

The Vilem Flusser Archive **ORG** owns a personal computer
production of a software titled "Flusser-Hypertext". This computer contains a rare
working copy of the software which is dependent on the obsolete authoring system
erCard **PRODUCT**. The disk image has been acquired from a **Apple** **ORG**
PRODUCT Performa 630 containing a 270Mb IDE disk. The goal of this use-case
to web-based access to the **Flusser** **PERSON**-Hypertext through the

BitCurator NLP

Mining Collections for NEs, Relationships, and Topics to Enrich Access

nlp4arc – February 3, 2017

Kam Woods

Research Scientist / BitCurator NLP Technical Lead

University of North Carolina at Chapel Hill

School of Information and Library Science



UNC
SCHOOL OF INFORMATION
AND LIBRARY SCIENCE

BitCurator NLP Overview

Andrew W. Mellon Foundation funded project (Oct 2016 – Oct 2018)

*“The BitCurator NLP project will produce software allowing institutions to extract, analyze, and produce reports about **relevant features found in open text** within digital materials held in collections. The software will rely on **existing NLP libraries** to **identify and report on those items likely to be relevant** to ongoing preservation, information organization, and access activities, including entities (e.g. **persons, places, and organizations**), potential relationships among entities (for example, by describing those entities that appear together within documents or set of documents), and topic models to provide insight into **how concepts are naturally clustered within the documents.**”*

It often starts the same way...



Source: "Digital Forensics and creation of a narrative." *Da Blog: ULCC Digital Archives Blog*.
<http://dablog.ulcc.ac.uk/2011/07/04/forensics/>

Core Approach

Assume (simulate or replicate) a wide range of archival collections

- Raw and forensically packaged disk images
- Heterogeneous collections of files (many file types, limited metadata)
- Use established corpora such as GovDocs1

First steps...extracting text from several dozen extremely common formats (disregard the long tail to begin with)

- No single tool appropriate for this task – use existing wrappers around mature tools

<https://textract.readthedocs.io/en/stable/>

- `.csv` via python builtins
- `.doc` via `antiword`
- `.docx` via `python-docx`
- `.eml` via python builtins
- `.epub` via `ebooklib`
- `.gif` via `tesseract-ocr`
- `.jpg` and `.jpeg` via `tesseract-ocr`
- `.json` via python builtins
- `.html` and `.htm` via `beautifulsoup4`
- `.mp3` via `SpeechRecognition` and `sox`
- `.msg` via `msg-extractor`
- `.odt` via python builtins
- `.ogg` via `SpeechRecognition` and `sox`
- `.pdf` via `pdftotext` (default) or `pdfminer.six`
- `.png` via `tesseract-ocr`
- `.pptx` via `python-pptx`
- `.ps` via `ps2text`
- `.rtf` via `unrtf`
- `.tiff` and `.tif` via `tesseract-ocr`
- `.txt` via python builtins
- `.wav` via `SpeechRecognition`
- `.xlsx` via `xlrd`
- `.xls` via `xlrd`

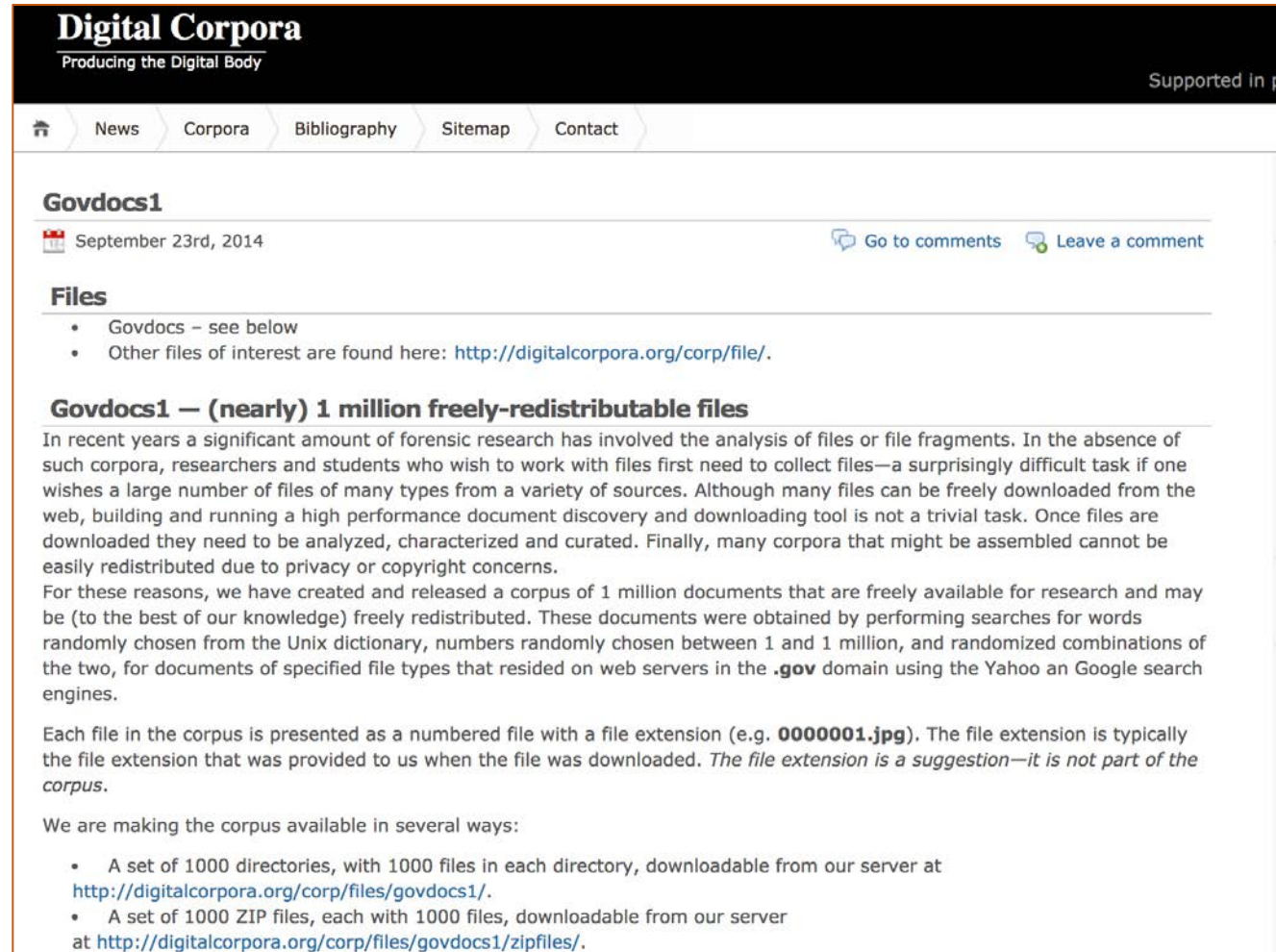
Core Approach

Advantages of using a corpus like GovDocs1:

- In many cases, these documents are **actual records** (publicly available on the web)
- Tests can be easily replicated, assessed by partners
- Partners often won't (or can't) give us collection data. Provides additional options for sharing.

Disadvantages:

- Excludes many legacy file types



The screenshot shows the website for Digital Corpora, with the tagline "Producing the Digital Body". The navigation menu includes Home, News, Corpora, Bibliography, Sitemap, and Contact. The main content area is titled "Govdocs1" and features a date of "September 23rd, 2014" and links for "Go to comments" and "Leave a comment". Under the "Files" section, there is a list of items: "Govdocs - see below" and "Other files of interest are found here: <http://digitalcorporas.org/corp/file/>".

Govdocs1 — (nearly) 1 million freely-redistributable files

In recent years a significant amount of forensic research has involved the analysis of files or file fragments. In the absence of such corpora, researchers and students who wish to work with files first need to collect files—a surprisingly difficult task if one wishes a large number of files of many types from a variety of sources. Although many files can be freely downloaded from the web, building and running a high performance document discovery and downloading tool is not a trivial task. Once files are downloaded they need to be analyzed, characterized and curated. Finally, many corpora that might be assembled cannot be easily redistributed due to privacy or copyright concerns.

For these reasons, we have created and released a corpus of 1 million documents that are freely available for research and may be (to the best of our knowledge) freely redistributed. These documents were obtained by performing searches for words randomly chosen from the Unix dictionary, numbers randomly chosen between 1 and 1 million, and randomized combinations of the two, for documents of specified file types that resided on web servers in the **.gov** domain using the Yahoo and Google search engines.

Each file in the corpus is presented as a numbered file with a file extension (e.g. **0000001.jpg**). The file extension is typically the file extension that was provided to us when the file was downloaded. *The file extension is a suggestion—it is not part of the corpus.*

We are making the corpus available in several ways:

- A set of 1000 directories, with 1000 files in each directory, downloadable from our server at <http://digitalcorporas.org/corp/files/govdocs1/>.
- A set of 1000 ZIP files, each with 1000 files, downloadable from our server at <http://digitalcorporas.org/corp/files/govdocs1/zipfiles/>.

Core Approach

Use [spaCy.io](https://spacy.io) for entity recognition, topic modeling, other tasks...

- Why spaCy?
 - Geared towards product development more than research (e.g. NLTK, openNLP)
 - High-performance (multi-threaded, runs in 64-bit Python stack)
 - Relatively simple API
 - Good pre-trained models for entity and item recognition
 - Integrates easily with machine learning platforms (e.g TensorFlow, Keras, Scikit-Learn, Gensim)
- Strive for simple stacks
 - In this instance, [Python](#) + [PIP](#) + [textract](#) + [spaCy](#), deploy on any platform
 - Provide flexible APIs but simplify basic use cases: “Text goes in, entity span comes out”

Core Approach

Why do it locally at all? (Why not GCloud Language API?)

- Pricing structure is modest but could be prohibitive for institutions working with large collections
- All results in JSON
- Many institutions restricted from running collections through this kind of workflow

PRICE PER 1,000 UNITS, BY MONTHLY USAGE				
FEATURE	0 - 5K UNITS/MONTH	5K+ - 1M UNITS/MONTH	1M+ - 5M UNITS/MONTH	5M+ - 20M UNITS/MONTH
Entity Recognition	FREE	\$1.00	\$0.50	\$0.25
Sentiment Analysis	FREE	\$1.00	\$0.50	\$0.25
Syntax Analysis	FREE	\$0.50	\$0.25	\$0.125

<https://cloud.google.com/natural-language/>

Generating entity views for the web

4.1 Using an EWF Image as Boot Disk

To evaluate the capabilities of our tools and workflow, we have chosen a real use case, demonstrating the image generation process, i.e. a technical generalization to be used as an appropriate emulator.

The Vilem Flusser Archive owns a personal computer associated with the production of a software titled “Flusser-Hypertext”. This computer contains a rare working copy of the software which is dependent on the obsolete authoring system HyperCard. The disk image has been acquired⁶ from an Apple Mac Performa 630 containing a 270 MB IDE disk. The goal was to enable web-based access to the Flusser-Hypertext through the archive’s web site.

Using the acquired disk image directly with an emulator failed. The original machine used a hardware-related extension (A/ROSE) that is not supported by the emulator used and prevented the system to start properly. A simple solution was to boot the system with all extensions disabled and to delete the A/ROSE extension file from the system’s extensions folder. The result of this process is an overlay-

Original text (shown here – clip of PDF)

Text extraction: textract
Entity ident: spaCy
Web display: displaCy API

The Vilem Flusser Archive **ORG** owns a personal computer associated with the production of a software titled “Flusser-Hypertext”. This computer contains a rare working copy of the software which is dependent on the obsolete authoring system HyperCard **PRODUCT**. The disk image has been acquired from a Apple **ORG** Mac **PRODUCT** Performa 630 containing a 270Mb IDE disk. The goal of this use-case is to enable web-based access to the Flusser **PERSON**-Hypertext through the archive’s web site.

Web rendering (autogen’d HTML + CSS)

```
<div class="entities"><mark data-entity="org">The Vilem Flusser Archive</mark> owns a personal computer associated<br>with the production of a software titled “FlusserHypertext”. <br>This computer contains a rare working copy of<br>the software which is dependent on the obsolete authoring<br>system <mark data-entity="product">HyperCard</mark>. The disk image has been acquired<sup>6</sup> from<br>an <mark data-entity="org">Apple</mark> <mark data-entity="product">Mac</mark> Performa 630 containing a 270 <mark data-entity="org">MB</mark> IDE disk.The goal was to enable web-based access to the FlusserHypertextthrough the archive’s web site.</div>
```

```
.entities { line-height: 2; }[data-entity] { padding: 0.25em 0.35em; margin: 0px 0.25em; line-height: 1; display: inline-block; border-radius: 0.25em; border: 1px solid; }[data-entity]::after { box-sizing: border-box; content: attr(data-entity); font-size: 0.6em; line-height: 1; padding: 0.35em; border-radius: 0.35em; text-transform: uppercase; display: inline-block; vertical-align: middle; margin: 0px 0px 0.1rem 0.5rem; }[data-entity][data-entity="person"] { background: rgba(166, 226, 45, 0.2); border-color: rgb(166, 226, 45); }
```


Generating entity views for the web

The Vilem Flusser Archive **ORG** owns a personal computer associated with the production of a software titled "Flusser-Hypertext". This computer contains a rare working copy of the software which is dependent on the obsolete authoring system HyperCard **PRODUCT**. The disk image has been acquired from a Apple **ORG** Mac **PRODUCT** Performa 630 containing a 270Mb IDE disk. The goal of this use-case is to enable web-based access to the Flusser **PERSON**-Hypertext through the archive's web site.

Not always as clean as we'd like...

```
HTML
of<br>the software which is dependent
on the obsolete authoring<br>system
<mark data-
entity="product">HyperCard</mark>.
The disk image has been acquired6
from<br>an <mark data-
entity="org">Apple</mark> <mark data-
entity="product">Mac</mark> Performa
630 containing a 270 <mark data-
entity="org">MB</mark> IDE disk.
2 The goal was to enable web-based
access to the FlusserHypertext
3 through the archive's web site.</div>

CSS
24 margin: 0px 0px 0.1rem 0.5rem;
25 }
26
27 [data-entity][data-entity="person"]
28 {
29     background: rgba(166, 226, 45,
30     0.2);
31     border-color: rgb(166, 226,
32     45);
33 }
34 [data-entity][data-
35 entity="person"]::after {
36     background: rgb(166, 226, 45);
37 }
```

The Vilem Flusser Archive **ORG** owns a personal computer associated with the production of a software titled “FlusserHypertext”.

This computer contains a rare working copy of

the software which is dependent on the obsolete authoring

system **HyperCard** **PRODUCT**. The disk image has been acquired6 from

an **Apple** **ORG** **Mac** **PRODUCT** Performa 630 **MB** **ORG** IDE disk. The goal was to enable web-based access to the FlusserHypertext through the archive’s web site.

```
2 The goal was to enable web-based
access to the FlusserHypertext
3 through the archive's web site.</div>
```

```
31
32 [data-entit
entity="per
33 backgr
```

The Vilem Flusser Archive **ORG** owns a personal computer associated with the production of a software titled “FlusserHypertext”.

This computer contains a rare working copy of

the software which is dependent on the obsolete authoring

system **HyperCard** **PRODUCT**. The disk image has been acquired⁶ from

an **Apple** **ORG** **Mac** **PRODUCT** Performa 630 containing a 270 **MB** **ORG** IDE disk

Hmmmm.....

Entity type	Description
PERSON	People
NORP	Nationalities, religious, and political groups.
FACILITY	Buildings, airports, highways, bridges, etc.
ORG	Companies, agencies, and institutions.
GPE	Countries, cities, and states.
LOC	Locations other than GPE (e.g. mountain ranges, bodies of water)
PRODUCT	Objects other than services (e.g. devices, foods)
EVENT	Historical events (e.g. cultural, weather, conflicts)
WORK_OF_ART	Titles of works of art
LANGUAGE	Named languages
Additional feature types	Description
DATE	Dates or periods (absolute / relative)
TIME	Time periods less than a day
PERCENT	Percentages (also marked by '%')
MONEY	Monetary values, including by unit
QUANTITY	Weight, distance, other measurements
ORDINAL	E.g 'first', 'second'
CARDINAL	Numeral identifiers other than those typed above

We expect this code to be deployed in real-world institutions – performance is a consideration.

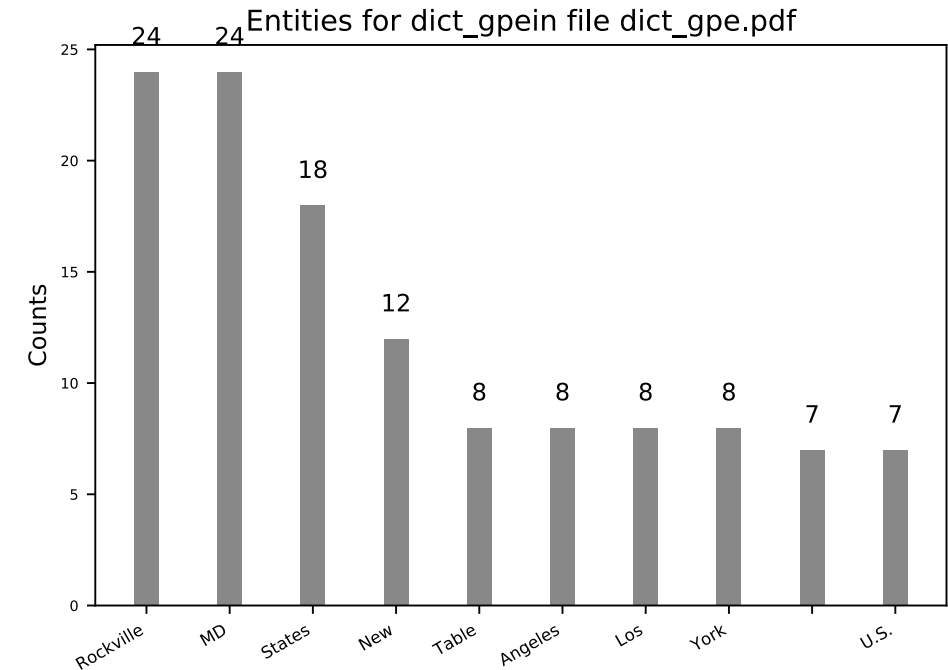
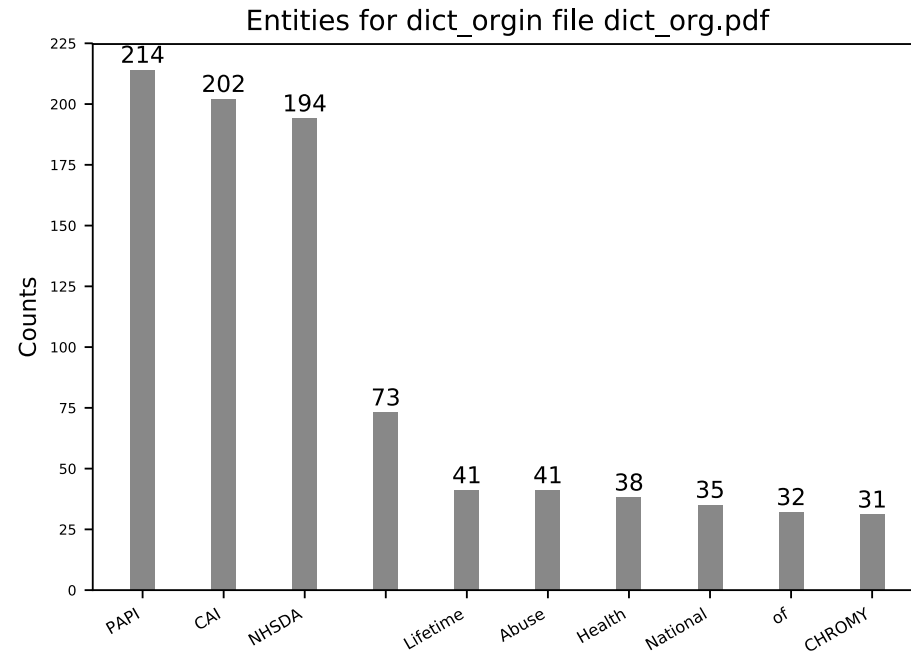
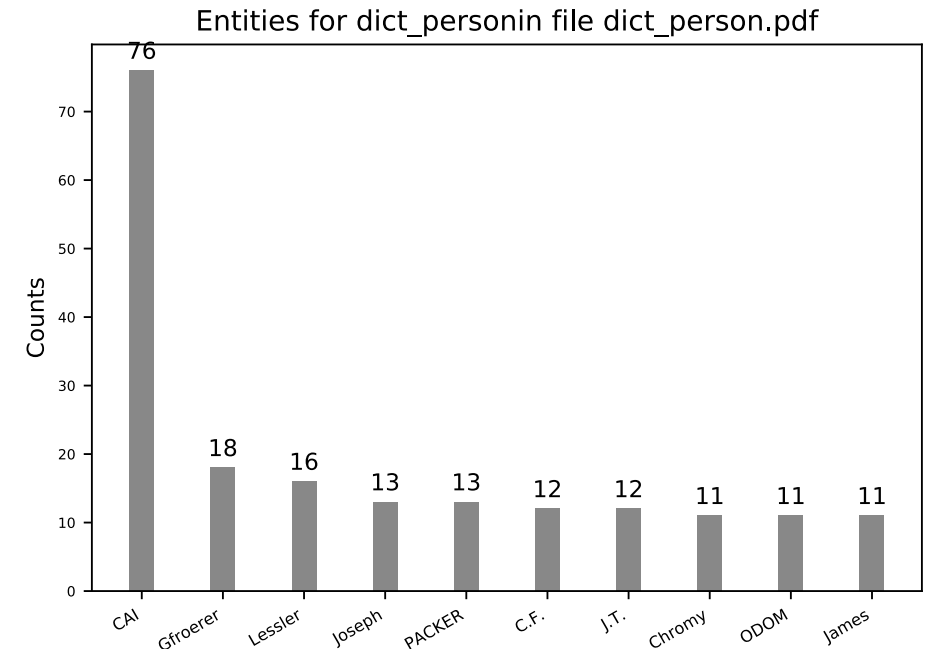
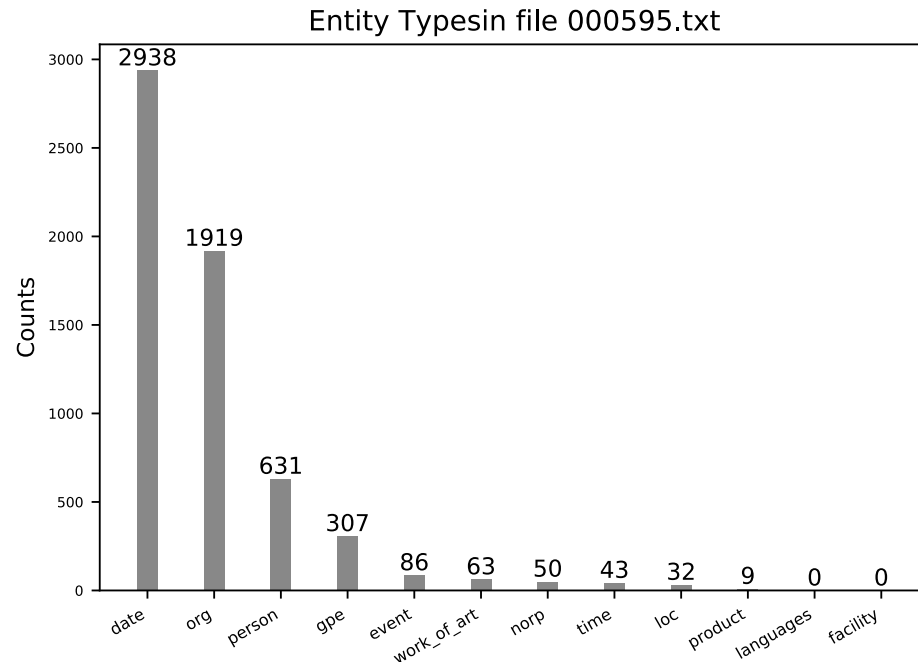
Baseline test on a circa-2014 Core i7 ThinkPad:

- 1336 files (approx. 1GB)
- Text extraction via textacy -> entity extraction via spaCy
- 52 minutes (including OCR of image formats)

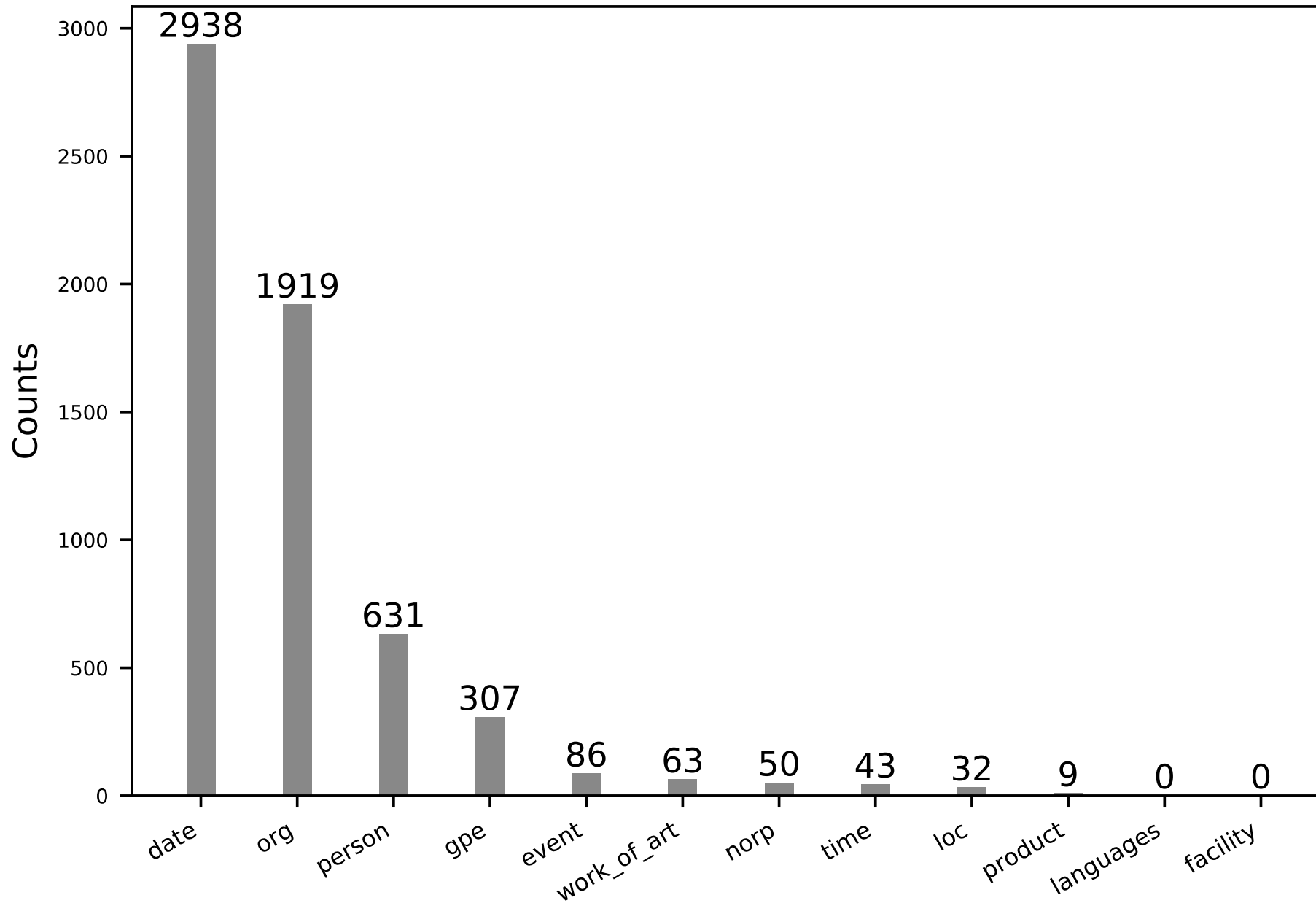
```
real 51m55.043s
user 46m25.511s
sys 1m37.768s
(venv)sunitha@sm-T440s:~/BC/NLP/displaCy$ ls indir | wc
 27  27  308
(venv)sunitha@sm-T440s:~/BC/NLP/displaCy$ ls indir
000000.swf  000004.doc  000007.doc  gif_files  new_infile.pdf  wp_files
000001.doc  000004.doc.span 000008.ppt  html_files  pdffiles000  xls_files
000002.doc  000005.doc  000009.pdf  infile.txt  ppt_files
000002.doc.span 000005.doc.span csv_files  infile.txt.span ps_files
000003.doc  000006.doc  dir1  jpg_files  txtfiles000
(venv)sunitha@sm-T440s:~/BC/NLP/displaCy$ cd indir
(venv)sunitha@sm-T440s:~/BC/NLP/displaCy/indir$ ls gif_files | wc
 46  46  621
(venv)sunitha@sm-T440s:~/BC/NLP/displaCy/indir$ ls html_files | wc
 362  362  5249
(venv)sunitha@sm-T440s:~/BC/NLP/displaCy/indir$ ls wp_files | wc
  2  2  20
(venv)sunitha@sm-T440s:~/BC/NLP/displaCy/indir$ ls pdffiles000 | wc
 200  200  2200
(venv)sunitha@sm-T440s:~/BC/NLP/displaCy/indir$ ls xls_files | wc
 124  124  1674
(venv)sunitha@sm-T440s:~/BC/NLP/displaCy/indir$ ls ppt_files | wc
  88  88  968
(venv)sunitha@sm-T440s:~/BC/NLP/displaCy/indir$ ls ps_files | wc
  30  30  370
(venv)sunitha@sm-T440s:~/BC/NLP/displaCy/indir$ ls txtfiles000 | wc
 283  283  3758
(venv)sunitha@sm-T440s:~/BC/NLP/displaCy/indir$ ls jpg_files | wc
 178  178  2403
(venv)sunitha@sm-T440s:~/BC/NLP/displaCy/indir$ ls csv_files | wc
  21  21  281
(venv)sunitha@sm-T440s:~/BC/NLP/displaCy/indir$ ls dir1 | wc
  2  2  14
```

For a sample set of several hundred files from the GovDocs corpus, in clockwise order from top left:

- Entity types
- Persons
- Organizations
- Geopolitical entities

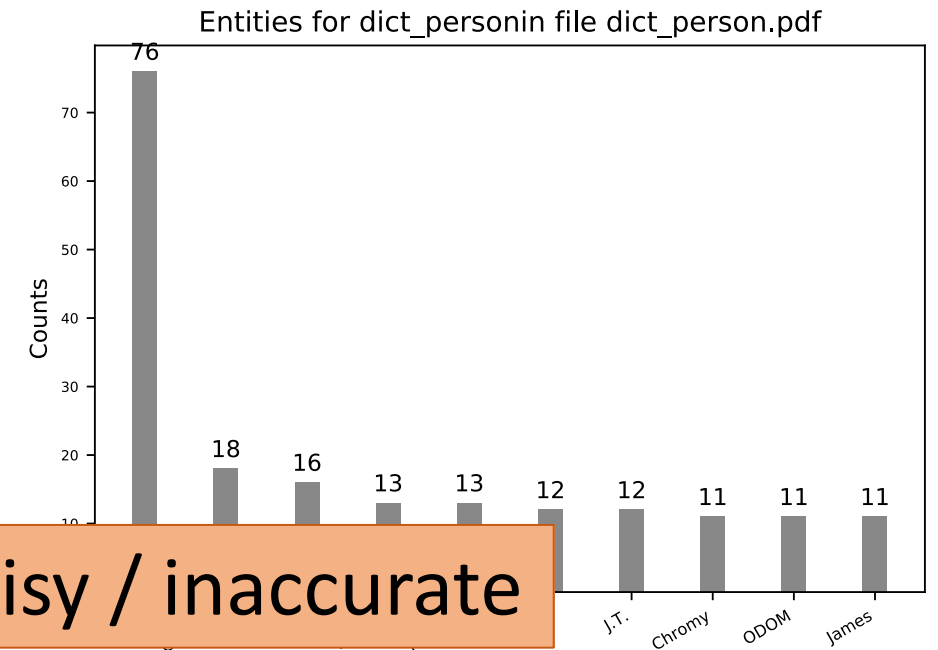
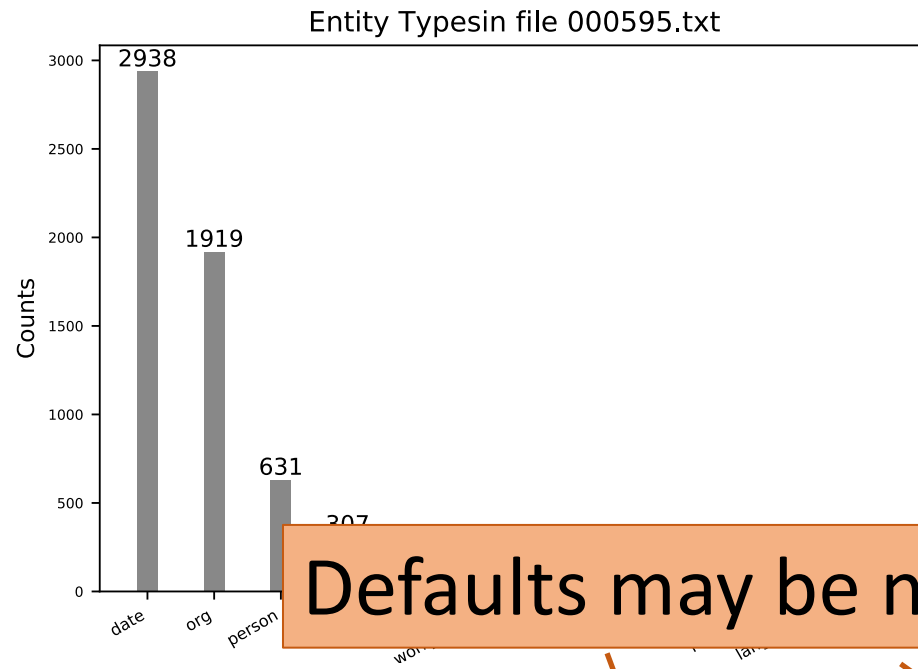


Entity Types in file 000595.txt

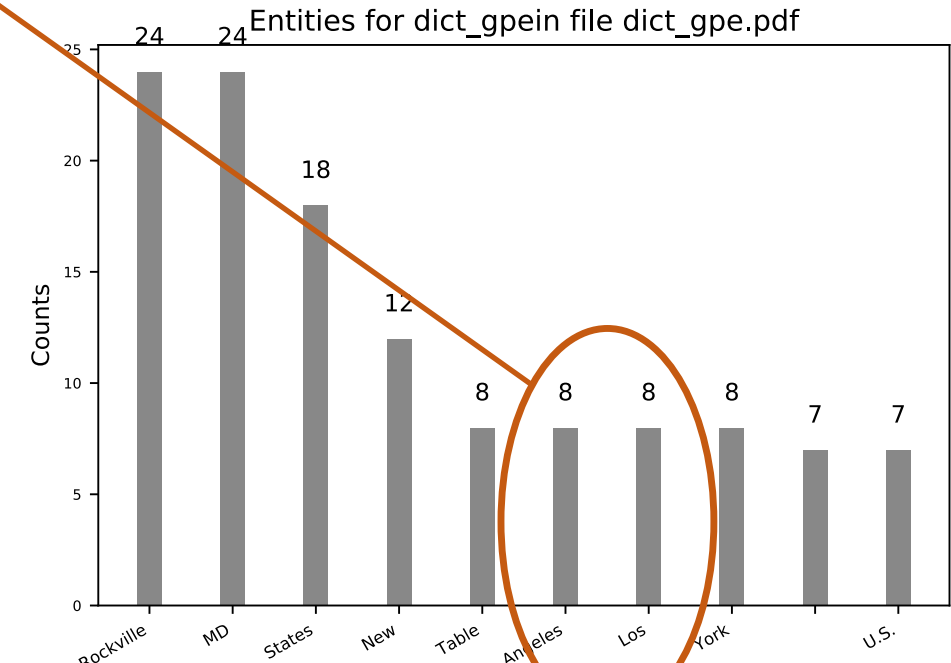
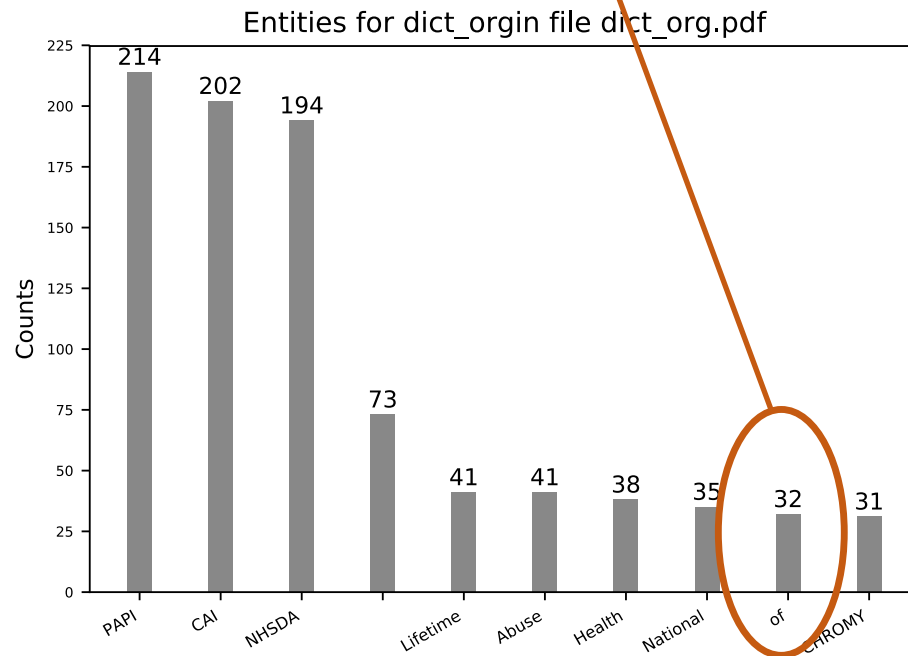


For a sample set of several hundred files from the GovDocs corpus, in clockwise order from top left:

- Entity types
- Persons
- Organizations
- Geopolitical entities



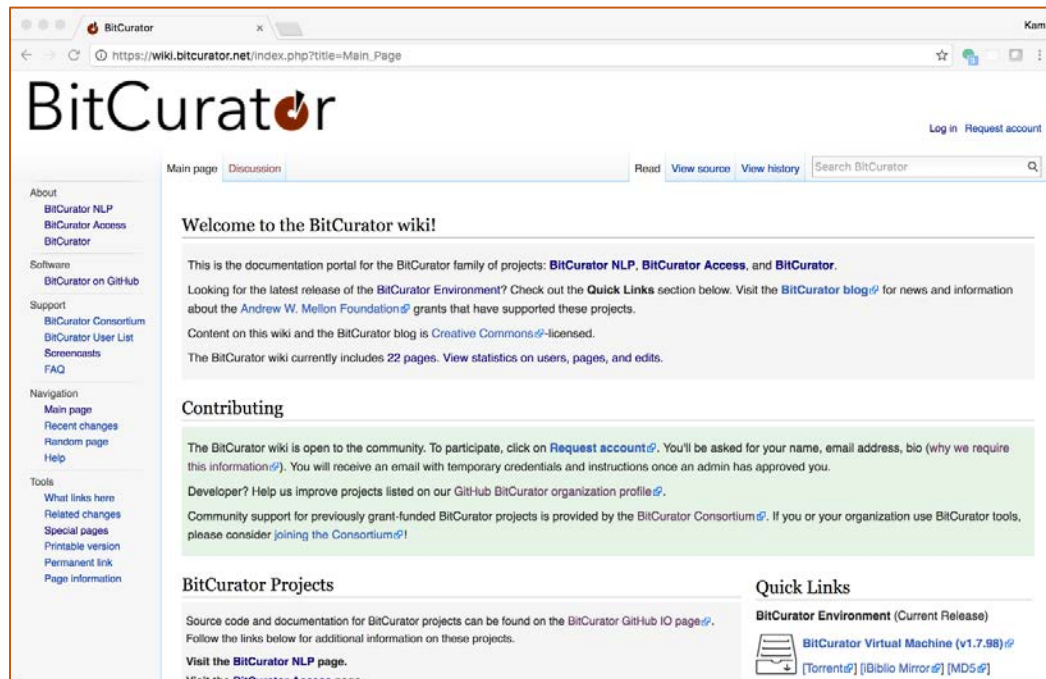
Defaults may be noisy / inaccurate



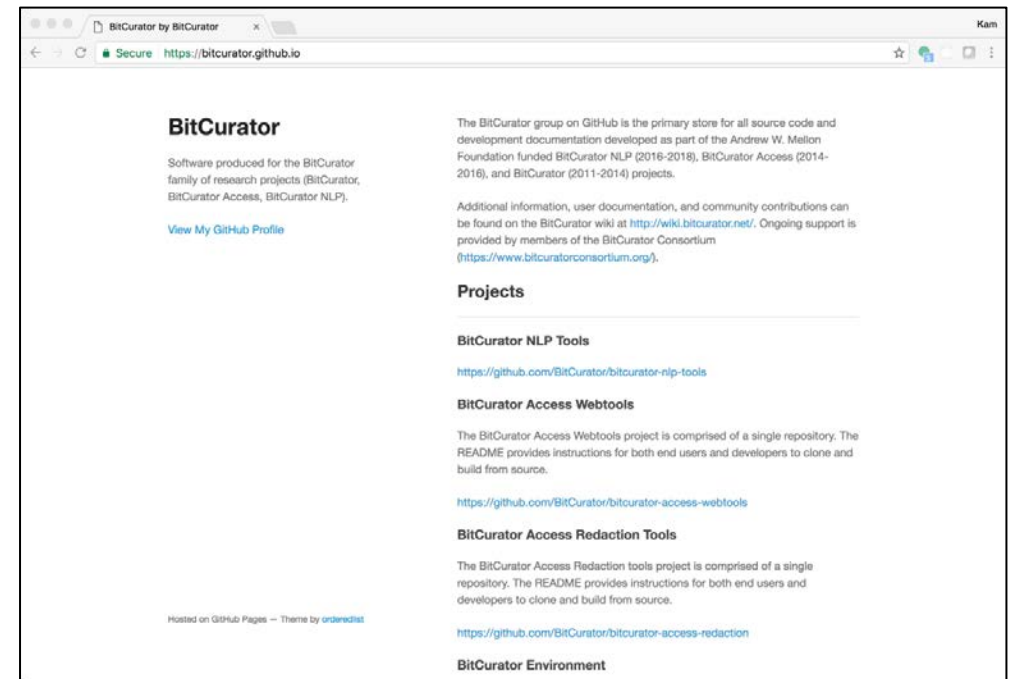
Development and Infrastructure Notes

- BitCurator team keeps in-development software on GitHub
 - <https://bitcurator.github.io>
 - <https://github.com/bitcurator/bitcurator-nlp-tools>
- Development and project documentation posted to wiki
 - <https://wiki.bitcurator.net/>
- In-house development servers:
 - azalea.ils.unc.edu (large)
 - dogwood.ils.unc.edu (small)
- We often have **publicly-available deployments** of the tools available on at least one machine...

Questions?



<https://wiki.bitcurator.net/>



<https://bitcurator.github.io/>