PLSC 476: Empirical Legal Studies

Christopher Zorn

April 27, 2021

Analyzing Text: Goals

Humans:

- Good at: Meaning, subtlety (irony, sarcasm, subtle negation, etc.), context, tone, etc.
- Bad at: Doing things quickly and consistently.

Computers:

- Good at: Doing things quickly and consistently.
- Bad at: Meaning, subtlety (irony, sarcasm, subtle negation, etc.), context, tone, etc.

Key: Use computers / humans for what they're good at...

Example: SEC "Litigation Releases"

- Short (2-4 paragraph) summaries of civil litigation matters involving the Securities and Exchange Commission (SEC)
- Issues include securities (and other) fraud, insider trading, illegal stock sales, etc.
- Summaries of settled suits, litigation / trial outcomes, etc. plus charges filed
- Available from 1995-2021; we'll focus on 2019...
- Links are available here

sec.gov/litigation/litreleases/2019/lr24640.htm

Search SEC Documents Go

Company Filings More Search Options

U.S. SECURITIES AND EXCHANGE COMMISSION

ABOUT DIVISIONS & OFFICES ENFORCEMENT REGULATION EDUCATION FILINGS NEWS

ENFORCEMENT

- Accounting and Auditing Enforcement Releases
- Administrative Proceedings
- ALJ Initial Decisions
- ALJ Orders
- Amicus / Friend of the Court Briefs
- Delinquent Filings
- Fair Funds
- Information for Harmed Investors
- Litigation Releases
- Opinions and Adjudicatory Orders
- Receiverships
- Stop Orders
- Trading Suspensions

SEC Obtains Sanctions Against Investment Adviser

Litigation Release No. 24640 / October 10, 2019

Securities and Exchange Commission v. Thomas Conrad, Jr. et al., No. 1:16-cv-2572-LMM (N.D. Ga.) (filed July 15, 2016)

On September 30, 2019, the United States District Court for the Northern District of Georgia entered a final judgment against Thomas Conrad, Jr. ("Conrad") of Alpharetta, Georgia. The SEC's complaint alleged that between 2010 and late 2014, Conrad directed preferential redemptions and other disbursements out of a hedge fund and its feeder funds operated by firms he controlled. According to the complaint, Conrad directed disbursements to himself, his extended family, and certain favored investors, while representing to other investors that redemptions were suspended. The complaint also alleged that Conrad failed to disclose conflicts of interest arising from loans the funds made to Conrad's family members, and from Conrad's appointment of himself as a sub-manager, for which he received a fee. As alleged, offering documents the firms gave to prospective investors touted Conrad's significant experience in the securities industry, but failed to disclose his disciplinary history, which included an industry bar that the SEC imposed on him in 1971.

The Court previously ruled that the SEC was entitled to summary judgment against Conrad and the two advisory firms he controlled on its fraud claims based on the defendants' fraudulent redemption practices and their failure to disclose Conrad's disciplinary history. After the Commission dismissed its remaining claims, the Court entered a final judgment against Conrad enjoining Conrad from future violations of Section 17(a) of the Securities Act of 1933, Section 10(b) of the Exchange Act of 1934 view-source:https://www.sec.gov/litigation/litreleases/2019/lr24640.htm

Update

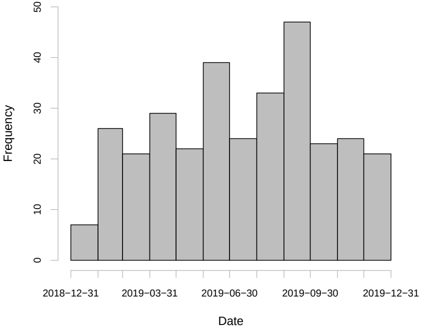
```
</div>
  <div id="main-content" class="grid_10 push_2">

    <!-- title spans main content area + right column> -->

    <h1 class="alphaheads">SEC Obtains Sanctions Against Investment Adviser</h1>

    <h2 class="alphaheads">Litigation Release No. 24640 / October 10, 2019</h2>

    <h2 class="alphaheads"><i>Securities and Exchange Commission v. Thomas Conrad, Jr. et al.</i>, No. 1:16-cv-2572-LMM (N.D. Ga.) (fi


    <div class="grid_7 alpha">

      <!-- MAIN CONTENT -->

      <!-- Text Goes here -->

      <p>On September 30, 2019, the United States District Court for the Northern District of Georgia entered a final judgment against

      <p>The Court previously <a href="/litigation/litreleases/2019/lr24390.htm">ruled</a> that the SEC was entitled to summary judgme

      <p>The SEC is represented by M. Graham Loomis, William P. Hicks, and Kristin W. Murnahan of the Atlanta Regional Office in this

<p>See also:  <a href="/litigation/litreleases/2019/lr24390.htm">Litigation Release No. 24390</a> and <a href="/litigation/litreleases

      <!-- End text -->
    </div>

      <!--Complaint-->

      <!--<div class="grid_3 omega">
        <ul class="bullet-1">
          <li><a href="/litigation/complaints/20xx/compxxxxx.pdf">SEC Complaint</a></li>
        </ul>
      </div>
-->

  </div>
```

Step One: Get Some Data

```
> url<-"https://www.sec.gov/litigation/litreleases/litrelarchive/litarchive2019.shtml"
> pg<-read_html(url)
> linx<-html_attr(html_nodes(pg, "a"), "href")
> mylinx<-linx[grep("/litigation/litreleases/2019/",linx)]
> mylinx<-mylinx[grep(".htm",mylinx)]
> mylinx
  [1] "/litigation/litreleases/2019/lr24701.htm"
  [2] "/litigation/litreleases/2019/lr24700.htm"
  [3] "/litigation/litreleases/2019/lr24699.htm"
  [4] "/litigation/litreleases/2019/lr24698.htm"
     .
     .
     .
[319] "/litigation/litreleases/2019/lr24383.htm"
[320] "/litigation/litreleases/2019/lr24382.htm"
[321] "/litigation/litreleases/2019/lr24381.htm"

> N<-length(mylinx)
> dir.create("2019") # make a folder for the files...
>
> for(i in 1:N){
>  url<-paste0("http://sec.gov",mylinx[i])
>  dest<-paste0("2019/SEC",i,".htm")
>  download.file(url,dest)
> }
```

(After some processing...)

```
> summary(SEC.df)

    doc_id           text               Title                  LRN             Date
 Min.   :  1   Length:321          Length:321          Min.   :24381   Min.   :2019-01-17
 1st Qu.: 81   Class :character    Class :character    1st Qu.:24461   1st Qu.:2019-04-25
 Median :161   Mode  :character    Mode  :character    Median :24540   Median :2019-07-18
 Mean   :161                                           Mean   :24541   Mean   :2019-07-15
 3rd Qu.:241                                           3rd Qu.:24621   3rd Qu.:2019-09-27
 Max.   :321                                           Max.   :24701   Max.   :2019-12-30
                                                       NA's   :5       NA's   :5
```



$\rightarrow$ Corpus $\rightarrow$ DTM

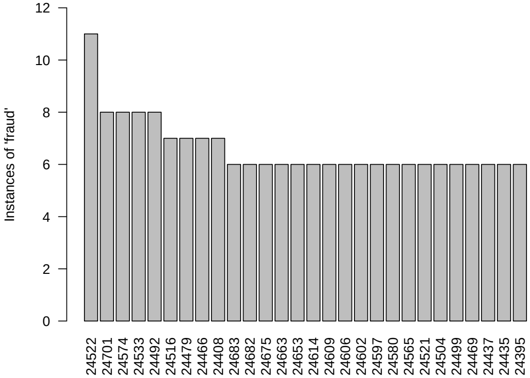
```
> SEC<-VCorpus(DataframeSource(SEC.df))

> SEC

<<VCorpus>>
Metadata:  corpus specific: 0, document level (indexed): 3
Content:  documents: 321

> SEC.DTM<-DocumentTermMatrix(SEC,
+               control=list(removePunctuation=TRUE,
+                            tolower=TRUE,
+                            stopwords=TRUE,
+                            removeNumbers=TRUE,
+                            stemming=TRUE))

> rownames(SEC.DTM)<-SEC.df$LRN # document IDs...

> SEC.DTM

<<DocumentTermMatrix (documents: 321, terms: 5200)>>
Non-/sparse entries: 42443/1626757
Sparsity           : 97%
Maximal term length: 61
Weighting          : term frequency (tf)
```

"Search"-Like Things: Fraud

```
> frauds<-SEC.DTM[,grepl("fraud",SEC.DTM$dimnames$Terms)]

> frauds
<<DocumentTermMatrix (documents: 321, terms: 10)>>
Non-/sparse entries: 605/2605
Sparsity           : 81%
Maximal term length: 18
Weighting          : term frequency (tf)

> inspect(frauds)
<<DocumentTermMatrix (documents: 321, terms: 10)>>
Non-/sparse entries: 605/2605
Sparsity           : 81%
Maximal term length: 18
Weighting          : term frequency (tf)
Sample             :
       Terms
Docs    antifraud defraud fraud fraud  fraudlitig fraudster fraudul fraudulentlyalt fraudulentlyobtain securitiesfraud
  24408         1       1     4      0          0         0       1               0                  0               0
  24466         2       0     5      0          0         0       0               0                  0               0
  24479         2       1     0      0          0         0       4               0                  0               0
  24492         1       0     4      0          0         0       3               0                  0               0
  24516         3       0     1      0          0         0       3               0                  0               0
  24522         8       1     2      0          0         0       0               0                  0               0
  24533         1       2     4      1          0         0       0               0                  0               0
  24574         1       1     4      0          0         0       1               1                  0               0
  24683         1       0     5      0          0         0       0               0                  0               0
  24701         2       0     2      0          0         0       4               0                  0               0
```

More Fraud...

```
> table(row_sums(frauds))

 0  1  2  3  4  5  6  7  8 11
33 59 53 74 40 34 19  4  4  1

# Pull the documents with the greatest incidence of "fraud"...
```



Weighting (TF v. TF-IDF)

A standard *document-term matrix* has:

- $i \in 1 \ldots N$ rows, corresponding to the N documents D in the corpus
- $j \in 1 \ldots J$ columns, corresponding to the J unique terms T in the corpus
- Cell entries N_{ij} that represent the number of times term j appears in document i

Term frequency:

$$N_{ij} = \text{The number of times term } j \text{ appears in document } i$$

Term frequency (normalized for document length):

$$TF_{ij} = \frac{N_{ij}}{\sum_{i=1}^{J} N_{ij}},$$

the fraction of all terms in document D_i that are term T_j.

Inverse document frequency (normalized):

$$IDF_j = \log_2 \frac{J}{J_j}$$

where J_j is the number of documents in which T_j occurs.

TF-IDF$_{ij}$ is then simply $\mathbf{TF_{ij}} \times \mathbf{IDF_j}$

TF-IDF: Examples

Three "documents":

$$
\begin{aligned}
A &= \{\text{red}, \text{blue}, \text{red}\} \\
B &= \{\text{green}, \text{blue}, \text{orange}\} \\
C &= \{\text{yellow}, \text{blue}, \text{yellow}\}
\end{aligned}
$$

Example one:

- In document A "red" appears twice ($TF_{ij} = 2$), and
- "red" is two of the three total terms in that document (normed $TF_{ij} = 2/3 = 0.67$)
- "red" appears in only one of the three documents ($IDF_i = log_2[3/1] = 1.6$)
- The TF-IDF for "red" in document A is $0.67 \times 1.6 = 1.1$

Example two:

- In document C "blue" appears once ($TF_{ij} = 1$), and
- "blue" is one of the three total terms in that document (normed $TF_{ij} = 1/3 = 0.33$)
- "blue" appears in all three documents ($IDF_i = log_2[3/3] = 0$)
- The TF-IDF for "blue" in document C is $0.33 \times 0 = 0$

TF-IDF: Intuition

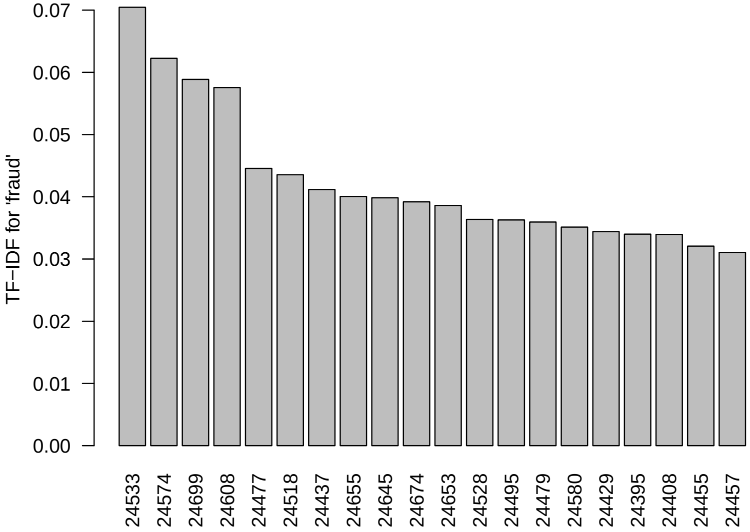
In general:

- (Normalized) TF indicates the prevalence of a term in a document
- IDF reflects how common or rare the word is across documents
- IDF is thus a measure of the level of "informativeness" (or "document-specificity") of a word
- TF-IDF is thus a measure of a term's **"importance"** (in some respects)
  - TF-IDF is zero if a term does not appear in a document at all
  - The TF-IDF is also always zero for any term that appears in *all* documents in a corpus
  - Values of TF-IDF are generally < 1, but need not be
  - Higher values of TF-IDF indicate more important / distinctive terms for that document

Making a TF-IDF Document-Term Matrix

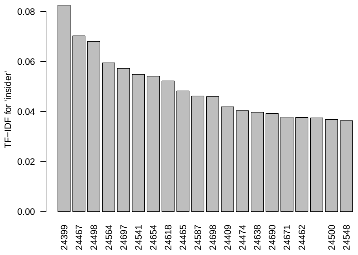
```
> SEC.TFIDF<-weightTfIdf(SEC.DTM) #TF-IDF weighting

> SEC.TFIDF
<<DocumentTermMatrix (documents: 321, terms: 5200)>>
Non-/sparse entries: 41801/1627399
Sparsity           : 97%
Maximal term length: 61
Weighting          : term frequency - inverse document frequency (normalized) (tf-idf)

> inspect(SEC.TFIDF)
<<DocumentTermMatrix (documents: 321, terms: 5200)>>
Non-/sparse entries: 41801/1627399
Sparsity           : 97%
Maximal term length: 61
Weighting          : term frequency - inverse document frequency (normalized) (tf-idf)
Sample             :
       Terms
Docs          advis client       final        fund      invest    investor    judgment        llc       stock      trade
  24381 0.000000000      0 0.003465235 0.000000000 0.000000000 0.000000000 0.003023097 0.00000000 0.000000000 0.02819580
  24384 0.000000000      0 0.000000000 0.000000000 0.012131326 0.009147762 0.007773678 0.00000000 0.000000000 0.00000000
  24410 0.000000000      0 0.000000000 0.010060054 0.000000000 0.000000000 0.000000000 0.00000000 0.000000000 0.00000000
  24413 0.000000000      0 0.022132793 0.000000000 0.000000000 0.000000000 0.019308814 0.00000000 0.000000000 0.00000000
  24473 0.000000000      0 0.000000000 0.014042774 0.013696658 0.020656237 0.000000000 0.00978427 0.000000000 0.00000000
  24482 0.000000000      0 0.000000000 0.006763539 0.000000000 0.019897674 0.000000000 0.00000000 0.000000000 0.00000000
  24494 0.011676977      0 0.000000000 0.000000000 0.000000000 0.007914356 0.000000000 0.00000000 0.000000000 0.00000000
  24592 0.000000000      0 0.025040751 0.000000000 0.003409168 0.012853607 0.030584034 0.00000000 0.000000000 0.00000000
  24621 0.000000000      0 0.005008150 0.000000000 0.000000000 0.000000000 0.004369148 0.00000000 0.000000000 0.00000000
  24693 0.008056209      0 0.000000000 0.000000000 0.000000000 0.000000000 0.000000000 0.00000000 0.006302286 0.04868694
```

Top “Fraud” Documents (by TF-IDF)

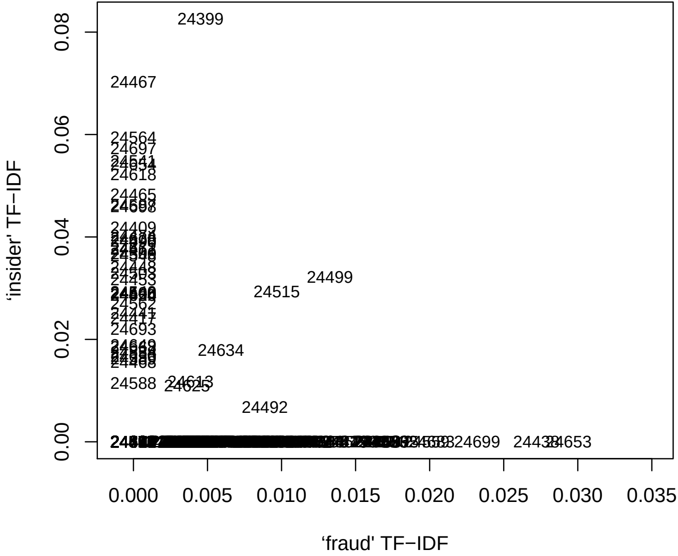


Another Term: “insider”

```
> IT<-SEC.TFIDF[,grepl("insid",SEC.DTM$dimnames$Terms)]

> inspect(IT)
<<DocumentTermMatrix (documents: 321, terms: 2)>>
Non-/sparse entries: 46/596
Sparsity           : 93%
Maximal term length: 11
Weighting          : term frequency - inverse document frequency (normalized) (tf-idf)
Sample             :
       Terms
Docs         insid insidertrad
  24399 0.08256048  0.00000000
  24465 0.04824811  0.00000000
  24467 0.07027875  0.00000000
  24498 0.03923289  0.02881117
  24541 0.05486277  0.00000000
  24564 0.05946664  0.00000000
  24587 0.04621592  0.00000000
  24618 0.05225026  0.00000000
  24654 0.05416388  0.00000000
  24697 0.05726417  0.00000000
```



“fraud” + “insider”: Relevant Docs



Moving Beyond...

- More advanced search
- Correlations among terms within documents
- Measurement models (e.g., for combining similar or related terms)
- Visualizing "outlier" observations
- Various natural language processing tools:
    - Entity recognition
    - Sentiment analysis
    - Topic models