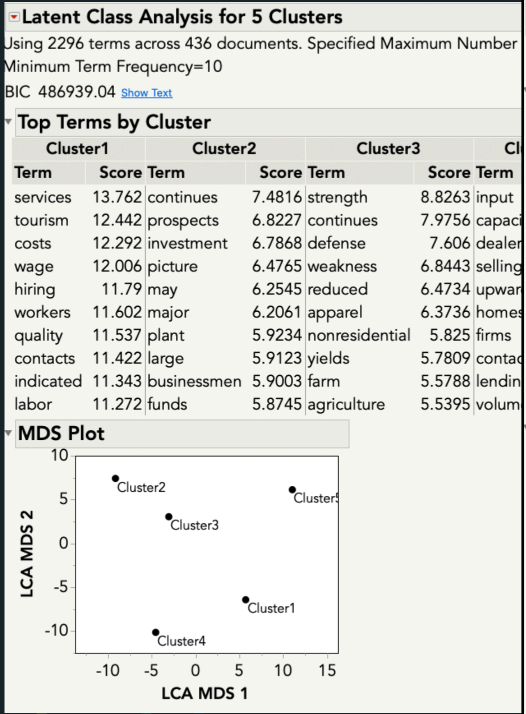
Expanding Our Text Mining Toolkit

Sentiment Analysis and Term Selection in JMP Pro 16

Ross Metusalem, PhD

JMP Systems Engineer

Text Explorer



Latent Class Analysis for 5 Clusters

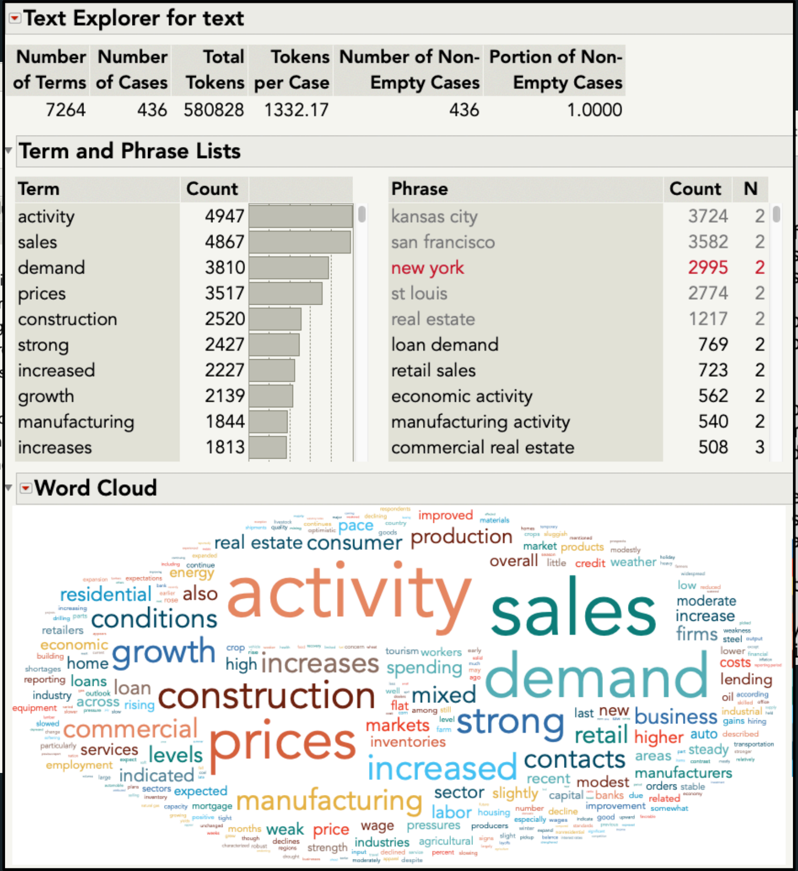
Using 2296 terms across 436 documents. Specified Maximum Number
Minimum Term Frequency=10

BIC 486939.04 Show Text

Top Terms by Cluster

Cluster1		Cluster2		Cluster3		Cl
Term	Score	Term	Score	Term	Score	Term
services	13.762	continues	7.4816	strength	8.8263	input
tourism	12.442	prospects	6.8227	continues	7.9756	capaci
costs	12.292	investment	6.7868	defense	7.606	dealer
wage	12.006	picture	6.4765	weakness	6.8443	selling
hiring	11.79	may	6.2545	reduced	6.4734	upwar
workers	11.602	major	6.2061	apparel	6.3736	homes
quality	11.537	plant	5.9234	nonresidential	5.825	firms
contacts	11.422	large	5.9123	yields	5.7809	contac
indicated	11.343	businessmen	5.9003	farm	5.5788	lendin
labor	11.272	funds	5.8745	agriculture	5.5395	volum

MDS Plot



Text Explorer for text

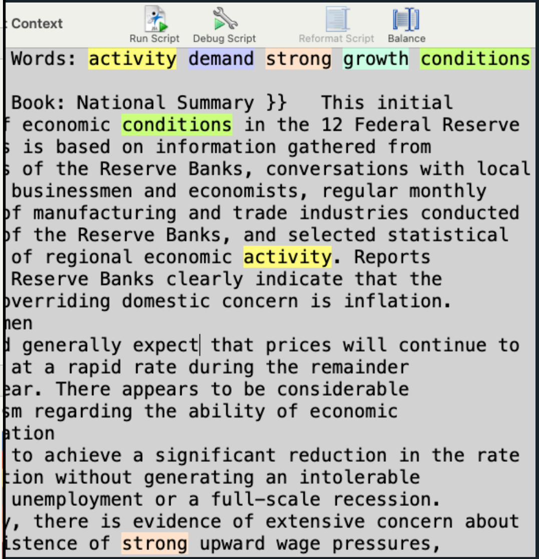
Number of Terms	Number of Cases	Total Tokens	Tokens per Case	Number of Non-Empty Cases	Portion of Non-Empty Cases
7264	436	580828	1332.17	436	1.0000

Term and Phrase Lists

Term	Count
activity	4947
sales	4867
demand	3810
prices	3517
construction	2520
strong	2427
increased	2227
growth	2139
manufacturing	1844
increases	1813

Phrase	Count	N
kansas city	3724	2
san francisco	3582	2
new york	2995	2
st louis	2774	2
real estate	1217	2
loan demand	769	2
retail sales	723	2
economic activity	562	2
manufacturing activity	540	2
commercial real estate	508	3

Word Cloud

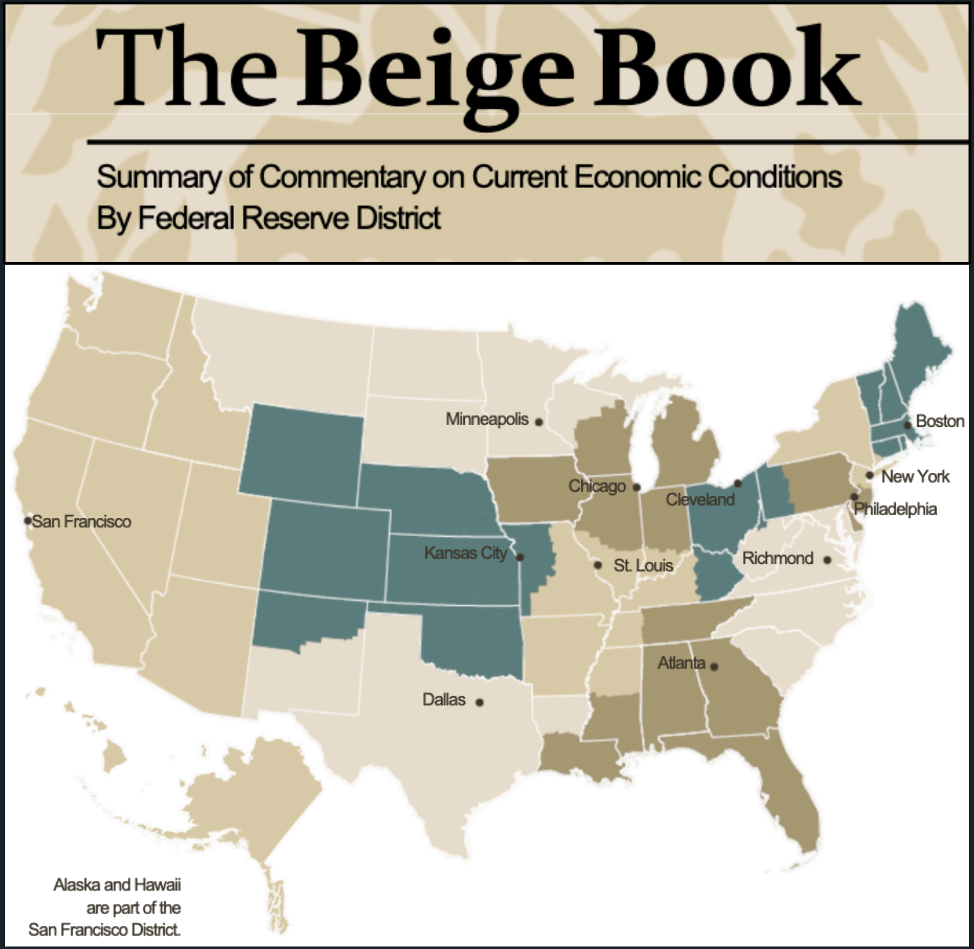


```
Words: activity demand strong growth conditions

Book: National Summary }}   This initial
f economic conditions in the 12 Federal Reserve
s is based on information gathered from
s of the Reserve Banks, conversations with local
businessmen and economists, regular monthly
of manufacturing and trade industries conducted
of the Reserve Banks, and selected statistical
 of regional economic activity. Reports
Reserve Banks clearly indicate that the
overriding domestic concern is inflation.
hen
d generally expect that prices will continue to
 at a rapid rate during the remainder
ear. There appears to be considerable
sm regarding the ability of economic
ation
 to achieve a significant reduction in the rate
tion without generating an intolerable
unemployment or a full-scale recession.
/, there is evidence of extensive concern about
istence of strong upward wage pressures,
```

Sentiment Analysis

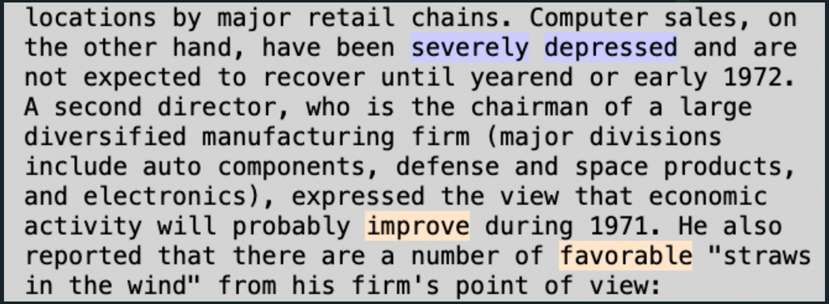
How emotionally positive or negative is a text?



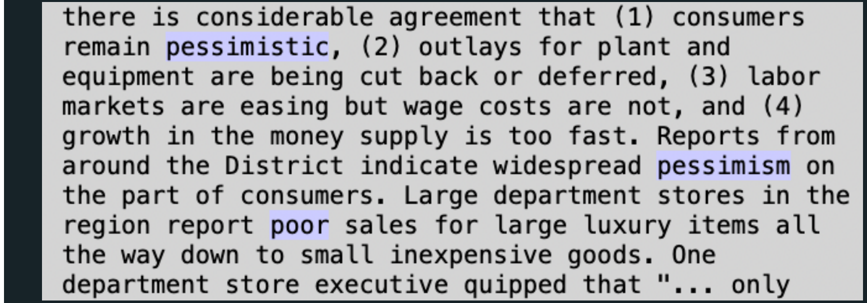
The Beige Book

Summary of Commentary on Current Economic Conditions
By Federal Reserve District

Alaska and Hawaii are part of the San Francisco District.



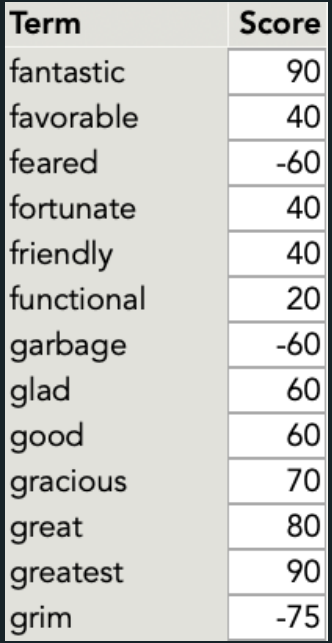
locations by major retail chains. Computer sales, on the other hand, have been severely depressed and are not expected to recover until yearend or early 1972. A second director, who is the chairman of a large diversified manufacturing firm (major divisions include auto components, defense and space products, and electronics), expressed the view that economic activity will probably improve during 1971. He also reported that there are a number of favorable "straws in the wind" from his firm's point of view:



there is considerable agreement that (1) consumers remain pessimistic, (2) outlays for plant and equipment are being cut back or deferred, (3) labor markets are easing but wage costs are not, and (4) growth in the money supply is too fast. Reports from around the District indicate widespread pessimism on the part of consumers. Large department stores in the region report poor sales for large luxury items all the way down to small inexpensive goods. One department store executive quipped that "... only

Sentiment Term Scores

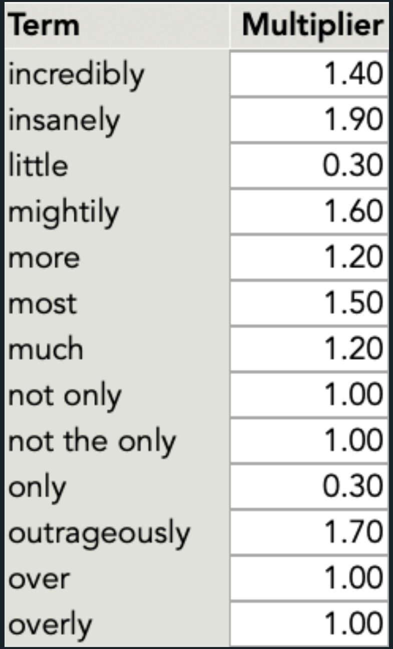
×



Term	Score
fantastic	90
favorable	40
feared	-60
fortunate	40
friendly	40
functional	20
garbage	-60
glad	60
good	60
gracious	70
great	80
greatest	90
grim	-75

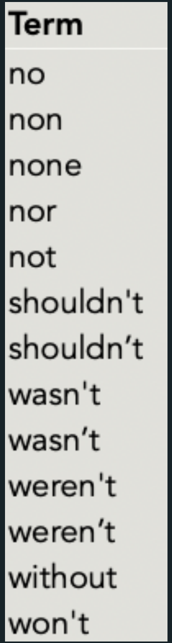
Intensifiers

×

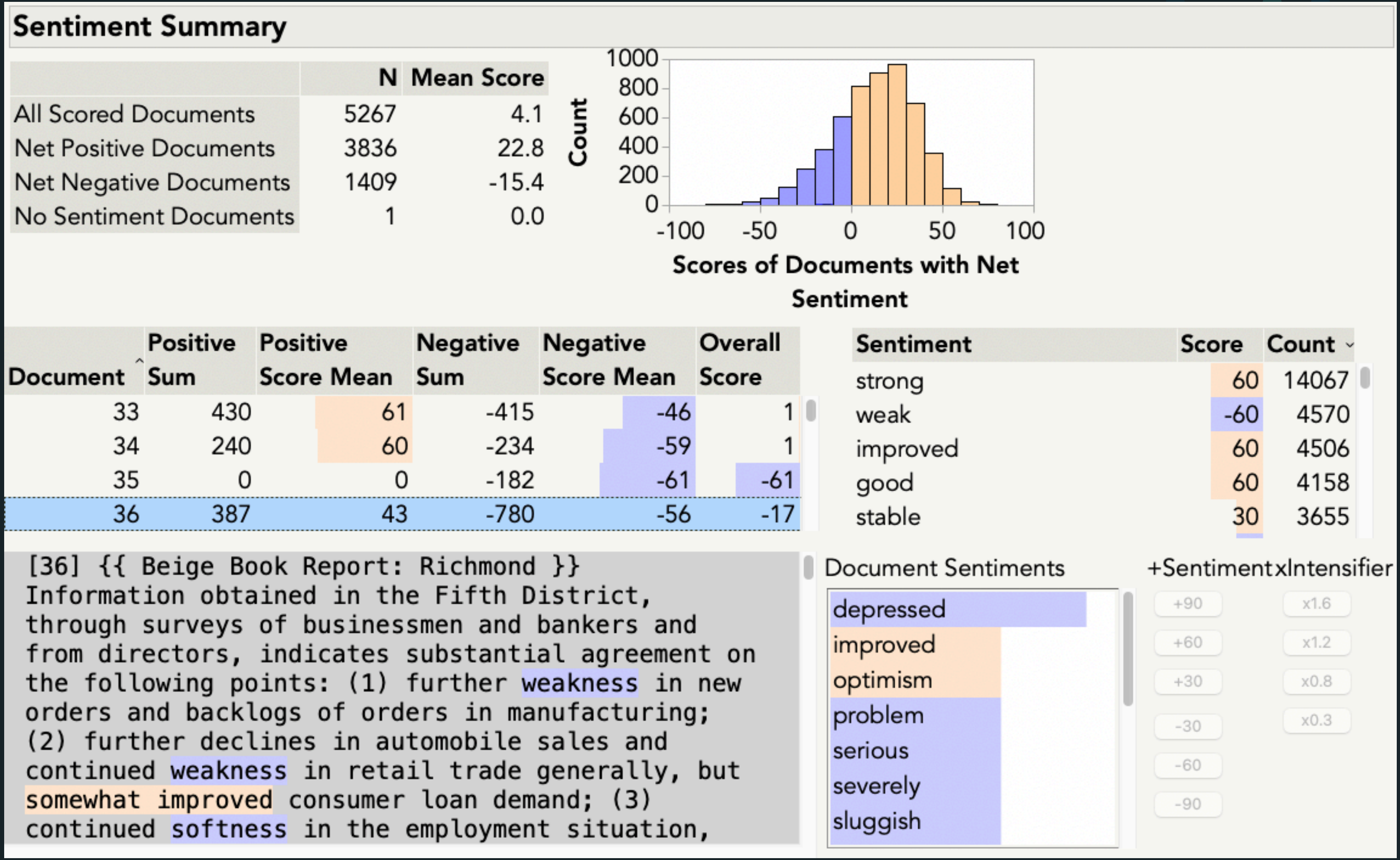


Term	Multiplier
incredibly	1.40
insanely	1.90
little	0.30
mightily	1.60
more	1.20
most	1.50
much	1.20
not only	1.00
not the only	1.00
only	0.30
outrageously	1.70
over	1.00
overly	1.00

Negators (-1)



Term
no
non
none
nor
not
shouldn't
shouldn’t
wasn't
wasn’t
weren't
weren’t
without
won't



Sentiment Summary

	N	Mean Score
All Scored Documents	5267	4.1
Net Positive Documents	3836	22.8
Net Negative Documents	1409	-15.4
No Sentiment Documents	1	0.0

Document	Positive Sum	Positive Score Mean	Negative Sum	Negative Score Mean	Overall Score
33	430	61	-415	-46	1
34	240	60	-234	-59	1
35	0	0	-182	-61	-61
36	387	43	-780	-56	-17

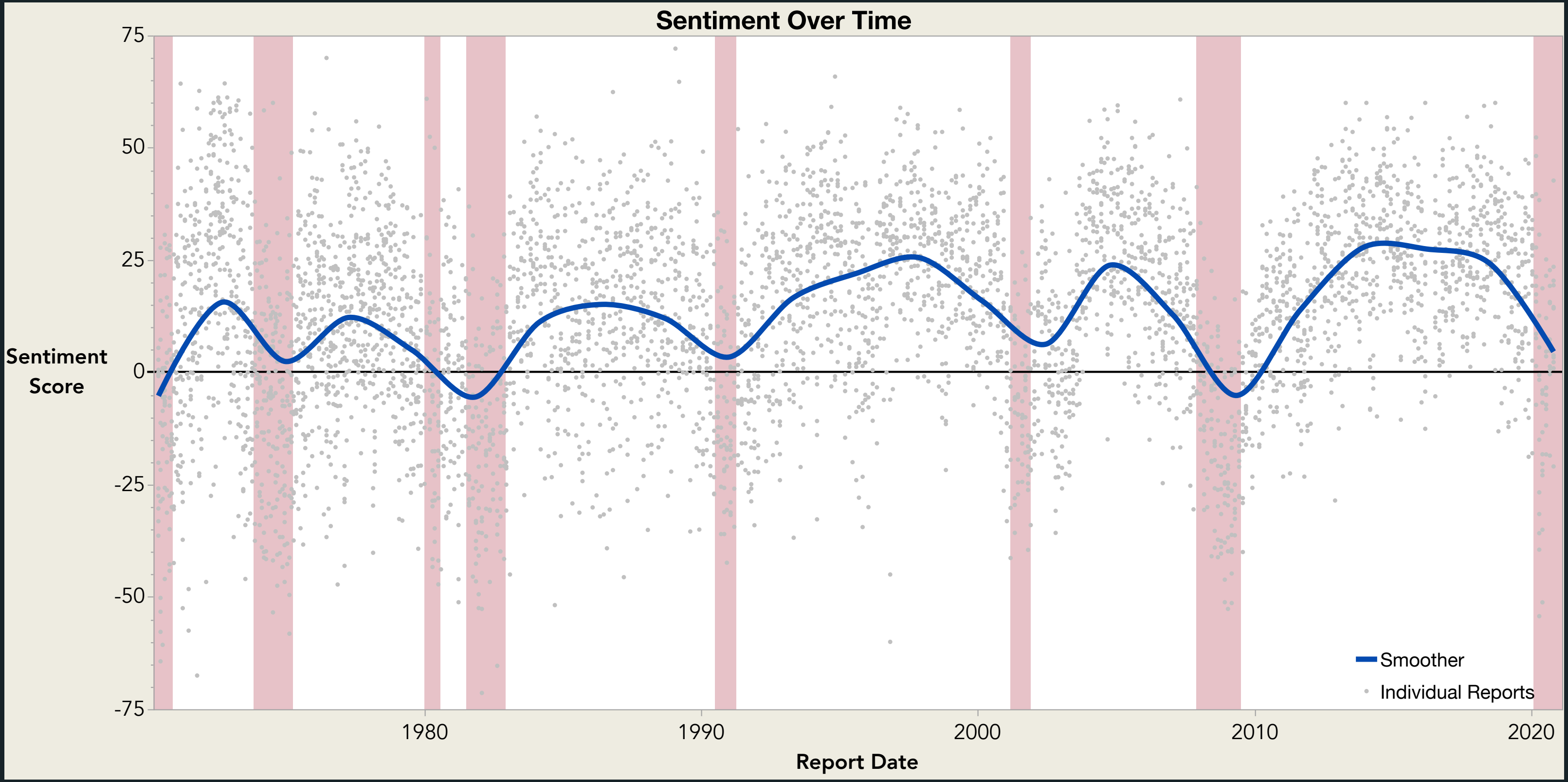
Sentiment	Score	Count
strong	60	14067
weak	-60	4570
improved	60	4506
good	60	4158
stable	30	3655

[36] {{ Beige Book Report: Richmond }}
Information obtained in the Fifth District, through surveys of businessmen and bankers and from directors, indicates substantial agreement on the following points: (1) further weakness in new orders and backlogs of orders in manufacturing; (2) further declines in automobile sales and continued weakness in retail trade generally, but somewhat improved consumer loan demand; (3) continued softness in the employment situation,

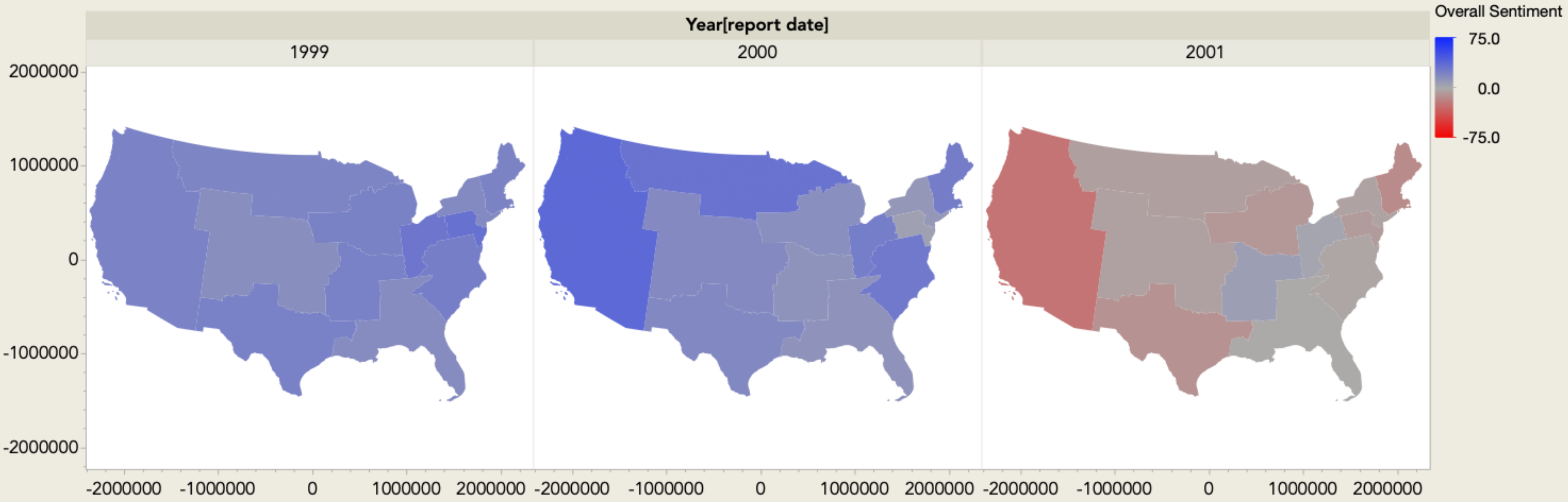
Document Sentiments

- depressed
- improved
- optimism
- problem
- serious
- severely
- sluggish

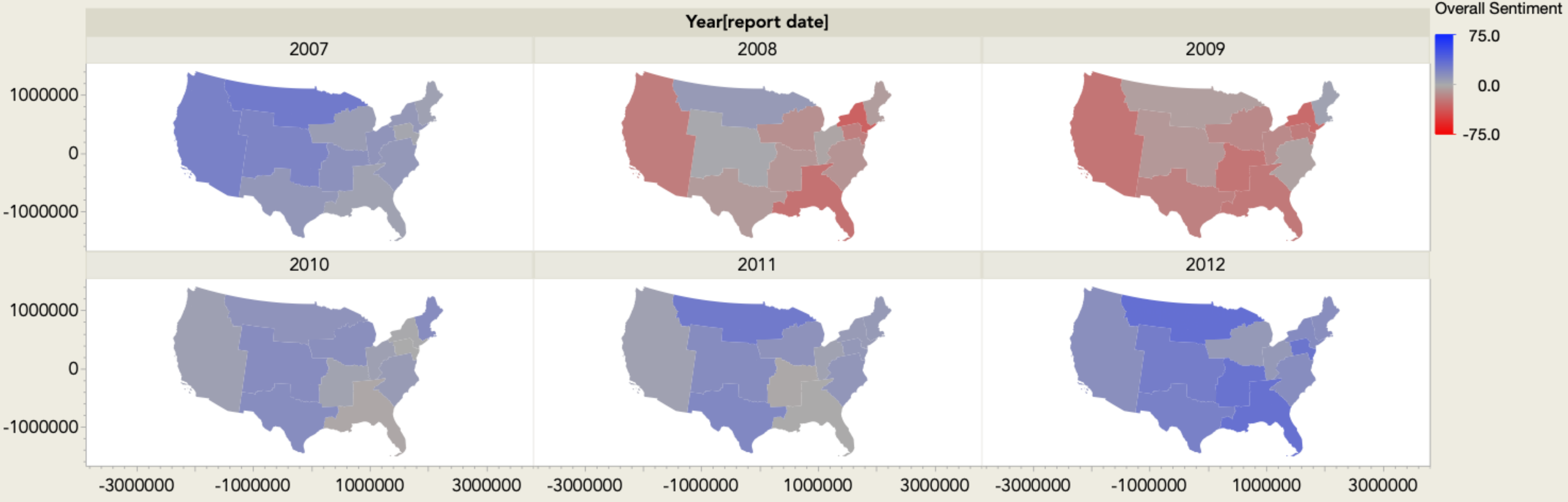
Leading indicator of recession?



The Dotcom Bust



The Great Recession



Some Applications of Sentiment Analysis

- Consumer research
- Product improvement
- Customer support
- Public policy

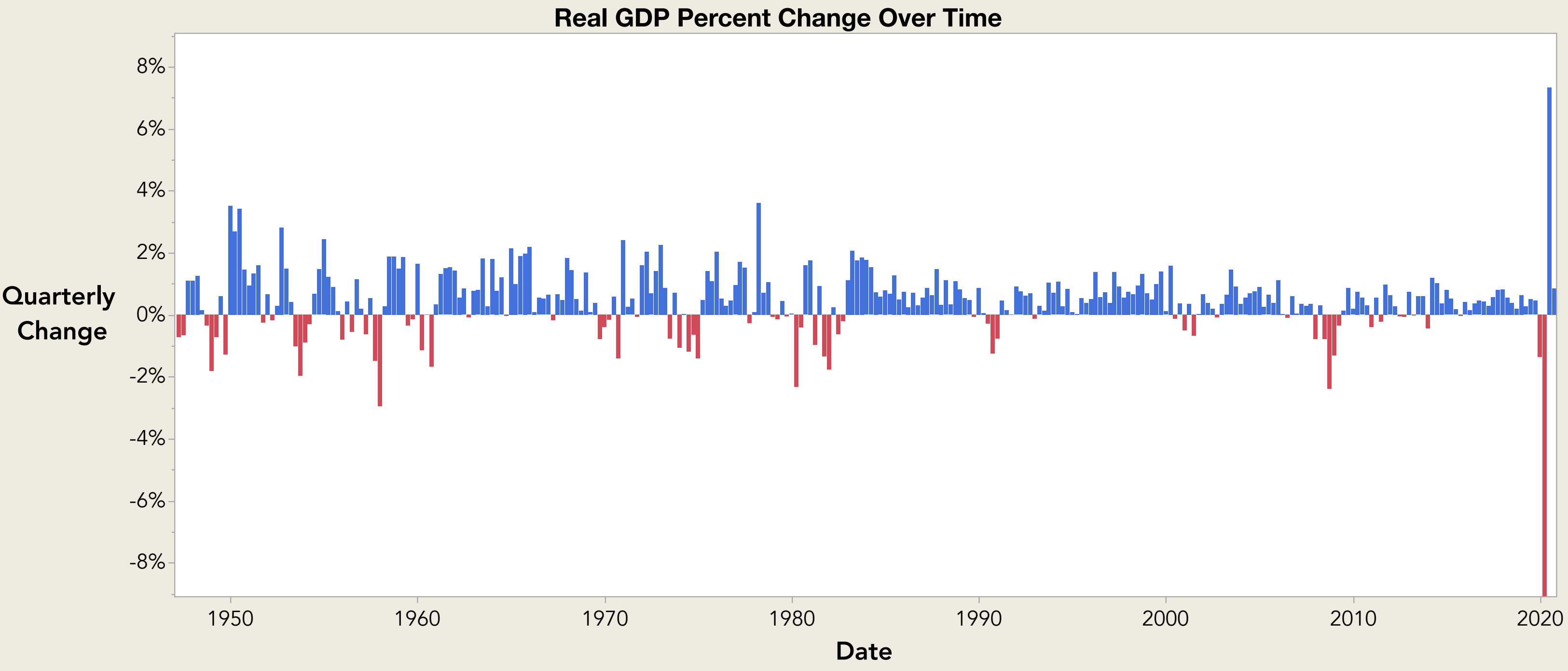
Let’s see it in JMP

Term Selection

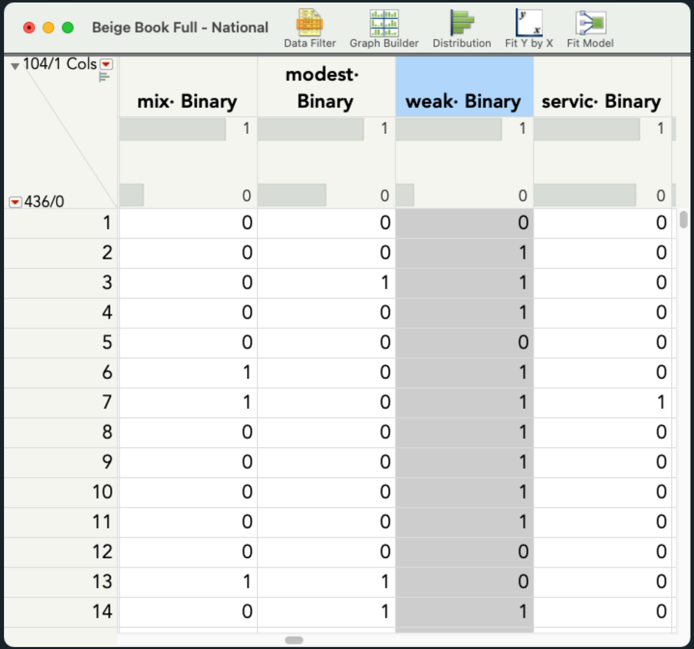
Which words are most strongly associated with an important variable?

Which words are associated with recessions?

The Beige Book



Document-Term Matrix



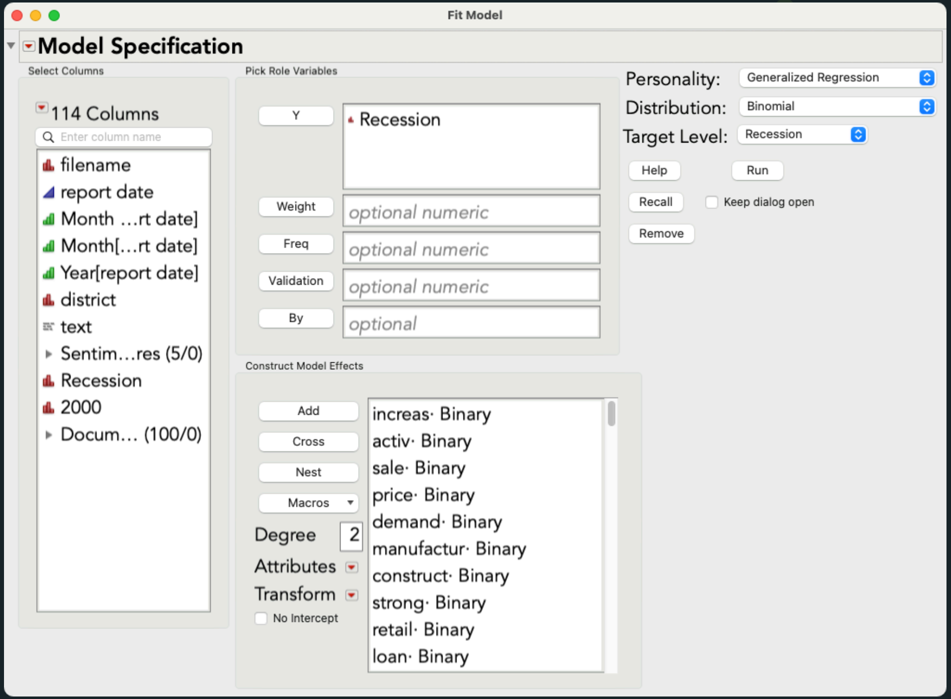
Beige Book Full - National

104/1 Cols

436/0

	mix· Binary	modest· Binary	weak· Binary	servic· Binary
1	0	0	0	0
2	0	0	1	0
3	0	1	1	0
4	0	0	1	0
5	0	0	0	0
6	1	0	1	0
7	1	0	1	1
8	0	0	1	0
9	0	0	1	0
10	0	0	1	0
11	0	0	1	0
12	0	0	0	0
13	1	1	0	0
14	0	1	1	0

Generalized Regression



Fit Model

Model Specification

Select Columns

114 Columns

Enter column name

- filename
- report date
- Month …rt date]
- Month[…rt date]
- Year[report date]
- district
- text
- Sentim…res (5/0)
- Recession
- 2000
- Docum… (100/0)

Pick Role Variables

Y: Recession

Weight: optional numeric

Freq: optional numeric

Validation: optional numeric

By: optional

Personality: Generalized Regression

Distribution: Binomial

Target Level: Recession

Construct Model Effects

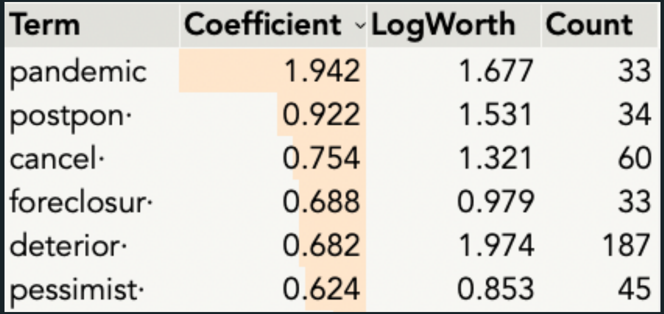
Degree: 2

- increas· Binary
- activ· Binary
- sale· Binary
- price· Binary
- demand· Binary
- manufactur· Binary
- construct· Binary
- strong· Binary
- retail· Binary
- loan· Binary

jmp STATISTICAL DISCOVERY FROM SAS

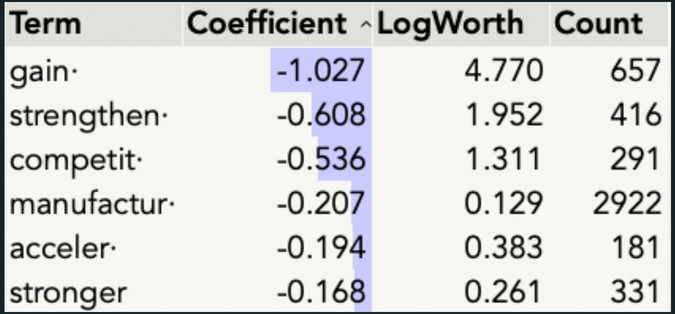
$$Log\left(\frac{P(Recession)}{P(No\ Recession)}\right) = -0.89 + (1.94 * "pandemic") - (1.02 * "gain") \ldots$$

Terms associated with recession

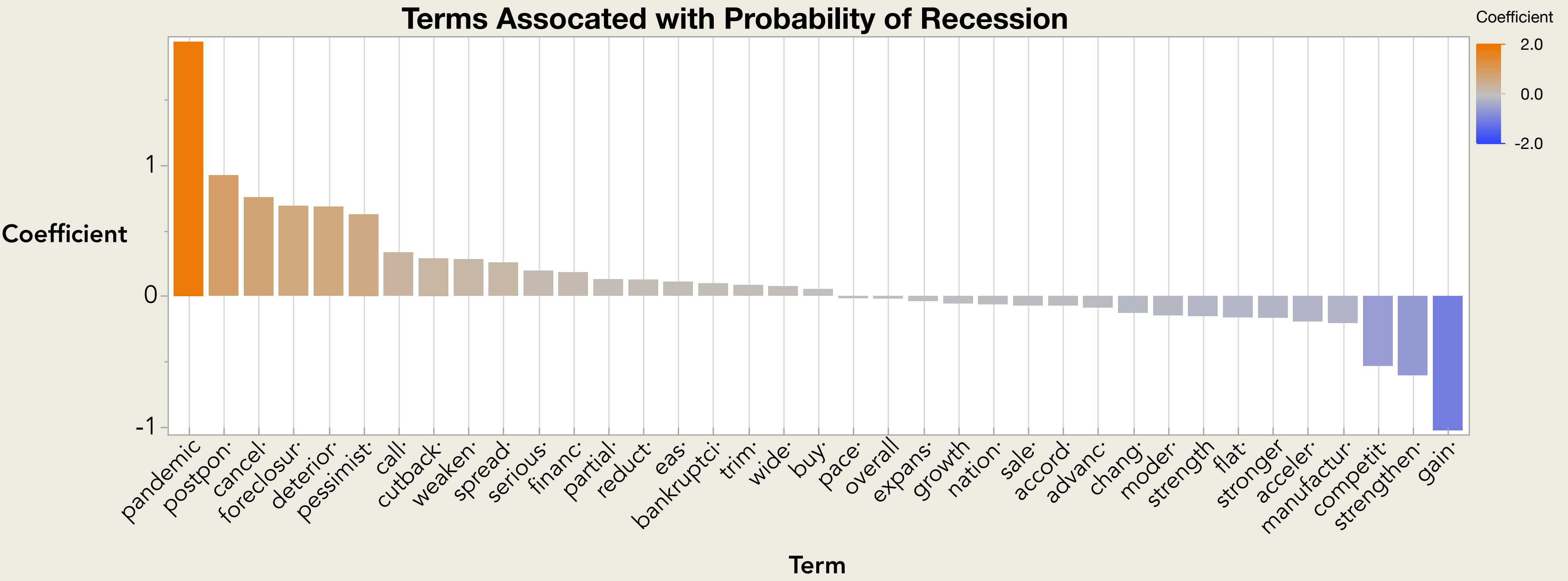


Term	Coefficient	LogWorth	Count
pandemic	1.942	1.677	33
postpon·	0.922	1.531	34
cancel·	0.754	1.321	60
foreclosur·	0.688	0.979	33
deterior·	0.682	1.974	187
pessimist·	0.624	0.853	45

Terms associated with NOT recession



Term	Coefficient	LogWorth	Count
gain·	-1.027	4.770	657
strengthen·	-0.608	1.952	416
competit·	-0.536	1.311	291
manufactur·	-0.207	0.129	2922
acceler·	-0.194	0.383	181
stronger	-0.168	0.261	331



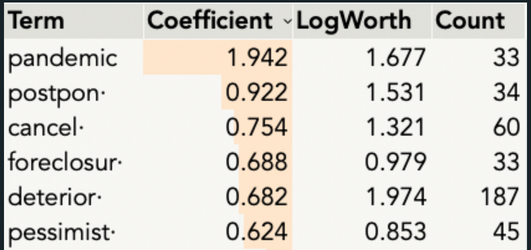
Terms Assocated with Probability of Recession

Let’s see it in JMP

Term Selection

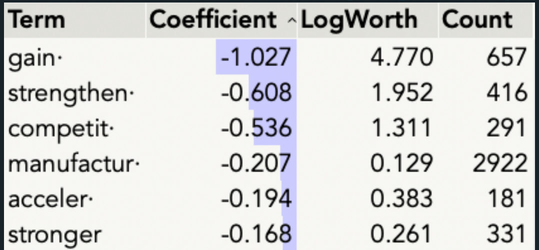
Identify words associated with an important variable

Terms associated with recession

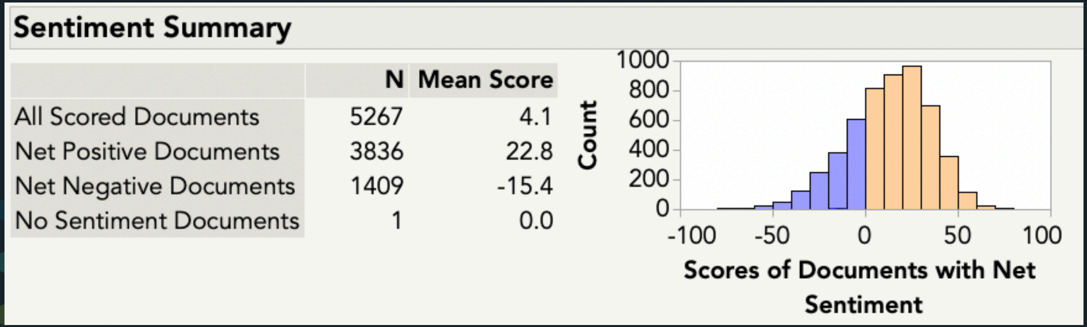


Term	Coefficient	LogWorth	Count
pandemic	1.942	1.677	33
postpon·	0.922	1.531	34
cancel·	0.754	1.321	60
foreclosur·	0.688	0.979	33
deterior·	0.682	1.974	187
pessimist·	0.624	0.853	45

Terms associated with NOT recession



Term	Coefficient	LogWorth	Count
gain·	-1.027	4.770	657
strengthen·	-0.608	1.952	416
competit·	-0.536	1.311	291
manufactur·	-0.207	0.129	2922
acceler·	-0.194	0.383	181
stronger	-0.168	0.261	331



Sentiment Summary

	N	Mean Score
All Scored Documents	5267	4.1
Net Positive Documents	3836	22.8
Net Negative Documents	1409	-15.4
No Sentiment Documents	1	0.0

Sentiment Analysis

Quantify positive-negative emotion in texts