Lecture: Overview of Text Analytics

Dr. Goutam Chakraborty

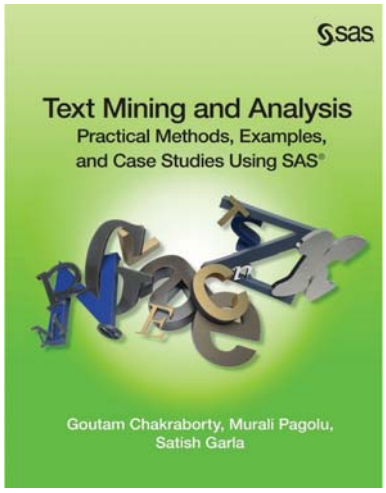
1

Outline

- **TEXT ANALYTICS** (Analyzing textual Data)
  - Explain structured versus unstructured data
  - Define text analytics
  - List examples of text analytic applications
  - Applications of text analytics using JMP Pro
    - Text parsing, stemming, …
    - Word cloud
    - Cluster and Topics

2

Before I begin, I must say…



3

Structured Vs. Unstructured Data

Customer	Age	Income	Gender	Response	…	Target
John	30	1200	M	No		0
Sarah	25	800	F	Yes		1
Sophie	52	2200	F	Yes		1
David	48	2000	M	No		0
Peter	34	1800	M	Yes		1

★★★★★ **Strongly recommended**, April 3, 2013

By REAL NAME (Annandale, VA, USA) - See all my reviews

This review is from: **A PRACTITIONER'S GUIDE TO BUSINESS ANALYTICS: Using Data Analysis Tools to Improve Your Organization's Decision Making and Strategy (Hardcover)**

There are a number of recent books urging managers to use quantitative analytics for better results, noting prominent examples of organizations that did so. There are a few books that explain some of the technical issues in applying analytics effectively, and a smaller number of these books might actually touch on how to select software. There might be one or two about how to integrate analytics into the organization for best results. As far as I can tell, there is only one book that does all of these things: this one.

And it does all three well, with a light, entertaining style that sugar-coats the very real and effective quantitative medicine. If you carefully read and think about all the real-life examples of how to (and, in a few cases, how not to) perform quantitative, mostly statistical analyses of business problems, you'll emerge wiser, not just with knowledge of new techniques and some cautions about how not to do it, but also with a more insightful way of looking at the world. Non-technical managers can learn much about how to utilize technical subordinates and consultants; techies can learn much about how to do better analyses and present them more cogently to management.

If you read just one book on business analytics this year, this should be it. Warning, though: it will stimulate you to learn some more. (Full disclosure: I reviewed a couple of chapters in draft and got acknowledged for helping.)

Help other customers find the most helpful reviews | Report abuse | Permalink

Was this review helpful to you? Yes No | Comment

4

Text generated per minute

470,000 tweets

510,000 posts

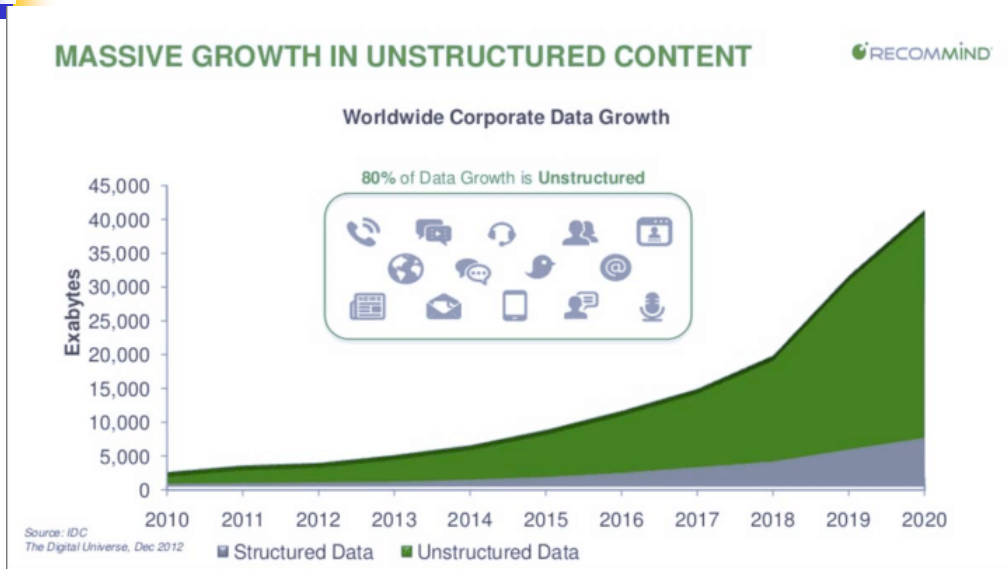
2,400,000 searches

16,000,000 texts

156,000,000 emails

Unstructured text is the largest human generated data source today

Growth of Unstructured Data Over Last Decade



Opportunities in Unstructured Data Space

Data Source	Structured	Unstructured
External	Mobile Phone / GPS, Credit History, Travel History, Real Estate Records, Census Data	Google+, Facebook, Twitter, Instagram, Pinterest, Blogs, External Sensor Data
Internal	CRM, Web Profiles, Sales Records, HR Records, Financials, Inventory	Online Forums, Web Feeds, SharePoint, Call Center, Sensor Data, Text Documents

Data Type

Source: Data and Technology Perspective by Ramesh Nair and Andy Narayanan

What is Text Analytics?

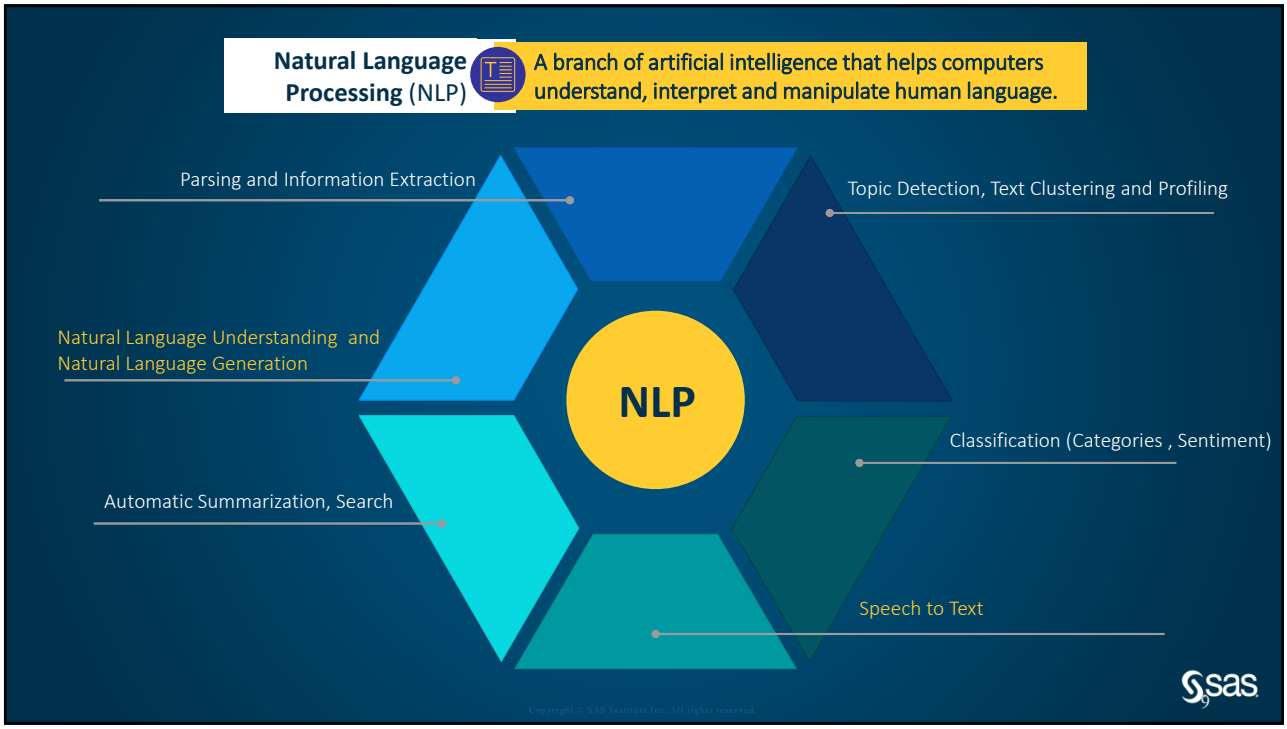
"The term **text analytics** describes a set of linguistic, statistical, and machine learning techniques that model and structure the information content of textual sources for business intelligence, exploratory data analysis, research, or investigation."*

*Source: Wikipedia

Or, put more simply…

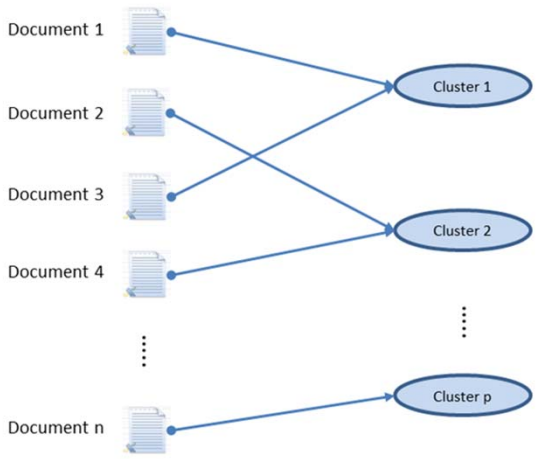
*"Using technology to **scale** the human acts of **reading, organizing, and quantifying** freeform text in meaningful ways."*

8



Common Application Areas of Text Analytics: Text Clustering

- Text Clustering:
  An unsupervised classification of text documents into mutually exclusive groups based on the presence of similar themes in the documents

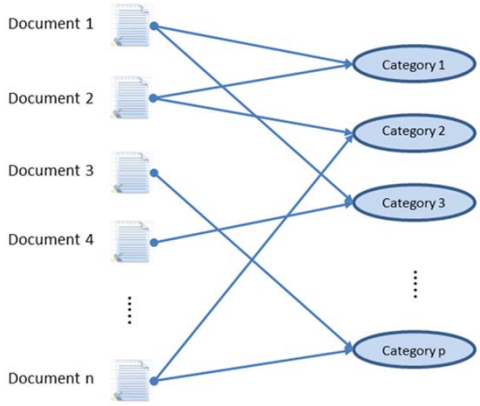


10

Common Application Areas of Text Analytics: Text Classification (Theme or Topic Based)

Text Classification (theme or topic based): The process of finding commonalities of documents in a corpus and grouping them into :

1. either pre-determined labels based on the topical themes exhibited by those documents
2. deriving the topics without having pre-determined labeled documents



11

Common Application Areas of Text Analytics: Predictive Models

- Predictive models can be used in text analytics in many ways, including
  - Classifying documents into groups based on models trained on labeled examples
  - Enhancing existing predictive models based on numerical data by augmenting with textual data and using the textual information in models.

12

Common Application Areas of Text Analytics: Sentiment Analysis

- Sentiment analysis deals with categorization (or classification) of opinions expressed in textual documents. Often such text units are classified into multiple categories such as positive, negative, or neutral based on the valence of the opinion expressed in those units.

The TV is wonderful. Great size, great picture, easy interface. It makes a cute little song when you boot it up and when you shut it off. I just want to point out that the 43" does not in fact play videos from the USB .This is really annoying because that was one of the major perks I wanted from a new TV. Looking at the product description now, I realize that the feature list applies to the X758 series as a whole, and that each model's capabilities are listed below. Kind of a dumb oversight on my part, but it's equally stupid to put a description that does not apply on the listing for a very specific model.

13

Text Parsing: First Step in any Text Analytics

- Converting unstructured text to spreadsheet (structured) type format for further analysis by algorithms

	X1	X2	X3	.....	Xp
Document 1	1	0	1	.....	1
Document 2	0	1	0	.....	0
Document 3	1	1	0	.....	1
Document 4	0	0	1	.....	1
.....	.....	.....	.....	.....	.....
Document n	1	0	0	.....	0

14

Demo: Basics of Text Parsing

Dr. Goutam Chakraborty

15

JMP Specific Terminologies

- A *term* or *token* is the smallest piece of text, similar to a word in a sentence. However, you can define terms in many ways, including through the use of regular expressions; the process of breaking the text into terms is called *tokenization*.
- •A *phrase* is a short collection of terms; the platform has options to manage phrases that are specified as terms in and of themselves.
- •A *document* refers to a collection of words; in a JMP data table, the unstructured text in each row of the text column corresponds to a document.
- •A *corpus* refers to a collection of documents.

16

Simplifying Analysis

- It is often desirable to exclude some common words from the analysis. These excluded words are called *stop words*. The platform has a default list of stop words, but you can also add specific words as stop words. Although stop words are not eligible to be terms, they can be used in phrases.
- *Stemming* is the process of combining words with identical beginnings (*stems*) by removing the endings that differ. This results in "jump", "jumped", and "jumping" all being treated as the term "jump
- You can also recode terms; this is useful for combining synonyms into one common term such as Car and Automobile.

17

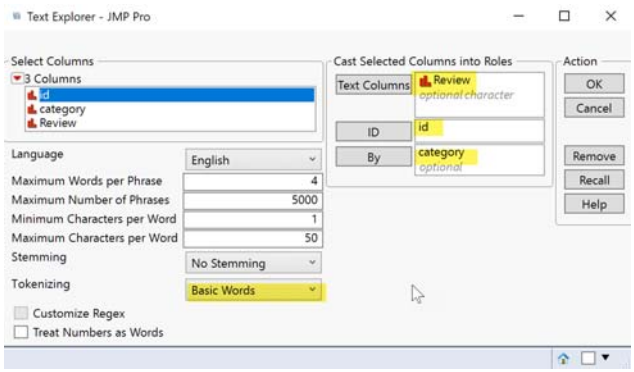
Android App Review

- An artificial data set of 500 reviews was created by modifying and anonymizing *actual customer reviews* that were posted online.
- Raw textual data was categorized into positive and negative groups based on five-star numerical ratings given by a consumer on the review site.
  - Comments greater than or equal to four stars are categorized as positive, and those less than or equal to two stars are categorized as negative for this analysis
- Data Set name: App_review_data
- Variables: ID (identification number), Category (positive or negative), Review (textual reviews by consumers)

18

Understand Tokenization

- Analyze > Text Explorer > Select Review as Text Columns, id as ID and Category as By > Change tokenizing to Basic Words > click OK



19

JMP Output

Text Explorer for Review category=negative

Number of Terms	Number of Cases	Total Tokens	Tokens per Case	Number of Non-empty Cases	Portion Non-empty per Case
1099	250	6835	27.34	250	1.0000

Term and Phrase Lists

Term	Count
product	173
products	76
device	63
update	63
skins	61
clock	57
time	55
weather	55
fix	41
now	39
please	34
work	34
even	26
like	26
battery	25
download	22
working	22
screen	20
since	20
last	19
paid	19
used	19
still	18
updates	18
just	17
new	17
system	17

Phrase	Count	N
please fix	29	2
last update	15	2
competitor products	11	2
weather product	9	2
super clock	8	2
since the last	7	3
waste of money	7	3
home screen	7	2
since the last update	6	4
clock freezes	6	2
great product	6	2
latest update	6	2
clock product	5	2
keeps freezing	5	2
stopped working	5	2
love this product	4	3
super clock product	4	3
anonymous statistics	4	2
battery life	4	2
battery product	4	2
every time	4	2
force closes	4	2
new skins	4	2
toggle products	4	2
wrong time	4	2
product on my device	3	4
discover known accounts	3	3

Token = a word or, any sequence of non-whitespace characters.
Terms = basic unit of analysis. Jmp makes a quick determination of useful terms from the tokens. The initial number of terms is 1099 but it will change as we work the data.
Phrases = A short collection of terms that appear more than once

Red triangle option > Save Document Term Matrix
Accept default settings. Look at data table

20

Term-by-Document Matrix

	Review	product Binary By ...	products Binary By ...	device Binary By category	update Binary By category
1	eather don't show on installed skin list.	0	1	0	0
2	ip in the installed skins list. Please fix it. it's buggy as hell, and some skins won't downloa...	0	0	0	0
3	device.	0	0	1	0
4	at same time. It's very annoying.	0	0	1	0
5		1	0	0	0
6	gest making two versions of this to avoid detrimental changes. Even use an activation co...	1	1	0	1
7	pening again. Making it useless once again.	0	0	0	0
8	zy addon and setting needs to be simplified	0	0	0	0

Likely too many terms – we should simplify by stemming, grouping etc.
Go ahead and Delete all of the newly added binary columns

21

Stemming : Red Triangle > Term Options >Stemming > Stem for Combining

Text Explorer for Review category=negative

Number of Terms	Number of Cases	Total Tokens	Tokens per Case	Number of Non-empty Cases	Portion Non-empty per Case
867	250	6835	27.34	250	1.0000

Term and Phrase Lists

Term	Count
product·	250
updat·	104
skin·	77
work·	69
devic·	67
clock·	64
time·	64
weather	55
fix·	54
use·	46
now	39
please	34
download·	33
like·	29
look·	27
even	26
battery	25
need·	24
show·	22
screen·	21
freez·	20
last·	20
load·	20
since	20
paid	19
chang·	18
still	18

Phrase	Count	N
please fix	29	2
last update	15	2
competitor products	11	2
weather product	9	2
super clock	8	2
since the last	7	3
waste of money	7	3
home screen	7	2
since the last update	6	4
clock freezes	6	2
great product	6	2
latest update	6	2
clock product	5	2
keeps freezing	5	2
stopped working	5	2
love this product	4	3
super clock product	4	3
anonymous statistics	4	2
battery life	4	2
battery product	4	2
every time	4	2
force closes	4	2
new skins	4	2
toggle products	4	2
wrong time	4	2
product on my device	3	4
discover known accounts	3	3

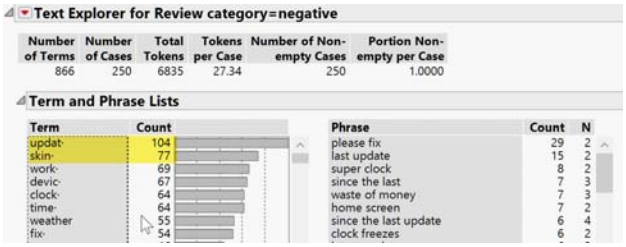
Right-click "freez." and select Show Text to see all of the documents where all versions of "freez." appear.

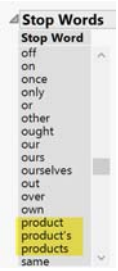
We can do the same with phrases such as weather product Or, Last update or, please fix – these help generate insights

22

Stop Word

- The term product or products seem to appear in almost every document.
  - It is probably not very meaningful to distinguish between documents
  - Right-click and Add Stop Word
  - Note changes in the term and phrase list.
  - Red triangle> Display options > Show Stop Words. Scroll down to see product and its variants added to the list of default stop words > Click and open Stop Words list

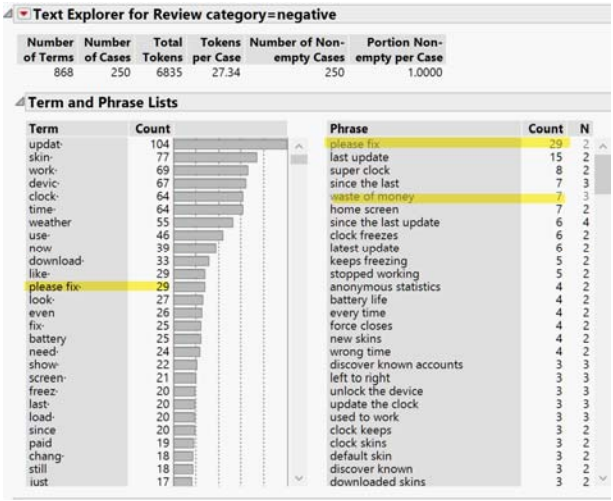




23

Working with Phrases

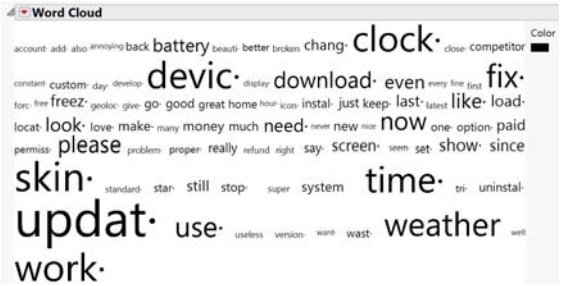
- Goal is to look at phrases and decide if we want to treat some phrases as terms and analyze them in that way.
- As an example, we will take two phrase, please fix and waste of money, and add that to the terms.
- Right click on each of the two phrases and select Add Phrase



24

Red Triangle > Display Options > Show Word Cloud

Red triangle next to Word Cloud allows you to change display from ordered to alphabetical To centered or colors



25

To Summarize,

- Text analytics involve a lot of user involvement to get a sense of the data.
- General recommended process as follows (before you begin analysis):
  - Recode all misspellings and synonyms
  - Use stemming
  - Use Regex (instead of basic words) for parsing
  - Examine phrases and specify which phrases (usually high frequency) you want to treat as terms
  - If needed,
    - Remove *least frequent* terms via stop words
    - Remove *most frequent* terms via stop words

26

Demo: More Text Analytics

Dr. Goutam Chakraborty

27

Outline

- Install Random Seed Add-In
- Do some data cleaning (as an example)
- Extract topics from documents (unsupervised)
- Classify documents into clusters (unsupervised)

28

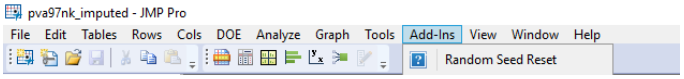
Install Random Seed Reset Add-in in JMP

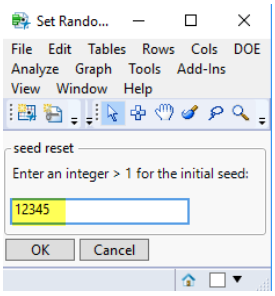
- Why we need this?
  - Many of the solutions are probabilistic (not deterministic)
  - If we don't fix the random seed – it may produce different results every time you run it

29

Random Seed Reset Add-In

- First, download "Random Seed Reset Add In".
- Double-click on the file and it should prompt you for installing the Add-In. Accept
- If things go right, in your JMP top-menu, you will see Random Seed Reset under Add-ins
- Open the Add-In and set seed to 12345 – Must do this every time you run SVD, LSA, LSI etc.

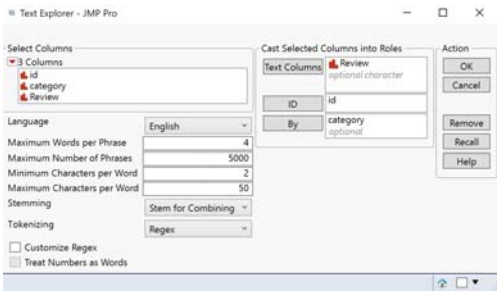


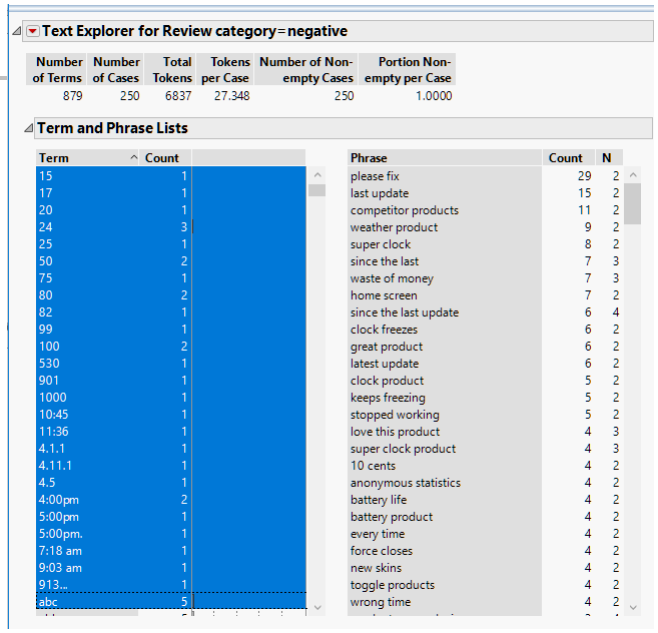


30

Cleaning up the Data

- First, let's redo Parsing, Analyze > Text Explorer > recall change Tokenizing to **Regex**, Minimum characters per word as **2** and **Stem for combining** > OK
- **Right-click** terms and phrase lists > select alphabetical order > **Mark** and all of the terms starting from top to abc > **Right-click** and select Add Stop Words





Text Explorer for Review category=negative

Number of Terms	Number of Cases	Total Tokens	Tokens per Case	Number of Non-empty Cases	Portion Non-empty per Case
879	250	6837	27.348	250	1.0000

Term and Phrase Lists

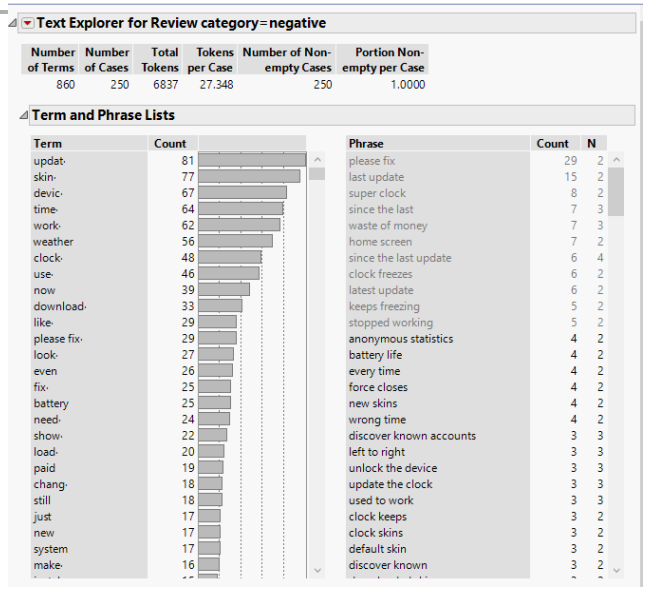
Term	Count
15	1
17	1
20	1
24	3
25	1
50	2
75	1
80	2
82	1
99	1
100	2
530	1
901	1
1000	1
10:45	1
11:36	1
4.1.1	1
4.11.1	1
4.5	1
4:00pm	2
5:00pm	1
5:00pm.	1
7:18 am	1
9:03 am	1
913...	1
abc	5

Phrase	Count	N
please fix	29	2
last update	15	2
competitor products	11	2
weather product	9	2
super clock	8	2
since the last	7	3
waste of money	7	3
home screen	7	2
since the last update	6	4
clock freezes	6	2
great product	6	2
latest update	6	2
clock product	5	2
keeps freezing	5	2
stopped working	5	2
love this product	4	3
super clock product	4	3
10 cents	4	2
anonymous statistics	4	2
battery life	4	2
battery product	4	2
every time	4	2
force closes	4	2
new skins	4	2
toggle products	4	2
wrong time	4	2

31

Cleaning up the Data (contd.)

- **Right-click** terms table > deselect alphabetical order
- **Right-click** product. and Add Stop Word
- **Right-click** all phrases with counts of 5 or more (please fix to stopped working) and select add phrase



Text Explorer for Review category=negative

Number of Terms	Number of Cases	Total Tokens	Tokens per Case	Number of Non-empty Cases	Portion Non-empty per Case
860	250	6837	27.348	250	1.0000

Term and Phrase Lists

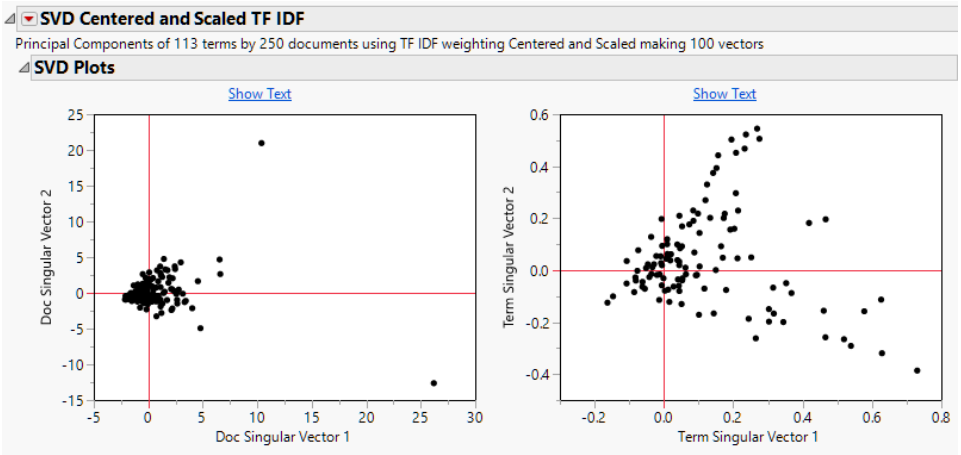
Term	Count
updat·	81
skin·	77
devic·	67
time·	64
work·	62
weather	56
clock·	48
use·	46
now	39
download·	33
like·	29
please fix·	29
look·	27
even	26
fix·	25
battery	25
need·	24
show·	22
load·	20
paid	19
chang·	18
still	18
just	17
new	17
system	17
make·	16

Phrase	Count	N
please fix	29	2
last update	15	2
super clock	8	2
since the last	7	3
waste of money	7	3
home screen	7	2
since the last update	6	4
clock freezes	6	2
latest update	6	2
keeps freezing	5	2
stopped working	5	2
anonymous statistics	4	2
battery life	4	2
every time	4	2
force closes	4	2
new skins	4	2
wrong time	4	2
discover known accounts	3	3
left to right	3	3
unlock the device	3	3
update the clock	3	3
used to work	3	3
clock keeps	3	2
clock skins	3	2
default skin	3	2
discover known	3	2

32

Topic Extraction via SVD

- Red triangle > Latent Semantic Analysis, SVD > accept all default options



33

Topics and Clusters

- Red Triangle>Topic Analysis Rotated SVD >Accept defaults
  - Understand topics extracted by default using terms
  - You may need to play with settings to get meaningful topics
  - 5 topics may give a clearer idea than 10 (default)
- Red triangle > Latent Class Analysis > Accept defaults
  - Understand clusters assigned by default
  - You may need to play with settings to get meaningful topics

34

Where do We Go from Here?

- Automatic topic extraction and cluster assignment is a just a starting point.
  - Trial-and-error is needed to get meaningful results
  - Also, once we get the basic ideas from default topics, for better insights, we design custom topics using terms that are meaningful from a domain perspective
- We can save the SVD scores for use in predictive modeling by combining with numeric data.
- Your turn –play with the data and see what insights you can gain from the positive comments.
  - Compare and contrast topics/clusters from positive versus negative comments

35

Lecture: Case Studies and Next Steps

Dr. Goutam Chakraborty

36

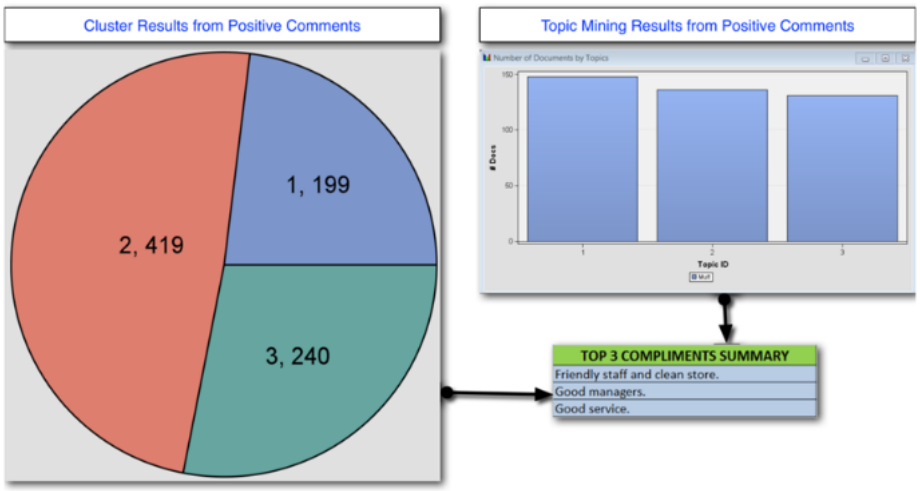
Oil and Gas Industry Case

- Data: 5,000+ text messages received during 1-month **test run** of a mobile app used by professional drivers.
- Company personnel manually read each message to understand and summarize content.
- Analysis goals:
  - Automatic generation of text themes that provide insights about message content in texts
  - Automatic prediction of sentiments expressed in texts

Opinion Mining and Geo-Positioning of Textual Feedback from Professional Drivers, by Mantosh Sarkar and Goutam Chakraborty, SAS Global Forum 2013,.

37

Oil and Gas Industry Case (contd.)

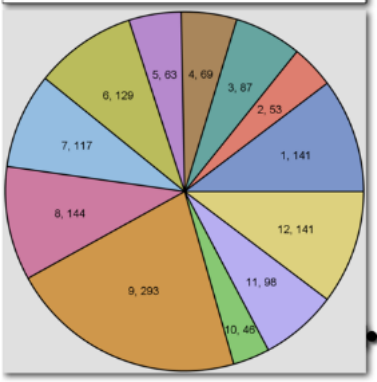


Three themes emerge from analyzing **positive** comments

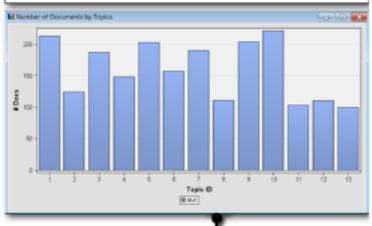
38

Oil and Gas Industry Case (contd.)

Cluster Results from Negative Comments



Topic Mining Results from Negative Comments



TOP 13 COMPLAINTS SUMMARY

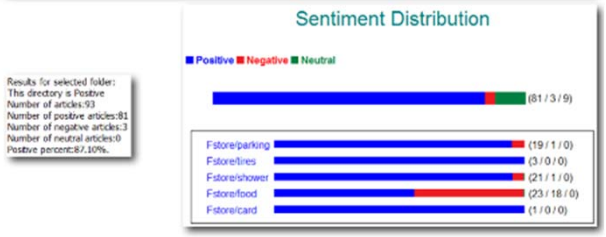
Dirty shower rooms. Especially floors, sink, and shower floor.
Trash all over the parking lot.
Small parking lot. Trucks parked on the road and even on the grass.
Waiting for shower (>1 Hour).
Unhygienic food preparation.
Rude employees and poor customer service.
Non-functioning card reader.
Poor management of Fuel Island. Drivers keep their trucks unattended on the Fuel Island and go for eating or taking shower.
Empty coffee pot and creamers.
Drivers are getting charged more than what they are supposed to even after being a platinum member. Discrepancy in price advertised and charged.
No hot water.
Over-priced tires and 2-5 hours of tire changing time.
Low water pressure.

Many themes emerge From analyzing **negative** comments

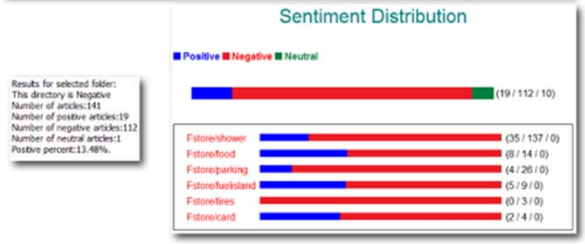
39

Oil and Gas Industry Case (contd.)

Positive directory results



Negative directory results



Overall accuracy of the automatic model for predicting positive sentiments : **87.1%**
Overall accuracy of the automatic model for predicting negative sentiments : **86.5%**

Value:

- **Save $** by deploying scoring model to score texts when the app is rolled out to **all customers**
- Improve store operations to **increase** customer **satisfaction**

40

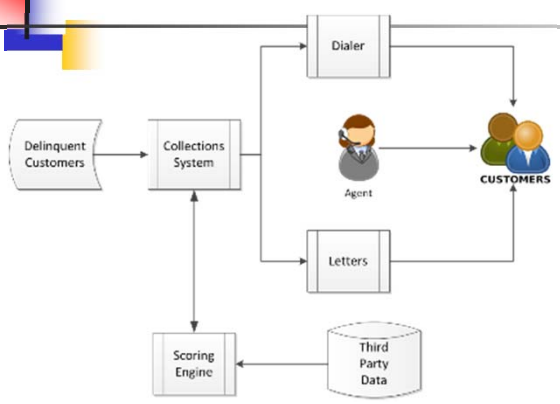
Financial Industry Case

- Company specializes in debt collections from delinquent customers
- Business Challenge: Current prediction model for a delinquent customer's propensity to pay is based on transaction and third party data.
  - Can this prediction model be improved by analyzing texts from agent-customer interaction over phone?

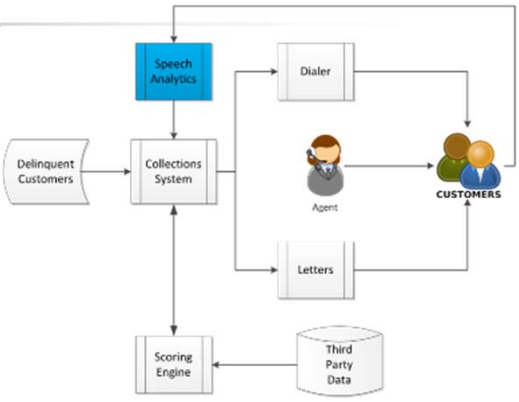
Speech Analytics Applications to Predictive Modeling by Dmitriy Khots and Goutam Chakraborty presented at the SAS Analytics 2013 Conference.

41

Financial Industry Case (contd.)



Current System: Creates a score-based treatment strategy for the life of the account (e.g. best scores get more calls)

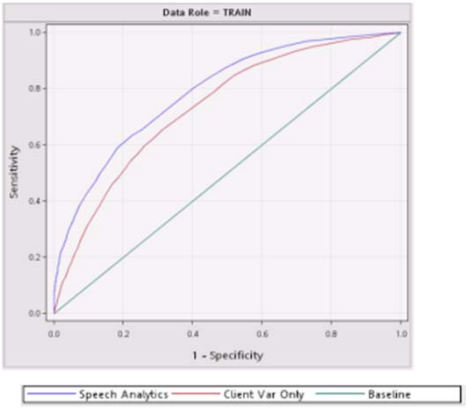


New System: Contact call details are recorded and analyzed via speech/text analytics, re-scoring is triggered if certain indicators appear on the account

42

Financial Industry Case (contd.)

- **Simplified Model Design using SAS EM**



- Both models used stepwise logistic regression
- Model with speech (text) analytics derived indicators provides a substantial lift over the model with customer attributes only.

Value:
- Increase in collections

43

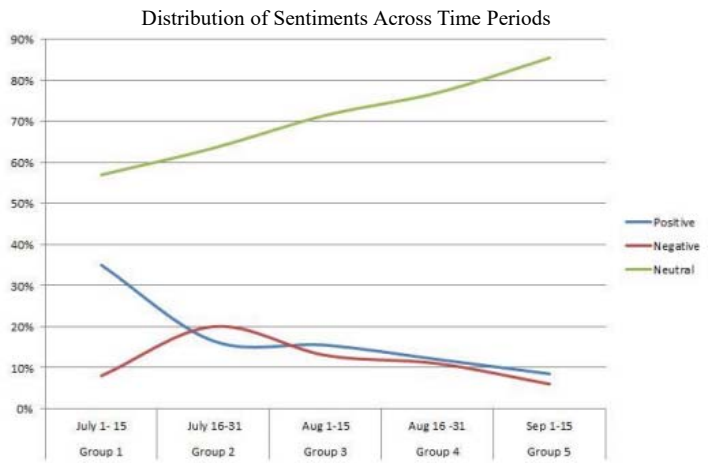
Fast Food Industry Case

- The second largest quick-service chicken restaurant chain in the United States, with over 1,700 locations.
- Business Issue: Company CEO made a strong statement about a controversial topic that went viral in the social media space giving rise to a lot of positive and comments on Twitter
  - What impact, if any, did the CEO's comment have on public sentiments?
  - Is the impact short-term or does it persist?

Analysis of Change in Sentiments towards Chick-fil-A after Dan Cathy's Statement about Same-Sex Marriage Using SAS® Text Miner and SAS® Sentiment Analysis Studio by Goutam Chakraborty, Jeffin Jacob, and Swati Grover, SAS Global Forum 2013

44

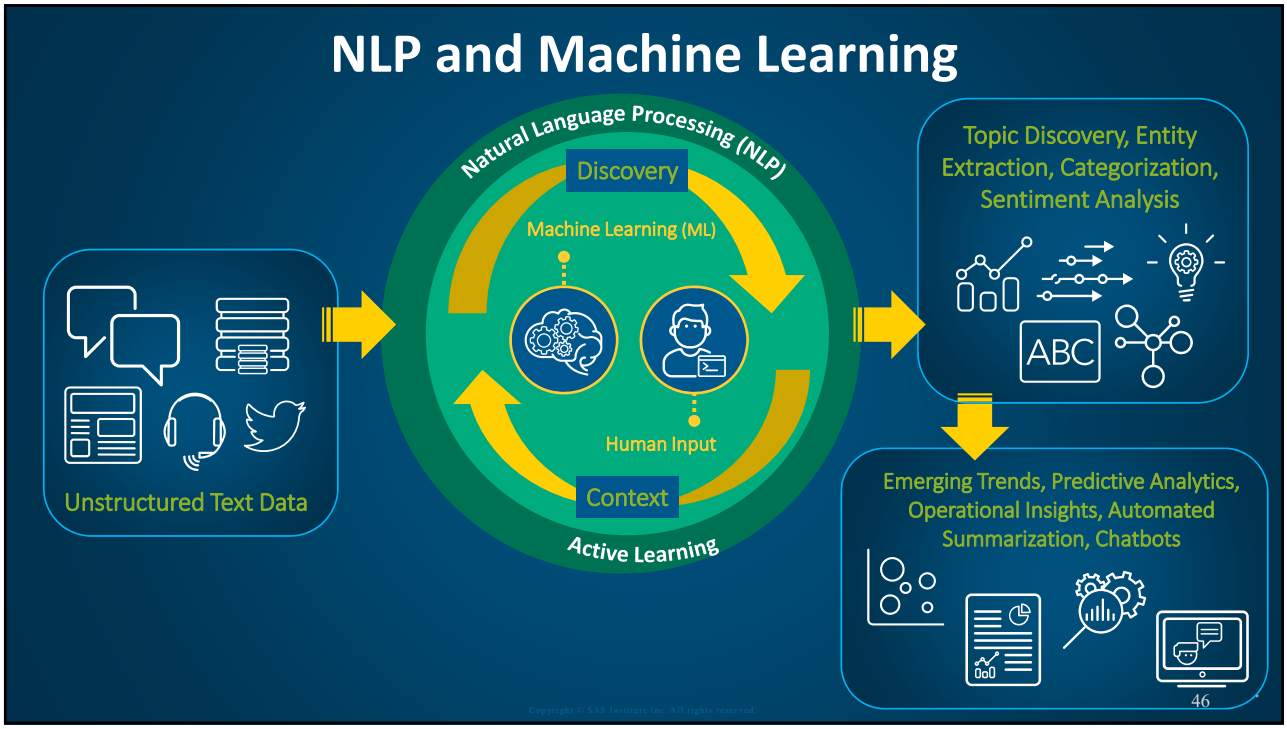
Fast Food Industry Case (contd.)



Value:

- Tracking sentiment allows company to spot shift (transient and permanent) and take actions

45



Hybrid Approach to Text Analytics

Rule Based

Segmentation
Tokenization
Lemmatization
Part-of-Speech tagging
Concept extraction
Sentiment analysis
Categorization

Supervised

Language modeling
Part-of-speech tagging
Named entity recognition
Sentiment analysis
Categorization
Rule generation
Text Summarization

Unsupervised

Topic Recognition
Component Analysis

Deep Learning

47

Lessons Learned from Case Studies

- No easy button in text analytics!
- What's needed for successful text analytics projects are:
  - Enterprise level software that allows seamless integration of numerical and text data
  - Time needed to play and learn!
  - Team with domain experts and trained text analysts
- Suggestions:
  - Start with internal text data for quick results
  - System to track and monitor external text data continuously

48

Questions?

Comments?

Thanks very much

49