Information Extraction for Legal Documents

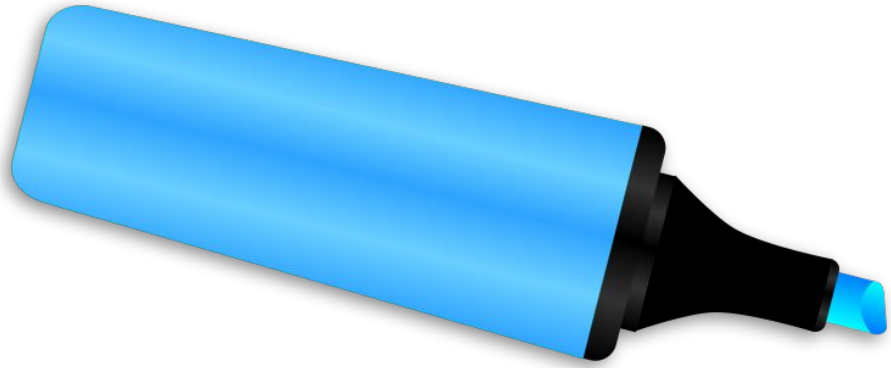
Halle Ritter

Data Source

- EPA CAFOs and ESAs
- Scraped **~12,000** OCR'd PDFs, **~11,500** converted to text strings
- Each 1 doc -> ~8 "pghs" = **~89,000** pghs

Why Extract Information?

- Docs are ~80+% boilerplate
- Time-consuming to scan by hand



UNITED STATES
ENVIRONMENTAL PROTECTION AGENCY
ATLANTA, GEORGIA

In the Matter of:

Nisus Corporation

Respondent.

Docket No.: FIFRA-04-2011-3022(b)

DOC TYPE

CONSENT AGREEMENT AND FINAL ORDER

I. Nature of the Action LAW

This is a civil penalty proceeding pursuant to Section 14(a) of the Federal Insecticide, Fungicide, and Rodenticide Act, as amended, 7 U.S.C. § 136*l*(a) (FIFRA), and pursuant to the Consolidated Rules of Practice Governing Administrative Assessment of Civil Penalties and the Revocation/Termination or Suspension of Permits (Consolidated Rules), 40 C.F.R. Part 22. Complainant is the Director of the Air, Pesticides, and Toxics Management Division, United States Environmental Protection Agency, Region 4 (EPA). Respondent is Nisus Corporation. PARTY 2

Complainant and Respondent have conferred for the purpose of settlement pursuant to 40 C.F.R. § 22.18 and desire to resolve this matter and settle the allegations described herein without a formal hearing. Therefore, without the taking of any evidence or testimony, the making of any argument, or the adjudication of any issue in this matter, and in accordance with 40 C.F.R. § 22.13(b), this Consent Agreement and Final Order (CAFO) will simultaneously commence and conclude this matter.

UNITED STATES ENVIRONMENTAL PROTECTION AGENCY
Region 6, P.O. Box 50625, Dallas, Texas 75250-0625

DOC TYPE

EXPEDITED SETTLEMENT AGREEMENT

FILED
2011 MAR -7 AM 10:49
EPA REGION VI

Docket Number: CWA-06-2010-1879, NPDES Facility Number: NMU001651

PARTY

New Mexico Department of Cultural Affairs ("Respondent") is a "person," within the meaning of Section 502(5) of the Clean Water Act ("Act"),33 U.S.C. § 1362(5), and 40 C.F.R. § 122.2.

LAW VIOLATED

Attached is an Expedited Settlement Offer Deficiencies Form ("Form"), which is incorporated by reference. By its signature, the Environmental Protection Agency, Region 6 ("EPA") finds that Respondent is responsible for the deficiencies specified in the Form.

ACT

Respondent's activities caused or resulted in the unauthorized discharge of pollutants carried by storm water during rainfall events that occurred during the months of July, September, October, 2009, and January 2010, into the Santa Fe River, in violation of Section 301(a) of the Act, 33 U.S.C. § 1311. Respondent failed to submit a Notice of Intent to obtain required permit coverage for activities conducted at the Center for New Mexico Archaeology construction site, located on Caja Del Rio Road, BLM Lot #23, Santa Fe, Santa Fe County, New Mexico. Respondent also failed to obtain coverage under a National Pollutant Discharge Elimination System ("NPDES") permit at the relevant times for the relevant activities.

EPA finds, and Respondent admits, that Respondent is subject to Section 301(a) of the Act, 33 U.S.C. § 1311, and that EPA has jurisdiction over any person who discharges pollutants from a "point source" to "waters of the United States." Respondent neither admits nor denies the deficiencies specified in the Form.

PENALTY

EPA is authorized to enter into this Expedited Settlement Agreement ("ESA") under the authority vested in the Administrator of EPA by Section 309(g)(2)(A) of the Act, 33 U.S.C. § 1319(g)(2)(A), and by 40 C.F.R. § 22.13(b). The parties enter into this ESA in order to settle the civil violation(s) alleged in the ESA for a penalty of two thousand dollars ($2,000). Respondent consents to the assessment of this penalty, and waives the right to: 1) contest the finding(s) specified in the Form; 2) a hearing pursuant to Section 309(g)(2) of the Act, 33 U.S.C. § 22.31(b). Within thirty (30) days of filing this ESA, Respondent shall submit via certified mail, a bank, cashiers, or certified check, with case name and docket number noted, for the amount specified above payable to the Treasurer, United States of America, to:

TIME FRAME

U.S. Environmental Protection Agency
Fines and Penalties
Cincinnati Finance Center
P.O. Box 979077
St. Louis, MO 63197-9000

This ESA settles and resolves EPA's civil penalty clai against Respondent for violations of the Act alleged in thi Agreement. EPA does not waive its right to take enforcemer action against Respondent for any other past, present, or futu civil or criminal violation of the Act, or for any other violation federal statute or regulation. EPA does not waive its right to iss a compliance order for any uncorrected deficiencies or violati described in the Form. EPA has determined this ESA to appropriate.

This ESA is binding on the parties signing below effective upon filing.

APPROVED BY EPA:

Debra A. Griffin Date: Feb 23, 2011

John Blevins
Director
Compliance Assurance and
Enforcement Division

APPROVED BY RESPONDENT:

Name (print): STUART A. ASHMAN

Title (print): CABINET SECRETARY

Signature: Date: 2/24/10

Information to Extract

Trivial Keyword Search

WHAT

- Doc Type
- Parties
  - "Respondent"
  - "Complainant"
- Penalty: "$"
- Date? "Days", "date"

HOW

Trivial find/regex exercise

Entity Extraction

WHAT

- Law
- Parties?
- Doc type?

HOW

spaCy

Autosummarization or ???

WHAT

- Act allegedly committed to violate law
- Other features

HOW

gensim summarization/keywords

???

Entity Extraction

In [96]:
```
for ent in train_smol1_spacy.ents:
    if ent.label_== 'LAW':
        print(ent, ent.label_)
```

```
Section 31 1(b)(3) of the LAW
Section 3 1 LAW
Section 311(a)(7 LAW
the Clean Water Act LAW
Section 31 LAW
Section 3 1 l(a)(l) of the Act LAW
Section 31 LAW
Section 502(7 LAW
Section 31 1(b)(3) of the Act LAW
Section 31 LAW
Region 4 6 1  Forsyth St. Atlanta LAW
```

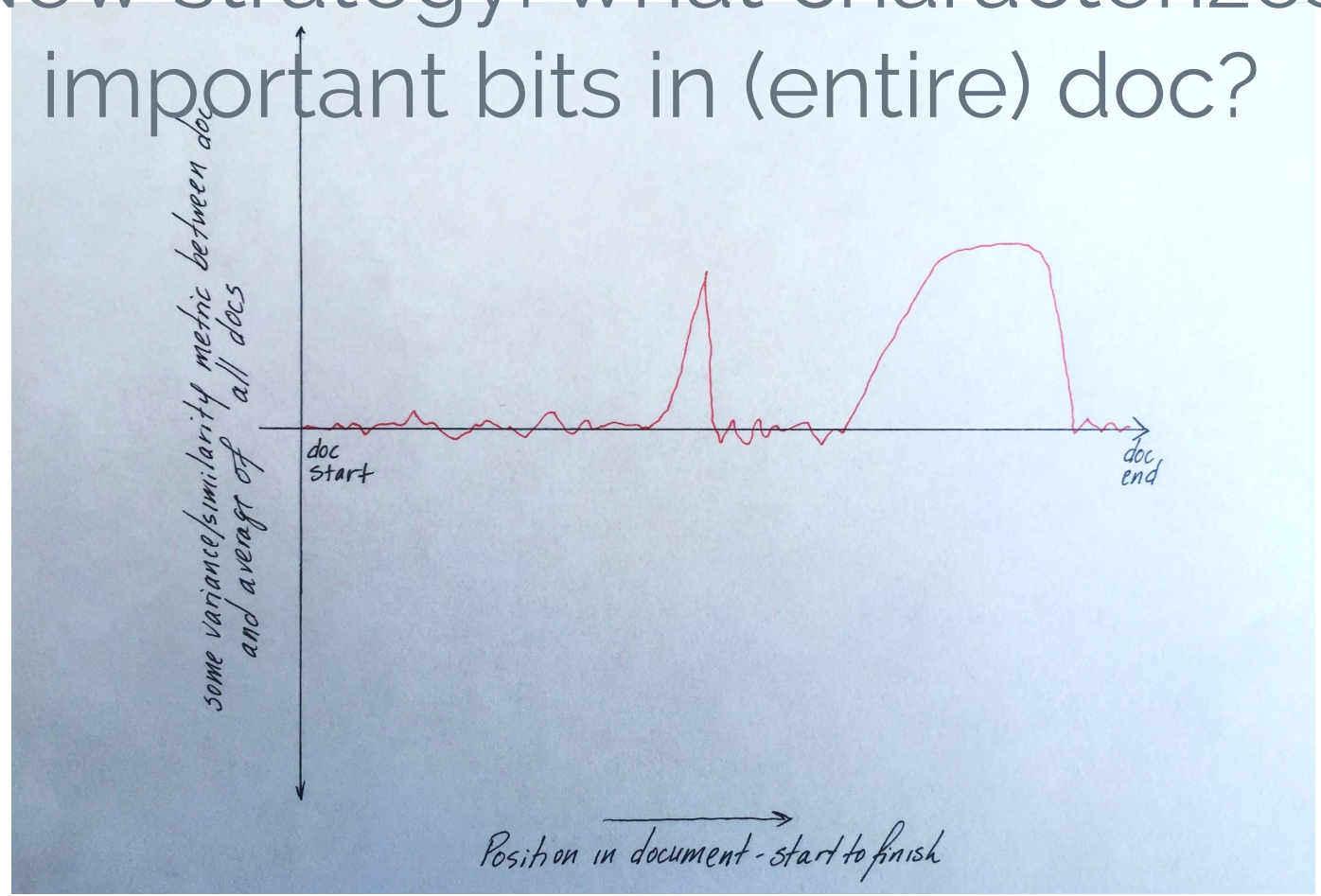
Autosummarization- gensim

```
In [42]: print(summarize(train_smol1, ratio=0.05))
```

```
UNITED STATES ENVIRONMENTAL PROTECTION AGENCY  REGION 4 ATLANTA FEDERAL CENTER 61 FORSYTH STREET ATLANTA, GEORGIA  30
303-8960 CERTIFIED MAIL RETURN RECEIPT REQUESTED Ms. Birdie  Simms, Office Manager Kelcas Petroleum Production  & Exp
loration P.O. Box 21345 Owensboro, KY 2304 SUBJ: Consent Agreement  and Final  Order: Docket No. CWA-04-2007-5002 Dea
r Ms. Simms: Enclosed is a copy  of the Consent Agreement and Final  Order (CAFO)  for the  above referenced matter.
." * ro On: July  8,2006 Time:18:30 ~n At: or near the intersection of KY Hy 1554 and Street Upon signing and returni
ng this  Expedited Spill  Settlement Road in Daviess County, Kentucky,  Kelcas Petroleum  Agreement to EPA, Responden
t  waives the opportunity for Production & Exploration (Respondent)  discharged 160 a hearing or appeal pursuant  to
 Section 3  11 of the Act, and  gallons of  oil in violation of  Section 31 1(b)(3) of the Clean consents  to EPA's a
pproval of the Expedited  Settlement Water Act (the  Act), as noted on the attached ALLEGED without further notice.
Docket No. CWA-04-2007-5002 CERTIFICATE OF  SERVICE The undersigned certifies that a  true and correct copy  of the a
ttached Consent Agreement and Final Order, in the Matter  of Kelcas Petroleum Production  & Exploration, Docket  No.
 CWA-04-2007-5002 (filed with the Regional  Hearing Clerk  on & 2 1 7flfl72007) was served on 1.m 2 1 m07 in the mann
er specified  to each of the persons set  forth below: Birdie Simms,  Office Manager Via Certified  Mail, Kelcas Petr
oleum  Production & Exploration Return Requested  P.O. Box 21345 Owensboro, KY 42304 Victor Weeks, Risk Management Pl
an  Coordinator Via EPA's Internal Mail  EPCRA Enforcement Section U.S. EPA, Region 4 6 1  Forsyth St. Atlanta, GA 30
303 Me1 Rechtman RCRA OPA  Enforcement & Compliance Branch U.S. EPA - Region 4 61 Forsyth Street Atlanta, GA 30303 Da
te: 347-07 Via EPA' s Internal Mail Patricia A.
```

(This strategy not very helpful)

New strategy: what characterizes important bits in (entire) doc?

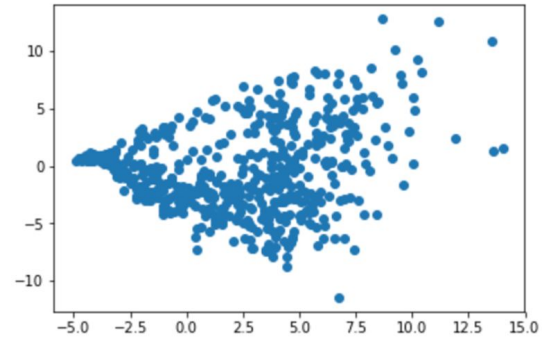


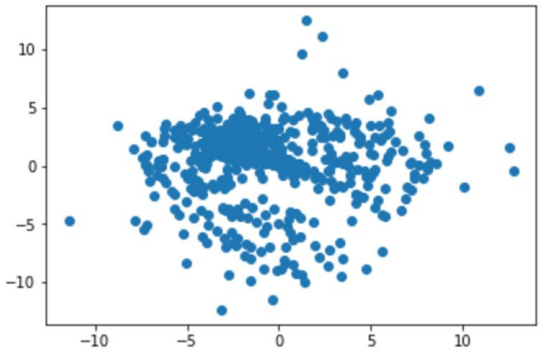
Procedure for doc “outlier detection”

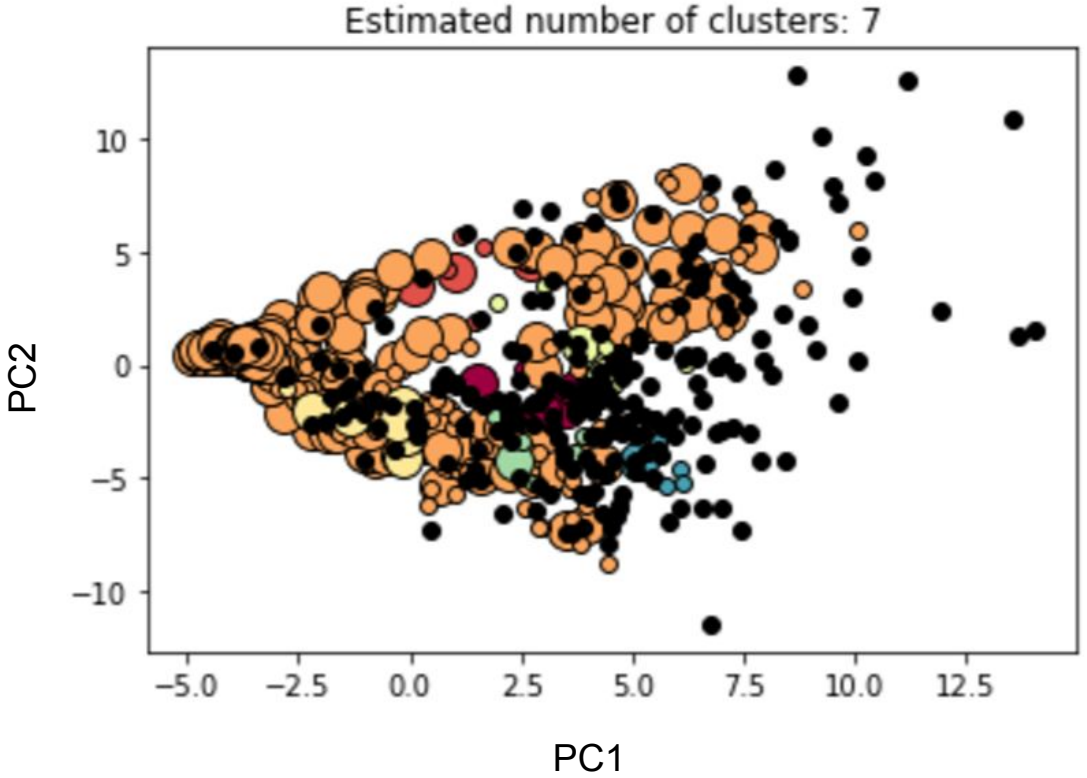
Preliminary Results

Small Scale - CV, PCA, DBScan

Full Scale- TFIDF, LSI, DBScan (in progress)







Future Directions

- Local outlier factor (LOF) algorithm
- Interactive visualization: mapping to individual doc “pghs”
- Closer to continuous rather than pghs/chunks
- Better PDF to text