CE706 - Information Retrieval 2021

Assigment 1

2004532

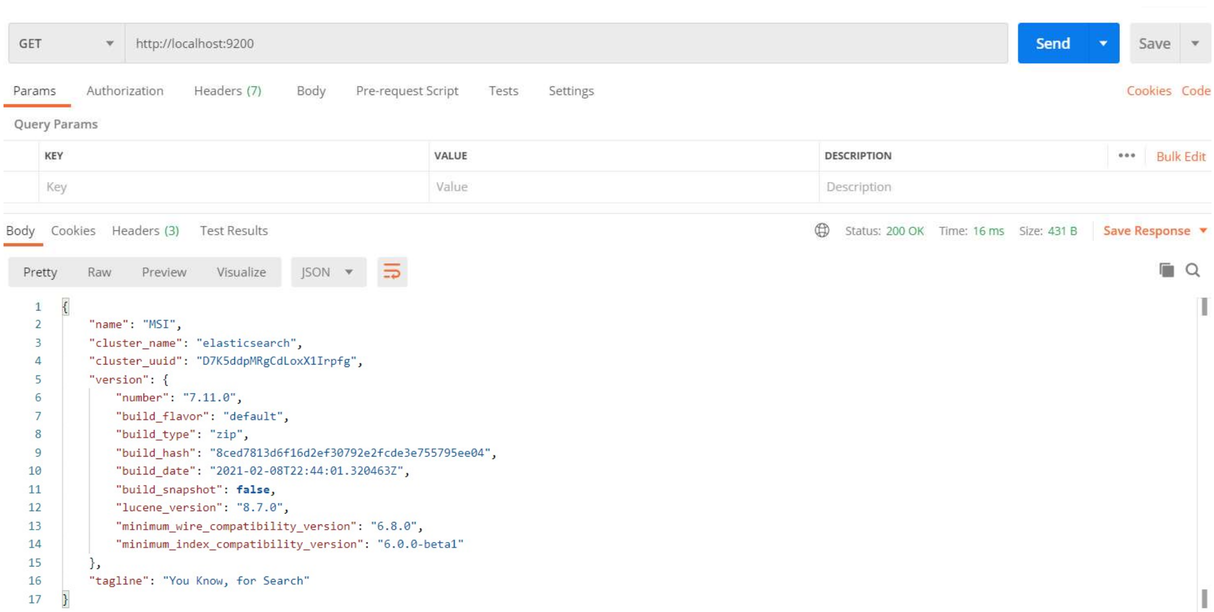
Instructions for running your system (Engineering a Complete System)

Elastic Search:

- Download Elastic Search for Windows

https://artifacts.elastic.co/downloads/elasticsearch/elasticsearch-7.11.1-windows-x86_64.zip

- Unzip the downloaded file
- Navigate to bin folder and run elasticsearch.bat
- Check http://localhost:9200/ to see elastic search running.

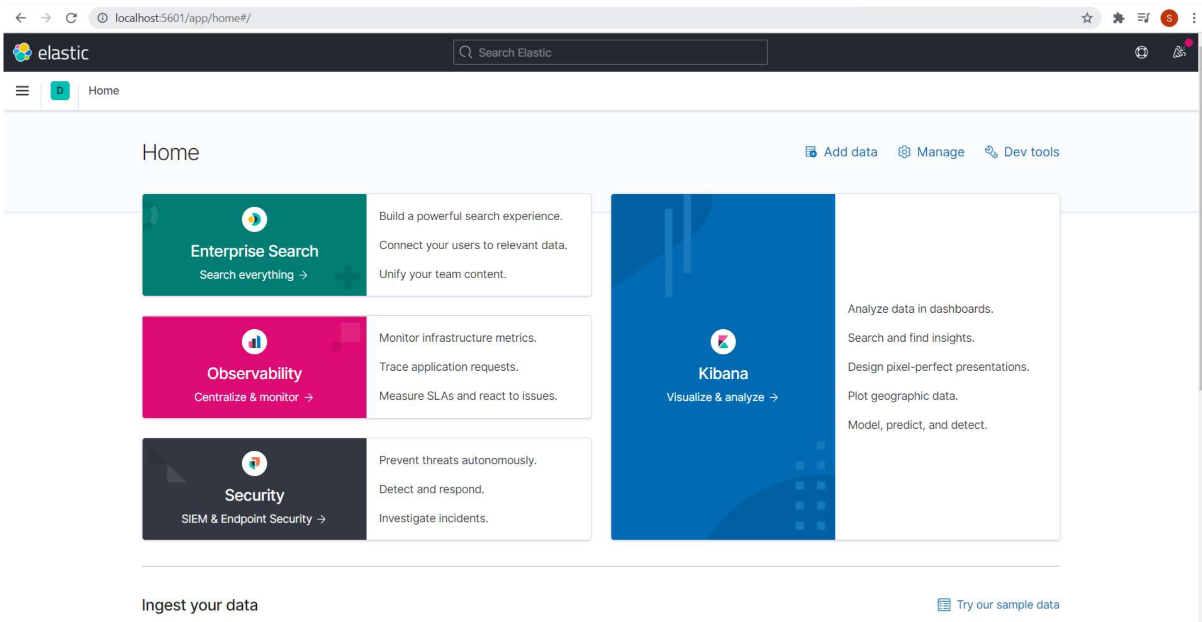


Kibana:

- Download Kibana for Windows3

https://artifacts.elastic.co/downloads/kibana/kibana-7.11.1-windows-x86_64.zip

- Unzip the downloaded file
- Navigate to bin folder and run kibana.bat
- Check http://localhost:5601/ to see kibana running.



Python Jupyter and Elastic Search:

- Make sure you have python and jupyter correctly set up.
- Install the following python libraries by running these conmands:

pip install -U scikit-learn

pip install -U pip setuptools wheel

pip install -U spacy

python -m spacy download en_core_web_sm

pip install elasticsearch

Indexing

An index can be thought of as an optimized collection of documents and each document is a collection of fields, which are the key-value pairs that contain your data. In this assignment I have set the index to value “covid” because we are dealing with covid articles. Indexing allows faster search of documents. An index consists of documents, and a document consists of fields.

I downloaded the dataset for covid articles i.e metadata.csv from Kaggle site. The dataset is in csv form. So before uploading the data to elastic search, I added two new fields to the dataset i.e. index and id. Specifying index allows us to dump data to elasticsearch to specific index. It is not necessary to add id field to the data because otherwise elasticsearch automatically generates an id for each document. However in the dataset there is a field called cord_uid so I set this as an id for each document. Moreover I found out that there are columns (mag_id, arxiv_id) that contain almost no data so I dropped these two columns from the dataset.

```
df.isnull().sum()
```

```
cord_uid                0
sha                299973
source_x                0
title                 223
doi                199517
pmcid              290607
pubmed_id          231259
license                 0
abstract           124997
publish_time          218
authors             13110
journal             30185
mag_id             450385
who_covidence_id   268359
arxiv_id           444490
pdf_json_files     299973
pmc_json_files     329971
url                180409
s2_id               41014
dtype: int64
```

Some of the values in title, abstract and authors column are missing so I excluded those documents.

```
df[pd.isna(df.title) | pd.isna(df.abstract) | pd.isna(df.authors)].shape
```

```
(128062, 17)
```

```
df=df[pd.notna(df.title) & pd.notna(df.abstract) & pd.notna(df.authors)]
```

Some of the values in publish_time field were also missing so we removed those records from the dataset as well.

```
df.publish_time=pd.to_datetime(df.publish_time)
```

```
df[pd.isna(df.publish_time)].shape
```

```
(188, 17)
```

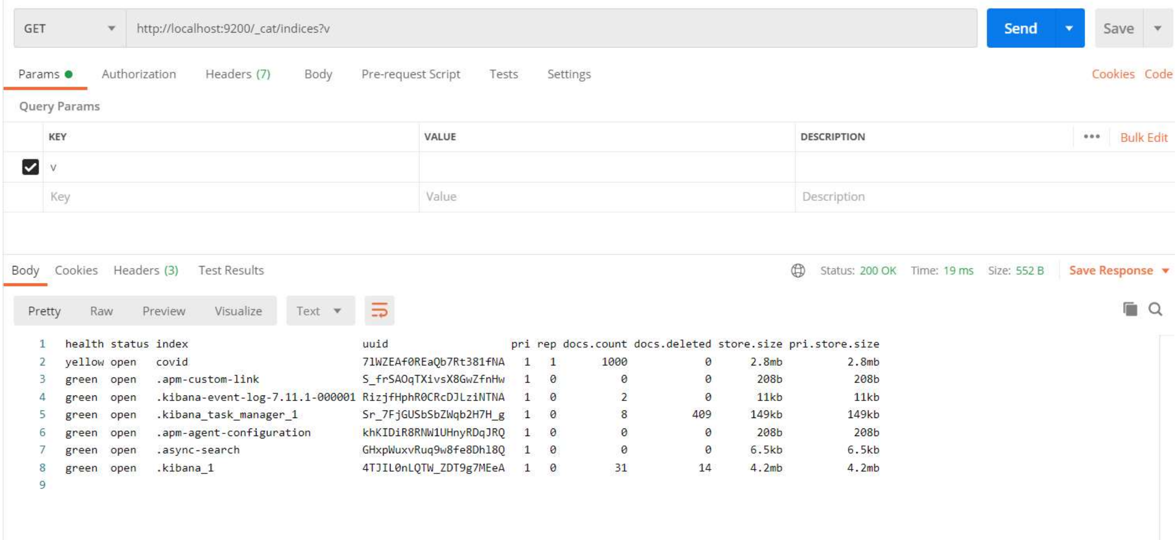
```
df=df[pd.notna(df.publish_time)]
```

After that I selected 1000 documents from the dataset and uploaded it to elasticsearch. Moreover I specified mapping for two fields i.e. authors and keywords in data as keyword type in elastic search. I created keywords field in data that contains keywords from textual data in the dataset. Specifying these fields as keyword type allows faster search when we query for these respective fields. I also specified publish_time field as date type in elastic search.

```
from elasticsearch import helpers
from elasticsearch import Elasticsearch

es=Elasticsearch()
es.indices.create(index="covid",ignore=400, body={
    "mappings": {
        "properties": {
          "authors":{
            "type": "keyword",
          },
          "keywords":{
            "type": "keyword",
          },
          "publish_time":{
              "type" : "date",
              "format" : "yyyy-MM-dd"
          }
        }
    }
})
```

After uploading and indexing the data we could verify it on postman:



Sentence Splitting, Tokenization and Normalization

In the dataset, there are three columns that contain textual data i.e. title, abstract and authors. For authors field, the name of authors are separated by ";", so I split the text in authors field with ";" delimiter and stored it as an array. In mapping I specified the author field as keyword type so that it would be easier to search for specific author publications. For title and abstract, I merged the content from these two fields to a single field. In the initial preprocessing I removed extra spaces, new line/tab characters, punctuations and numbers/digits from the text using python and regex expressions. After that I used spacy library in python to tokenize the text and for removing stop words.

Before:

```
Debate: Transfusing to normal haemoglobin levels will not improve outcome Recent evidence suggests that critically ill patien
ts are able to tolerate lower levels of haemoglobin than was previously believed. It is our goal to show that transfusing to
a level of 100 g/l does not improve mortality and other clinically important outcomes in a critical care setting. Although ma
ny questions remain, many laboratory and clinical studies, including a recent randomized controlled trial (RCT), have establi
shed that transfusing to normal haemoglobin concentrations does not improve organ failure and mortality in the critically ill
patient. In addition, a restrictive transfusion strategy will reduce exposure to allogeneic transfusions, result in more effi
cient use of red blood cells (RBCs), save blood overall, and decrease health care costs.
```

After initial preprocessing:

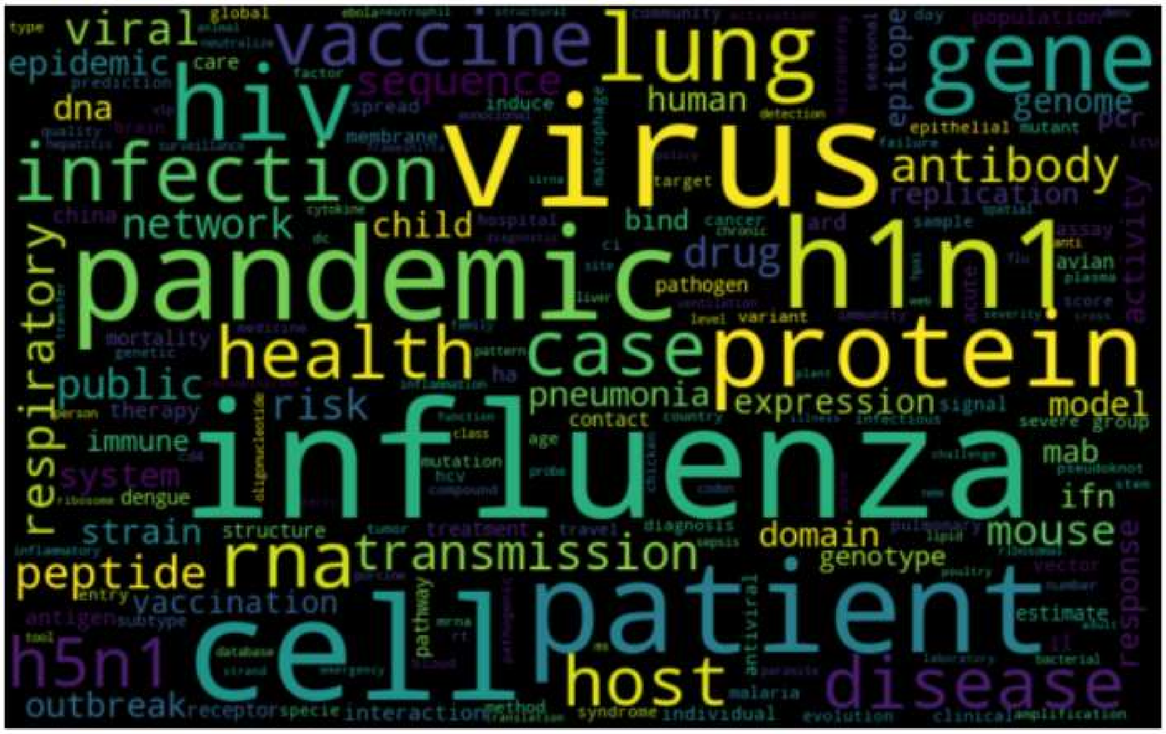
```
Debate Transfusing to normal haemoglobin levels will not improve outcome Recent evidence suggests that critically ill patient
s are able to tolerate lower levels of haemoglobin than was previously believed It is our goal to show that transfusing to a
level of g l does not improve mortality and other clinically important outcomes in a critical care setting Although many ques
tions remain many laboratory and clinical studies including a recent randomized controlled trial RCT have established that tr
ansfusing to normal haemoglobin concentrations does not improve organ failure and mortality in the critically ill patient In
addition a restrictive transfusion strategy will reduce exposure to allogeneic transfusions result in more efficient use of r
ed blood cells RBCs save blood overall and decrease health care costs
```

After spaCy preprocessing:

```
['debate', 'transfuse', 'normal', 'haemoglobin', 'level', 'improve', 'outcome', 'recent', 'evidence', 'suggest', 'criticall
y', 'ill', 'patient', 'able', 'tolerate', 'low', 'level', 'haemoglobin', 'previously', 'believe', 'goal', 'transfusing', 'lev
el', 'g', 'l', 'improve', 'mortality', 'clinically', 'important', 'outcome', 'critical', 'care', 'setting', 'question', 'rema
in', 'laboratory', 'clinical', 'study', 'include', 'recent', 'randomize', 'controlled', 'trial', 'rct', 'establish', 'transfu
se', 'normal', 'haemoglobin', 'concentration', 'improve', 'organ', 'failure', 'mortality', 'critically', 'ill', 'patient', 'a
ddition', 'restrictive', 'transfusion', 'strategy', 'reduce', 'exposure', 'allogeneic', 'transfusion', 'result', 'efficient',
'use', 'red', 'blood', 'cell', 'rbc', 'save', 'blood', 'overall', 'decrease', 'health', 'care', 'cost']
```

Selecting Keywords

After preprocessing the text data, I removed all the unnecessary text from data and only left with words that can be keyword. For selecting keyword I used tf-idf (term frequency–inverse document frequency). TF-IDF allows us to compute a score against each word in a document that shows how important a word is in a document in a collection of documents. Lower tf-idf score shows that the word is more common in different documents and high tf-idf score shows that a word can be a keyword of document. I only selected top 10 keywords for each document and stored it with respective document in the dataset. The wordcloud for our keywords in all documents is shown below:



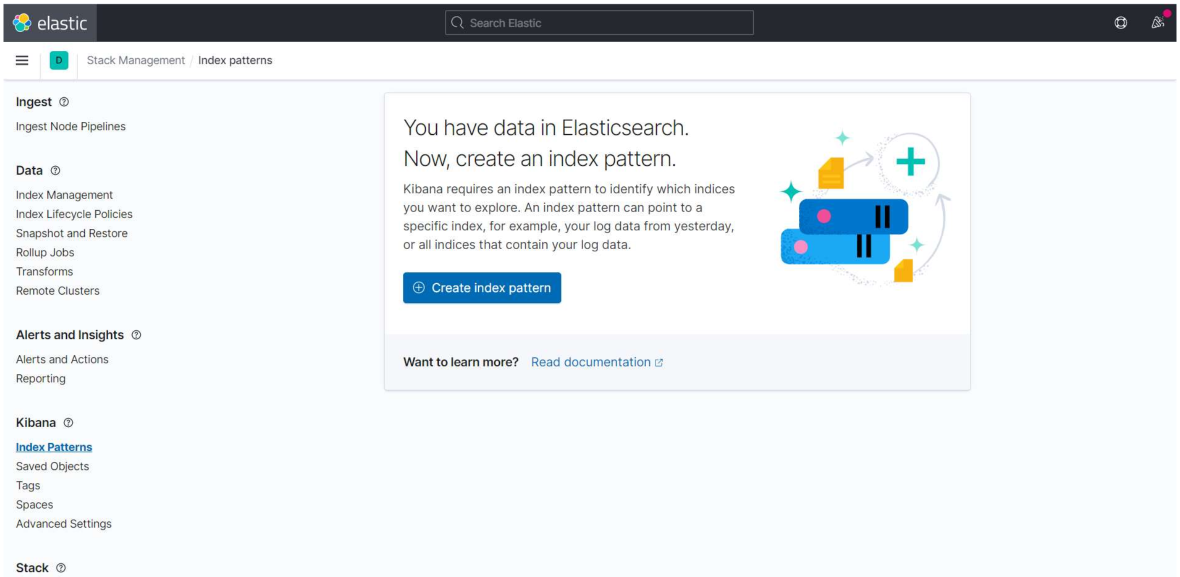
Stemming or Morphological Analysis

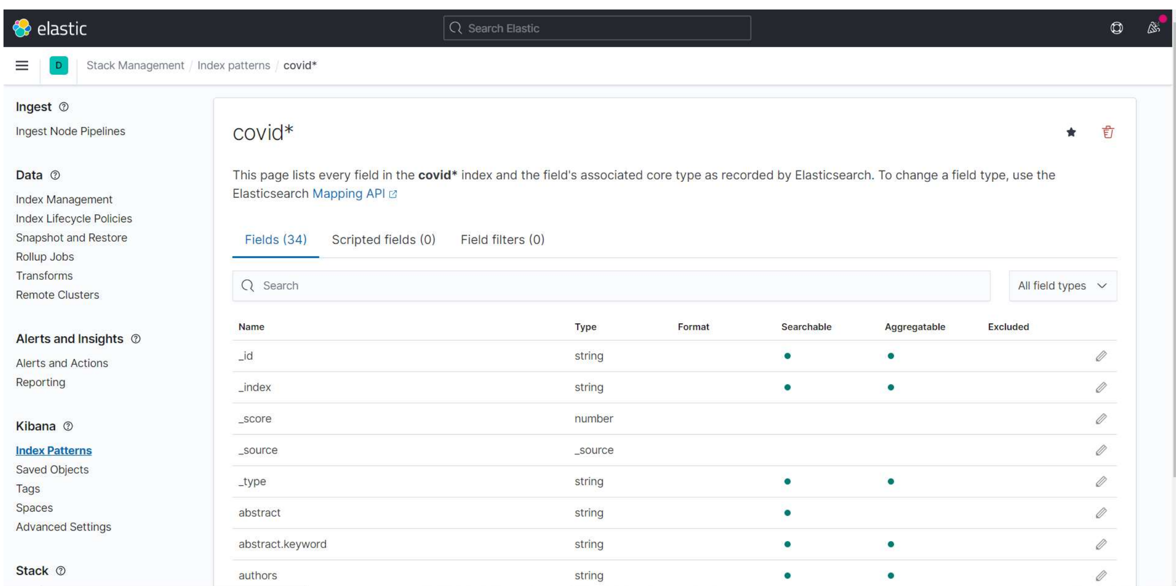
Stemming and Lemmatization both generate the root form of the words. Moreover stemming might produce words that are not in English vocabulary but with lemmatization we don't need to worry about that. spaCy doesn't contain any function for stemming so for morphological analysis I used lemmatization that is supported by spaCy and works very well. Here are few examples of lemmatization in our dataset:

```
Before Lemmatization
['Managing', 'emerging', 'infectious', 'diseases', 'federal', 'system', 'impediment', 'effective', 'laws', 's', 's', 'HIV', 'AI
DS', 'emerging', 'infectious', 'disease', '2004', 'saw', 'emergence', 'SARS', 'Avian', 'influenza', 'Anthrax', 'man', 'form',
'bioterrorism', 'Emergency', 'powers', 'legislation', 'Australia', 'patchwork', 'Commonwealth', 'quarantine', 'laws', 'State',
'Territory', 'based', 'emergency', 'powers', 'public', 'health', 'legislation', 'time', 'review', 'legislation', 'time', 'consi
deration', 'efficacy', 'legislation', 'country', 'wide', 'perspective', 'age', 'consider', 'possibility', 'mass', 'outbreaks',
'communicable', 'diseases', 'ignore', 'jurisdictional', 'boundaries']
After Lemmatization
['manage', 'emerge', 'infectious', 'disease', 'federal', 'system', 'impediment', 'effective', 'law', 's', 's', 'hiv', 'aids',
'emerge', 'infectious', 'disease', '2004', 'see', 'emergence', 'sars', 'avian', 'influenza', 'anthrax', 'man', 'form', 'bioterr
orism', 'emergency', 'power', 'legislation', 'australia', 'patchwork', 'commonwealth', 'quarantine', 'law', 'state', 'territor
y', 'base', 'emergency', 'power', 'public', 'health', 'legislation', 'time', 'review', 'legislation', 'time', 'consideration',
'efficacy', 'legislation', 'country', 'wide', 'perspective', 'age', 'consider', 'possibility', 'mass', 'outbreak', 'communicabl
e', 'disease', 'ignore', 'jurisdictional', 'boundary']
```

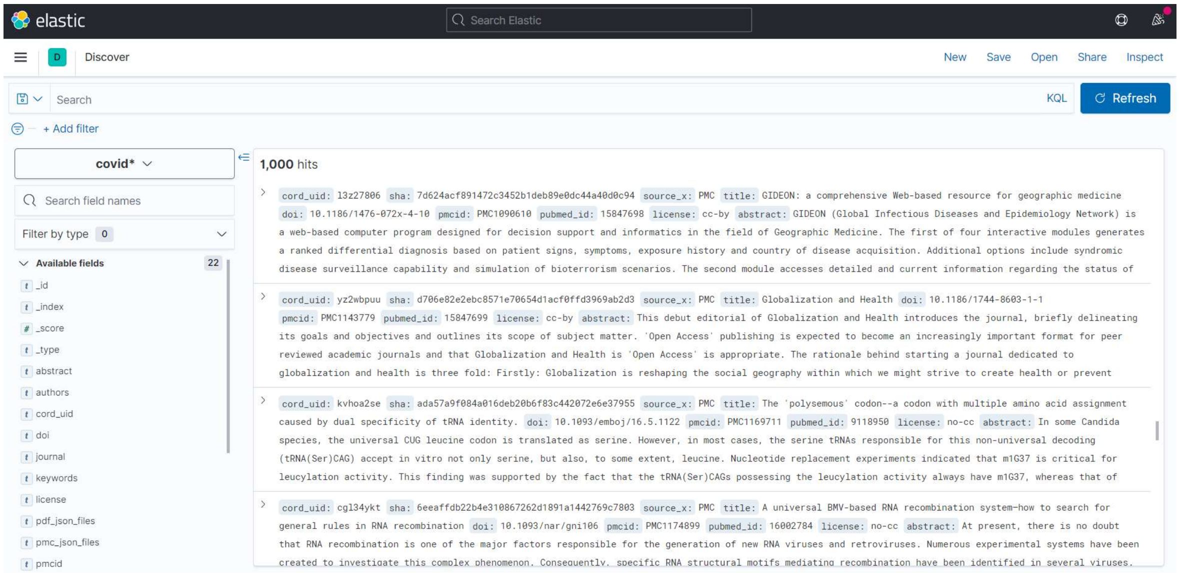
Searching

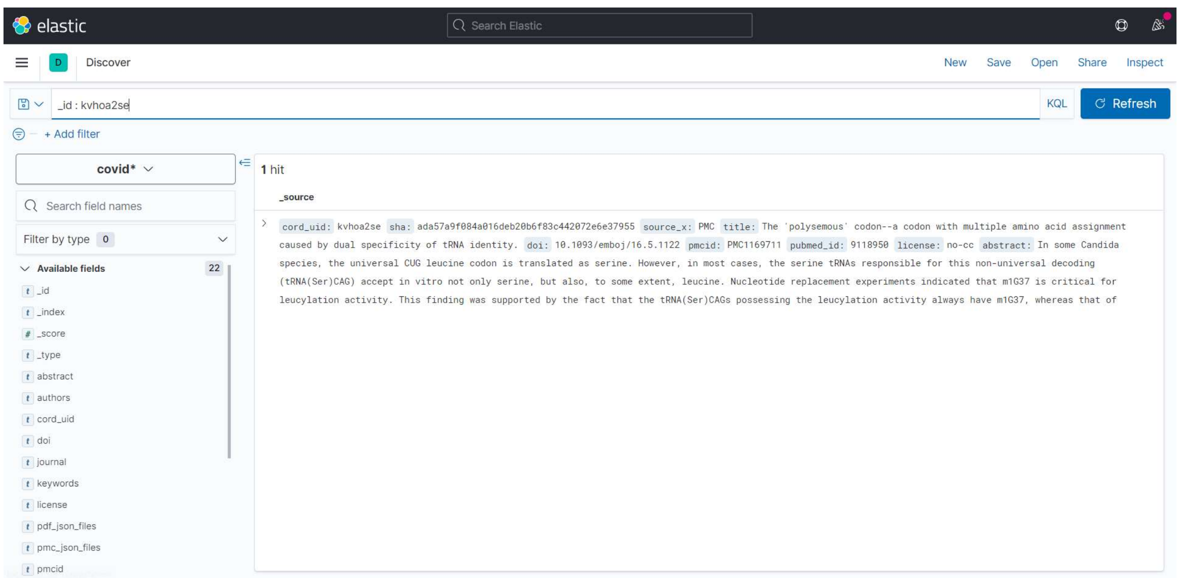
I used Kibana for searching the indexed documents that we uploaded to elastic search. Kibana gives a gui for better visualization and selection of fields in our data.

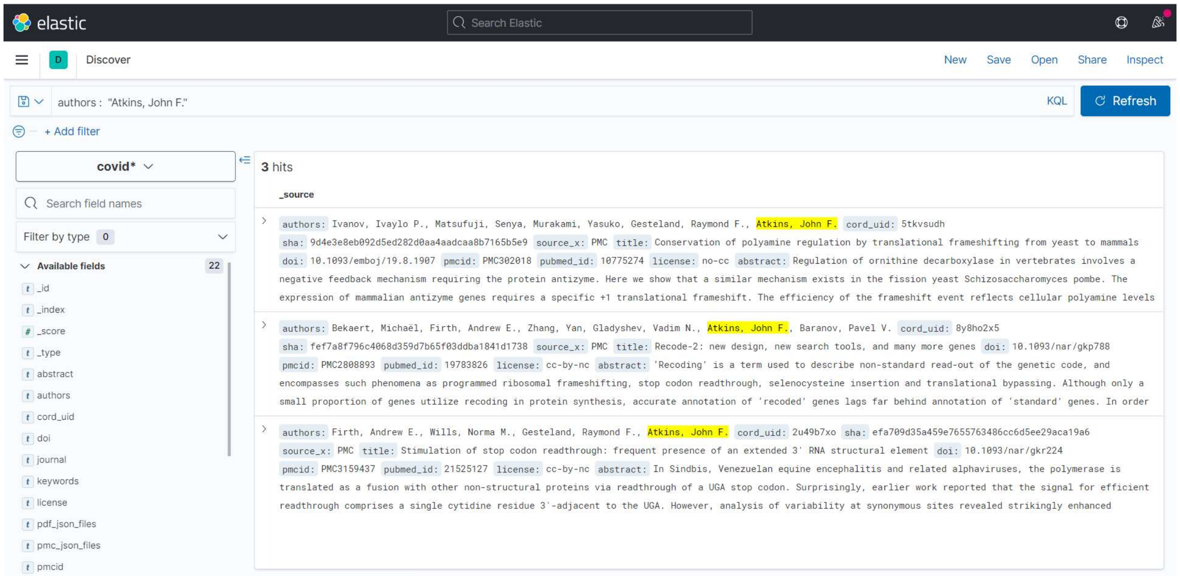


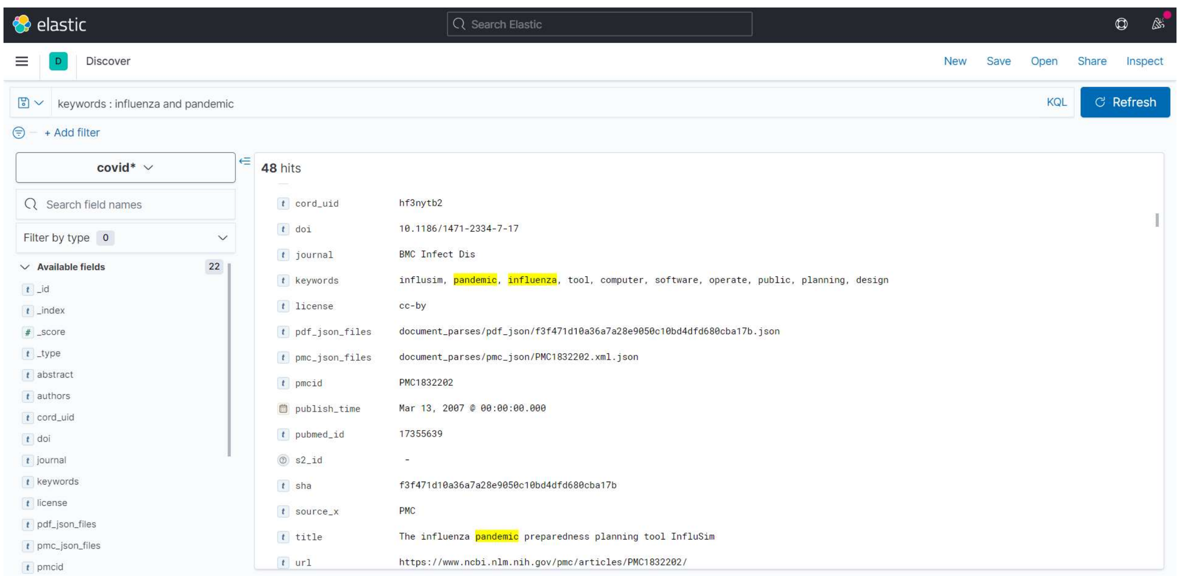


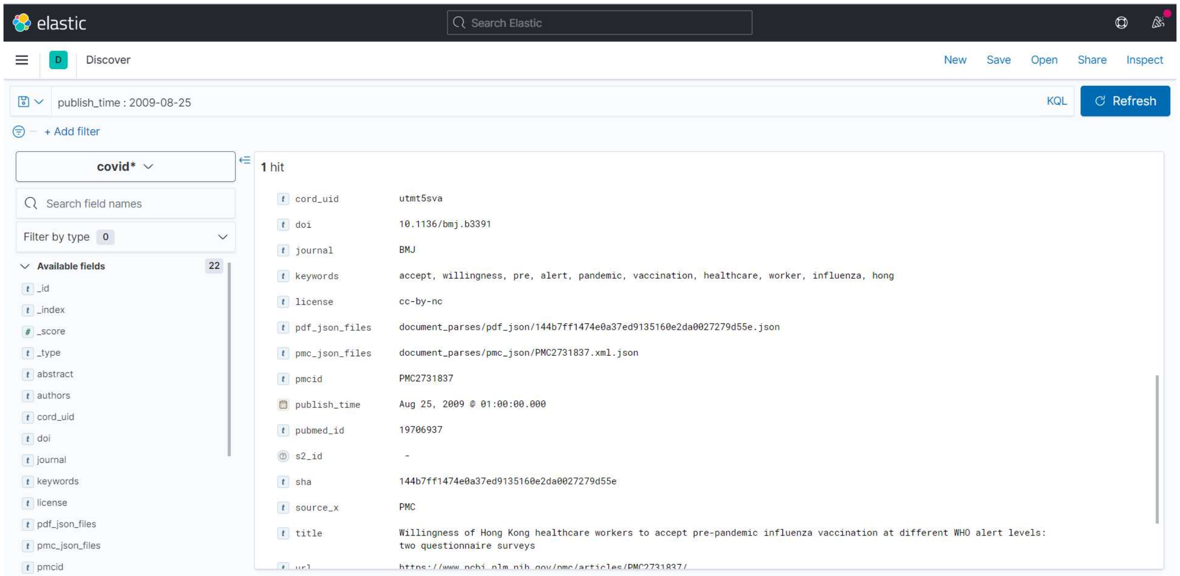
The search results for different fields are given below:


elastic
Search Elastic
Discover
New
Save
Open
Share
Inspect
Search
KQL
Refresh
+ Add filter
covid*
Search field names
Filter by type 0
Available fields 22
_id
_index
_score
_type
abstract
authors
cord_uid
doi
journal
keywords
license
pdf_json_files
pmc_json_files
pmcid
1,000 hits
cord_uid: 13z27806 sha: 7d624acf891472c3452b1deb89e0dc44a40d0c94 source_x: PMC title: GIDEON: a comprehensive Web-based resource for geographic medicine doi: 10.1186/1476-072x-4-10 pmcid: PMC1090610 pubmed_id: 15847698 license: cc-by abstract: GIDEON (Global Infectious Diseases and Epidemiology Network) is a web-based computer program designed for decision support and informatics in the field of Geographic Medicine. The first of four interactive modules generates a ranked differential diagnosis based on patient signs, symptoms, exposure history and country of disease acquisition. Additional options include syndromic disease surveillance capability and simulation of bioterrorism scenarios. The second module accesses detailed and current information regarding the status of
cord_uid: yz2wbpuu sha: d706e82e2ebc8571e70654d1acf0ffd3969ab2d3 source_x: PMC title: Globalization and Health doi: 10.1186/1744-8603-1-1 pmcid: PMC1143779 pubmed_id: 15847699 license: cc-by abstract: This debut editorial of Globalization and Health introduces the journal, briefly delineating its goals and objectives and outlines its scope of subject matter. 'Open Access' publishing is expected to become an increasingly important format for peer reviewed academic journals and that Globalization and Health is 'Open Access' is appropriate. The rationale behind starting a journal dedicated to globalization and health is three fold: Firstly: Globalization is reshaping the social geography within which we might strive to create health or prevent
cord_uid: kvhoa2se sha: ada57a9f084a016deb20b6f83c442072e6e37955 source_x: PMC title: The 'polysemous' codon--a codon with multiple amino acid assignment caused by dual specificity of tRNA identity. doi: 10.1093/emboj/16.5.1122 pmcid: PMC1169711 pubmed_id: 9118950 license: no-cc abstract: In some Candida species, the universal CUG leucine codon is translated as serine. However, in most cases, the serine tRNAs responsible for this non-universal decoding (tRNA(Ser)CAG) accept in vitro not only serine, but also, to some extent, leucine. Nucleotide replacement experiments indicated that m1G37 is critical for leucylation activity. This finding was supported by the fact that the tRNA(Ser)CAGs possessing the leucylation activity always have m1G37, whereas that of
cord_uid: cgl34ykt sha: 6eeaffdb22b4e310867262d1891a1442769c7803 source_x: PMC title: A universal BMV-based RNA recombination system—how to search for general rules in RNA recombination doi: 10.1093/nar/gni106 pmcid: PMC1174899 pubmed_id: 16002784 license: no-cc abstract: At present, there is no doubt that RNA recombination is one of the major factors responsible for the generation of new RNA viruses and retroviruses. Numerous experimental systems have been created to investigate this complex phenomenon. Consequently, specific RNA structural motifs mediating recombination have been identified in several viruses.


elastic
Search Elastic
Discover
New
Save
Open
Share
Inspect
_id : kvhoa2se
KQL
Refresh
+ Add filter
covid*
Search field names
Filter by type 0
Available fields 22
_id
_index
_score
_type
abstract
authors
cord_uid
doi
journal
keywords
license
pdf_json_files
pmc_json_files
pmcid
1 hit
_source
cord_uid: kvhoa2se sha: ada57a9f084a016deb20b6f83c442072e6e37955 source_x: PMC title: The 'polysemous' codon--a codon with multiple amino acid assignment caused by dual specificity of tRNA identity. doi: 10.1093/emboj/16.5.1122 pmcid: PMC1169711 pubmed_id: 9118950 license: no-cc abstract: In some Candida species, the universal CUG leucine codon is translated as serine. However, in most cases, the serine tRNAs responsible for this non-universal decoding (tRNA(Ser)CAG) accept in vitro not only serine, but also, to some extent, leucine. Nucleotide replacement experiments indicated that m1G37 is critical for leucylation activity. This finding was supported by the fact that the tRNA(Ser)CAGs possessing the leucylation activity always have m1G37, whereas that of


elastic
Search Elastic
Discover
New
Save
Open
Share
Inspect
authors : "Atkins, John F."
KQL
Refresh
+ Add filter
covid*
Search field names
Filter by type 0
Available fields 22
_id
_index
_score
_type
abstract
authors
cord_uid
doi
journal
keywords
license
pdf_json_files
pmc_json_files
pmcid
3 hits
_source
authors: Ivanov, Ivaylo P., Matsufuji, Senya, Murakami, Yasuko, Gesteland, Raymond F., Atkins, John F. cord_uid: 5tkvsudh
sha: 9d4e3e8eb092d5ed282d0aa4aadcaa8b7165b5e9 source_x: PMC title: Conservation of polyamine regulation by translational frameshifting from yeast to mammals
doi: 10.1093/emboj/19.8.1907 pmcid: PMC302018 pubmed_id: 10775274 license: no-cc abstract: Regulation of ornithine decarboxylase in vertebrates involves a
negative feedback mechanism requiring the protein antizyme. Here we show that a similar mechanism exists in the fission yeast Schizosaccharomyces pombe. The
expression of mammalian antizyme genes requires a specific +1 translational frameshift. The efficiency of the frameshift event reflects cellular polyamine levels
authors: Bekaert, Michaël, Firth, Andrew E., Zhang, Yan, Gladyshev, Vadim N., Atkins, John F., Baranov, Pavel V. cord_uid: 8y8ho2x5
sha: fef7a8f796c4068d359d7b65f03ddba1841d1738 source_x: PMC title: Recode-2: new design, new search tools, and many more genes doi: 10.1093/nar/gkp788
pmcid: PMC2808893 pubmed_id: 19783826 license: cc-by-nc abstract: 'Recoding' is a term used to describe non-standard read-out of the genetic code, and
encompasses such phenomena as programmed ribosomal frameshifting, stop codon readthrough, selenocysteine insertion and translational bypassing. Although only a
small proportion of genes utilize recoding in protein synthesis, accurate annotation of 'recoded' genes lags far behind annotation of 'standard' genes. In order
authors: Firth, Andrew E., Wills, Norma M., Gesteland, Raymond F., Atkins, John F. cord_uid: 2u49b7xo sha: efa709d35a459e7655763486cc6d5ee29aca19a6
source_x: PMC title: Stimulation of stop codon readthrough: frequent presence of an extended 3' RNA structural element doi: 10.1093/nar/gkr224
pmcid: PMC3159437 pubmed_id: 21525127 license: cc-by-nc abstract: In Sindbis, Venezuelan equine encephalitis and related alphaviruses, the polymerase is
translated as a fusion with other non-structural proteins via readthrough of a UGA stop codon. Surprisingly, earlier work reported that the signal for efficient
readthrough comprises a single cytidine residue 3'-adjacent to the UGA. However, analysis of variability at synonymous sites revealed strikingly enhanced


elastic
Search Elastic
Discover
New
Save
Open
Share
Inspect
keywords : influenza and pandemic
KQL
Refresh
+ Add filter
covid*
Search field names
Filter by type 0
Available fields 22
_id
_index
_score
_type
abstract
authors
cord_uid
doi
journal
keywords
license
pdf_json_files
pmc_json_files
pmcid
48 hits
cord_uid hf3nytb2
doi 10.1186/1471-2334-7-17
journal BMC Infect Dis
keywords influsim, pandemic, influenza, tool, computer, software, operate, public, planning, design
license cc-by
pdf_json_files document_parses/pdf_json/f3f471d10a36a7a28e9050c10bd4dfd680cba17b.json
pmc_json_files document_parses/pmc_json/PMC1832202.xml.json
pmcid PMC1832202
publish_time Mar 13, 2007 @ 00:00:00.000
pubmed_id 17355639
s2_id -
sha f3f471d10a36a7a28e9050c10bd4dfd680cba17b
source_x PMC
title The influenza pandemic preparedness planning tool InfluSim
url https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1832202/


elastic
Search Elastic
Discover
New
Save
Open
Share
Inspect
publish_time : 2009-08-25
KQL
Refresh
+ Add filter
covid*
Search field names
Filter by type 0
Available fields 22
_id
_index
_score
_type
abstract
authors
cord_uid
doi
journal
keywords
license
pdf_json_files
pmc_json_files
pmcid
1 hit
cord_uid utmt5sva
doi 10.1136/bmj.b3391
journal BMJ
keywords accept, willingness, pre, alert, pandemic, vaccination, healthcare, worker, influenza, hong
license cc-by-nc
pdf_json_files document_parses/pdf_json/144b7ff1474e0a37ed9135160e2da0027279d55e.json
pmc_json_files document_parses/pmc_json/PMC2731837.xml.json
pmcid PMC2731837
publish_time Aug 25, 2009 @ 01:00:00.000
pubmed_id 19706937
s2_id -
sha 144b7ff1474e0a37ed9135160e2da0027279d55e
source_x PMC
title Willingness of Hong Kong healthcare workers to accept pre-pandemic influenza vaccination at different WHO alert levels: two questionnaire surveys