Building a File Observatory for Secure Parser Development

LangSec 2021

May 27, 2021

Jet Propulsion Laboratory
California Institute of Technology

The Team

Chris Mattmann
PI; Division Mgr,
AI AID Org

Tim Allison
Files and Search

Wayne Burke
Cognizant Engineer

Michael Fedell
Data Scientist

Anastasia Menshikova
Data Scientist

Phil Southam
Trouble (Fun?)
Maker

Ryan Stonebraker
Data Scientist
Alaskan

Debts of Gratitude

Sergey Bratus

Peter Wyatt and Duff Johnson, PDF Association

Dan Becker, John Kansky and team at Kudu Dynamics

Trail of Bits, Galois, BAE and SRI

Outline

1. Motivation and Corpora
2. Fuzzy matching/advanced search
3. API/UI
4. Classification of Creator Tool
5. Next Steps

Motivation and Corpora

Motivation

- Inducing grammars
- Devtesting parsers during development
- Testing/profiling/tracing existing parsers
  - Literal files
  - Seeds for fuzzing

Corpora

- Common Crawl
- Bug tracker corpora

Common Crawl

- Monthly open source crawls of large portions of the web: for May 2021, 2.6 billion pages (280 TB).
- Available via Amazon Web Services Public Datasets
- Searchable indexes available

https://commoncrawl.org

Common Crawl -- known limitations

- Files truncated at 1MB
- Coverage: ~convenience sample

Bugtrackers -- November 2020 Crawl

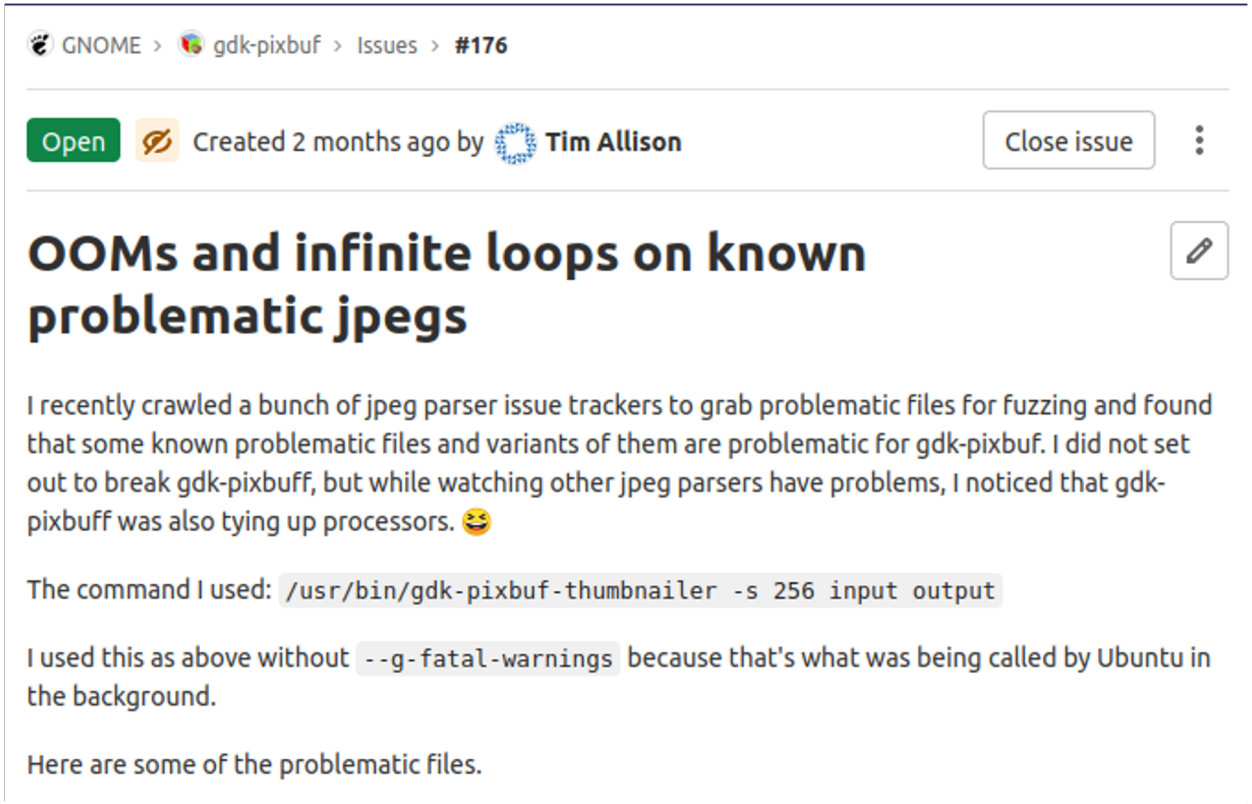
- 35 issue trackers
- 32 tools (3 tools have 2 issue trackers -- legacy and current)
- 1.2 million files (311 GB)
- PDF-centric subset ~33k files

https://corpora.tika.apache.org/base/docs/bug_trackers/

https://corpora.tika.apache.org/base/packaged/pdfs/pdfs_202011/

Bugtrackers -- March 2021 Crawl (jpegs)

- 13 issue trackers
- 7k files (3.3 GB)



GNOME › gdk-pixbuf › Issues › #176

Open Created 2 months ago by Tim Allison Close issue

OOMs and infinite loops on known problematic jpegs

I recently crawled a bunch of jpeg parser issue trackers to grab problematic files for fuzzing and found that some known problematic files and variants of them are problematic for gdk-pixbuf. I did not set out to break gdk-pixbuff, but while watching other jpeg parsers have problems, I noticed that gdk-pixbuff was also tying up processors. 😆

The command I used: `/usr/bin/gdk-pixbuf-thumbnailer -s 256 input output`

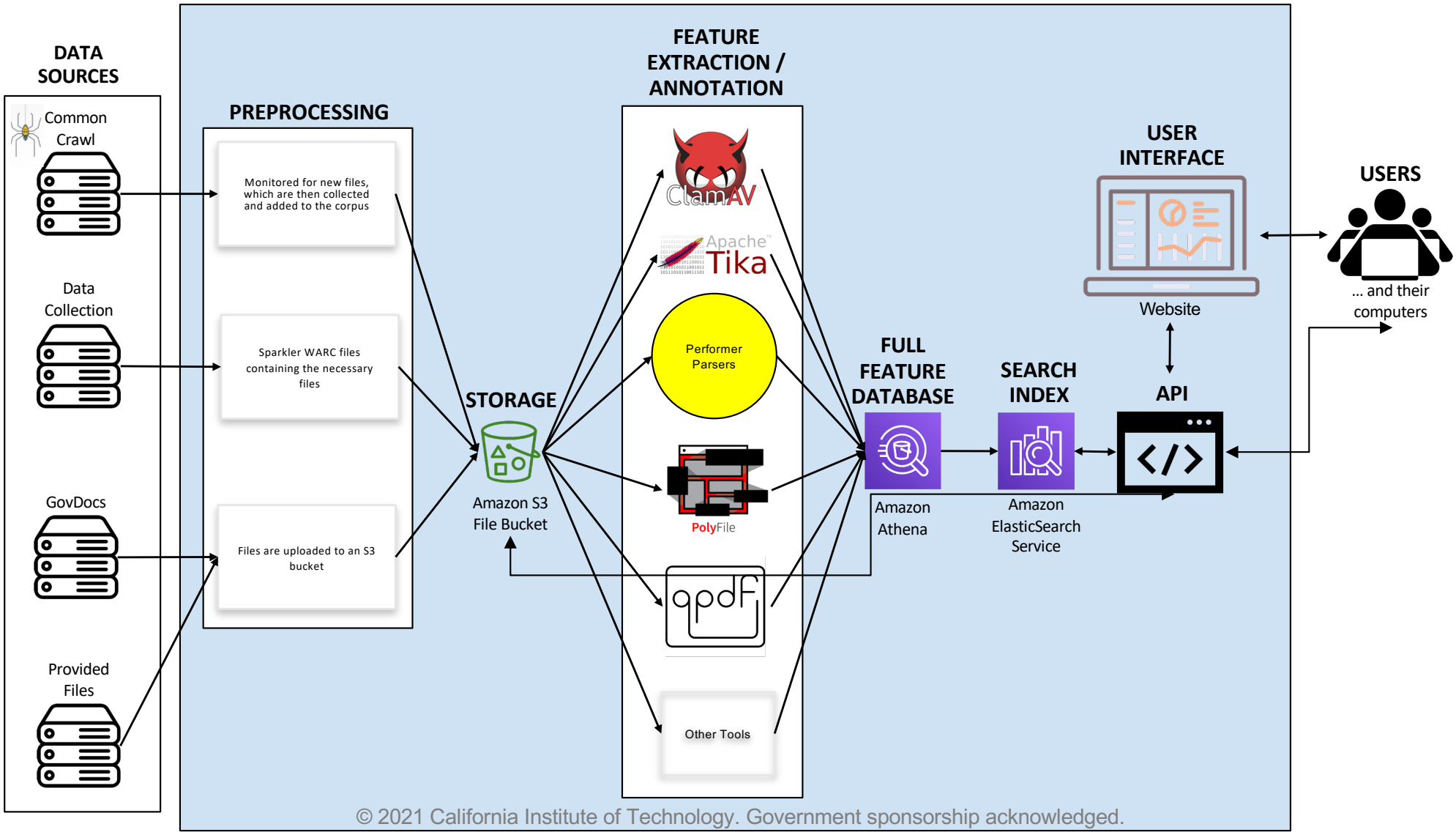
I used this as above without `--g-fatal-warnings` because that's what was being called by Ubuntu in the background.

Here are some of the problematic files.

Observatory at 2 scales

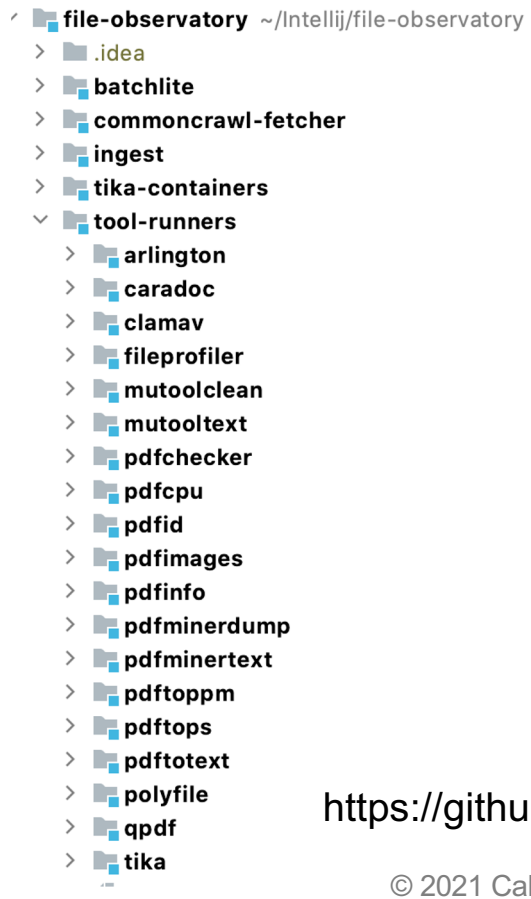
Various Scales

- Full Cloud -- AWS components (Athena, Kinesis…), S3 for storage -> Elasticsearch
- Desktop -- Postgresql, local files -> Elasticsearch
- Hybrid -- hosted Postgresql, EC2 instance for processing, S3 for storage -> Elasticsearch


DATA SOURCES
Common Crawl
Data Collection
GovDocs
Provided Files
PREPROCESSING
Monitored for new files, which are then collected and added to the corpus
Sparkler WARC files containing the necessary files
Files are uploaded to an S3 bucket
STORAGE
Amazon S3 File Bucket
FEATURE EXTRACTION / ANNOTATION
ClamAV
Apache Tika
Performer Parsers
PolyFile
pdf
Other Tools
FULL FEATURE DATABASE
Amazon Athena
SEARCH INDEX
Amazon ElasticSearch Service
API
USER INTERFACE
Website
USERS
... and their computers
© 2021 California Institute of Technology. Government sponsorship acknowledged.

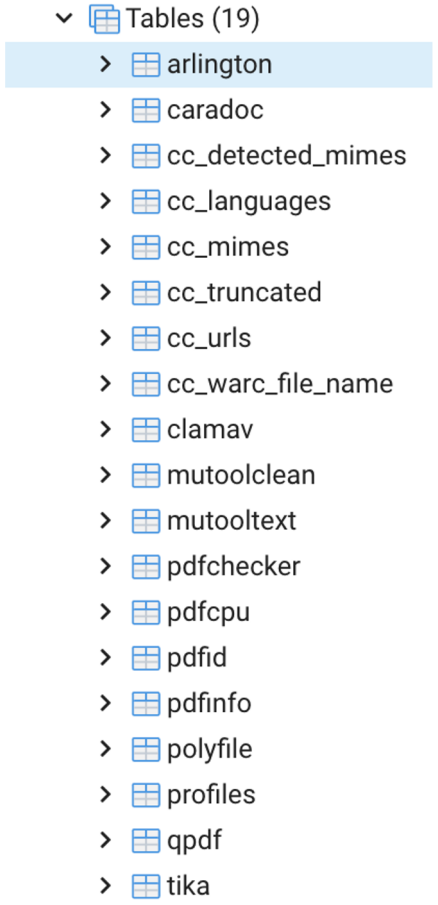
File Observatory in a Box

Public github repository



https://github.com/tballison/file-observatory

Standard Metadata Output Per Tool



Standard table for each tool

Explain

Query Editor

```
1  select path, exit_value, process_time_ms, stdout, stderr from caradoc
```

e "EXPLAIN (FORMAT JSON) [QUERY]".

Data Output

	path [PK] character varying (500)	exit_value integer	process_time_ms bigint	stdout character varying (20000)	stderr character varying (20000)
1	EvalTwo/dangerous/01b6856a7e6...	0	58	Version : 1.4	Warning : Flate/Zlib stream with appended newline in o...
2	EvalTwo/dangerous/023e918b4e1...	2	53	Version : 1.4	PDF error : Invalid stream filter : AHx in object 4 at entry ...
3	EvalTwo/dangerous/0291076dcac...	2	56	Not a PDF file	PDF error : Xref does not start at object 0 in object traile...
4	EvalTwo/dangerous/019734dc8e7...	2	24	Version : 1.5	Warning : Flate/Zlib stream with appended newline in o...
5	EvalTwo/dangerous/025aecde10b...	2	30	Version : 1.4	PDF error : Invalid stream filter : AHx in object 4 at entry ...
6	EvalTwo/dangerous/026929a0433...	2	296	Not a PDF file	PDF error : Lexing error : unexpected word at offset 493 ...
7	EvalTwo/dangerous/00af2c403b20...	2	56	Version : 1.7	Type error : Invalid type : expected s"metadata_stream" i...
8	EvalTwo/dangerous/00ee0b8dba2...	2	113	Not a PDF file	PDF error : Lexing error : unexpected word at offset 53 [...
9	EvalTwo/dangerous/007e86e378e...	2	14	Version : 1.4	Type error : Unexpected entry /ITXT in instance of class ...
10	EvalTwo/dangerous/02935e7a7e2...	2	20	Version : 1.4	PDF error : Invalid stream filter : AHx in object 4 at entry ...
11	EvalTwo/dangerous/02fc5ed61c7e...	0	39	Version : 1.3	Warning : Flate/Zlib stream with appended newline in o...
12	EvalTwo/dangerous/02ed8c4bc27...	2	20	Version : 1.4	PDF error : Invalid stream filter : AHx in object 4 at entry ...
13	EvalTwo/dangerous/00eb20ea546...	2	29	Not a PDF file	PDF error : Lexing error : unexpected character : 0x25 at...
14	EvalTwo/dangerous/02aedcb47ea...	2	85	Version : 1.7	Type error : Invalid variant type : expected (c"pagenode" ...

Discovery with built-in Elasticsearch features

Search – Information Leakage

Using Elasticsearch's Prefix Completion

```
GET file-observatory-202102/_search
{
  "_source": false,
  "suggest": {
    "YOUR_SUGGESTION": {
      "prefix": "/Rotate->",
      "completion": {
        "field": "q_keys_and_values
          .completion",
        "size": 2000,
        "skip_duplicates": true

      }
    }
  }
}
```


```
"options" : [
  {
    "text" : "/Rotate->-10",
    "_index" : "file-observatory-202102",
    "_type" : "_doc",
    "_id" : "s3://safedocs-dev/bugtrackers/PDFBOX/PDFBOX-3353-0.pdf",
    "_score" : 1.0
  },
  {
    "text" : "/Rotate->-13",
    "_index" : "file-observatory-202102",
    "_type" : "_doc",
    "_id" : "s3://safedocs-dev/bugtrackers/PDFIUM/PDFIUM-1380-1.pdf",
    "_score" : 1.0
  },
  {
    "text" : "/Rotate->-180",
    "_index" : "file-observatory-202102",
    "_type" : "_doc",
    "_id" : "s3://safedocs-dev/bugtrackers/GHOSTSCRIPT/GHOSTSCRIPT-689418-0.zip-0
      .pdf",
    "_score" : 1.0
  },
  {
    "text" : "/Rotate->-90",
    "_index" : "file-observatory-202102",
    "_type" : "_doc",
    "_id" : "s3://safedocs-dev/bugtrackers/PDFBOX/PDFBOX-3785-0.pdf",
    "_score" : 1.0
```

Using Elasticsearch's Autosuggest

```
GET file-observatory-202102/_search
{
  "_source": false,
  "suggest": {
    "my-suggest1": {
      "text": "/Subtype",
      "term": {
        "field": "q_keys",
        "suggest_mode":"always",
        "sort":"frequency",
        "size":100,
        "max_edits":2,
        "min_word_length":2,
        "max_term_freq":2000000
      }

    }
  }
}
```

```
"options" : [
  {
    "text" : "/SubType",
    "score" : 0.875,
    "freq" : 147
  },
  {
    "text" : "/Subtype2",
    "score" : 0.875,
    "freq" : 9
  },
  {
    "text" : "/Subtyp",
    "score" : 0.85714287,
    "freq" : 3
  },
  {
    "text" : "/Pubtype",
    "score" : 0.875,
    "freq" : 2
  },
```

```
  {
    "text" : "/Subtypd",
    "score" : 0.875,
    "freq" : 2
  },
  {
    "text" : "/Subtypg",
    "score" : 0.875,
    "freq" : 2
  },
  {
    "text" : "/subtype",
    "score" : 0.875,
    "freq" : 1
  },
  {
    "text" : "/SUbtyPe",
    "score" : 0.75,
    "freq" : 1
  }
```

Significant Terms Query for /SubType

```
GET file-observatory-202103/_search
{
  "query" : {
    "match" : {
      "q_keys" : {
        "query":"\/SubType"
      }
    }
  },
  "size" :0,
  "aggregations" : {
    "significant_queries" : {
      "significant_text" : {
        "field" : "tk_creator_tool",
        "size": 10,
        "chi_square": {
          "background_is_superset" : false
        }
      }
    }
  }
}
```

```
22      "buckets" : [
23        {
24          "key" : "arcmap",
25          "doc_count" : 23,
26          "score" : 3740.5023984417203,
27          "bg_count" : 36
28        },
29        {
30          "key" : "esri",
31          "doc_count" : 23,
32          "score" : 3677.4677551236864,
33          "bg_count" : 37
34        },
35        {
36          "key" : "6.0",
37          "doc_count" : 50,
38          "score" : 2755.4940054732288,
39          "bg_count" : 322
40        },
41        {
42          "key" : "prinectsignastation",
43          "doc_count" : 6,
44          "score" : 1073.003716686631,
45          "bg_count" : 8
```

Significant keys for creator tools “arcmap” or “esri”

```
GET file-observatory-202103/_search
{
  "query" : {
    "match" : {
      "tk_creator_tool" : {
        "query":"arcmap OR esri"
      }
    }
  },
  "size" :0,
  "aggregations" : {
    "significant_queries" : {
      "significant_text" : {
        "field" : "q_keys",
        "size": 50,
        "chi_square": {
          "background_is_superset" : false
        }
      }
    }
  }
}

GET file-observatory-202103/_search
```

```
    "bg_count" : 31219,
    "buckets" : [
      {
        "key" : "/OC0",
        "doc_count" : 19,
        "score" : 10762.751337739273,
        "bg_count" : 25
      },
      {
        "key" : "/WKT",
        "doc_count" : 15,
        "score" : 8431.315225385713,
        "bg_count" : 20
      },
      {
        "key" : "/OC4",
        "doc_count" : 14,
        "score" : 8033.2824452632785,
        "bg_count" : 18
      },
      {
        "key" : "/ESRINorth",
        "doc_count" : 13,
        "score" : 7917.152340761275,
        "bg_count" : 15
      }
```

■ ■ ■

```
      "key" : "/SubType",
      "doc_count" : 23,
      "score" : 4766.166123231515,
      "bg_count" : 122
```

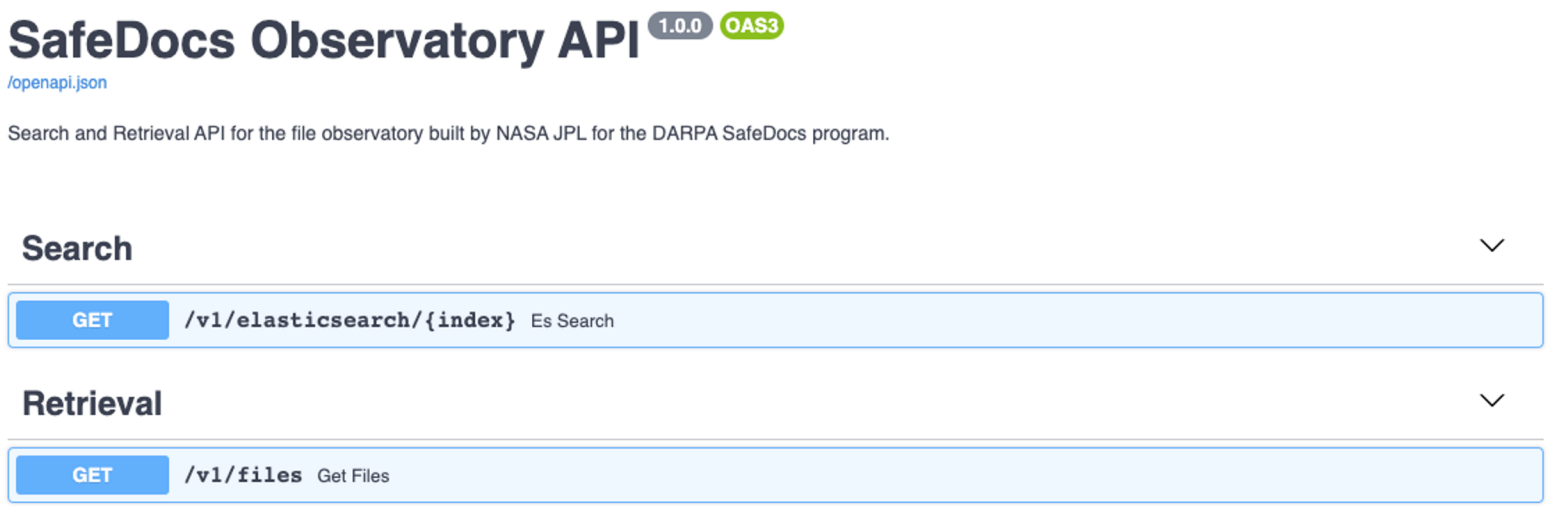
Spelling Variants

Scripting autosuggest!

/Subtype 8798	
/Subtype	8798
/SubType	41
/Subtype2	4
/subtype	1
/CapHeight 6489	
/CapHeight	6489
/CapHieght	8
/CVHeight	2
/ColorSpace 5748	
/ColorSpace	5748
/Colorspace	1
/FirstChar 5650	
/FirstChar	5650
/FirtsChar	1
/Widths 5646	
/Widths	5646
/Width 4625	
/WXdths	1

Observatory API

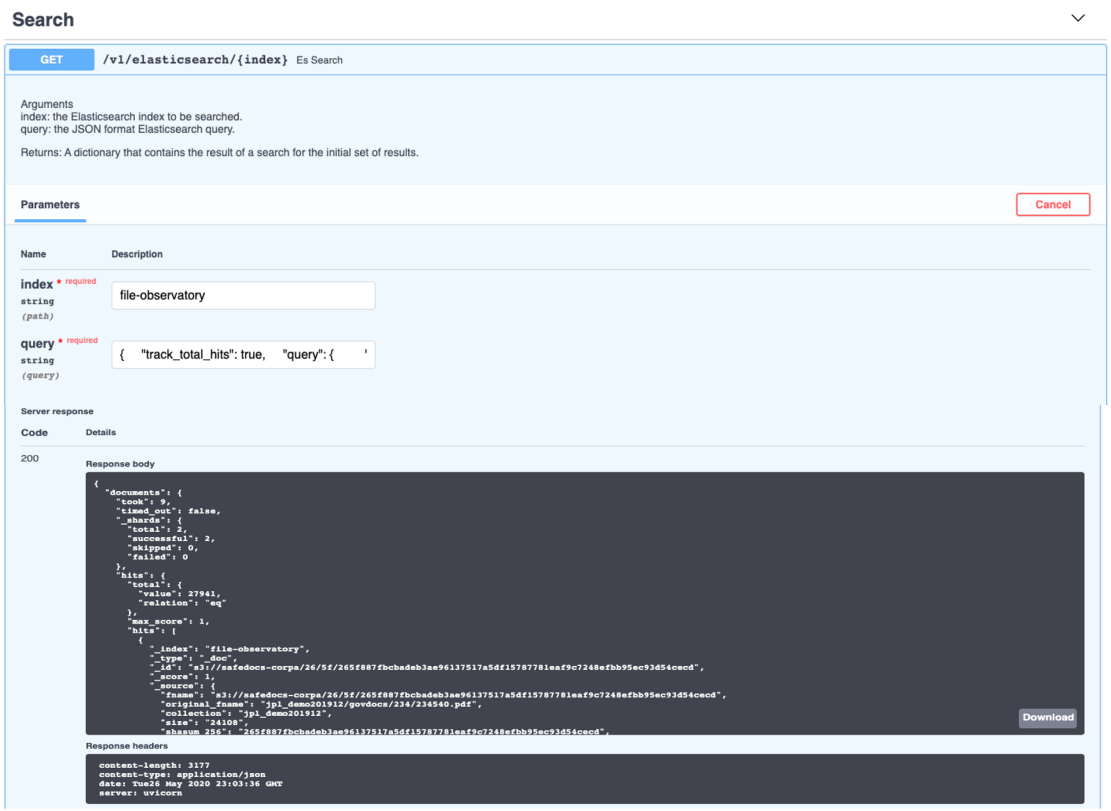
Observatory API



- Consists of two endpoints;
  - Elasticsearch pass-through
  - File Retrieval

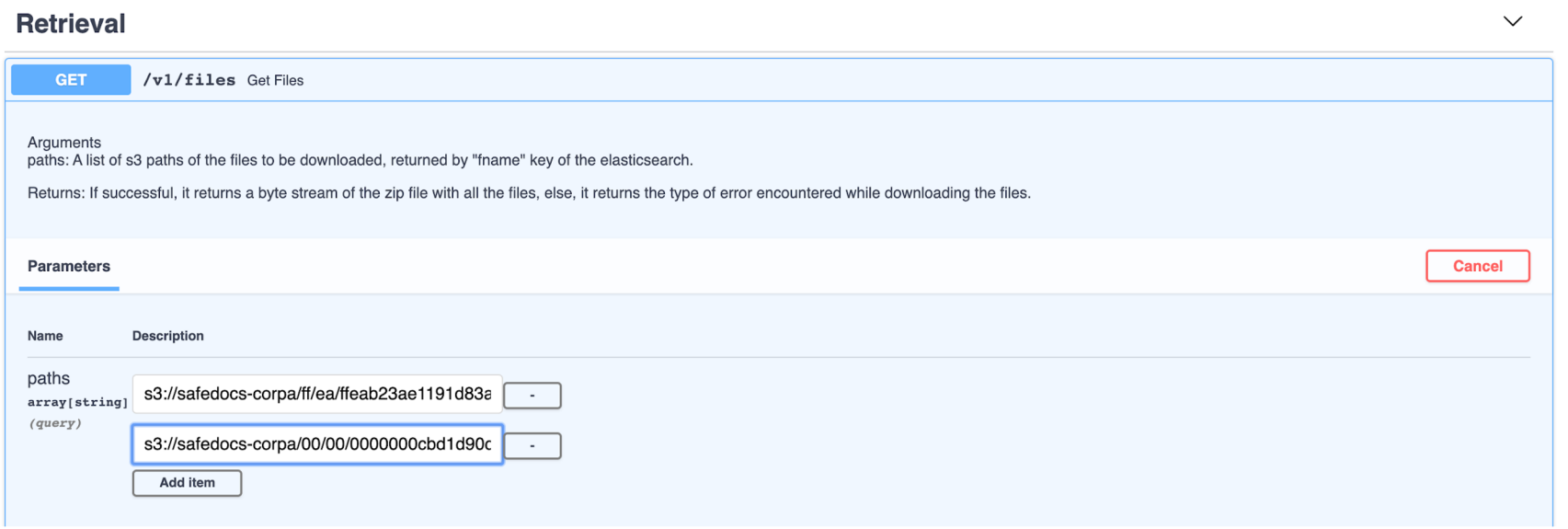
Observatory API – Elasticsearch pass through

- Acts as a pass-through for any query through the API's query parameter
- Returns the total number of document matches
- Upcoming:
  - Will also be accepting json queries through the request body



Observatory API – s3 file retrieval

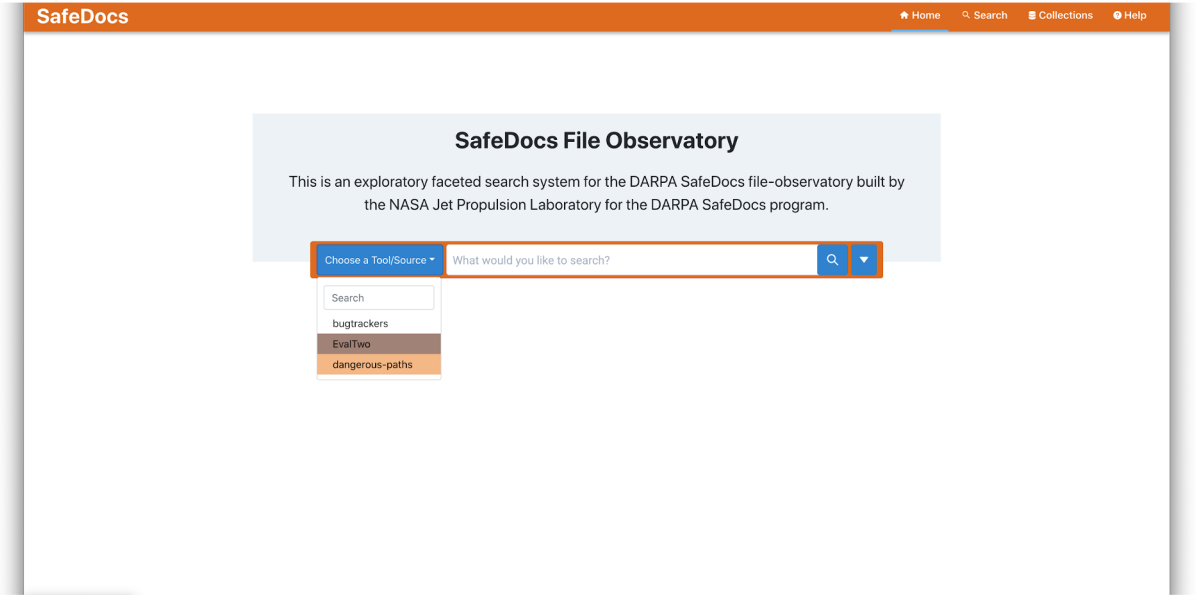
- Input: List of S3 urls returned by the “fname” parameter of ES.
- Returns a zip folder containing all the requested files



Search and Analytics at Scale: Observatory UI

Search UI

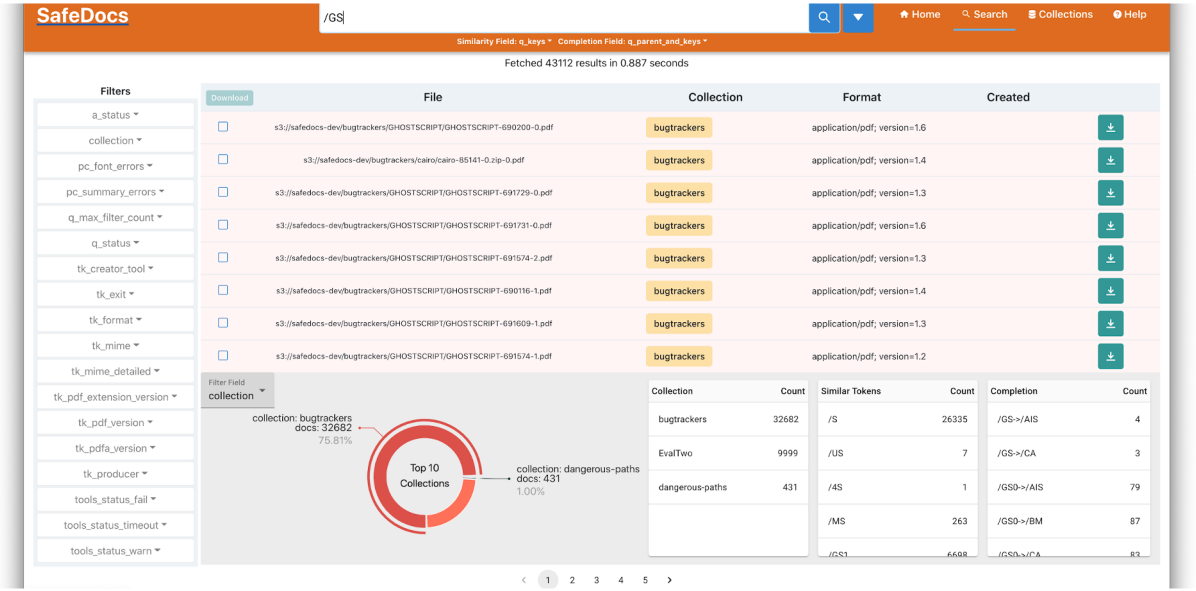
- Interactive way to search through the file-observatory, visualize content, and download files of interest
- Built with a large-volume of data in mind
- Currently live and working



Search UI

Search/Results Page

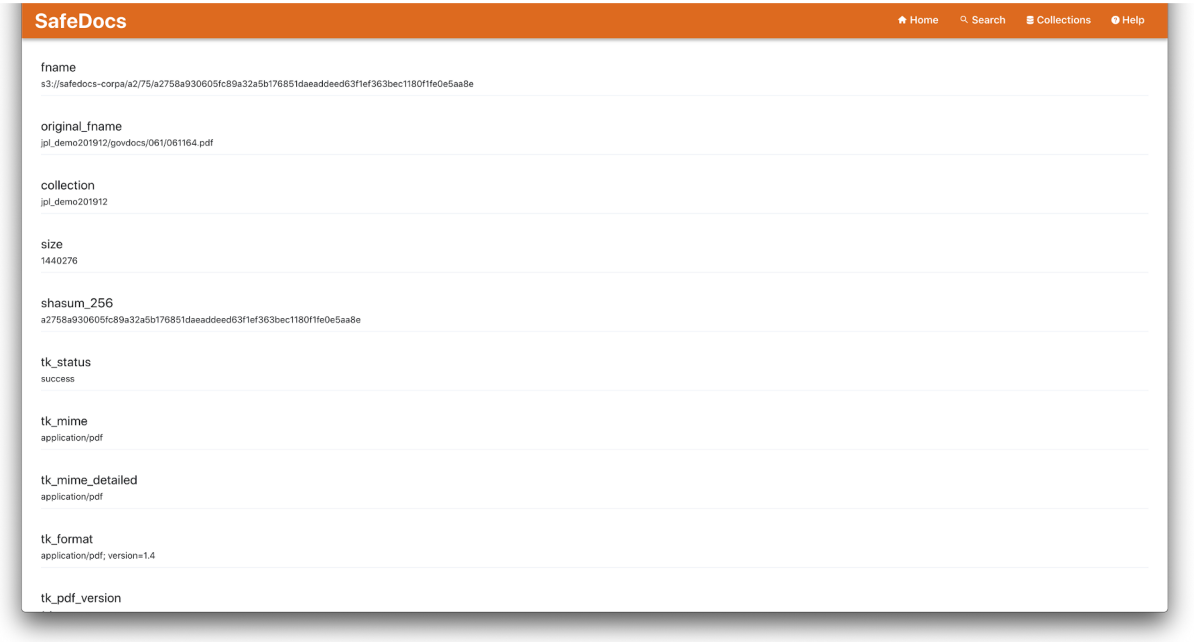
- Dynamic filtering based on 18 different fields and search text
- Allows for downloading of individual or multiple PDFs
- Provides table of similar tokens within a 2 character distance
- Provides table of suggested completions and the number of documents in which they occur
- Dynamic filter field breakdown visualization



Search UI

File View

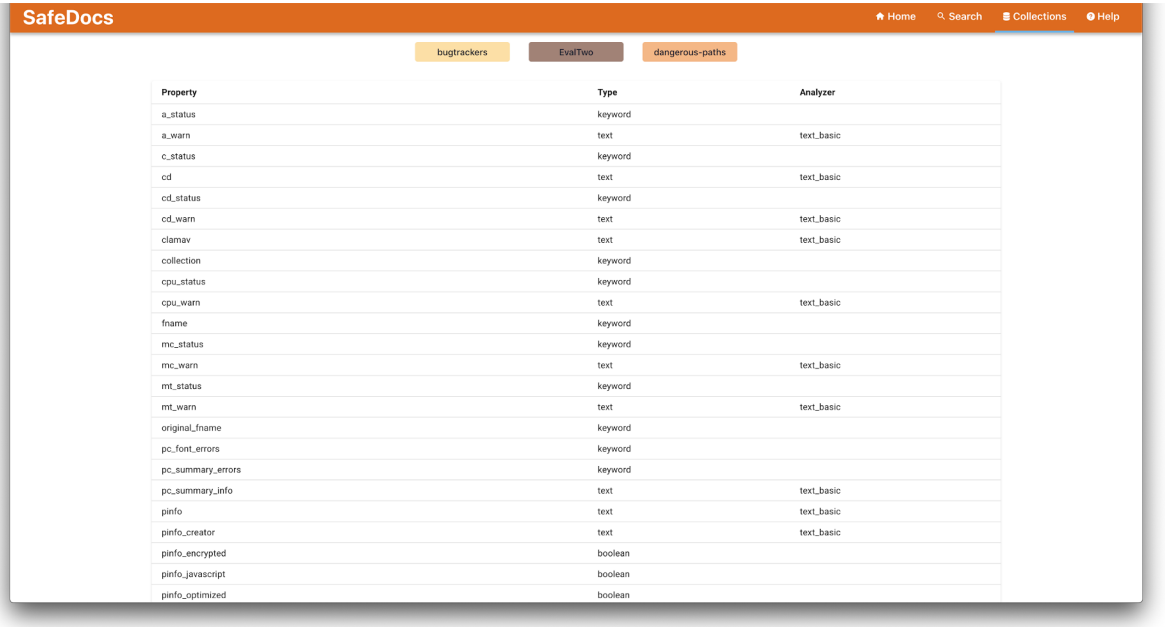
- Provides in-depth view of all fields stored in elasticsearch for a document by clicking on name



Search UI

Collections Page

- Provides a real-time view of all collections in the file-observatory
- Pulls the latest ElasticSearch file-observatory mapping schema with all available fields



Property	Type	Analyzer
a_status	keyword	
a_warn	text	text_basic
c_status	keyword	
cd	text	text_basic
cd_status	keyword	
cd_warn	text	text_basic
clamav	text	text_basic
collection	keyword	
cpu_status	keyword	
cpu_warn	text	text_basic
fname	keyword	
mc_status	keyword	
mc_warn	text	text_basic
mt_status	keyword	
mt_warn	text	text_basic
original_fname	keyword	
pc_font_errors	keyword	
pc_summary_errors	keyword	
pc_summary_info	text	text_basic
pinfo	text	text_basic
pinfo_creator	text	text_basic
pinfo_encrypted	boolean	
pinfo_javascript	boolean	
pinfo_optimized	boolean	

Document Classification Task

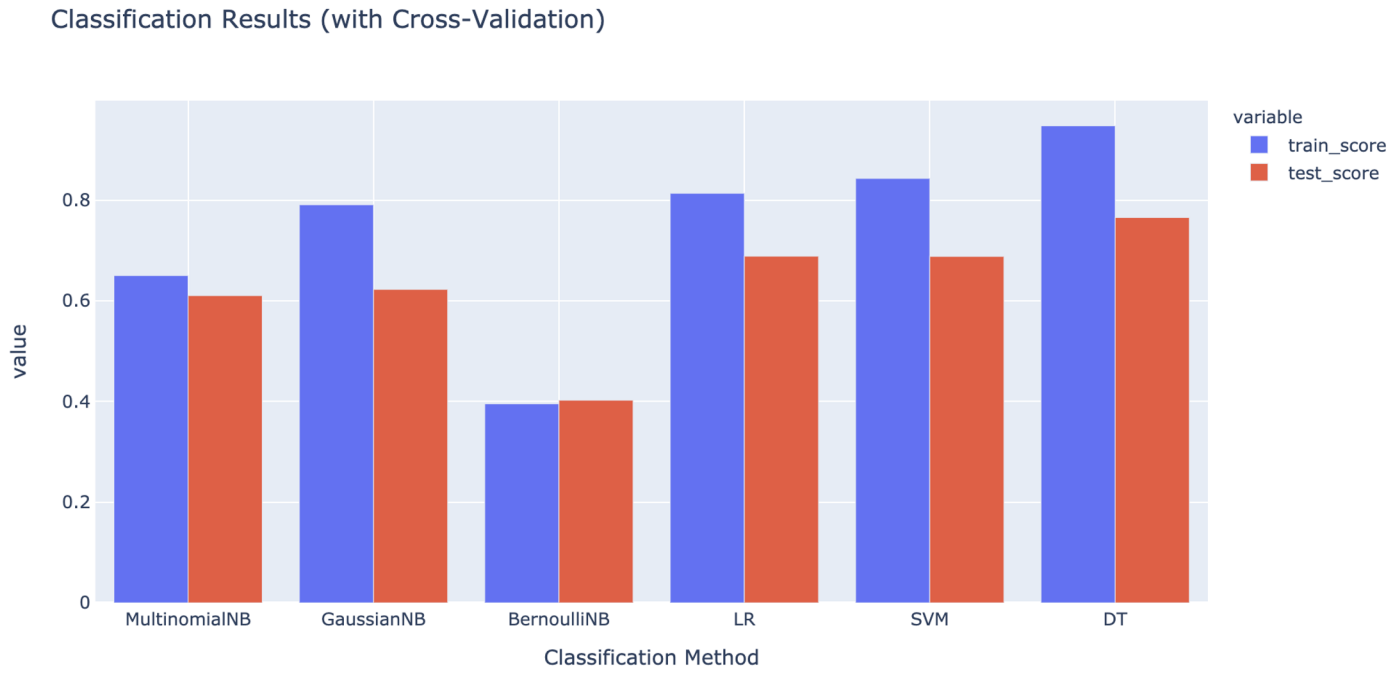
Document Classification Task: Setup

- Classifying the ***creator tool*** based on ***keys***
  - 2021 dataset: q keys, key-value pairs, parent-key pairs
- Top 10 most occurring creator tools
- Splitting the dataset 80:20
- Label = creator tool, features = keys
- Using CountVectorizer to vectorize the data
- Testing out different classifiers (with cross-validation):
  - MultinomialNB
  - GaussianNB
  - Bernoulli NB
  - Logistic Regression
  - SVM
  - Decision Trees

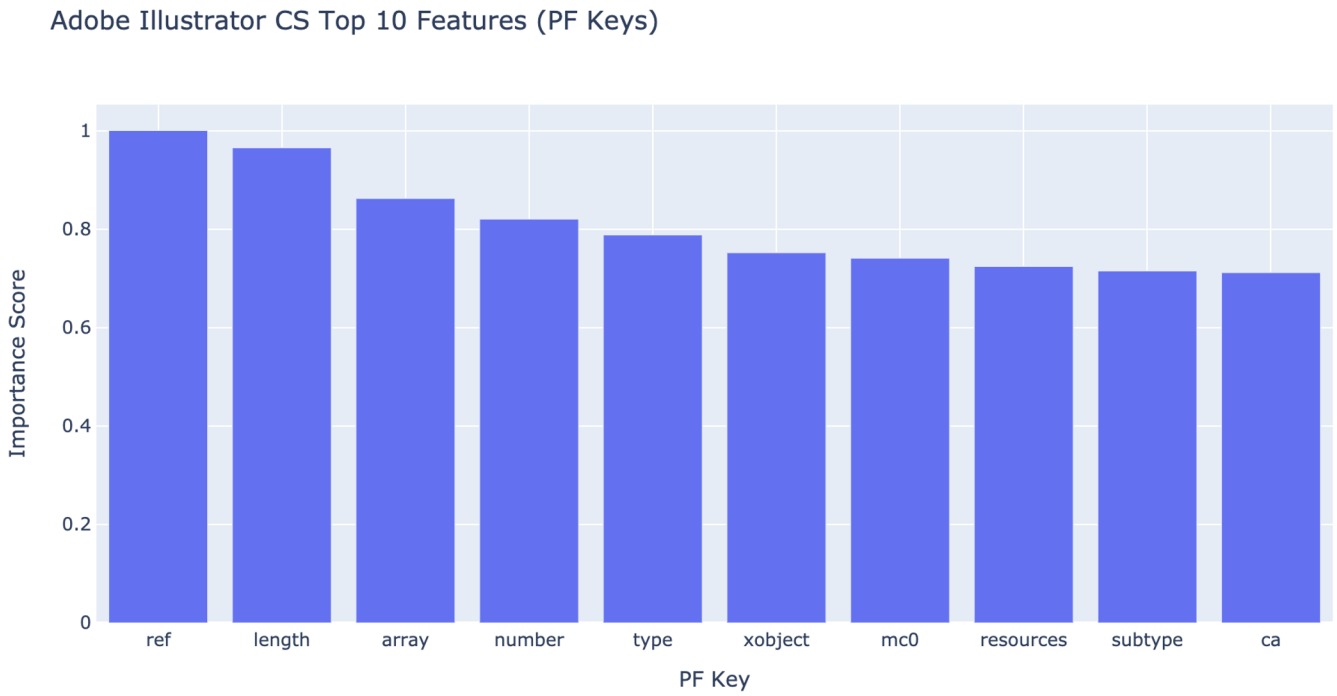
Document Classification Task: Data

	id	tool	keys
0	bugtrackers/GHOSTSCRIPT/GHOSTSCRIPT-688979-0.pdf	PScriptdll	[[', A',', Annots',', Ascent',', Author',', BG...
2	bugtrackers/GHOSTSCRIPT/GHOSTSCRIPT-696715-0.pdf	PScriptdll	[[', Ascent',', Author',', BaseFont',', BitsPe...
5	bugtrackers/GHOSTSCRIPT/GHOSTSCRIPT-690548-2.r...	PScriptdll	[[', Ascent',', Author',', BaseFont',', BitsPe...
9	bugtrackers/GHOSTSCRIPT/GHOSTSCRIPT-691045-0.b...	Adobe InDesign CS	[[', Ascent',', BaseFont',', BitsPerComponent'...
14	bugtrackers/GHOSTSCRIPT/GHOSTSCRIPT-692988-0.pdf	Adobe Illustrator CS	[[', Ascent',', BaseFont',', BitsPerComponent'...
...	...	...	...
1224	bugtrackers/GHOSTSCRIPT/GHOSTSCRIPT-699947-0.t...	TeX	[[', Ascent',', BaseFont',', BitsPerComponent'...
1240	bugtrackers/GHOSTSCRIPT/GHOSTSCRIPT-693588-1.pdf	Microsoft Office Word	[[', Ascent',', Author',', BaseFont',', CapHei...
1253	bugtrackers/GHOSTSCRIPT/GHOSTSCRIPT-695334-0.pdf	PScriptdll	[[', Ascent',', Author',', BaseFont',', CapHei...
1263	bugtrackers/GHOSTSCRIPT/GHOSTSCRIPT-697350-0.pdf	Adobe InDesign CS	[[', Author',', BaseFont',', BitsPerComponent'...
1266	bugtrackers/GHOSTSCRIPT/GHOSTSCRIPT-695390-0.pdf	TeX	[[', ColorSpace',', Contents',', Count',', Cre...

Document Classification Task: Results



Document Classification Task: Feature Importance
(an example, for MultinomialNB)



Next Steps

Next Steps

Settling on features for extraction

Simplifying code to enable faster integration of new tools

Releasing “observatory in a box”

Jet Propulsion Laboratory
California Institute of Technology

jpl.nasa.gov

Extras

Map of PDFs in the July 2020 Common Crawl Data -- GeoLocation of URL/IPs via MaxMind's GeoIP City Database

jpl.nasa.gov

Top 10 “Top Level Domains” for PDFs in the July 2020 Common Crawl Data