Topic Extraction

Richard Careaga

March 9, 2019

Email discovery in litigation

Rule 34 of The Federal Rules of Civil Procedure permits parties to civil litigation to require opposing parties and, in some cases, third parties, to produce evidence or materials likely to lead to the production of evidence that is stored in electronic form. Regulatory and law enforcement agencies can also require or subpoena electronic information, including email. Given the adverse interests, time and cost constraints, and the sheer volume of electronic records, including email, both sides struggle with making the process manageable. The civil courts require parties to engage in a cooperative process (see the Sedona Conference Report, the report of the Advisory Committee on the most recent changes to Rule 34 and the ESI Guidelines and ESI Checklist issued by the U.S. District Court for the Northern District of California).

Among the questions of greatest important is how to separate the relevant content from the large universe of what is available. For that, participants usually turn to search terms – keywords and phrases. (*See* recent article pointing keyword limitations on effective discovery.)

As the author of the keyword limitations article points out, words have synonmyms, private meanings, multiple meanings and meanings that depend critically on context. Less appreciated, perhaps, is that most people fail to distinguish between styles of language. While it's easy to tell the difference between an inaugeral address and a sports color commentator's styles in oral communication and to distinguish a scientific paper from a young adult novel in formal written conversation, email falls in the separate sphere of informal written language.[1]

> Reading email is eavesdropping.

Information sought

Prior to the investigation that resulted in the an initial staff report, FERC received a report from the California agency responsible for operating the electric energy exchange that FERC was concerned that the market had been manipulating.[2], which will be referred to as the [Wolak report].

The [Wolak report] was prepared before FERC began its investigation, did not rely on the Enron email corpus[3] and provides a lengthy analysis of the indicia of the exercise of market power in the California electricity wholesale market. A topical and keyword digest was prepared, as discussed below, to identify subjects of interest in the corpus.

Wolak keywords

A latent Dirichlet allocation model specifying classification into six topics was prepared for the [Wolak Report]. As descibed in the classic paper by David Blei, Andrew Ng and Michael Jordan, *Latent Dirichlet Allocation* (*J. Machine Learning Res.* 3 (2003). 993-1022), this machine learning technique is a generative probabalistic model. Each discrete object in a collectikon (such as words in a document) is "modeled as a finixed mixture over an underlying sete of topics" (*id.*). Topics on the other hand are "modeled as an infinite mixture over an underlying set of topic probabilities." (*id.*) A list of some of the keywords in each modeled topic are

[1] *See* McWhorter.

[2] Frank A. Wolak, Chairman, Market Surveillance Committee of California Independent System Operator, *Report on Redesign of California Real-Time Energy and Ancillary Services Markets* dated October 18, 1999

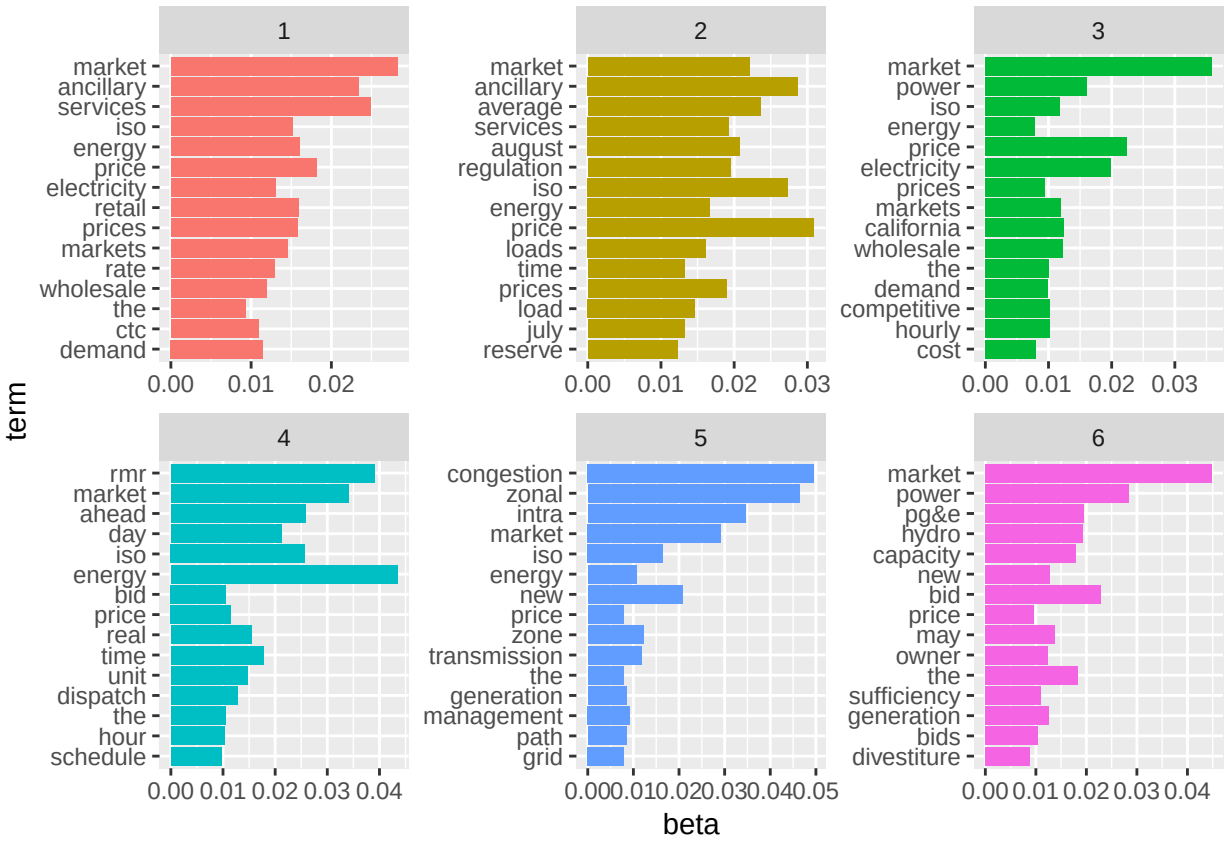
[3] The report does not even name Enron.

```
pander(w_top_terms)
```

topic	term	beta
1	market	0.02821
1	services	0.02491
1	ancillary	0.02335
1	price	0.01811
1	energy	0.01599
1	retail	0.01587
1	prices	0.01578
1	iso	0.01516
1	markets	0.0145
1	electricity	0.01306
1	rate	0.01293
1	wholesale	0.01196
1	demand	0.01149
1	ctc	0.01088
1	the	0.009315
2	price	0.03093
2	ancillary	0.02869
2	iso	0.02735
2	average	0.02361
2	market	0.02211
2	august	0.02082
2	regulation	0.01952
2	services	0.01932
2	prices	0.01892
2	energy	0.01671
2	loads	0.01616
2	load	0.01466
2	july	0.01323
2	time	0.01318
2	reserve	0.01223
3	market	0.03584
3	price	0.02235
3	electricity	0.01986
3	power	0.01611
3	california	0.0124
3	wholesale	0.01232
3	markets	0.01194
3	iso	0.01176
3	competitive	0.01029
3	hourly	0.01021
3	the	0.01012
3	demand	0.009934
3	prices	0.009473
3	cost	0.008033
3	energy	0.007764
4	energy	0.04342
4	rmr	0.03901
4	market	0.03401
4	ahead	0.02584
4	iso	0.02572

topic	term	beta
4	day	0.02117
4	time	0.01777
4	real	0.01548
4	unit	0.01466
4	dispatch	0.01276
4	price	0.01151
4	bid	0.01059
4	the	0.01046
4	hour	0.0103
4	schedule	0.009777
5	congestion	0.04962
5	zonal	0.04637
5	intra	0.03461
5	market	0.02918
5	new	0.02072
5	iso	0.01648
5	zone	0.01226
5	transmission	0.01181
5	energy	0.01065
5	management	0.009227
5	path	0.008538
5	generation	0.008436
5	grid	0.007981
5	price	0.007843
5	the	0.007821
6	market	0.04482
6	power	0.02835
6	bid	0.02293
6	pg&e	0.01944
6	hydro	0.01926
6	the	0.01833
6	capacity	0.01797
6	may	0.01378
6	new	0.0128
6	generation	0.01263
6	owner	0.01235
6	sufficiency	0.01099
6	bids	0.01032
6	price	0.009676
6	divestiture	0.008746

```
w_top_terms                                %>%
  mutate(term = reorder(term, beta))       %>%
  ggplot(aes(term, beta, fill = factor(topic))) +
  geom_col(show.legend = FALSE)                 +
  facet_wrap(~ topic, scales = "free")          +
  coord_flip()
```


1
market
ancillary
services
iso
energy
price
electricity
retail
prices
markets
rate
wholesale
the
ctc
demand
0.00 0.01 0.02
2
market
ancillary
average
services
august
regulation
iso
energy
price
loads
time
prices
load
july
reserve
0.00 0.01 0.02 0.03
3
market
power
iso
energy
price
electricity
prices
markets
california
wholesale
the
demand
competitive
hourly
cost
0.00 0.01 0.02 0.03
4
rmr
market
ahead
day
iso
energy
bid
price
real
time
unit
dispatch
the
hour
schedule
0.00 0.01 0.02 0.03 0.04
5
congestion
zonal
intra
market
iso
energy
new
price
zone
transmission
the
generation
management
path
grid
0.000.010.020.030.040.05
6
market
power
pg&e
hydro
capacity
new
bid
price
may
owner
the
sufficiency
generation
bids
divestiture
0.00 0.01 0.02 0.03 0.04
term
beta

Guidelines