kibana

TML

- Designed expecting format for text summarization, cluster information, sentiment analysis and NER data
- Tested a subset of text summarization data ingested into Elasticsearch

Achievements

CME

99.8%

30,925 Electronic Thesis Documents ingested including metadata and full text.

- Fully searchable documents
- Can be filtered and sorted.
- Prepared automated script for addition of new documents

CMT

99.9%

5,595,936 Tobacco Settlement Documents metadata ingested (81 failed); including 100,000 metadata and full text.

- Fully searchable documents
- Can be filtered and sorted.
- Prepared automated script for addition of new documents.

TML

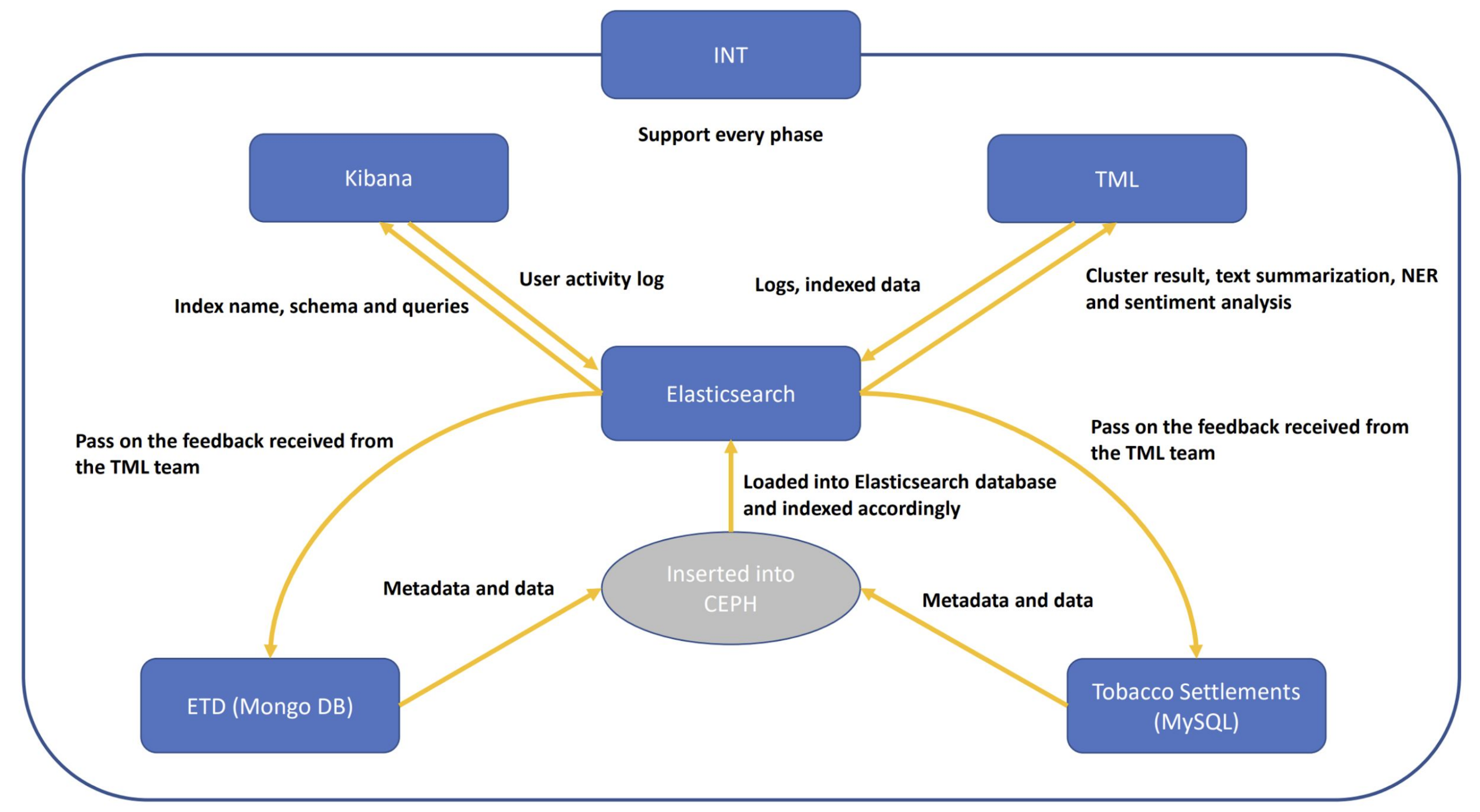
In Progress

Text Summarization, Sentiment Analysis, Named-Entity Recognition, Cluster Data

- Tested the text summarisation format.
- Receiving data from TML.
- Work in progress.

DESIGN & IMPLEMENTATION

Concept Map for Elasticsearch



INT
Support every phase
Kibana
TML
User activity log
Index name, schema and queries
Logs, indexed data
Cluster result, text summarization, NER and sentiment analysis
Elasticsearch
Pass on the feedback received from the TML team
Pass on the feedback received from the TML team
Loaded into Elasticsearch database and indexed accordingly
Metadata and data
Inserted into CEPH
Metadata and data
ETD (Mongo DB)
Tobacco Settlements (MySQL)

Tobacco Data Schema

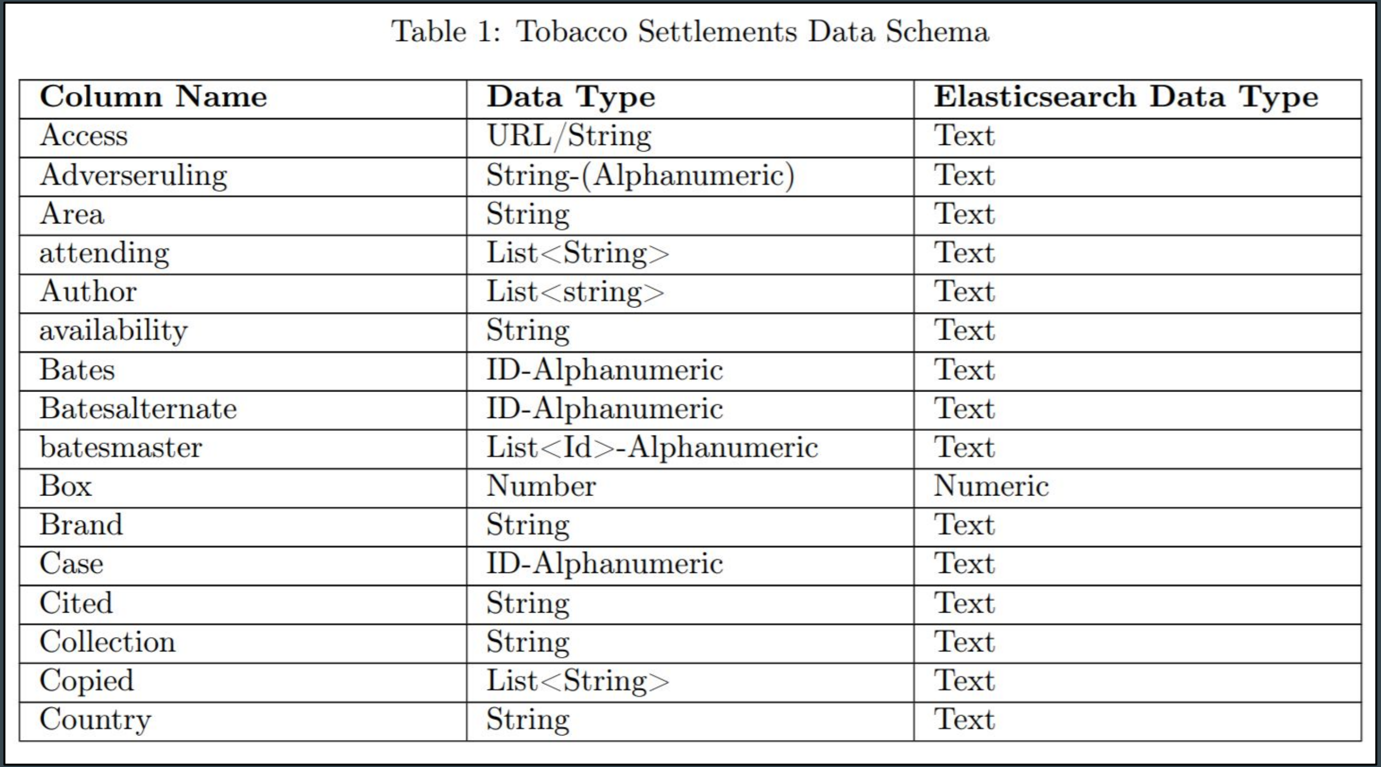


Table 1: Tobacco Settlements Data Schema

Column Name	Data Type	Elasticsearch Data Type
Access	URL/String	Text
Adverseruling	String-(Alphanumeric)	Text
Area	String	Text
attending	List<String>	Text
Author	List<string>	Text
availability	String	Text
Bates	ID-Alphanumeric	Text
Batesalternate	ID-Alphanumeric	Text
batesmaster	List<Id>-Alphanumeric	Text
Box	Number	Numeric
Brand	String	Text
Case	ID-Alphanumeric	Text
Cited	String	Text
Collection	String	Text
Copied	List<String>	Text
Country	String	Text

ETD Data Schema

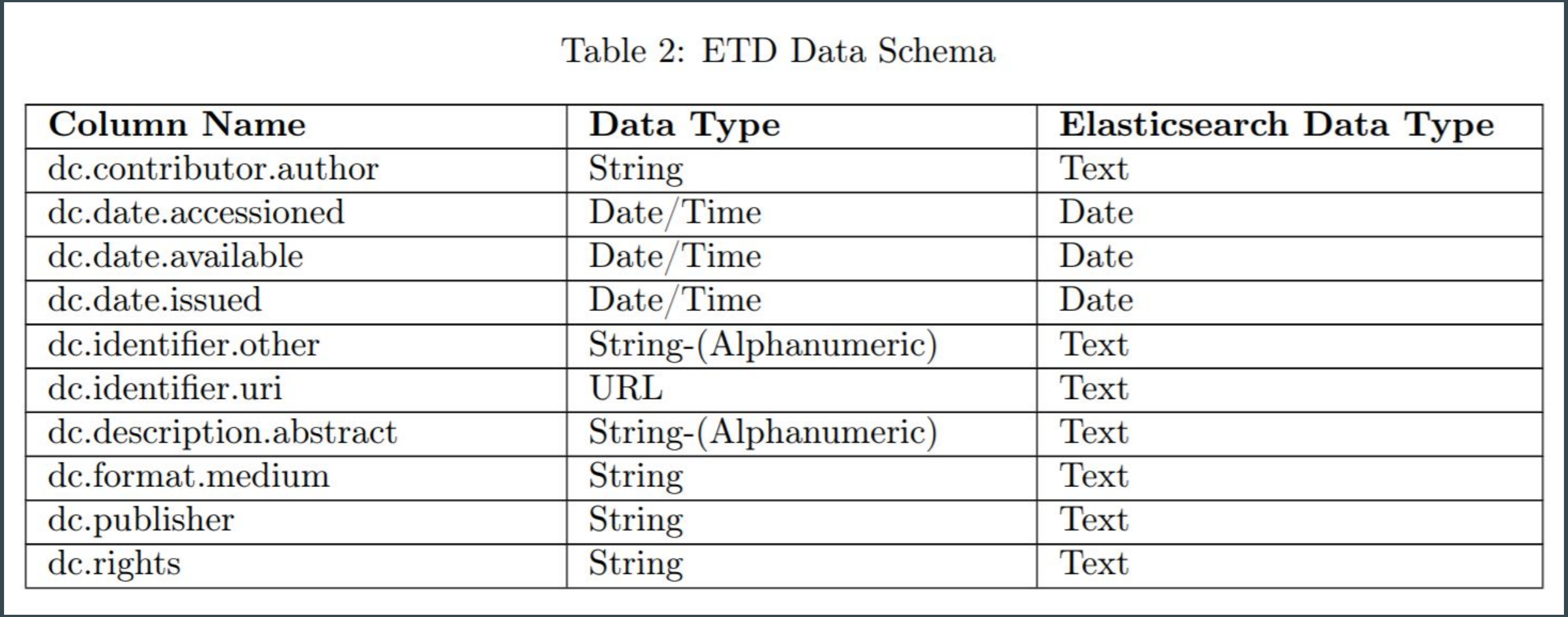
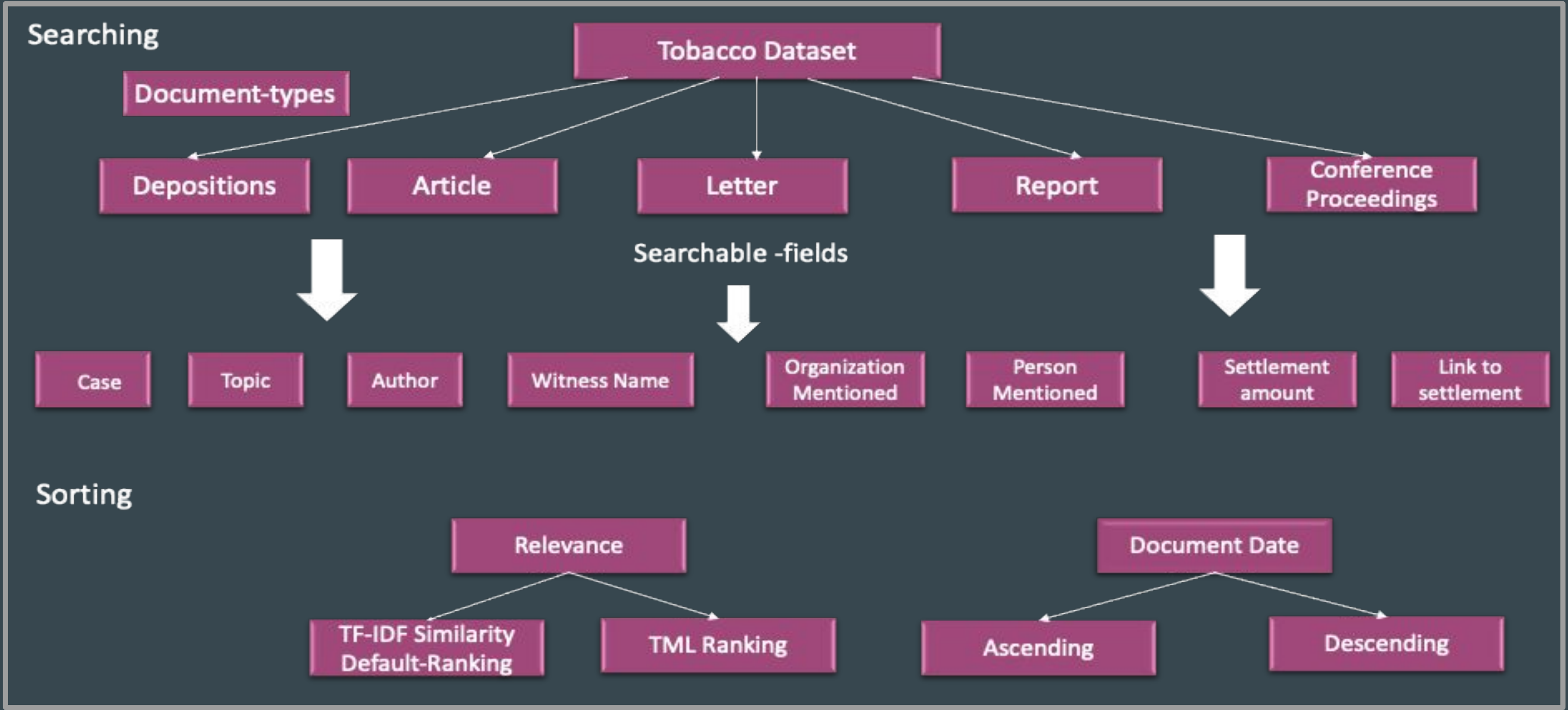


Table 2: ETD Data Schema

Column Name	**Data Type**	**Elasticsearch Data Type**
dc.contributor.author	String	Text
dc.date.accessioned	Date/Time	Date
dc.date.available	Date/Time	Date
dc.date.issued	Date/Time	Date
dc.identifier.other	String-(Alphanumeric)	Text
dc.identifier.uri	URL	Text
dc.description.abstract	String-(Alphanumeric)	Text
dc.format.medium	String	Text
dc.publisher	String	Text
dc.rights	String	Text

Tobacco Data Structure



Searching

Document-types

Tobacco Dataset

Depositions

Article

Letter

Report

Conference Proceedings

Searchable -fields

Case

Topic

Author

Witness Name

Organization Mentioned

Person Mentioned

Settlement amount

Link to settlement

Sorting

Relevance

Document Date

TF-IDF Similarity Default-Ranking

TML Ranking

Ascending

Descending

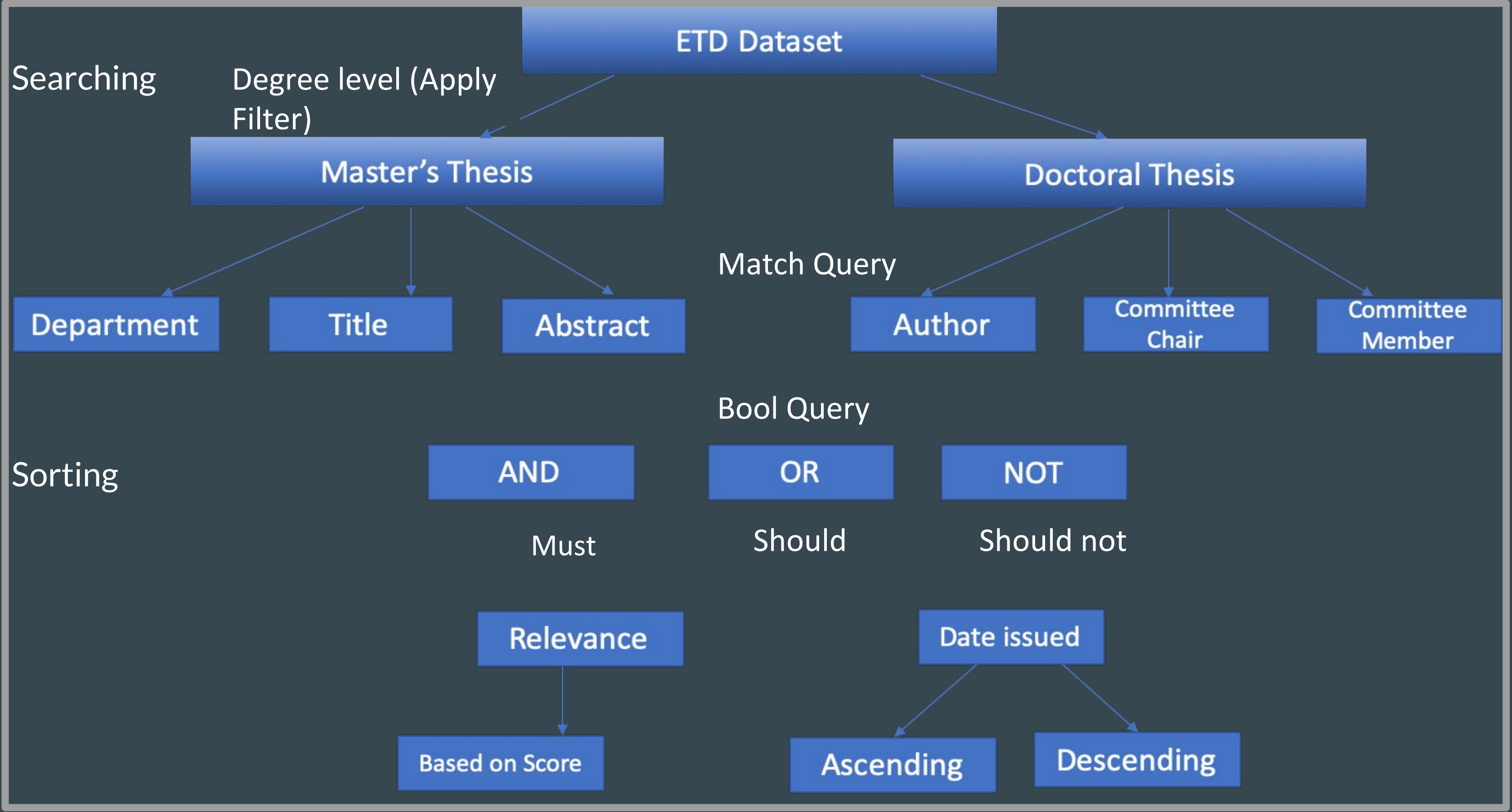
Fields for Searching and Filtering:

TOBACCO SETTLEMENT DOCUMENTS

For all field types of ‘Text’, use ***field_name*** for searching and ***field_name.keyword*** for filtering or sorting

Field Name	Field Type	Field Demo
Case	text	Minnesota v. Philip Morris Inc.
Brands	text	Marlboro
Witness_Name	text	"Wyant, Timothy (affiliation: Decipher; expertise: Statistical analysis; job_title:
Topic	text	advertising; health effects
Person_Mentioned	text	Burns, David Michael, M.D
Organization_Mentioned	text	R.J. Reynolds Tobacco Co.
Description	text	"The plaintiffs expert witness, a statistician, was deposed”
Title	text	"Deposition of TIMOTHY S. WYANT, Ph.D., August 19, 1997
Date_Added_UCSF	text	20 January 2006
Document_Date	text	19 August 1997
Cluster	text/keyword	321
page	text/keyword	5
content	text/keyword	Paper details

ETD Data Structure



Fields for Searching and Filtering:

ETDs

For all field types of ‘Text’, use ***field_name*** for searching and ***field_name.keyword*** for filtering or sorting

Field Name	FIeld Type	Field Demo
degree-level	text	masters
contributor-department	text	Computer science
contributor-author	text	Tony Stark
Contributor-committee chair	text	John wick
Contributor-committee co-chair	text	Chris scott
Contributor-committee member	text	David knight
date-available	date	2017-01-23
date-issued	date	2018-02-21
degree-name	text	MS or P.hD
description-abstract	text	This field conveys the abstract of the thesis in 10-15 lines
Author Email	text	tony_s@stark.com
subject-none	text	Soils -- Aluminum content Cations
title-none	text	Hydrolysis of aluminum in synthetic cation exchange
type-none	text	Dissertation

Indexing Methods

Metadata

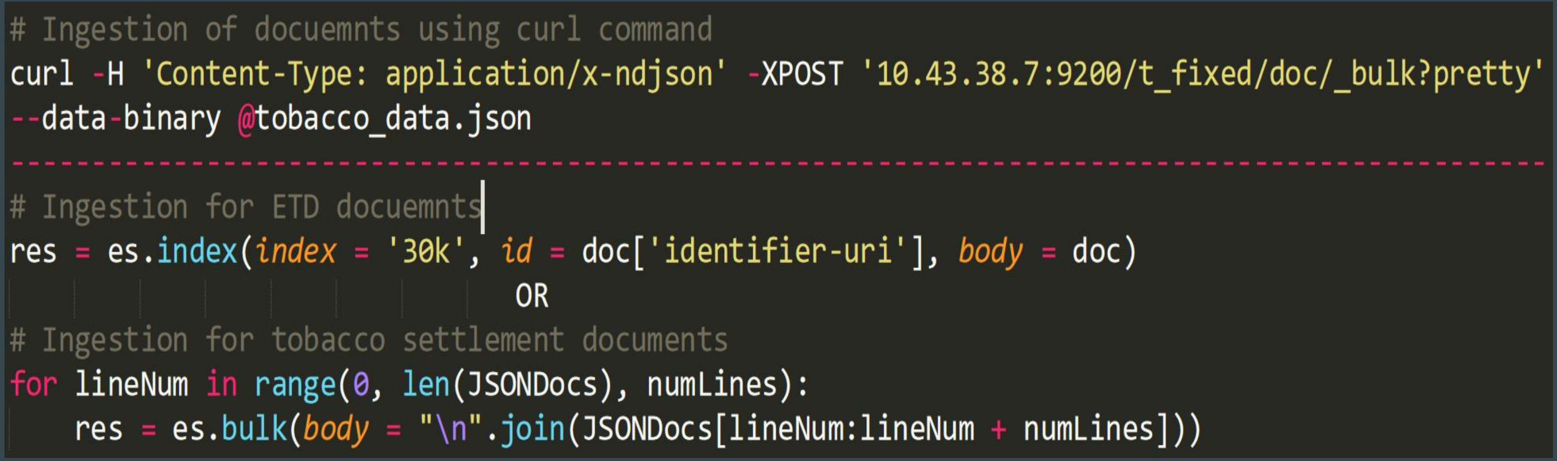
Stores the records detail that describes and gives information about the source data

Data

Stores the text content of the ETD and tobacco settlement datasets (page-wise)

Generated Data

Data generated by the TML team consists of cluster ID, text summary, sentiment analysis and NER keywords



```
# Ingestion of docuemnts using curl command
curl -H 'Content-Type: application/x-ndjson' -XPOST '10.43.38.7:9200/t_fixed/doc/_bulk?pretty'
--data-binary @tobacco_data.json
-------------------------------------------------------------------------------------------------

# Ingestion for ETD docuemnts
res = es.index(index = '30k', id = doc['identifier-uri'], body = doc)
                                OR
# Ingestion for tobacco settlement documents
for lineNum in range(0, len(JSONDocs), numLines):
    res = es.bulk(body = "\n".join(JSONDocs[lineNum:lineNum + numLines]))
```

Executable python script on ceph in els directory

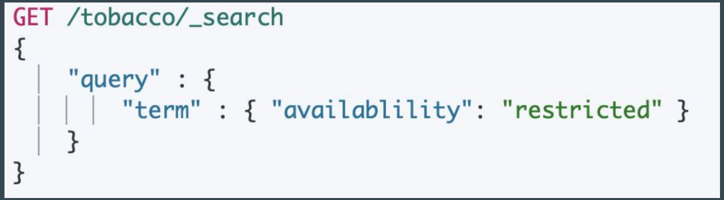
Ingesting by Elasticsearch-Python Client

Parsing files into designed format for ingesting

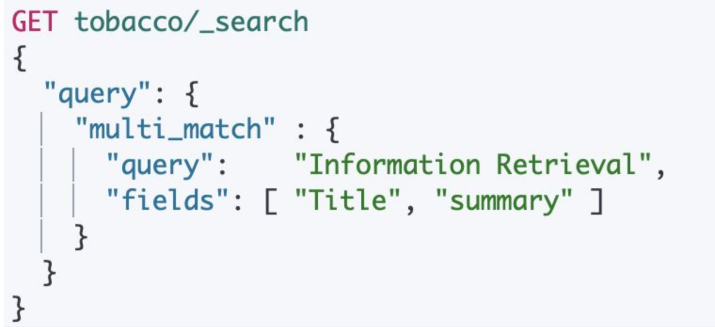
Assign the ID and the name of index

Logging errors (document ID and error messages)

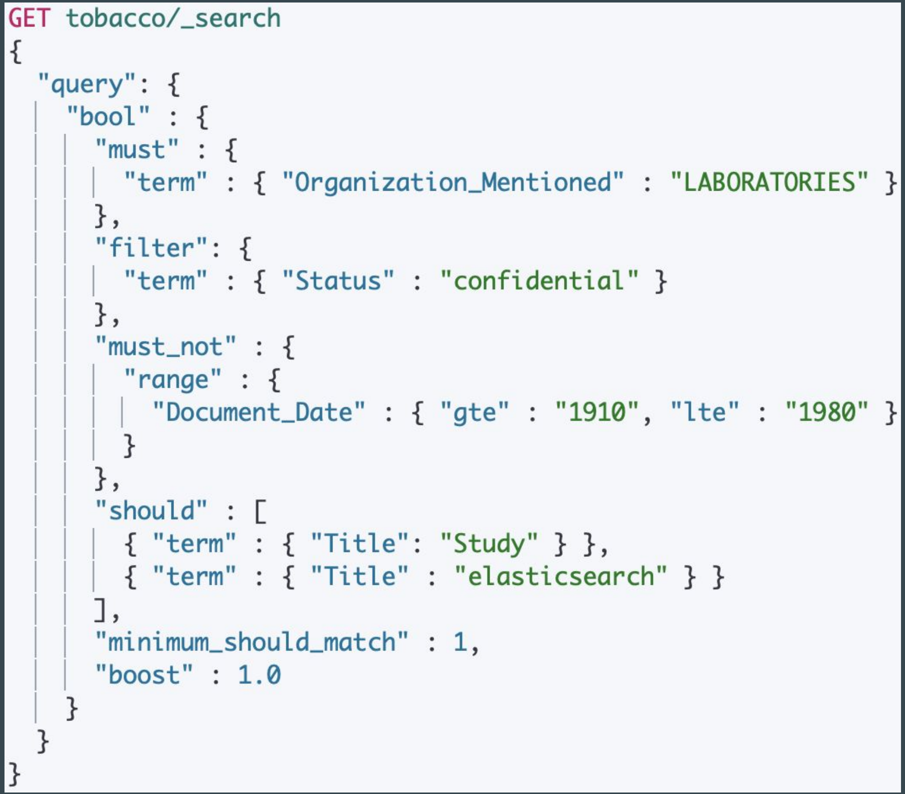
Searching Query



```
GET /tobacco/_search
{
    "query" : {
        "term" : { "availablility": "restricted" }
    }
}
```

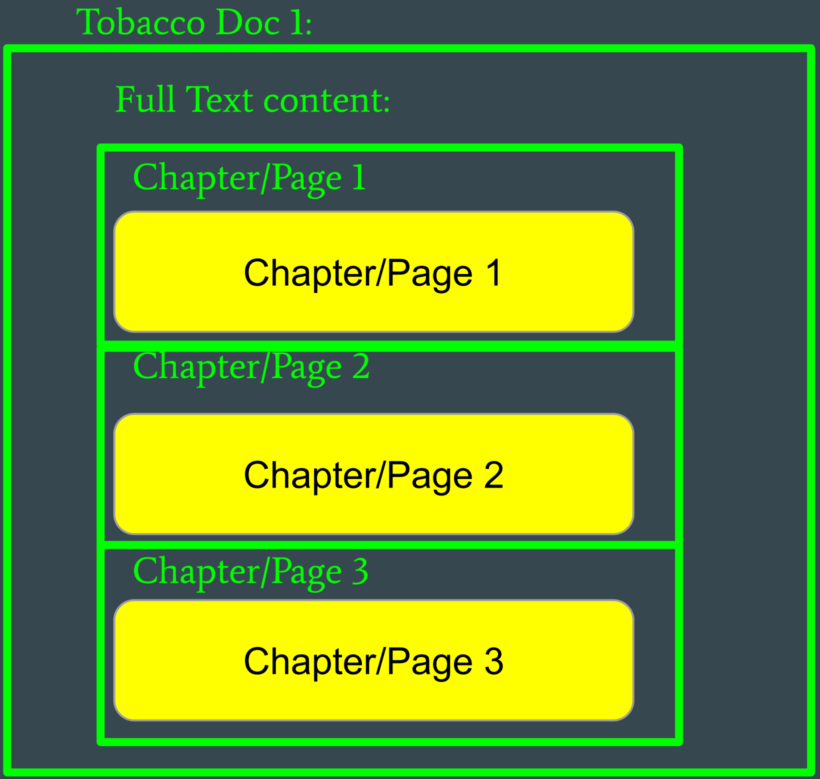


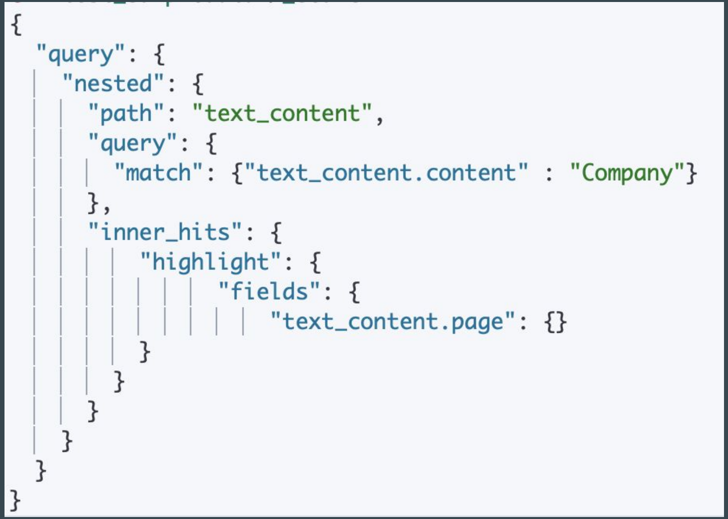
```
GET tobacco/_search
{
  "query": {
    "multi_match" : {
      "query":    "Information Retrieval",
      "fields": [ "Title", "summary" ]
    }
  }
}
```



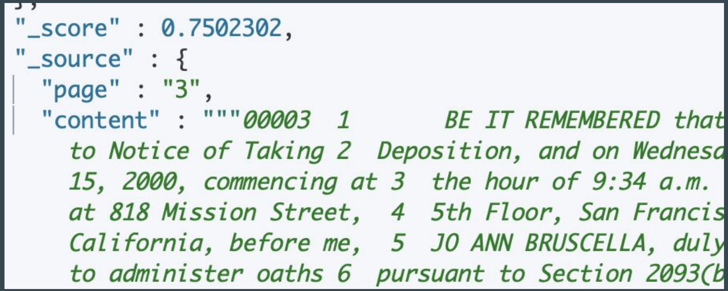
```
GET tobacco/_search
{
  "query": {
    "bool" : {
      "must" : {
        "term" : { "Organization_Mentioned" : "LABORATORIES" }
      },
      "filter": {
        "term" : { "Status" : "confidential" }
      },
      "must_not" : {
        "range" : {
          "Document_Date" : { "gte" : "1910", "lte" : "1980" }
        }
      },
      "should" : [
        { "term" : { "Title": "Study" } },
        { "term" : { "Title" : "elasticsearch" } }
      ],
      "minimum_should_match" : 1,
      "boost" : 1.0
    }
  }
}
```

Full Text Search: Nested Query





```
{
  "query": {
    "nested": {
      "path": "text_content",
      "query": {
        "match": {"text_content.content" : "Company"}
      },
      "inner_hits": {
        "highlight": {
          "fields": {
            "text_content.page": {}
          }
        }
      }
    }
  }
}
```



```
"_score" : 0.7502302,
"_source" : {
  "page" : "3",
  "content" : """00003  1       BE IT REMEMBERED that
    to Notice of Taking 2  Deposition, and on Wednesd
    15, 2000, commencing at 3  the hour of 9:34 a.m.
    at 818 Mission Street,  4  5th Floor, San Francis
    California, before me,  5  JO ANN BRUSCELLA, duly
    to administer oaths 6  pursuant to Section 2093(b
```

Search Preference: Boosting

Elasticsearch rank searching results based on a designed score.

The scores are calculated by a similarity model based on Term Frequency (TF) and Inverse Document Frequency (IDF) as well as using the Vector Space Model (VSM) for multi-term queries.

Search Preference: Boosting

Field 1, with no boost

Field 1

Field 2, with boost weight = 2

Field 2

Field 3, with boost weight = 0.5

Field 3

Score = field_1 + 2 * field_2 + 0.5 * field_3

```
{ETD Doc 1: field_1: A,
            field_2: None,
            field_3: None}
{ETD Doc 2: field_1: None,
            field_2: A,
            field_3: None}
{ETD Doc 3: field_1: None,
            field_2: None,
            field_3: A}
```

Searching for A:
score_2 > score_1 > score_3

Logging

User Logs:

User-oriented information: username, timestamp, query content, IP, cookie, user-agent, etc.

Recommendation, detecting malicious user behaviors, website data analysis.

Index: .logging-yyyy/mm/dd

Logging

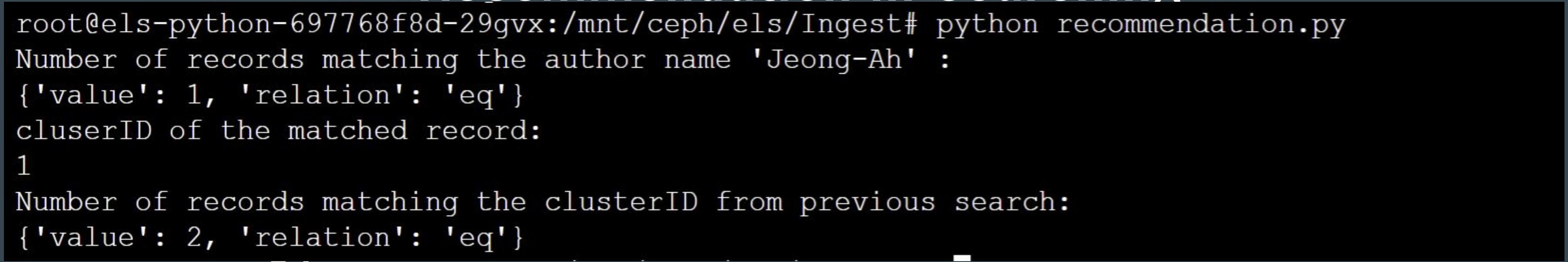
System Logs:

Event/request recording: timestamp, cluster.name, node.name, cluster.uuid, request/event message.

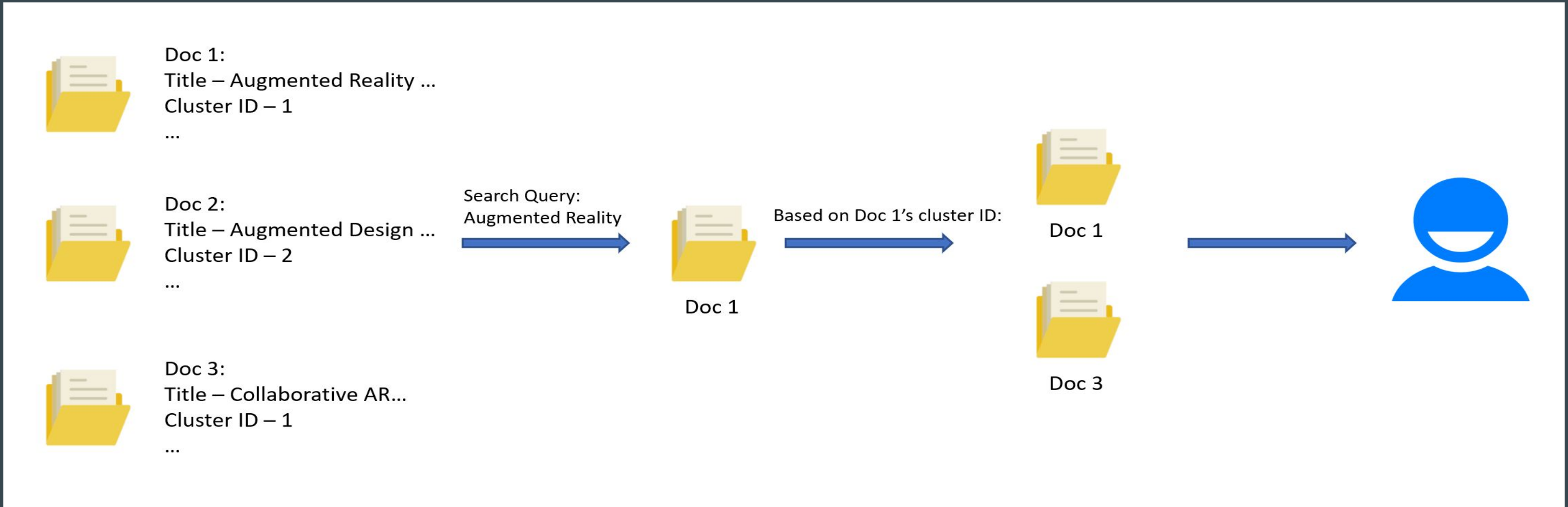
```
PUT /tobacco/_settings
{
    "index.search.slowlog.threshold.query.warn": "1s",
    "index.search.slowlog.threshold.query.info": "1s",
    "index.search.slowlog.threshold.query.debug": "2s",
    "index.search.slowlog.threshold.query.trace": "500ms",
    "index.search.slowlog.threshold.fetch.warn": "1s",
    "index.search.slowlog.threshold.fetch.info": "800ms",
    "index.search.slowlog.threshold.fetch.debug": "500ms",
    "index.search.slowlog.threshold.fetch.trace": "200ms",
    "index.search.slowlog.level": "info"
}
```

```
          {"type": "index_search_slowlog", "timestamp":
"2019-12-04T01:09:09,002Z", "level": "WARN", "component": "i.s.s.query",
"cluster.name": "elasticsearch", "node.name": "elasticsearch-master-0",
"message": "[30k][0]", "took": "930.9ms", "took_millis": "930", "total_hits": "19
hits", "stats": "[]", "search_type": "QUERY_THEN_FETCH", "total_shards":
"1", "source":
"{\"query\":{\"term\":{\"title-none\":{\"value\":\"data\",\"boost\":1.0}}}}",
"cluster.uuid": "M7gJSQVkSYi3THDYCTvIew", "node.id":
"nXkX9qONS2y0g5WB8NGezQ" }
          {"type": "index_search_slowlog", "timestamp":
"2019-12-04T01:17:14,635Z", "level": "WARN", "component": "i.s.s.fetch",
"cluster.name": "elasticsearch", "node.name": "elasticsearch-master-1",
"message": "[tobacco][0]", "took": "1.5ms", "took_millis": "1", "total_hits": "446
hits", "stats": "[]", "search_type": "QUERY_THEN_FETCH", "total_shards":
"1", "source":
"{\"query\":{\"term\":{\"Topic\":{\"value\":\"health\",\"boost\":1.0}}}}",
"cluster.uuid": "M7gJSQVkSYi3THDYCTvIew", "node.id":
"iLagChv6S8OxTzRhY9yLFQ" }
```

Recommendation in Searching



```
root@els-python-697768f8d-29gvx:/mnt/ceph/els/Ingest# python recommendation.py
Number of records matching the author name 'Jeong-Ah' :
{'value': 1, 'relation': 'eq'}
cluserID of the matched record:
1
Number of records matching the clusterID from previous search:
{'value': 2, 'relation': 'eq'}
```



Doc 1:
Title – Augmented Reality ...
Cluster ID – 1
...

Doc 2:
Title – Augmented Design ...
Cluster ID – 2
...

Doc 3:
Title – Collaborative AR...
Cluster ID – 1
...

Search Query:
Augmented Reality

Doc 1

Based on Doc 1's cluster ID:

Doc 1

Doc 3

Incorporating TML Data

We are able to modify, update the desired field in an existing index because we pre-configured the following fields in both datasets as plain text fields.

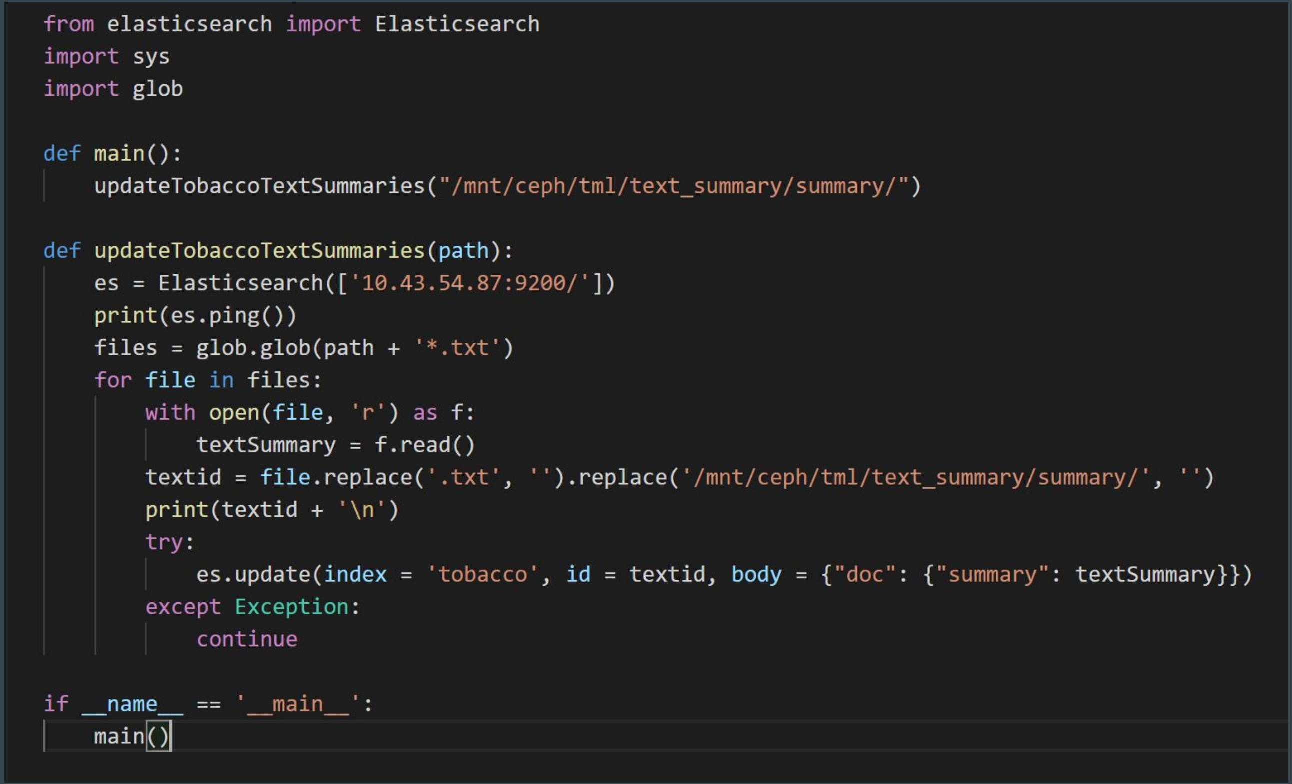
1. Text Summarization (97,484 for tobacco settlement documents)
2. Sentiment Analysis (765,530*, for tobacco settlement documents)
3. Named-Entity Recognition (213,883 for tobacco settlement documents)
4. Cluster Data (N/A, only for ETDs)

As of 03:14 AM, 12/10/2019

Incorporating TML Data - cont.

The data files can be processed as:

- Plain text file.
- Named after document ID



```python
from elasticsearch import Elasticsearch
import sys
import glob

def main():
    updateTobaccoTextSummaries("/mnt/ceph/tml/text_summary/summary/")

def updateTobaccoTextSummaries(path):
    es = Elasticsearch(['10.43.54.87:9200/'])
    print(es.ping())
    files = glob.glob(path + '*.txt')
    for file in files:
        with open(file, 'r') as f:
            textSummary = f.read()
        textid = file.replace('.txt', '').replace('/mnt/ceph/tml/text_summary/summary/', '')
        print(textid + '\n')
        try:
            es.update(index = 'tobacco', id = textid, body = {"doc": {"summary": textSummary}})
        except Exception:
            continue

if __name__ == '__main__':
    main()
```

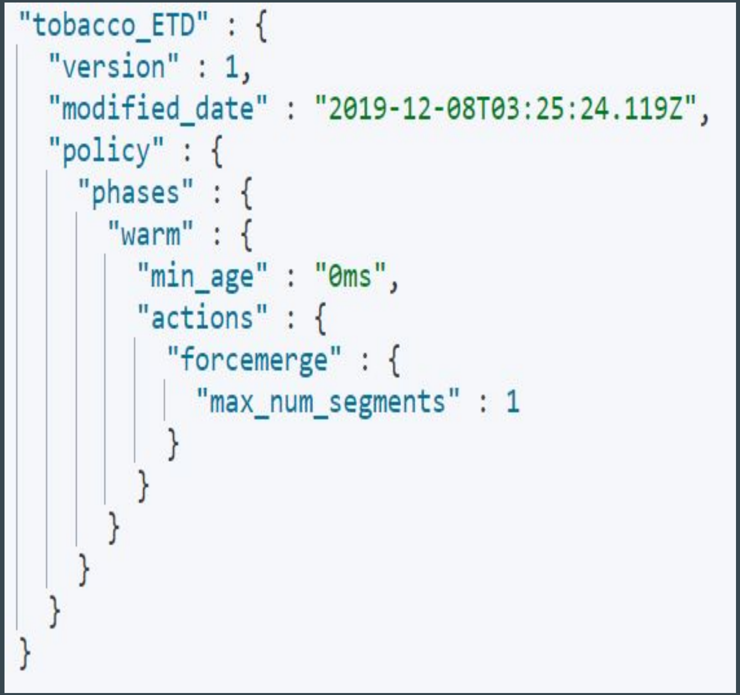
Index Lifecycle Management

- Indices should be properly managed over time.
- Different indices should be managed differently given their nature
  - Tobacco Settlement Documents: constantly queried, seldom updated
  - ETDs: constantly queried, periodically updated
  - Logs: periodically queried, extensively updated

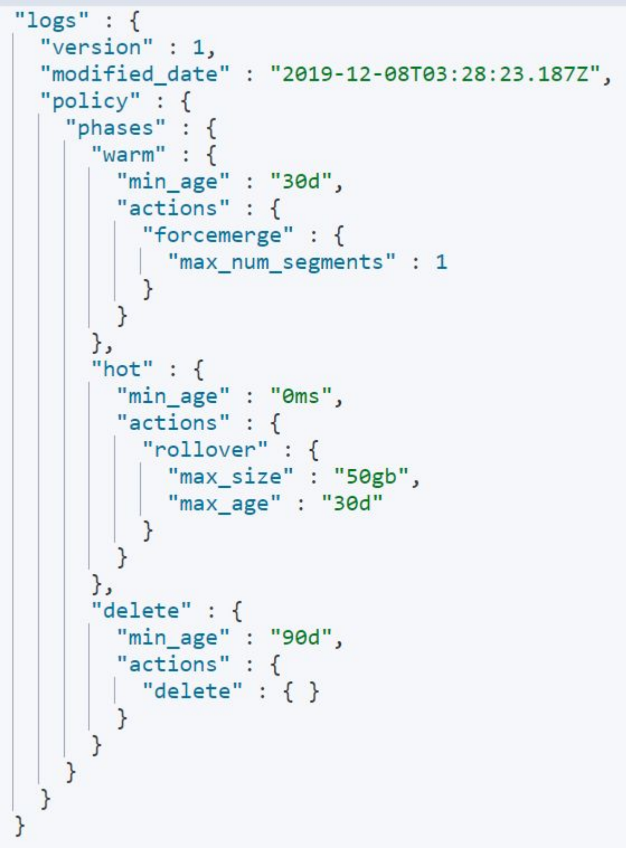
Index Lifecycle Management - cont.

- Determine appropriate policy for different dataset
  - Tobacco Settlement Documents - Stay in *warm* stage as long as possible and keep in one segment
  - ETDs - Stay in *warm* stage as long as possible and keep in one segment
  - Logs - Stay in *hot* stage, with a limited size of storage and limited life span

Index Lifecycle Management - cont.

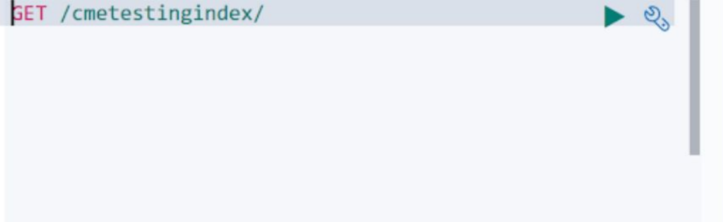


```
"tobacco_ETD" : {
  "version" : 1,
  "modified_date" : "2019-12-08T03:25:24.119Z",
  "policy" : {
    "phases" : {
      "warm" : {
        "min_age" : "0ms",
        "actions" : {
          "forcemerge" : {
            "max_num_segments" : 1
          }
        }
      }
    }
  }
}
```

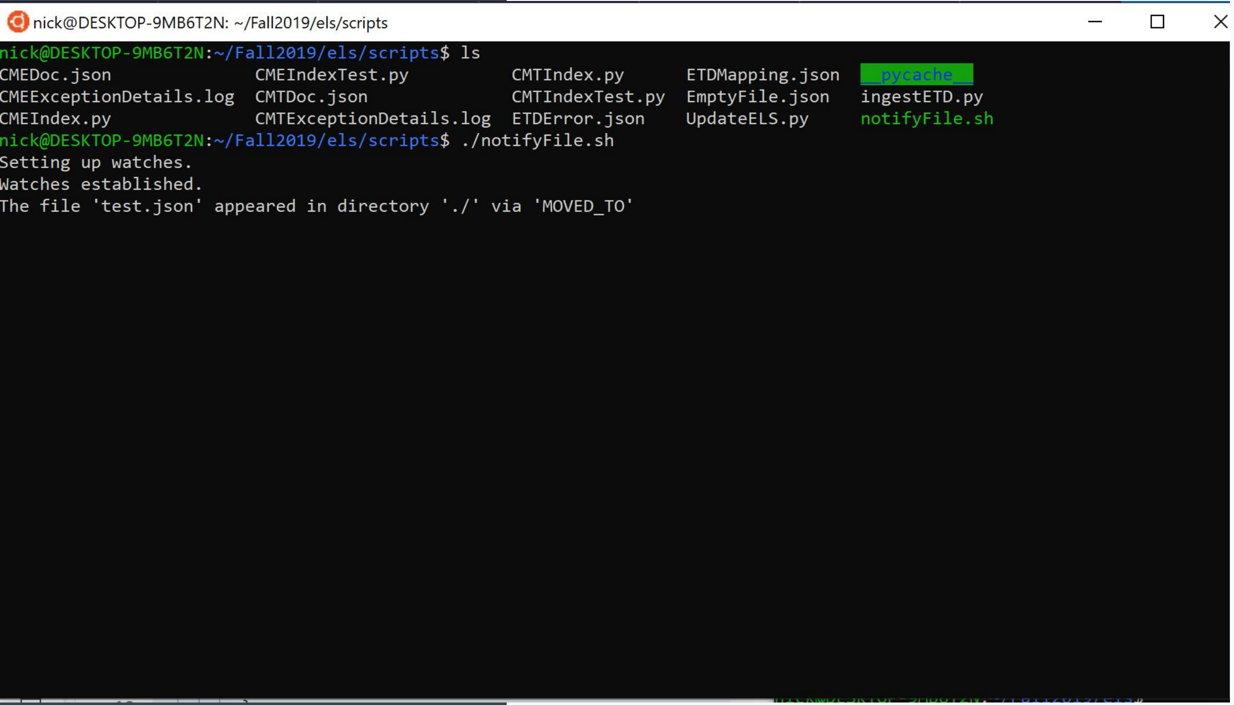


```
"logs" : {
  "version" : 1,
  "modified_date" : "2019-12-08T03:28:23.187Z",
  "policy" : {
    "phases" : {
      "warm" : {
        "min_age" : "30d",
        "actions" : {
          "forcemerge" : {
            "max_num_segments" : 1
          }
        }
      },
      "hot" : {
        "min_age" : "0ms",
        "actions" : {
          "rollover" : {
            "max_size" : "50gb",
            "max_age" : "30d"
          }
        }
      },
      "delete" : {
        "min_age" : "90d",
        "actions" : {
          "delete" : { }
        }
      }
    }
  }
}
```

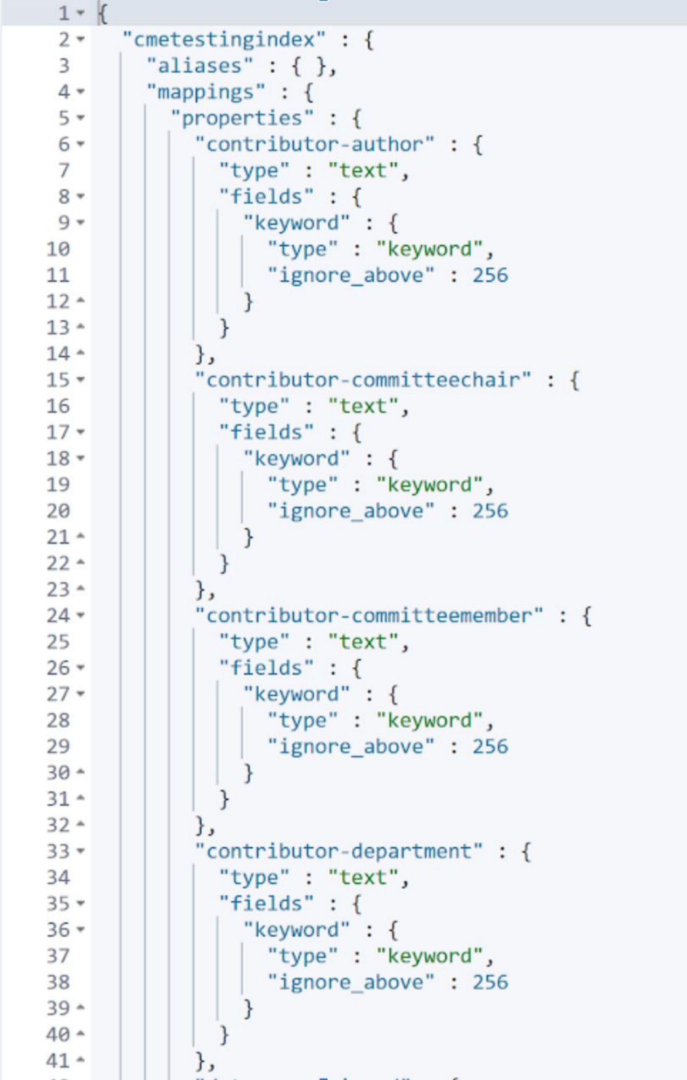
Automatic Script



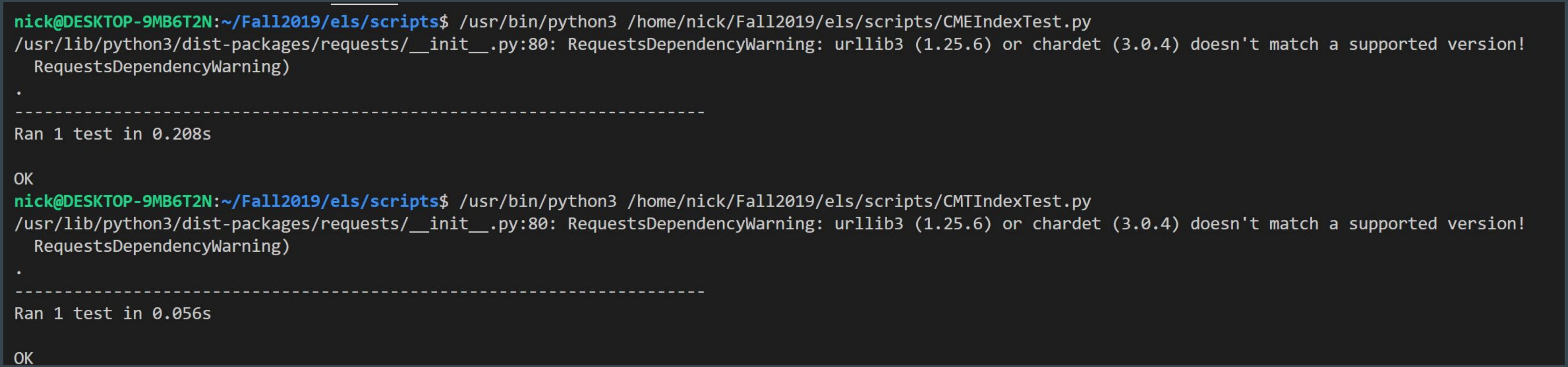
```
GET /cmetestingindex/
```



```
nick@DESKTOP-9MB6T2N:~/Fall2019/els/scripts$ ls
CMEDoc.json               CMEIndexTest.py           CMTIndex.py      ETDMapping.json  __pycache__
CMEExceptionDetails.log   CMTDoc.json               CMTIndexTest.py  EmptyFile.json   ingestETD.py
CMEIndex.py               CMTExceptionDetails.log   ETDError.json    UpdateELS.py     notifyFile.sh
nick@DESKTOP-9MB6T2N:~/Fall2019/els/scripts$ ./notifyFile.sh
Setting up watches.
Watches established.
The file 'test.json' appeared in directory './' via 'MOVED_TO'
```



```
{
  "cmetestingindex" : {
    "aliases" : { },
    "mappings" : {
      "properties" : {
        "contributor-author" : {
          "type" : "text",
          "fields" : {
            "keyword" : {
              "type" : "keyword",
              "ignore_above" : 256
            }
          }
        },
        "contributor-committeechair" : {
          "type" : "text",
          "fields" : {
            "keyword" : {
              "type" : "keyword",
              "ignore_above" : 256
            }
          }
        },
        "contributor-committeemember" : {
          "type" : "text",
          "fields" : {
            "keyword" : {
              "type" : "keyword",
              "ignore_above" : 256
            }
          }
        },
        "contributor-department" : {
          "type" : "text",
          "fields" : {
            "keyword" : {
              "type" : "keyword",
              "ignore_above" : 256
            }
          }
        },
```



```
nick@DESKTOP-9MB6T2N:~/Fall2019/els/scripts$ /usr/bin/python3 /home/nick/Fall2019/els/scripts/CMEIndexTest.py
/usr/lib/python3/dist-packages/requests/__init__.py:80: RequestsDependencyWarning: urllib3 (1.25.6) or chardet (3.0.4) doesn't match a supported version!
  RequestsDependencyWarning)
.
----------------------------------------------------------------------
Ran 1 test in 0.208s

OK
nick@DESKTOP-9MB6T2N:~/Fall2019/els/scripts$ /usr/bin/python3 /home/nick/Fall2019/els/scripts/CMTIndexTest.py
/usr/lib/python3/dist-packages/requests/__init__.py:80: RequestsDependencyWarning: urllib3 (1.25.6) or chardet (3.0.4) doesn't match a supported version!
  RequestsDependencyWarning)
.
----------------------------------------------------------------------
Ran 1 test in 0.056s

OK
```

Unit Tests

CONCLUSIONS AND FUTURE WORK

Deliverables

1. Data schema for ETD and tobacco datasets has been provided to the FEK, TML, CMT and CME teams
2. 30k - index for ETD dataset
3. Tobacco - index for tobacco settlements dataset
4. Facet names, field types, usage recommendation, and field examples provided to the FEK team for filtering, searching and visualization
5. Search query format with example
   a. Ordinary search (FEK)
   b. Nested search with page hit (FEK)
   c. Boosting
   d. Recommendation script (FEK)
6. Automated scripts
   a. Shell script for monitoring new files
   b. Python script for Ingestion and updating
7. Search log (Slow log) on Kibana
8. Unit test scripts
9. Ingesting and indexing data received from the TML team (ClusterID, summary, sentiment, NER)

Future Work

- Continue to ingest the rest of the documents into Elasticsearch
  - Increase space in Elasticsearch

- Improve the recommendations by working with TML team
  - Text Summaries
  - Sentiment Analysis
  - NER
  - Clustering

- Add support for user logs and recommendations
  - User-Specific Logs with FEK team
  - Index Logs / Store in CEPH

ACKNOWLEDGEMENTS

Thank you!